Structure, Function and Motion in Transmembrane Proteins: Computational Studies with Receptors, Transporters and Channels.

Thesis submitted for the degree
"Doctor of Philosophy"

By Sarel J. Fleishman

Submitted to the Senate of Tel-Aviv University
July, 2006

This work was carried out under the supervision of
Prof. Nir Ben-Tal

Acknowledgements

My gratitude to Nir Ben-Tal, who has been both a scientific mentor and a friend. His insight, sensitivity, and respect for the other's opinion made everyday work a superb experience of learning and discourse.

Two collaborators, Ofer Yifrach and Vinzenz Unger, deserve special thanks for their patience in guiding me through the intricacies of structure-function analysis, and especially for a wonderful collaboration.

Many thanks to Dan Graur and Lewi Stone, who took their time and invested great effort in showing me how to put ideas to the test in fields, in which I had no experience.

My thanks to members of the Ben-Tal laboratory, Amit Kessel, Dalit Shental-Bechor, Meytal Landau, Guy Nimrod, and Yanay Ofran, who have been both enduring friends and colleagues, and to all other members of the Ben-Tal group.

I am grateful to the Clore Israel Foundation for providing a Doctoral Fellowship and travel support.

This dissertation is dedicated to my parents, Tamar and Pniel, and to my wife Dana, who has supported me and been my partner.

Table of Contents

Abstract

Transmembrane (TM) proteins constitute 15-30% of the genome, but less than 1% of the structures in the Protein Data Bank. This discrepancy is disturbing and emphasizes that structure determination of TM proteins remains challenging. The challenge is greatest for proteins from eukaryotes, whose structures remain intractable, despite tremendous advances made towards the structure determination of bacterial TM proteins. Notably, over half of the membrane-protein families in eukaryotes lack bacterial homologs. It is therefore conceivable that many more years will elapse before atomic-resolution structures of eukaryotic TM proteins will emerge. Until then, integrated approaches, combining biochemical and computational analyses with low-resolution structures, will probably play increasingly important roles in providing frameworks for the mechanistic understanding of membrane protein structure and function. I have thus developed several methods for structure prediction in TM proteins, which were based on phylogenetic analysis, biochemical and low-resolution structural data. The computational methods relied on evolutionary conservation to predict which amino-acid positions are buried in the protein core, and a method for detecting correlations in the substitutions of amino-acid sites in order to predict interactions between residues. These and other methods were applied the study of the structure and mechanism of action of the receptor tyrosine kinase ErbB2, the gap junction intercellular channel, and the small multidrug resistance antiporter from *E. coli* EmrE. In all of these cases, significant agreement was found between published experimental results and the model structures, and additional insight was provided into the proteins' mechanisms of action, including in the two former cases, their involvement in disease. We postulated several hypotheses on structural

relationships between amino-acid residues in the gap junction TM domain. These hypotheses were tested experimentally and provide partial validation of this model.

Introduction

This Introduction is based on review articles in Trends Biochem. Sci. (M11:Fleishman et al., 2006) and Curr. Op. Struct Biol. (M12:Fleishman and Ben-Tal, 2006).

It is estimated that transmembrane (TM) proteins constitute some 15-30% of eukaryotic genomes (Krogh et al., 2001; Liu and Rost, 2001; Mitaku et al., 1999; Rost et al., 1996). Due to their strategic localization at the interfaces between the interior and exterior of the cell and between cellular compartments, membrane proteins play pivotal roles in many cellular processes, including cell-to-cell signaling events, solute transport, and cellular organization. For this reason, membrane proteins are by far the most attractive targets for drug discovery. Despite their significance, however, only a few distinct folds of TM proteins have been solved to date by high-resolution methods such as X-ray crystallography (White, 2004) and NMR (Opella and Marassi, 2004); accordingly, TM-protein structures constitute less than 1% of the entries in the Protein Data Bank (PDB). Disturbingly, only two of the current entries represent a membrane protein from human origin (MacKenzie et al., 1997; Murata et al., 2000), while the large majority of entries are of bacterial membrane proteins (Fig. 1).

Part of the reason why progress has been faster for bacterial membrane proteins stems from the fact that they can more easily be expressed in large quantities in bacterial hosts, and that they lack many of the posttranslational modifications that potentially complicate crystallization. Moreover, the fast pace at which bacterial genomes are sequenced provides an almost unlimited repertoire of target proteins including homologs from thermophilic bacteria that are often more stable during detergent solubilization, purification, and crystallization. In contrast, eukaryotic membrane proteins are more difficult to express than their bacterial homologs, are

subject to posttranslational modifications, and often only few candidate genes are available for screens to identify the ideal target protein. It thus comes as no surprise that over the past few years, efforts have been focused on identifying bacterial homologs of eukaryotic membrane proteins, and to pursue their structure determination by "brute-force" approaches, utilizing sometimes, thousands of combinations of homologs of the protein and different crystallization conditions (Chang et al., 1998). As evident in Figure 1, this strategy has begun to bear fruit and many exciting structures should be expected to emerge over the next several years (White, 2004). However, this approach does not represent the ultimate solution because a large number of eukaryotic membrane proteins do not have bacterial homologs. In fact, a search in the Pfam-A database of protein families (Bateman et al., 2004) shows that only 47% of the eukaryotic TM-protein families have bacterial or archaeal homologs.

It should nevertheless be stressed that even bacterial membrane proteins are difficult targets for structure analysis. One of the crucial steps in the crystallization of membrane proteins relies on the use of detergents and the extraction of lipid. It has been demonstrated, however, that this step can distort the protein fold (Lee et al., 2005). In this connection, substantial evidence exists suggesting that a small number of published structures of membrane proteins from x-ray crystallography do not represent physiologically relevant conformations (Davidson and Chen, 2005; M12:Fleishman and Ben-Tal, 2006; Lee et al., 2005; Tate, 2006). One such case that has been a subject of part of this doctoral work concerns the small multidrug resistance protein from *E. coli*, EmrE.

With such a large fraction of proteins in eukaryotes, whose direct structure determination will have to await possibly many years, and due to the challenges that

still face the structure determination of bacterial membrane proteins, we have developed several data-based modeling approaches that rely on inferences derived from biochemical, computational, evolutionary, and intermediate-resolution structural data.

Architecture of helical TM proteins

A simple rule that has guided many of the approaches to modeling helical TM proteins is the two-stage model of folding (Engelman et al., 2003) (Fig. 2). According to this model, hydrophobic segments are first inserted into the plasma membrane in the form of helices, which engage the polar carbonyl and amide groups on the peptide chain's backbone in hydrogen bonds, and shield them from the hydrophobic lipid bilayer. Next, these helices associate with one another to shape the protein's tertiary structure. One of the implications of the two-stage model for computational modeling is that each of the hydrophobic segments comprising the TM domain can be approximated as an energetically stable canonical α-helix, whose polar backbone and N- and C- termini are shielded from the membrane environment. Hence, TM-protein structure prediction can concentrate on the relative configurations of preformed α-helices. This constraint reduces considerably the number of degrees of freedom that must be explored computationally.

This relatively simple picture of TM-protein architecture was supported by the first few membrane proteins to be solved (Baldwin, 1993; Deisenhofer et al., 1995; Henderson et al., 1990; Kuhlbrandt and Wang, 1991) (e.g., Fig. 3). Moreover, the extramembrane loops are relatively short in these proteins dictating that consecutive domains in the sequence are proximal in the 3-dimensional (3D) structure (Bowie, 1997). However, this simplistic picture collapsed when the first ion-channel structures

revealed that helices need not span the entire width of the bilayer (Doyle et al., 1998), and can be extremely long and highly tilted with respect to the membrane normal (Dutzler et al., 2002) (Fig. 4a,b). Recent transporter structures have also shown marked deviations from α-helicity; these deviations were suggested to play a role in the conformational changes underlying transporter functions by destabilizing the structures (Fig. 4c) (Abramson et al., 2003). All of these structural features are still beyond what can be reliably predicted by computational methods, raising the question of how many membrane domains might have gone unnoticed. More importantly, however, the observation that not all consecutive hydrophobic domains formed physical contact (Doyle et al., 1998; Dutzler et al., 2002) heralded the end of naïve modeling of TM proteins, and underscored the importance of a joint experimental-computational approach to structure prediction. Over the past several years, mostly two sources of experimental data have proven valuable in aiding modeling exercises of polytopic membrane proteins: low-resolution structures obtained by cryo-electron microscopy (cryo-EM) and mutational analyses of structure-function relationships.

Cryo-EM of 2D crystals of TM proteins

In contrast to the difficulties usually experienced in obtaining 3D crystals of TM proteins, in some cases, membrane proteins readily form 2-dimensional (2D) arrays in the membrane (e.g., bacteriorhodopsin (Unwin and Henderson, 1975), photosystem II (Rhee et al., 1998), the gap junction (Unger et al., 1999), the bacterial translocon complex secYEG (Breyton et al., 2002), and the bacterial multidrug resistance transporter EmrE (Ubarretxena-Belandia et al., 2003)). Added advantages of 2D crystals are that they mimic the native environment of the protein more closely than does the 3D crystal, including interactions with the surrounding lipid molecules, which sometimes play important roles in determining the physiological structure

(Fujiyoshi, 1998). For instance, substantial differences were observed between the cryo-EM map of EmrE (Ubarretxena-Belandia et al., 2003) and two structures of that protein derived from x-ray analysis of 3D crystals (Ma and Chang, 2004; Pornillos et al., 2005). Since the detergent-solubilized form of EmrE that was used in the cryo-EM analysis binds substrate with the same high affinity as the protein in its native membrane (Tate et al., 2003), the cryo-EM structure is thought to be the more native-like structure (Tate, 2006). Another demonstration of the importance of maintaining a membrane-like environment is provided by the differences between two recent x-ray structures of the voltage-gated potassium channel (Jiang et al., 2003; Long et al., 2005), the latter of which was crystallized in the presence of lipids. Lastly, it is sometimes possible to induce crystal formation in 2D, even when the proteins are dispersed in the membrane (Hasler et al., 1998), and even relatively small and poorly ordered crystals can be used to derive data in the 5-10 Å resolution range thanks to digital image-processing protocols that allow crystals to be corrected for translational disorder (Amos et al., 1982; Henderson et al., 1990; Henderson et al., 1986).

However, cryo-EM of 2D crystals usually produces structures at limited resolutions (typically, above 4.5 Å in the plane of the membrane), where individual amino-acid sidechains, connecting loops, and extramembrane domains are not resolved, although a recent case of a high-resolution structure from cryo-EM analysis of an aquaporin family member is a notable exception (Gonen et al., 2004). Moreover, the resolution in the direction vertical to the lipid bilayer is lower than the in-plane resolution. This reduced resolution entails an uncertainty regarding the actual length of each helical segment, and may obscure the helical register. The lower vertical resolution may also limit the detection of helices that do not span the entire bilayer. In the case of the aquaporin-1 water channel for instance, an initial map at 6Å in-plane

resolution (Walz et al., 1997) did not reveal the surprising architecture of the channel, whereby two half-helices meet midway through the membrane (Fig. 4d), and misleadingly, these half-helices appeared as one. A subsequent cryo-EM map at 4.5 Å resolution uncovered the two half-helices (Mitsuoka et al., 1999), and allowed a combination of sequence-based methods to be used to predict a model structure (de Groot et al., 2000; Heymann and Engel, 2000), which was found to be in agreement with the high-resolution structure (Murata et al., 2000). The initially incorrect interpretation underscores the importance of improving resolution even marginally within the intermediate-resolution range in order to ascertain the general architecture of the protein.

Despite these shortcomings of intermediate-resolution maps, the fact that they provide an overall description of the protein architecture and the approximate packing of TM helices tremendously reduces the degrees of freedom for conformational search and the extent of uncertainty in constructing model structures. In fact, by assuming that ideal α-helices occupy the locations observed in the map, one can limit the conformation search for the backbone positions to identifying the native-state orientation of each helix around its principal axis (M4:Fleishman et al., 2004).

Building on this realization, and using further constraints obtained from multiple-sequence alignments as well as biochemical data, Baldwin et al. pioneered a structure-based modeling approach to derive a first model of the G-protein coupled receptor (GPCR) rhodopsin (Baldwin et al., 1997) based on a structure at 7 Å in-plane resolution (Unger et al., 1997). The essential feature of their modeling was the expectation that evolutionarily conserved amino-acid positions are packed inside the core of the helix bundle, whereas variable residues face the outside. Figure 3 provides an example of a high-resolution structure of a TM protein showing this pattern of

evolutionary conservation. While very rough, the model of rhodopsin's TM domain served as a template for modeling other GPCRs, which then provided a framework for interpreting the effects of mutations in the context of the receptor structure (see e.g. refs. (Latronico et al., 1998; Scheer et al., 2000)). Three years later, a first high-resolution structure of rhodopsin was solved by x-ray crystallography of 3D crystals (Palczewski et al., 2000), and showed that the previous model approximated the native-state structure to within 3.2Å root-mean-square deviation (RMSd). As Figure 5 illustrates, the orientations of all of the helices were predicted quite accurately by Baldwin et al., and the main structural differences are due to deviations in the positioning of the kinked helices.

The successful combination of cryo-EM and computational methods for the modeling of rhodopsin encouraged us to develop tools for modeling based on phylogenetic analysis and intermediate-resolution structures (M4:Fleishman et al., 2004; M5:Fleishman et al., 2004). These tools were subsequently used to model two membrane proteins based on cryo-EM maps, the vertebrate gap junction (M7:Fleishman et al., 2004) and the bacterial small multidrug resistance protein EmrE (M9:Fleishman et al., 2006).

Our principal aim in modeling TM-protein structures is to provide a platform for the experimental study of structure-function relationships (M11:Fleishman et al., 2006). Although the models are approximate, and do not contain sidechain atoms, they can nevertheless be used in order to plan and interpret biochemical experiments. Thus, we have used the model structure of the gap junction TM domain (M7:Fleishman et al., 2004) in order to design experiments that probed the stability of this domain (M11:Fleishman et al., 2006). Our experimental collaborators at Karen Avraham's laboratory (School of Medicine, Tel-Aviv University) then identified two

putative salt bridges and one pair of residues that is involved in packing interactions, in which one disease-causing mutation suppressed the effects of another. These results provide the first experimental data on interactions between residues in the gap junction.

Modeling pairs of tightly packed α-helices

As part of my Master's studies, I developed a methodology for predicting the structures of small systems consisting of pairs of tightly packed α-helices (A1:Fleishman and Ben-Tal, 2002). Each of the TM domains was modeled as an α-helix, and the interactions between a pair of helices were scored according to rules of preferred association that were inferred from biochemical studies (M11:Fleishman et al., 2006). The main features of the score function were a preference for locating small and polar amino-acid residues in the interface of the helix pair and the exclusion of large residues from there. This score function, though simple, was able to predict correctly the conformations of several pairs of TM helices from solved structures (A1:Fleishman and Ben-Tal, 2002). At the start of my doctoral studies I used this methodology in order to study the association of ErbB2 (or HER2) monomers that form dimers in the membrane (M1:Fleishman et al., 2002). ErbB2 is an oncogene that was implicated in 30% of breast cancers. The results provided evidence for a hypothesis on a mechanism of rotation-coupled receptor activation, whereby the TM helices rotate between two different states: an active (low-affinity) and an inactive (high-affinity) state. This suggested conformational change was recently recapitulated jointly with our collaborators using more sophisticated conformational sampling methods (M2:Enosh et al., 2006). In addition, the implications of the rotation-coupled activation of ErbB2 for the cytoplasmic kinase domain were studied (M3:Landau et

al., 2004). Although the conformation score function (A1:Fleishman and Ben-Tal, 2002) was useful in the study of pairs of helices, it has the significant drawback that it assumes that the pairs of helices under study are closely packed (< 9 Å separation between the principal axes of the helices), thus in effect precluding its applicability to most polytopic proteins (A1:Fleishman and Ben-Tal, 2002). This limitation provided the impetus for the development of other tools for structure prediction in TM proteins that were mentioned in the previous section (M6:Enosh et al., 2004; M4:Fleishman et al., 2004; M5:Fleishman et al., 2004).

Organization of the Thesis

This thesis is organized as an article dissertation. Appendix A1 provides a paper that summarizes the results of my Master's studies, during which I developed a method for predicting the orientations of tightly packed α-helices (A1:Fleishman and Ben-Tal, 2002), and is therefore not part of the main body of this thesis. It is, however, the methodological basis for the work, in which I predicted the putative stable conformations and a pathway linking these conformations for the receptor tyrosine kinase (RTK) ErbB2 (M1:Fleishman et al., 2002). This conformational change was also identified in joint work with our collaborators using a more advanced method for sampling conformations (M2:Enosh et al., 2006). Further work in the Ben-Tal laboratory, in which I participated, explored the implications of the suggested conformations of ErbB2 for the activation of the cytoplasmic kinase domain (M3:Landau et al., 2004). As mentioned above, the method for predicting the structures of tightly packed α-helices (A1:Fleishman and Ben-Tal, 2002) is not capable of treating most of the polytopic membrane proteins. I therefore developed two additional methodologies for structure prediction, one of which uses evolutionary

conservation and low-resolution structural data (M4:Fleishman et al., 2004), and the other identifies pairs of amino-acid residues that show a correlated pattern of substitutions in the evolutionary history of the protein family (M5:Fleishman et al., 2004). It is of note that the method that utilizes evolutionary conservation (M4:Fleishman et al., 2004) makes extensive use of methodological aspects that were developed in my Master's thesis and are provided as Appendix A (A1:Fleishman and Ben-Tal, 2002). We also developed a methodology for detecting correlated substitutions in the evolutionary history of protein families, which can be used for predicting whether residue pairs are proximal in space (M5:Fleishman et al., 2004). The paper describing this methodology also summarizes an analysis of the inter-relationships between three of the structural domains of the voltage-gated potassium channel. In another joint work with our collaborators, we developed a new methodology for assigning TM segments to α-helices observed in low-resolution cryo-EM structures based on geometric considerations (M6:Enosh et al., 2004). Based on some of these methodological developments, I next predicted the structure of the gap-junction TM domain (M7:Fleishman et al., 2004). In a paper that has been submitted for publication, but has not yet been published, I used the model of the gap junction TM domain to suggest which residues in this domain interact, and experimental assays by our collaborators in Karen Avraham's laboratory (School of Medicine, Tel-Aviv University) tested these hypotheses (M8:Fleishman et al., 2006). In another unpublished paper, I predicted the structure and mechanism of substrate translocation of the small multidrug resistance antiporter from *E. coli*, EmrE (M9:Fleishman et al., 2006). Another paper (M10:Shental-Bechor et al., 2006) summarizes a critique of an approach to derive an experimental hydrophobicity scale

that was proposed and applied in the beginning of 2005, and was subsequently used to study the voltage-gated potassium channel.

We recently published two reviews on structure prediction in TM proteins. The first of which provides a retrospective analysis of methods that have been used to predict TM-protein structures, and contrasts the predictions with experimental atomic-resolution structures that were subsequently solved, thus identifying weaknesses and strengths of currently employed methods (M11:Fleishman et al., 2006). The second is focused on methodological developments and new insights into TM-protein folding obtained over the past few years, attempting to delineate productive venues for future research (M12:Fleishman and Ben-Tal, 2006). The Introduction of this thesis is based in part on these two reviews.

An additional appendix presents an article that I published during the course of my doctoral studies, but is not directly related to structure prediction providing a critique of a commonly used method for identifying motifs in protein- and gene-interaction networks (A2:Artzy-Randrup et al., 2004).

Figure Legends

Figure 1: **Number of new helical membrane protein folds solved in recent years.** Tremendous progress has been made over the past few years in crystallization of TM proteins from bacteria. However, crystallization of eukaryotic TM proteins still lags far behind, and only a handful of structures has been obtained. Figures 1 and 3-5 are taken from (M11:Fleishman et al., 2006).

Figure 2: **Two stages of TM-protein folding.** TM-protein folding is thought to proceed in two stages (Popot and Engelman, 1990): the folding of individual TM

segments into helices (top) followed by helix packing (bottom). The topology of the protein is often determined by the positive-inside rule (von Heijne and Gavel, 1988), with the cytoplasmic loops tending to be enriched by positively charged residues compared to the extracellular loops. Figure 2 is taken from (M12:Fleishman and Ben-Tal, 2006).

Figure 3: **Evolutionary conservation can aid the orientation of transmembrane helices.** Evolutionary conservation is projected on the bacteriorhodopsin structure viewed from the direction vertical to the membrane plane showing that the core of the protein (within the yellow ellipse) is more conserved than its periphery. The observation of this correlation between evolutionary conservation and structure served as the principal means in this doctoral work for predicting the orientations of helices. An algorithm that identified conformations, in which conserved amino-acid positions are packed inside the protein core and variable residues face the outside, was successfully validated on TM proteins of solved structure (M4:Fleishman et al., 2004). It was then applied to predict the structure of the gap junction intercellular channel (M7:Fleishman et al., 2004) and the small multidrug resistance transporter from *E. coli* EmrE (M9:Fleishman et al., 2006). Conservation was computed using the *ConSurf* webserver (http://consurf.tau.ac.il/) (Glaser et al., 2003). This and all other molecular representations were generated using MolScript (Kraulis, 1991), and rendered with Raster3D (Merritt and Bacon, 1997).

Figure 4: **Recent structures reveal many discrepancies from the view that TM helices are canonical and span the entire lipid bilayer. (a)** For clarity, only three of the four monomers comprising the K+ ion channel are shown (Doyle et al., 1998). Blue cylinders represent the pore helix, which spans only half of the membrane width.

(b) A monomer of the ClC Cl- channel (Dutzler et al., 2002). The blue cylinders represent the locations of helices B and J, which are highly tilted with respect to the membrane normal, comprising approximately 35 amino acids each. **(c)** Structure of the transporter lac permease (Abramson et al., 2003). Some of the helices are kinked. A lactose analog is shown in orange spheres. **(d)** Structure of the aquaporin 1 water channel (Sui et al., 2001). Blue and red cylinders represent two half helices that meet midway in the membrane. Since all of these structural features are still beyond the capabilities of predictive methods, in this doctoral work, I have relied on intermediate-resolution structural data to constrain the search space and to engender more native-like model structures.

Figure 5: **Comparison of the hypothetical and high-resolution structures of rhodopsin.** The crystal and the hypothetical structures of rhodopsin are superimposed (yellow and green, respectively). The hypothetical structure was modeled on the basis on a cryo-EM map at 7 Å in-plane resolution (Baldwin et al., 1997). The hypothetical and x-ray structures deviate by 3.2 Å RMS. Spheres are aids to the eye in identifying identical positions in the hypothetical and crystal structures. The orientations of all of the helices are very similar, and the main differences are in the locations of the helices within the plane of the membrane, particularly in the kinked helices F and G.

Figures

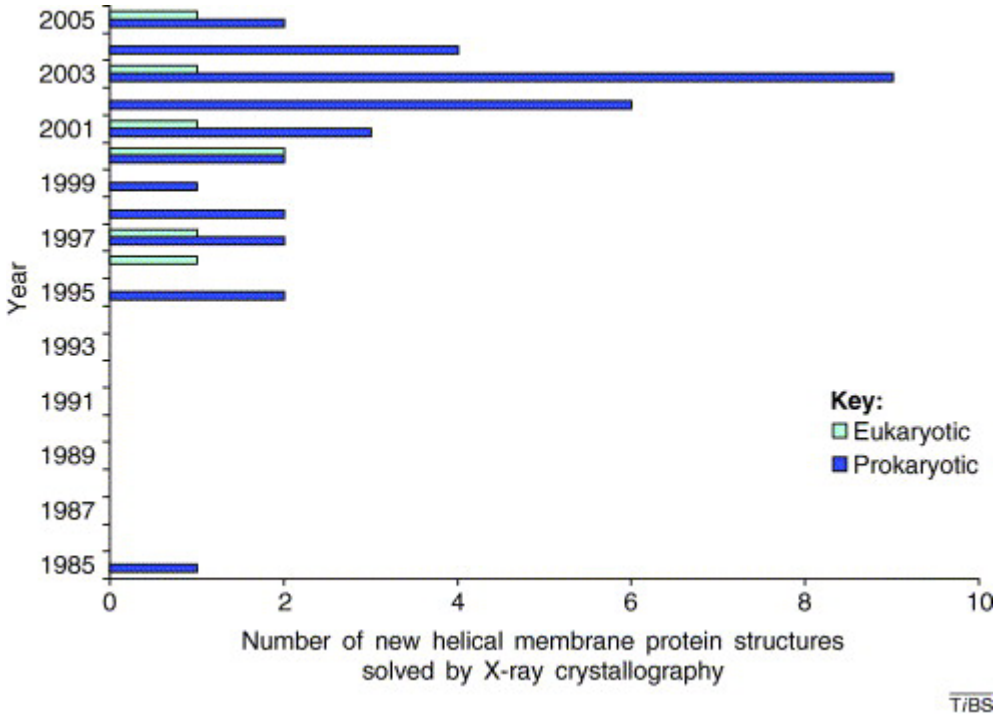


Figure 1

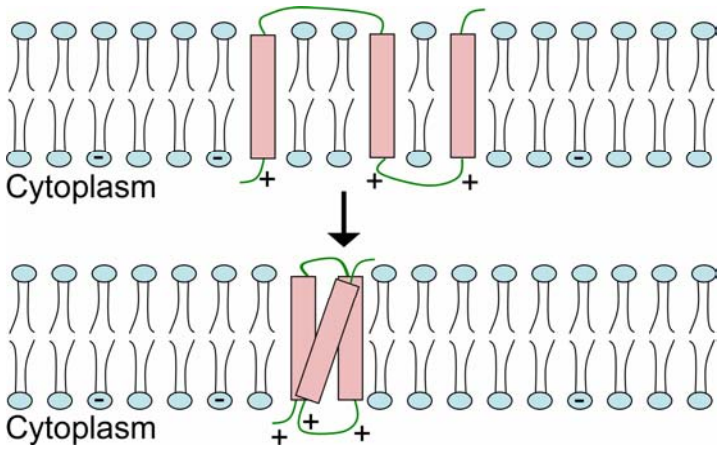


Figure 2

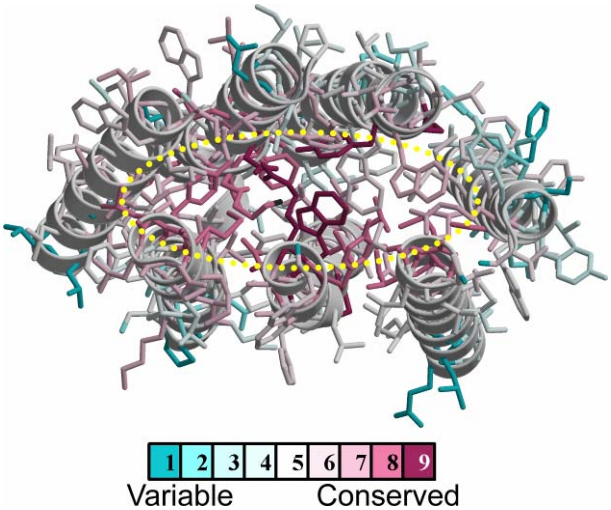

1
2
3
4
5
6
7
8
9
Variable
Conserved

Figure 3

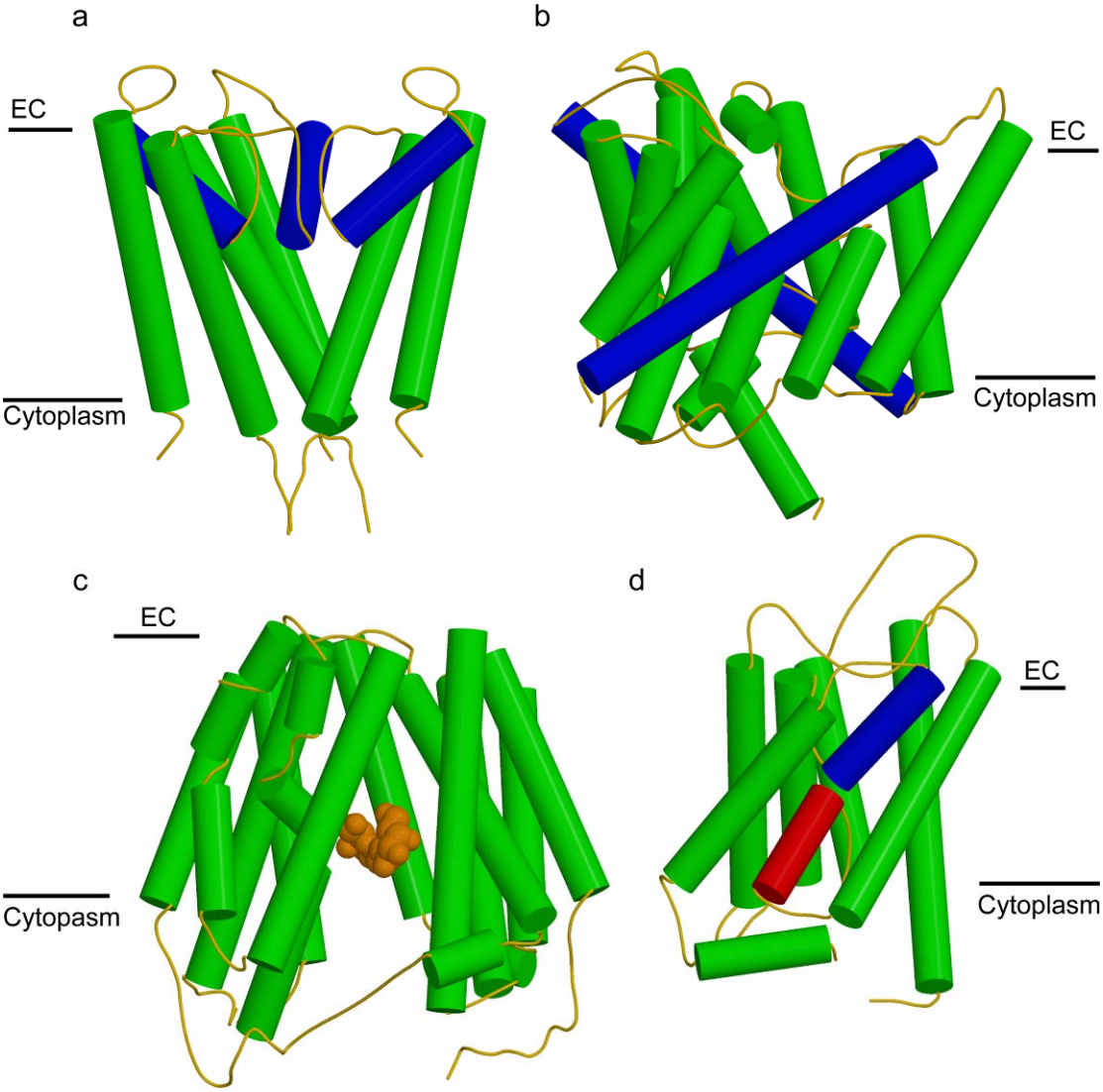

a
EC
Cytoplasm
b
EC
Cytoplasm
c
EC
Cytopasm
d
EC
Cytoplasm

Figure 4

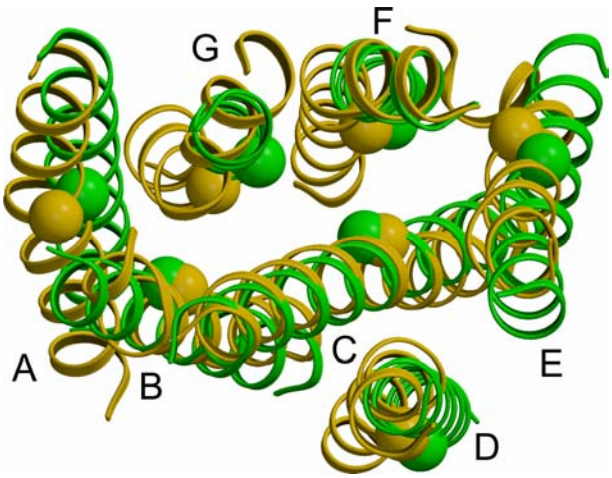


Figure 5

Bibliography

Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., and Iwata, S. (2003). Structure and mechanism of the lactose permease of Escherichia coli. Science *301*, 610-615.
Amos, L. A., Henderson, R., and Unwin, P. N. (1982). Three-dimensional structure determination by electron microscopy of two-dimensional crystals. Prog. Biophys. Mol. Biol. *39*, 183-231.
Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks".[comment]. Science *305*, 1107; author reply 1107.
Baldwin, J. M. (1993). The probable arrangement of the helices in G protein-coupled receptors. EMBO J. *12*, 1693-1703.
Baldwin, J. M., Schertler, G. F., and Unger, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J. Mol. Biol. *272*, 144-164.
Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L.*, et al.* (2004). The Pfam protein families database. Nucleic Acids Res. *32 Database issue*, D138-141.
Bowie, J. U. (1997). Helix packing in membrane proteins. J. Mol. Biol. *272*, 780-789.
Breyton, C., Haase, W., Rapoport, T. A., Kuhlbrandt, W., and Collinson, I. (2002). Three-dimensional structure of the bacterial protein-translocation complex SecYEG. Nature *418*, 662-665.
Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T., and Rees, D. C. (1998). Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. Science *282*, 2220-2226.
Davidson, A. L., and Chen, J. (2005). Structural biology. Flipping lipids: is the third time the charm?[comment]. Science *308*, 963-965.
de Groot, B. L., Heymann, J. B., Engel, A., Mitsuoka, K., Fujiyoshi, Y., and Grubmuller, H. (2000). The fold of human aquaporin 1. J. Mol. Biol. *300*, 987-994.
Deisenhofer, J., Epp, O., Sinning, I., and Michel, H. (1995). Crystallographic refinement at 2.3 A resolution and refined model of the photosynthetic reaction centre from Rhodopseudomonas viridis. J. Mol. Biol. *246*, 429-457.
Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., and MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science *280*, 69-77.
Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T., and MacKinnon, R. (2002). X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. Nature *415*, 287-294.
Engelman, D. M., Chen, Y., Chin, C. N., Curran, A. R., Dixon, A. M., Dupuy, A. D., Lee, A. S., Lehnert, U., Matthews, E. E., Reshetnyak, Y. K.*, et al.* (2003). Membrane protein folding: beyond the two stage model. FEBS Lett. *555*, 122-125.
Enosh, A., Fleishman, S. J., Ben-Tal, N., and Halperin, D. (2004). Assigning transmembrane segments to helices in intermediate-resolution structures. Bioinformatics *20 Suppl 1*, I122-I129.

Enosh, A., Fleishman, S. J., Ben-Tal, N., and Halperin, D. (2006). Prediction and simulation of motion in pairs of transmembrane alpha-helices. Bioinformatics *in press*.
Fleishman, S. J., and Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. J. Mol. Biol. *321*, 363-378.
Fleishman, S. J., and Ben-Tal, N. (2006). Progress in structure prediction of alpha-helical membrane proteins. Curr. Opin. Struc. Biol. *in press*.
Fleishman, S. J., Dagan, T., and Graur, D. (2003). pANT: a method for the pairwise assessment of nonfunctionalization times of processed pseudogenes. Mol Biol Evol *20*, 1876-1880. Epub 2003 Jul 1828.
Fleishman, S. J., Harrington, S., Friesner, R. A., Honig, B., and Ben-Tal, N. (2004). An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data. Biophys. J. *87*, 3448-3459.
Fleishman, S. J., Harrington, S. E., Enosh, A., Halperin, D., Tate, C. G., and Ben-Tal, N. (2006). Cryo-EM-based model structure of the bacterial multidrug transporter EmrE. in preparation.
Fleishman, S. J., Sabag, A. D., Ophir, E., Avraham, K. A., and Ben-Tal, N. (2006). The Structural Context of Disease-causing Mutations in Gap Junctions. in preparation.
Fleishman, S. J., Schlessinger, J., and Ben-Tal, N. (2002). A putative activation switch in the transmembrane domain of erbB2. Proc. Natl. Acad. Sci. USA *99*, 15937-15940.
Fleishman, S. J., Unger, V. M., and Ben-Tal, N. (2006). Transmembrane protein structures without X-rays. Trends Biochem. Sci. *31*, 106-113.
Fleishman, S. J., Unger, V. M., Yeager, M., and Ben-Tal, N. (2004). A C-alpha model for the transmembrane alpha-helices of gap-junction intercellular channels. Mol. Cell *15*, 879-888.
Fleishman, S. J., Yifrach, O., and Ben-Tal, N. (2004). An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J. Mol. Biol. *340*, 307-318.
Fujiyoshi, Y. (1998). The structural study of membrane proteins by electron crystallography. Advances in Biophysics *35*, 25-80.
Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163-164.
Gonen, T., Sliz, P., Kistler, J., Cheng, Y., and Walz, T. (2004). Aquaporin-0 membrane junctions reveal the structure of a closed water pore. Nature *429*, 193-197.
Hasler, L., Heymann, J. B., Engel, A., Kistler, J., and Walz, T. (1998). 2D crystallization of membrane proteins: rationales and examples. J. Struct. Biol. *121*, 162-171.
Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H. (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. J. Mol. Biol. *213*, 899-929.
Henderson, R., Baldwin, J. M., Downing, K. H., Kepault, J., and Zemlin, F. (1986). Structure of purple membrane from Halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5A resolution. Ultramicroscopy *19*, 147-187.
Heymann, J. B., and Engel, A. (2000). Structural clues in the sequences of the aquaporins. J. Mol. Biol. *295*, 1039-1053.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T., and MacKinnon, R. (2003). X-ray structure of a voltage-dependent K+ channel. Nature *423*, 33-41.
Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. *24*, 946-950.
Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567-580.
Kuhlbrandt, W., and Wang, D. N. (1991). Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature *350*, 130-134.
Landau, M., Fleishman, S. J., and Ben-Tal, N. (2004). A Putative Mechanism for Downregulation of the Catalytic Activity of the EGF Receptor via Direct Contact between Its Kinase and C-Terminal Domains. Structure (Camb) *12*, 2265-2275.
Latronico, A. C., Abell, A. N., Arnhold, I. J., Liu, X., Lins, T. S., Brito, V. N., Billerbeck, A. E., Segaloff, D. L., and Mendonca, B. B. (1998). A unique constitutively activating mutation in third transmembrane helix of luteinizing hormone receptor causes sporadic male gonadotropin-independent precocious puberty. Journal of Clinical Endocrinology & Metabolism *83*, 2435-2440.
Lee, S. Y., Lee, A., Chen, J., and MacKinnon, R. (2005). Structure of the KvAP voltage-dependent K+ channel and its dependence on the lipid membrane. Proc. Natl. Acad. Sci. USA *102*, 15441-15446.
Liu, J., and Rost, B. (2001). Comparing function and structure between entire proteomes. Protein Sci. *10*, 1970-1979.
Long, S. B., Campbell, E. B., and Mackinnon, R. (2005). Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. Science *309*, 897-903.
Ma, C., and Chang, G. (2004). Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli. Proc. Natl. Acad. Sci. USA *101*, 2852-2857.
MacKenzie, K. R., Prestegard, J. H., and Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. Science *276*, 131-133.
Magidovich, E., Fleishman, S. J., and Yifrach, O. (2006). Intrinsically disordered C-terminal segments of voltage-activated potassium channels: a possible fishing rod-like mechanism for channel binding to scaffold proteins. Bioinformatics *in press*.
Merritt, E. A., and Bacon, D. J. (1997). Raster3D photorealistic molecular graphics. Methods Enzymol. *277*, 505-524.
Mitaku, S., Ono, M., Hirokawa, T., Boon-Chieng, S., and Sonoyama, M. (1999). Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system. Biophys. Chem. *82*, 165-171.
Mitsuoka, K., Murata, K., Walz, T., Hirai, T., Agre, P., Heymann, J. B., Engel, A., and Fujiyoshi, Y. (1999). The structure of aquaporin-1 at 4.5-A resolution reveals short alpha-helices in the center of the monomer. J. Struct. Biol. *128*, 34-43.
Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B., Engel, A., and Fujiyoshi, Y. (2000). Structural determinants of water permeation through aquaporin-1. Nature *407*, 599-605.
Opella, S. J., and Marassi, F. M. (2004). Structure determination of membrane proteins by NMR spectroscopy. Chem Rev *104*, 3587-3606.
Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E.*, et al.* (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. Science *289*, 739-745.
Popot, J. L., and Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. Biochemistry *29*, 4031-4037.

Pornillos, O., Chen, Y. J., Chen, A. P., and Chang, G. (2005). X-ray structure of the EmrE multidrug transporter in complex with a substrate. Science *310*, 1950-1953.
Rhee, K. H., Morris, E. P., Barber, J., and Kuhlbrandt, W. (1998). Three-dimensional structure of the plant photosystem II reaction centre at 8 A resolution. Nature *396*, 283-286.
Rost, B., Fariselli, P., and Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci. *5*, 1704-1718.
Scheer, A., Costa, T., Fanelli, F., De Benedetti, P. G., Mhaouty-Kodja, S., Abuin, L., Nenniger-Tosato, M., and Cotecchia, S. (2000). Mutational analysis of the highly conserved arginine within the Glu/Asp-Arg-Tyr motif of the alpha(1b)-adrenergic receptor: effects on receptor isomerization and activation. Mol. Pharmacol. *57*, 219-231.
Shental-Bechor, D., Fleishman, S. J., and Ben-Tal, N. (2006). Has the code of protein translocation been broken? Trends Biochem. Sci. *31*, 192-196.
Sui, H., Han, B. G., Lee, J. K., Walian, P., and Jap, B. K. (2001). Structural basis of water-specific transport through the AQP1 water channel. Nature *414*, 872-878.
Tate, C. G. (2006). Comparison of three structures of the multidrug transporter EmrE. Curr Opin Struct Biol *in press*.
Tate, C. G., Ubarretxena-Belandia, I., and Baldwin, J. M. (2003). Conformational changes in the multidrug transporter EmrE associated with substrate binding. J. Mol. Biol. *332*, 229-242.
Ubarretxena-Belandia, I., Baldwin, J. M., Schuldiner, S., and Tate, C. G. (2003). Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer. EMBO J. *22*, 6175-6181.
Unger, V. M., Hargrave, P. A., Baldwin, J. M., and Schertler, G. F. (1997). Arrangement of rhodopsin transmembrane alpha-helices. Nature *389*, 203-206.
Unger, V. M., Kumar, N. M., Gilula, N. B., and Yeager, M. (1999). Three-dimensional structure of a recombinant gap junction membrane channel. Science *283*, 1176-1180.
Unwin, P. N., and Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. J. Mol. Biol. *94*, 425-440.
von Heijne, G., and Gavel, Y. (1988). Topogenic signals in integral membrane proteins. Eur. J. Biochem. *174*, 671-678.
Walz, T., Hirai, T., Murata, K., Heymann, J. B., Mitsuoka, K., Fujiyoshi, Y., Smith, B. L., Agre, P., and Engel, A. (1997). The three-dimensional structure of aquaporin-1. Nature *387*, 624-627.
White, S. H. (2004). The progress of membrane protein structure determination. Protein Sci. *13*, 1948-1949.

A putative molecular-activation switch in the transmembrane domain of erbB2

Sarel J. Fleishman[†], Joseph Schlessinger[‡], and Nir Ben-Tal[†§]

[†]Department of Biochemistry, The George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat-Aviv 69987, Israel; and [‡]Department of Pharmacology, Yale University School of Medicine, New Haven, CT 06520-8066

Contributed by Joseph Schlessinger, October 22, 2002

Overexpression of the receptor tyrosine kinase (RTK) erbB2 (also designated *neu* or HER2) was implicated in causing a variety of human cancers, including mammary and ovarian carcinomas. Ligand-induced receptor dimerization is critical for stimulation of the intrinsic protein tyrosine kinase (PTK) of RTKs. It was therefore proposed that PTK activity is stimulated as a result of the reorientation of the cytoplasmic domains within receptor dimers, leading to transautophosphorylation and stimulation of enzymatic activity. Here, we propose a molecular mechanism for rotation-coupled activation of the erbB2 receptor. Using a computational exploration of conformation space of the transmembrane (TM) segments of an erbB2 homodimer, we found two stable conformations of the TM domain. We suggest that these conformations correspond to the active and inactive states of erbB2, and that the receptor molecules may switch from one conformation to the other without crossing exceedingly unfavorable states. This model provides an explanation for the biochemical and oncogenic properties of erbB2, such as the effects of erbB2 overexpression on kinase activity and cell transformation. Furthermore, the opposing effects of the *neu activating oncogenic point mutation and the Val-655→Ile single-nucleotide polymorphism shown to be linked to reduced risk of breast cancer are explained in terms of shifts in the equilibrium between the active and inactive states of erbB2 *in vivo*.**

The epidermal growth factor-receptor (EGFR) family of receptor tyrosine kinases (erbB1, erbB2, erbB3 and erbB4) plays a critical role in the control of many physiological processes (reviewed in refs. 1–3). Moreover, overexpression of or dysfunction in the activity of EGFR and other members of the family has been implicated in the cause of a variety of human cancers (i.e., lung, brain, mammary, and ovarian). erbB1 and other members of the family are composed of a ligand-binding domain that is connected, via a single transmembrane (TM) helix, to a cytoplasmic domain endowed with intrinsic protein tyrosine kinase (PTK) activity flanked by regulatory sequences that are subject to autophosphorylation and phosphorylation by heterologous protein kinases. Ligand binding to the extracellular domain induces the formation of homo- and heterodimers of different members of the EGFR family, followed by stimulation of PTK activity by transautophosphorylation. In addition to their key regulatory role in the control of PTK activity, tyrosine autophosphorylation sites in RTKs serve as docking sites for recruitment and activation of cellular signaling proteins that mediate the pleiotropic responses induced by growth factor stimulation.

Despite an extensive search over more than a decade, a physiological ligand of erbB2 has not yet been identified (1, 3). It has therefore been proposed that erbB2 does not have a specific ligand, and that it functions as a preferred partner for heterodimerization with other members of the EGFR family (4–6). Indeed, strong activation of the PTK activity of erbB2 was shown to be induced by overexpression of erbB2, even without ligand stimulation (reviewed in ref. 1). Moreover, overexpression and activation of erbB2 have been detected in a large fraction of mammary and ovarian cancers. There is reliable evidence that the TM domain of erbB2 plays an active role in erbB2 dimerization and activation (7–11). A point mutation in the TM domain of the rat homologue *neu* (Val-664→Glu) induces PTK activation and oncogenic transformation (7, 9). The Val-664 residue is located within a consensus sequence in the TM segment's N terminus that is known to induce TM helix dimerization (11). This sequence motif is shared by the TM domains of all members of the EGFR family (Fig. 1). In addition to the N-terminal dimerization motif, erbB2 contains a second related GxxxG motif in the C terminus of its TM segment (12, 13). Each of these motifs mediate dimerization of the TM domain of erbB2 in the cell membrane (14).

In this report, we present a model for the activation of erbB2 that is based on two states of its TM domain. The conformational space of an erbB2 TM homodimer is explored by using a computational tool for predicting conformations of pairs of α-helices in TM domains of membrane proteins (15). The method is based on structural and thermodynamic considerations and consists of an exhaustive search for a structure that is likely to allow a pair of helices to pack tightly. Our computations retrieve empirical results, indicating that the TM domain of erbB2 may undergo dimerization via either one of the two dimerization motifs (14). We further show that receptor dimers are capable of switching between these two conformations. We propose that the balance between the two states may play a role in the control of the activity of erbB2 and its various mutants, both under normal conditions and in pathological states.

Methods

Calculating Scores for Helix-Pair Conformations. A detailed explanation of the method is presented in ref. 15. The essence of the score function consists of two contributions according to the simple rule "small residues go inside:" a negative contribution from residue pairs that form contacts in the given conformation and are known to allow helix pairs to tightly pack in TM proteins; and a positive term for the burial of bulky residues in the helix pair's interface. Thus, a conformation favoring tight packing of helices is expected to have a negative score. Based on the available structural data (16), helices were assumed to be canonical. The interhelical distance was assumed to be $\approx$7.5 Å, corresponding to the interhelical distance in the tightly packed TM homodimer glycophorin A, which has been used as a model for the dimerization of TM domains (17).

Global Search. A global search for an optimal conformation of the erbB2 homodimer was carried out on a five-dimensional lattice (Fig. 2 *Right*). To find the most optimal conformation for the helix dimer, we explored a very large part of the conformation space by modulating x between -10 and 10 Å, with a step size of 0.5 Å; z between -10 and 10 Å, with a step size of 1 Å; α and β from 0° to 360°, with a step size of 9°; and ψ between $-75°$ and 75°, with a step size of 3.75°. We thus examined more than 50 million different conformations of the helix pair.

Abbreviations: EGFR, epidermal growth factor receptor; RTK, receptor tyrosine kinase; TM, transmembrane; PTK, protein tyrosine kinase.

[§]To whom correspondence should be addressed. E-mail: bental@ashtoret.tau.ac.il.

BIOCHEMISTRY

```
erbB1  646 IATGMVGALLLLLVVALGIGLFM 668
erbB2  653 SIVSAVVGILLVVVLGVVFGILI 675
erbB2  653 SIISAVVGILLVVVLGVVFGILI 675
erbB3  644 MALTVIAGLVVIFMMLGGTFL   664
erbB4  653 IAAGVIGGLFILVIVGLTFAVYV 675
```

Fig. 1. Multiple sequence alignment of the TM domains of the human members of the EGFR family. Highlights indicate dimerization motifs in TM domains (12, 14). Yellow corresponds to Sternberg–Gullick motifs (11) and green to motifs that are related to the GxxxG motif (12). The transforming *neu** Val-664→Glu mutation in rats corresponds to a Val-659→Glu mutation in humans (shown in red). All family members except for human erbB3 contain two known dimerization motifs separated by seven positions, thus placing the two motifs on the same ridge of amino acid residues on a model α-helix (18) (Fig. 3). Position 655 in the human erbB2 (blue) exhibits a Val/Ile single-nucleotide polymorphism. The Ile variant is linked to reduced risk of contracting breast cancer (25).

Restricted Search. We also used finer resolution to map the erbB2 homodimer's conformation space by imposing symmetry and restricting the crossing angle Ψ to $-35°$ (Fig. 2 *Right*), which is typical of class 4–4 ridges-into-grooves helix packing (18). α was modulated throughout its potential range with a step size of 5°, and x was modulated between -15 and 15 Å with a step size of 0.5 Å.

Results and Discussion

We conducted a global search of the erbB2 TM homodimer's conformation space without imposing symmetry. We found a conformation, where the C-terminal GxxxG motif mediates dimerization, to have a minimal score in erbB2's TM conformation space. We therefore consider it to be optimal for tight packing of this TM helix pair.

Similar to the GxxxG motif mediating the dimerization of glycophorin A (12, 13, 19), the two dimerization motifs in erbB2 contain two critical residues that are separated by three residues in the amino acid sequence of the TM helix (Fig. 1). It is thus reasonable to assume that interactions between the motifs on two different helices are accommodated by class 4-4 ridges-into-grooves helix packing (18). We therefore conducted a restricted, although higher-resolution, search, assuming that the crossing angle (Ψ) between the two monomers is $-35°$ (Fig. 2 *Right*), a value typical for this class of helix packing (18).

Our results show that the TM domains of an erbB2 homodimer are stable in either of two distinct dimerization modes. These modes correspond to two minima in the score surface shown in Fig. 2 *Left*. The deeper minimum (white ellipse) corresponds to dimerization of the TM domain via the C-terminal dimerization motif, and the shallower minimum (yellow ellipse) corresponds to contact formation via the N-terminal dimerization motif. Notably, the two minima are connected through a saddle-point in the score surface (red ellipse in Fig. 2 *Left*), indicating that the dimer is capable, in theory, of switching between the two dimerization modes without moving through excessively unfavorable conformations. The movement consists of sliding along a ridge formed by amino acid residues (18) and a large 120° rotation of each monomer with respect to the other (Figs. 2 and 3 and Movie 1, which is published as supporting information on the PNAS web site, www.pnas.org).

In light of these results, we propose a molecular-switch model for the activation of erbB2, other members of the EGFR family, and possibly other RTKs. According to the model, the structure of the TM segment in erbB2 allows the receptor dimers to exist

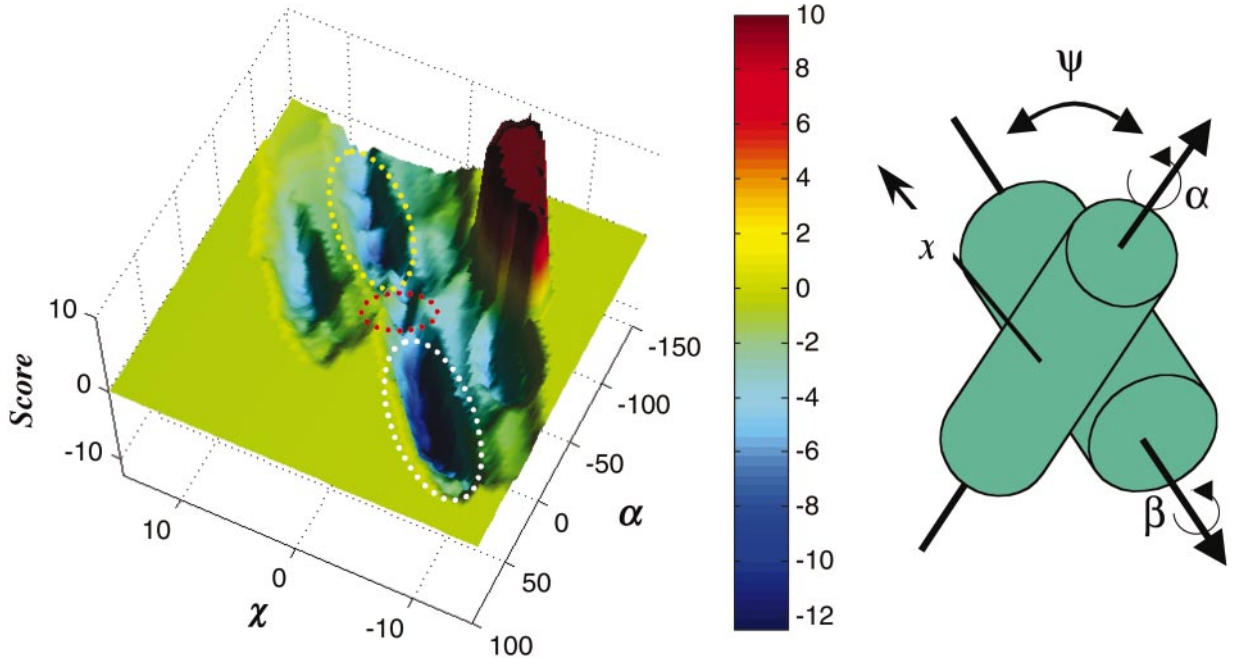


Fig. 2. A score potential surface of a homodimer corresponding to erbB2's TM domain at a crossing angle of $-35°$. (*Left*) Each coordinate on the surface represents a unique conformation of the helix pair. Two minima are colored in deep blue, corresponding to two dimerization modes (14), in which either of the dimerization motifs mediates contact between the TM domains. The deeper minimum (white ellipse) corresponds to conformations where the C-terminal dimerization motif (Fig. 1) mediates contact, whereas the shallower minimum (yellow ellipse) corresponds to conformations where the N-terminal motif mediates dimerization (Fig. 3). The minima are not disconnected (red ellipse), and movement is likely between the two dimerization modes (Movie 1). Score is given in arbitrary units. (*Right*) Different conformations of the helix pair are tested by modulating α and β corresponding to rotations (°) of the monomers around their principal axes, and x to a sliding movement (Å) of one helix across the face of the other. In the global search method, the crossing angle Ψ (°) and z (Å), corresponding to movement across the face of the opposing helix along an axis perpendicular to x (not shown) are also modulated. In the restricted search method symmetry is enforced, so that $\alpha = \beta$ and $z = 0$. Also, Ψ is set to $-35°$, corresponding to a typical crossing angle for helices in the 4–4 class of helix packing (18).

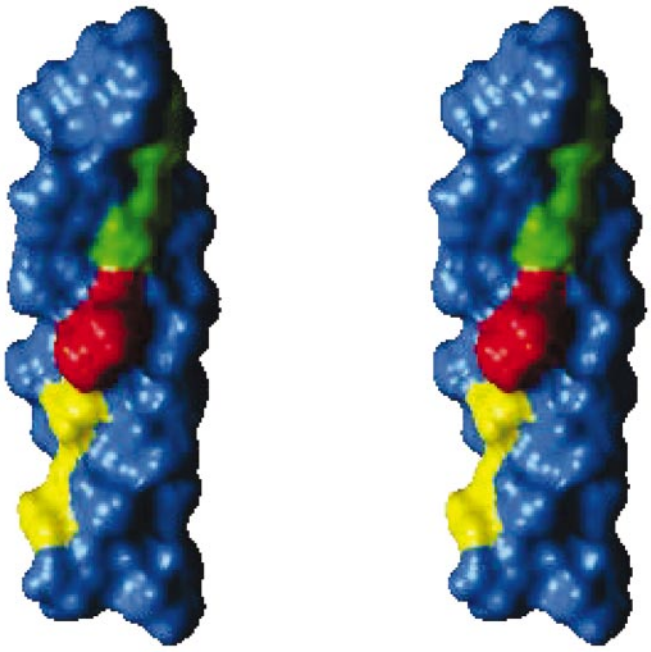


Fig. 3. Stereoview of the ideal α-helix model of the TM domain of erbB2 used for the calculations presented in Fig. 2. Ser-656 and Gly-660 of the N-terminal dimerization motif (11) (Fig. 1) are yellow; Gly-668 and Gly-672 of the C-terminal dimerization motif (12) are green; and Val-664 is red. The monomers pack through either of the two motifs (14). The structural basis that stabilizes the two conformations is that the two motifs form relatively even surfaces on the helical face. Thus they form grooves (18) into which the other monomer may pack. Val-664 (red) is situated between the two motifs on the same ridge (18). Tight packing of this residue in the transition between the two dimerization modes (Movie 1) forms the saddle-point in Fig. 2 *Left*.

in one of two states (Fig. 3). Our calculations show that the dimer mediated by the C-terminal dimerization motif is more stable than the dimer formed by the N-terminal motif (Fig. 2). We propose that the more stable conformation corresponds to the receptor's inactive state, which does not stimulate its PTK activity (20). At normal levels of erbB2 expression, its monomers are at equilibrium with dimers mediated by the more stable C-terminal dimerization motif, i.e., inactive dimers. Transition to the active state is caused by a conformational switch consisting of 120° rotation and movement to dimerization via the N-terminal dimerization motif (Movie 1). According to the model, contact formation via the N-terminal motif of erbB2 causes a reorientation in the cytoplasmic domains of the two juxtaposing catalytic domains of erbB2 (20, 21), resulting in transautophosphorylation and stimulation of the receptor's PTK activity.

The molecular-switch model helps explain the biochemical effects of mutations in the TM domain of erbB2 that were described in the literature (8, 14). Table 1, which is published as supporting information on the PNAS web site, summarizes the effects of various known erbB2 mutations on cell transformation and other properties of erbB2. The various mutations summarized in Table 1 modify either one or both dimerization motifs or leave them intact. Effects on focus formation and dimerization are explained in terms of the molecular-switch model.

It has been proposed that both active and inactive receptor dimers coexist on the cell surface at a normal level of receptor expression (22), and that overexpression increases the amount of active dimers resulting in enhanced PTK activity and cell transformation (1, 3). According to the model presented here, overexpression of erbB2 does not change the overall ratio of active-to-inactive receptor dimers. Rather, it suggests that the enhanced PTK activity, even in cases where an external signal is not received, is a direct consequence of the increase in the absolute number of erbB2 molecules that undergo dimerization via the active (N-terminal) dimerization motif.

The molecular-switch model for erbB2 activation may provide an explanation for seemingly contradictory properties of known erbB2 mutants or naturally occurring variants. It was proposed that the Val-664→Glu mutation in the oncogenic form of *neu*, known as *neu**, facilitates hydrogen-bond formation between neighboring TM domains (10), resulting in enhanced *neu** dimerization and autophosphorylation (23). However, it was recently demonstrated, by using an assay for receptor dimerization, that the activating mutation of *neu** does not enhance receptor dimerization (14). The model presented in this report may be used to explain both phenomena. Dimerization via the C-terminal motif would result in the exposure of a polar group on the side chain of Glu-664 to the hydrophobic lipid environment, which is energetically unfavorable. Thus, the set of inactive dimeric conformations in *neu* (forming contact via the C-terminal dimerization motif) are, in essence, inaccessible to the mutated receptor, which in turn results in a decrease in receptor dimerization (14). However, the amount of active dimers, mediated by the N-terminal dimerization motif, will be increased because of hydrogen bond formation (23), resulting in increased autophosphorylation, PTK activation, and cell transformation.

The model may also explain why the Ile variant of the single-nucleotide polymorphism at position 655 in humans (24) (Fig. 1) exhibits a reduced risk for contracting mammary carcinomas (25). We propose that substitution of Val for a bulkier Ile residue in this position of the TM domain will destabilize the formation of active erbB2 dimers that are mediated by the N-terminal dimerization motif. Consequently, receptor activation caused by overexpression of erbB2 will be reduced even at high levels of erbB2 overexpression. In other words, the activating Val-664→Glu mutation will shift the equilibrium toward the active dimeric form, whereas the Val-655→Ile variant will destabilize the formation of the active dimeric form, resulting in reduced PTK activity, even under conditions of erbB2 overexpression.

Two evolutionary arguments were raised to explain the formation of inactive receptor dimers in the membrane (20). First, formation of inactive receptors on the cell surface would allow more rapid initiation of signal transduction as compared with activation of monomeric receptors that must undergo dimerization for activation to take place. Moreover, the higher stability of inactive compared with active dimers on the cell surface may act as a safe-lock mechanism by decreasing inadvertent dimerization and activation caused by spontaneous collisions between laterally diffusing surface receptors (20).

The mechanism proposed in this report for the activation of erbB2 may apply to other members of the EGFR family and other RTKs (26). It is noteworthy that erbB3 contains only the N-terminal dimerization motif (Fig. 1). Unlike other members of the EGFR family, erbB3 possesses an inactive PTK domain and may serve as a preferred substrate of the other members of the EGFR family (27). Therefore, it may not need the safe-lock mechanism that exists in receptors endowed with active PTK domains. A similar mechanism may also apply for the activation of insulin and insulin-like growth factor (IGF)1 receptors by insulin and IGF1, respectively. The insulin and IGF1 receptors are expressed on the cell surface as disulfide-linked inactive dimers (1, 3). Insulin binding induces a conformational change in the dimeric insulin receptor, resulting in stimulation of the intrinsic PTK activity.

It would be interesting to test and quantify the phenotypic importance of the switching mechanism we have proposed here *in vivo* beyond the documented mutations of Table 1. For instance, mutating the TM segment to the effect that the

N-terminal motif exhibits a greater dimerization propensity than the C-terminal motif (e.g., Ser-656→Gly, Gly-668→Phe double mutant) may have a phenotypic effect similar to the Val-664→Glu substitution in *neu** (Fig. 1) (7, 9). Another interesting possibility would be to leave both motifs intact and to alter the pathway between them, e.g., by deleting residues from the sequence connecting the two motifs (positions 661–667 of erbB2; Fig. 1). Our calculations show that a deletion mutation such as this would disconnect the pathway between the conformations mediated by either of the two dimerization motifs. This mutation would retain the receptor's dimerization characteristics but slow down the kinetics of the switching mechanism (data not shown).

We have used a recently developed computational method for predicting TM helix conformations (15) and combined its results with a large body of past and recent experimental information on the sequence–structure–function relationships of the erbB2 TM domains. Based on these results, we suggested a model for the activation of erbB2 receptors in molecular detail. The model clarifies previously described clinical and biochemical information on erbB2 receptors (Table 1). Finally, targeting of this mechanism by a novel class of lipid soluble inhibitors may offer new therapeutic strategies for cancers caused by overexpression of erbB2.

We acknowledge helpful discussions with Mark A. Lemmon. We thank Sharron Bransburg-Zabary for assistance in producing the accompanying animation. The reported computations were conducted on infrastructure supplied by the Bioinformatics Unit and the Computation Center at Tel Aviv University. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund.

1. Prenzel, N., Fischer, O. M., Streit, S., Hart, S. & Ullrich, A. (2001) *Endocr. Rel. Cancer* **8,** 11–31.
2. Ullrich, A. & Schlessinger, J. (1990) *Cell* **61,** 203–212.
3. Schlessinger, J. (2000) *Cell* **103,** 211–225.
4. Graus-Porta, D., Beerli, R. R. & Hynes, N. E. (1995) *Mol. Cell. Biol.* **15,** 1182–1191.
5. Graus-Porta, D., Beerli, R. R., Daly, J. M. & Hynes, N. E. (1997) *EMBO J.* **16,** 1647–1655.
6. Karunagaran, D., Tzahar, E., Beerli, R. R., Chen, X., Graus-Porta, D., Ratzkin, B. J., Seger, R., Hynes, N. E. & Yarden, Y. (1996) *EMBO J.* **15,** 254–264.
7. Bargmann, C. I., Hung, M. C. & Weinberg, R. A. (1986) *Cell* **45,** 649–657.
8. Cao, H., Bangalore, L., Bormann, B. J. & Stern, D. F. (1992) *EMBO J.* **11,** 923–932.
9. Segatto, O., King, C. R., Pierce, J. H., Di Fiore, P. P. & Aaronson, S. A. (1988) *Mol. Cell. Biol.* **8,** 5570–5574.
10. Sternberg, M. J. & Gullick, W. J. (1989) *Nature* **339,** 587.
11. Sternberg, M. J. & Gullick, W. J. (1990) *Protein Eng.* **3,** 245–248.
12. Russ, W. P. & Engelman, D. M. (2000) *J. Mol. Biol.* **296,** 911–919.
13. Senes, A., Gerstein, M. & Engelman, D. M. (2000) *J. Mol. Biol.* **296,** 921–936.
14. Mendrola, J. M., Berger, M. B., King, M. C. & Lemmon, M. A. (2002) *J. Biol. Chem.* **277,** 4704–4712.
15. Fleishman, S. J. & Ben-Tal, N. (2002) *J. Mol. Biol.* **321,** 363–378.
16. Smith, S. O., Smith, C. S. & Bormann, B. J. (1996) *Nat. Struct. Biol.* **3,** 252–258.
17. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997) *Science* **276,** 131–133.
18. Chothia, C., Levitt, M. & Richardson, D. (1981) *J. Mol. Biol.* **145,** 215–250.
19. Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E. & Engelman, D. M. (1992) *J. Biol. Chem.* **267,** 7683–7689.
20. Jiang, G. & Hunter, T. (1999) *Curr. Biol.* **9,** R568–R571.
21. Bell, C. A., Tynan, J. A., Hart, K. C., Meyer, A. N., Robertson, S. C. & Donoghue, D. J. (2000) *Mol. Biol. Cell* **11,** 3589–3599.
22. Burke, C. L., Lemmon, M. A., Coren, B. A., Engelman, D. M. & Stern, D. F. (1997) *Oncogene* **14,** 687–696.
23. Weiner, D. B., Liu, J., Cohen, J. A., Williams, W. V. & Greene, M. I. (1989) *Nature* **339,** 230–231.
24. Papewalis, J., Nikitin, A. & Rajewsky, M. F. (1991) *Nucleic Acids Res.* **19,** 5452.
25. Xie, D., Shu, X. O., Deng, Z., Wen, W. Q., Creek, K. E., Dai, Q., Gao, Y. T., Jin, F. & Zheng, W. (2000) *J. Natl. Cancer Inst.* **92,** 412–417.
26. Moriki, T., Maruyama, H. & Maruyama, I. N. (2001) *J. Mol. Biol.* **311,** 1011–1026.
27. Carraway, K. L., III, & Cantley, L. C. (1994) *Cell* **78,** 5–8.

BIOINFORMATICS

Vol. 23 ECCB 2006, pages e212–e218
doi:10.1093/bioinformatics/btl325

Prediction and simulation of motion in pairs of transmembrane α-helices

Angela Enosh[1,*], Sarel J. Fleishman[2], Nir Ben-Tal[2] and Dan Halperin[1]

[1]School of Computer Science and [2]Department of Biochemistry Tel Aviv University, Ramat Aviv 69978, Israel

ABSTRACT

Motivation: Motion in transmembrane (TM) proteins plays an essential role in a variety of biological phenomena. Thus, developing an automated method for predicting and simulating motion in this class of proteins should result in an increased level of understanding of crucial physiological mechanisms. We have developed an algorithm for predicting and simulating motion in TM proteins of the α-helix bundle type. Our method employs probabilistic motion-planning techniques to suggest possible collision-free motion paths. The resulting paths are ranked according to the quality of the van der Waals interactions between the TM helices. Our algorithm considers a wide range of degrees of freedom (dofs) involved in the motion, including external and internal moves. However, in order to handle the vast dimensionality of the problem, we employ some constraints on these dofs in a way that is unlikely to rule out the native motion of the protein. Our algorithm simulates the motion, including all the dofs, and automatically produces a movie that demonstrates it.

Results: Overexpression of the RTK ErbB2 was implicated in causing a variety of human cancers. Recently, a molecular mechanism for rotation-coupled activation of the receptor was suggested. We applied our algorithm to investigate the TM domain of this protein, and compared our results with this mechanism. A motion pathway that was similar to the proposed mechanism ranked first, and motions with partial overlap to this pathway followed in rank order. In addition, we conducted a negative-control computational-experiment using Glycophorin A. Our results confirmed the immobility of this TM protein, resulting in degenerate paths comprising native-like conformations.

Supplementary information: Supplementary data are available at http://www.cs.tau.ac.il/~angela/EGFR.html

Contact: angela@post.tau.ac.il

1 INTRODUCTION

In total, approximately 20–30% of proteins encoded by the genome are transmembrane (TM). They form pumps and channels that control and guide the transportation of ions and metabolites across the membrane. Other TM proteins function as receptors and are responsible for molecular recognition of hormones and neurotransmitters. Despite recent advances, it is extremely difficult to crystallize these proteins, and even when a high-resolution structure is determined, much effort is required to elucidate the protein's mechanism of action. So far, cartoon-resolution mechanisms have been suggested for only a few TM-proteins, e.g. the lactose permease (Abramson *et al*., 2003) and ErbB2 (Fleishman *et al*., 2002). However, molecular details for these mechanisms are not defined yet. These molecular details include, for instance, the following questions: What exactly are the conformational changes that occur in each step along the reaction coordinate? Whether, and to what extent do the helices move as rigid bodies? Which torsion angles and side-chains alter during the conformational change? Thus, one of the challenging tasks in computational studies of TM-protein structures is to define these molecular details as continuous motion that goes beyond the cartoon-level resolution published so far in order to gain insight into these mechanisms.

Proteins display a broad range of motions, from the fast and localized motions (e.g. side-chain movements) to the slow large-scale motions (e.g. domain movements). An important characteristic of biomolecules is that the different types of motion are interdependent and coupled to one another. Thus, in the investigation of slow large-scale motions as we propose to find, ignoring the fast small-scale motions might obscure the overall conformational changes.

Many large-scale motions take place on time scales beyond the accessibility of time-dependent methods, such as molecular dynamics (MD) (Karplus *et al*., 2002). Normal-mode analysis (NMA), the Gaussian Network Model (GNM) and the Anisotropic Network Model (ANM) (Bahar *et al*., 2005) are fast time-independent methods used for computing vibrational modes and estimating the flexibility of the protein. However, these techniques are not ideally suited to deal with energy barriers and multiple minima in the potential-energy surface. Monte Carlo simulations provide a useful alternative, but to the best of our knowledge, they were not used to study large-scale motions in TM proteins.

Motion planning (MP) is a fundamental problem, originally studied in robotics and computational geometry, but with implications in numerous other fields (Latombe, 1991, 1999; Sharir, 2004). The MP problem can be stated as follows: given a robot in an environment with obstacles, find a collision-free path connecting the current (start) configuration of the robot to a desired (goal) configuration. A class of randomized-path planning methods, known as Probabilistic Road Map (PRM) methods have been successfully applied to complicated high-dimensional problems (Kavraki *et al*., 1996; Hsu *et al*., 1999; Choset *et al*., 2005). PRM techniques sample the robot's configuration space at random, and retain the collision-free samples as milestones. Then, pairs of milestones are connected with local paths that serve as collision-free connectors of the generated milestones. The result is an undirected graph, called a probabilistic roadmap, whose nodes are the milestones and the edges are the local paths.

A distinction exists between multi-query strategies (e.g. Kavraki *et al*., 1999) and single-query ones (e.g. Hsu *et al*., 1999). In a single-query strategy the goal is typically to find a collision-free path between the two query configurations by exploring as little space as possible. Single-query strategies often build a new road map for each query by growing trees of sampled milestones rooted at the initial and goal configurations (Hsu *et al*., 1999). Rapidly-exploring Random Trees (RRT) (LaValle *et al*., 2001;

*To whom correspondence should be addressed.

LaValle, 2006), briefly described in Section 3.1, have been recognized as a very useful tool for designing efficient single-query paths in highly constrained spaces.

Probabilistic techniques combined with optimization and clustering have been used to sample conformational spaces of ligands and identify their low-energy conformations (Finn *et al*., 1996). Randomized path-planning methods were used successfully in computational biology by replacing the collision detection, used in robotic applications, with a molecular force field. Singh *et al*. (1999) applied PRM techniques to the ligand-binding problem. Apaydin *et al*. (2001) and Amato *et al*. (2003) applied PRM techniques to study protein folding. Recently, Cortes *et al*. (2005) developed an algorithm to compute large-amplitude motions in flexible molecular models. They applied RRTs to compute protein loop conformational changes and ligand trajectories.

We extend the RRT framework to predict TM α-helix bundle motions and the conformational changes of the helices in the bundle. Eukaryotic TM proteins form predominantly α-helix bundles in the membrane. Considering the α-helices as rigid bodies may reduce the conformational space substantially. However, owing to the large spectrum of motion scales, we do not assume that the helices are completely rigid. Therefore, in addition to movements of the helices as rigid bodies in three-dimensional (3D) space, we consider also changes in torsion angles and side-chain flexibility within these helices, while using constraints on these degrees of freedom (dofs) in a way that the conformational space will not exceed reasonable computational limits. Our algorithm is divided into two main stages. The first stage filters out many infeasible pathways using purely geometric considerations resulting in collision-free paths. In the second stage, these paths are analyzed using an energy-based criterion. The direct output of the algorithm is several movies that simulate the feasible paths that can be further examined, while taking into account functional data on the protein under study.

We tested the effectiveness of the algorithm with an application to the receptor tyrosine kinase (RTK) ErbB2 and Glycophorin A. Our results comply with previous data on these proteins. It is encouraging to note that motion paths for ErbB2 suggested by our algorithm are similar to the mechanism proposed by Fleishman *et al*. (2002) although we used very different methods to suggest and simulate the motion path.

2 A TM PROTEIN MODEL

A protein can be described as a long linkage with side-chains attached to the C_α atoms on its backbone. Using a standard modeling assumption for proteins, bond lengths and angles are often treated as fixed during motion. However, torsion angles can change significantly when the protein's conformation changes. Thus, in our model, a protein is considered as an articulated mechanism with revolute joints corresponding to the torsion angles along the protein backbone.

TM proteins of the α-helix bundle type comprise helices that are embedded in the membrane. Although helices are often considered as rigid bodies, for motion prediction purposes we cannot treat them as entirely rigid. Thus, when moving from one conformation to another, there might be slight changes in the (ϕ, ψ) torsion angles of amino acids in the helices. We model a helix as a kinematic chain using the chain tree hierarchy introduced by Lotan *et al*. (2004). In the chain tree hierarchy, the rotatable bonds, around which the (ϕ, ψ) torsion angles are defined, cut the protein backbone into rigid groups of atoms, called links. There are two types of links. The first includes the C_{i-1}, O_{i-1} and N_i atoms, where i stands for the position of amino acids along the protein backbone. The second group includes $C\alpha_i$ and all side-chain atoms attached to it (Fig. 1). A reference frame is attached to each link in the chain and the relative location of consecutive frames is defined by a homogeneous transformation matrix, which is a function of the torsion-angle between them. As the conformations of a helix change, the update of the torsion angles of its backbone is done quickly by updating the matrices corresponding to these torsion angles instead of updating the Cartesian coordinates of the atoms. Collision detection with R rigid links, takes $O(R^{\frac{4}{3}})$ time, which is not optimal in the worst case, but performs well in practice.

The algorithm of Lotan *et al*. (2004) assumes that the side-chains are rigid, whereas in our implementation, under some criteria (as explained below), we do allow side-chains to move.

2.1 Structural constraints

On the one hand, one of the driving forces behind motion in TM proteins is to keep the helices tight together in a way that the interactions between these helices do not decrease dramatically. On the other hand, the helices cannot pack so closely as to generate steric clashes between atoms. A steric clash occurs, when the distance between the centers of two non-bonded atoms is significantly smaller than the sum of these atoms' van der Waals (vdW) radii. We partly allow penetration between atoms using a cutoff parameter $\mathcal{K}$, which is the percentage of the vdW radii, namely the centers of two non-bonded atoms of vdW radii r_1 and r_2 must be at least $\mathcal{K}(r_1{+}r_2)$ apart. For our experiments, we used $\mathcal{K} = 60\%$. Thus, a fine combination of the two contradicting forces, tightness and steric-clash avoidance, is considered in our model.

2.2 Problem statement

Given a set of helices represented as kinematic chains and an initial spatial conformation of these helices, we aim to find a feasible motion path (or paths) that simulate the native motion towards goal conformations (that may not be given in advance). We denote the set of n TM helices by $\{h_1 \ldots h_n\}$. Each helix has six dofs corresponding to its position and orientation.

2.3 Relaxations applied to the TM helices

If a helix h_i has m_i torsion angles, the dimensionality of the configuration space in our problem is enormous with $6n + \sum_{i=1}^{n} (m_i - 1)$ dofs, where n is the number of helices. In addition, we consider side-chain flexibility, leading to more dofs. However, we may use some relaxations on the dimensionality of the problem when considering TM helices. The relaxations we use are as follows: (1) The TM helices cannot be fully buried in the membrane and therefore their axes are limited to maximal tilt angles of 50° with respect to the membrane normal. (2) The lateral movements of the helices as a group in the membrane is not considered by our motion analysis, implying that a specific rigid link of one helix can be placed at a fixed location in 3D. (3) Canonical helices have (ϕ=−60, ψ=−40) torsion angles along the backbone. Since we want to limit helix distortion, we allow each angle to deviate by less than ±10° from torsion angles of a canonical helix. (4) Side-chain movements may be important players in the motion-prediction problem. However, for the purposes

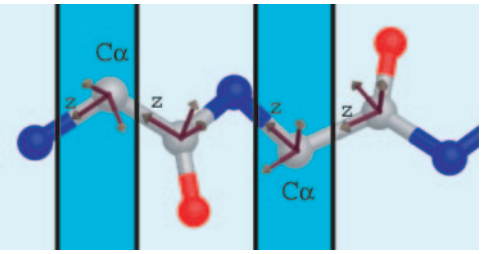


Fig. 1. The backbone degrees of freedom represented on a diglycine peptide. The two-color background shows the partition of the atoms into links. Reference frames are attached to each link origin at the $C\alpha$ and C atoms of the backbone. The z-axis of each frame is the vector along the rotatable bond; the other two axes complete the frame to form an orthogonal right-hand coordinate system.

of obtaining an approximation of the large-scale motions of the protein, it seems reasonable to consider side-chain movements only when they interfere with the way to a desired conformation. Thus, each time we derive motion from one conformation to another, we allow movements only in side-chains that are in conflict with this motion.

3 THE ALGORITHM

We have developed a motion-planning algorithm to predict motion in TM α-helix bundles. For a set of TM helices in 3D space, a conformation of an α-helix bundle comprises all the geometric information related to these helices, namely, the six dofs of helix positions and orientations in 3D space, the torsion angles of each amino acid and the conformations of the side-chains within these helices. The conformation space, C_{space}, is the union of all these possible conformations. C_{space} is divided into feasible, C_{feasible}, and forbidden, C_{forbid}, regions. C_{forbid} contains all the conformations that involve steric clashes between atoms (both within and between helices). In addition, C_{forbid} contains conformations that involve low vdW interactions between the helices. C_{feasible} is simply $C_{\text{space}} \backslash C_{\text{forbid}}$.

Our algorithm proceeds in two stages: Growing RRT—construction of a tree (RRT) that contains the set of feasible collision-free pathways emerging from a given initial conformation, using the constraints described in Section 2.1 applied to the TM helices. This stage is followed by Energy Analysis—assigning weights to the generated nodes and edges in the RRT, corresponding to the energy of a conformation (see Section 3.2 for details) and the energy associated with the move from one conformation to another, respectively. The rationale behind this division is that the first stage uses purely geometric terms to efficiently filter out unlikely pathways and reduces the search space on which the more intricate energy analysis should be applied. Following the two-stage algorithm, several weighted RRTs are built and clustering is performed on the emerging pathways. The energetically favorable pathways are chosen to produce movies.

3.1 Growing RRT

In its general form, the RRT algorithm is based on growing a conformation-space tree $\mathcal{T}$ rooted at the initial conformation q_{init}. $\mathcal{T}$ is incrementally grown to efficiently explore the feasible conformation space in order to find a feasible path connecting q_{init} to a goal conformation. In each iteration, a random conformation, q_{rand}, is generated and the nearest node, q_{near}, in $\mathcal{T}$ (according to some appropriate distance metric M) is expanded towards q_{rand}. If no collision is found on the way towards the random conformation, then q_{rand} becomes a new vertex in the tree and an edge is added between q_{near} and q_{rand}. Otherwise, q_{near} expands as close as possible towards q_{rand}. In this case, the last feasible conformation (unless it is too close to q_{near}) becomes a vertex in $\mathcal{T}$ and an edge is added between q_{near} and the new vertex (Fig. 2). It was shown (LaValle *et al*., 2001) that this method leads to Voronoi-biased growth of $\mathcal{T}$. This means that vertices with large Voronoi cells[1] have a larger probability of being extended. This is a useful property as large Voronoi cells represent unexplored areas of the conformation space.

In our implementation, each node in the tree represents an α-helix bundle conformation. In the beginning, the tree contains a given initial conformation q_{init}. During the expansion process, new conformations are sampled uniformly at random while satisfying the relaxations stated in Section 2.3. While growing an edge from q_{near} towards q_{rand} a forbidden conformation, q_{forbid}, may occur. q_{forbid} is either a conformation with steric clashes, or it contains highly remote helices, i.e. the distance between the helix axes are above a given cutoff $\mathcal{D}$ (we use $\mathcal{D} = 14$ Å in the experiments reported below). In the latter case the expansion is stopped and the algorithm continues as usual. However, when collision between side chains occurs during the expansion toward the sampled conformation, the algorithm tries to adopt a new conformation only for the colliding side-chains that obstruct the way to q_{rand}, in a way that the adopted conformation will be free of collisions. In case of a success, q_{near} continues to expand towards q_{rand}. Otherwise, a new node is generated for the last feasible conformation that was found.

Using the chain-tree hierarchy, the colliding side-chain can easily be detected and examined. We employed a fairly simple procedure that finds the set of collision free rotamers using the backbone-dependent rotamer library from Dunbrack *et al*. (1994), considering rotamers in the range $[-50, -70]$ for ϕ and $[-30, -50]$ for ψ. The backbone-dependent rotamer library evaluates each rotamer by a probability term. Our algorithm preferentially selects high-probability rotamers, while keeping the conformation free of clashes. This step can be computationally expensive, but the number of colliding side-chains in each iteration is relatively small. The algorithm continues to grow the tree till a stopping criterion is fulfilled. In our algorithm, the stopping criterion is reached if novel conformations are not added to the tree after several iterations. In other words, if the algorithm fails to expand $\mathcal{T}$ for a threshold number of consecutive iterations, it implies that the sampled conformations in $\mathcal{T}$ cover C_{feasible} sufficiently, and the expansion of $\mathcal{T}$ is stopped.

When a goal conformation is given, RRT strategies often try to grow two trees rooted at the initial and goal conformations (LaValle, 2006). However, we anticipate that, owing to the paucity of structural information regarding TM proteins, we may often encounter a case whereby only one conformation is known, and so a goal conformation is unavailable. Therefore, after the generation of the tree, our

[1] A Voronoi cell of a vertex v is the set of all points in space that are closer to v than to any other vertex, under the given metric.

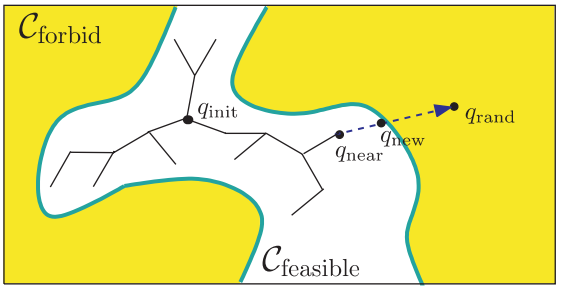


Fig. 2. Expansion of $\mathcal{T}$ using an RRT-based algorithm. The edge from q_{near} travels toward q_{rand} up to the boundary of the C_{forbid} region.

algorithm suggests a goal conformation as well as the path that leads to it.

3.2 Energy analysis

So far, we have considered only geometric constraints imposed on the motion of TM helices, resulting in a tree with collision-free paths. Our next goal is to incorporate energetic considerations into the generation of the tree. It has been suggested that tight packing of α-helices in TM proteins plays a considerable role in stabilizing these proteins (Curran and Engelman, 2003), implying that vdW forces are important descriptors of inter-helix interactions. We calculated the vdW interactions between the helices using the Lennard-Jones (LJ) 6–12 potential. The vdW energy of an α-helix bundle conformation was calculated as

$$E_{\text{vdW}} = \sum_{i>j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \qquad (1)$$

where r_{ij} is the distance between atoms i and j, ϵ_{ij} is the energy-well depth and σ_{ij} is the atomic radii sums. The parameters were taken from CHARMM19 (Neria *et al*., 1996). Thus, a weight was assigned to each node in $\mathcal{T}$, based on the LJ potential of its respective conformation. In the same manner, we added a penalty-weight to each edge between two conformations that corresponds to the maximal LJ potential observed along the local path between them.

Given a weighted RRT, we wish to find paths that minimize the weights along the pathway, and more importantly, lead to a goal conformation that is associated with a low value of the potential. We rely on a common assumption that a pathway may have some energetically unfavorable conformations that may lead to a more favorable conformation, and our aim is to capture these goal conformations. We define two different energy functions for each path: a pathway function $\mathcal{P}$ that equals to the highest value of the potential that is observed along the nodes and edges in the pathway, and a goal function $\mathcal{G}$ that corresponds to the value of the potential of the last conformation in the path, which we refer to as the goal conformation. Formally, for a path $\pi=\{v_0, e_0, v_1, e_1 \ldots e_{k-1}, v_k\}$, where v_i stands for a node and e_j for an edge, $\mathcal{P}(\pi)= \max_{0\le i\le k, 0\le j\le k-1}\{\mathcal{W}(v_i),\mathcal{W}(e_j)\}$, where $\mathcal{W}$ is the weight of the nodes or edges in $\mathcal{T}$, and $\mathcal{G}(\pi) = \mathcal{W}(v_k)$.

3.2.1 Path clustering Different sequences of randomly sampled conformations lead to different trees (RRTs). Thus, instead of growing one tree, several RRTs have been grown in the same way as described in Section 3.1, and clustering is performed on the paths derived from these trees. Each cluster comprises a set of paths that end with the same goal conformation [i.e. the root-mean-square deviation (rmsd) between the atoms of any two goal conformations in a cluster is below a predefined cutoff $\mathcal{Q}$; in our experiments we use $\mathcal{Q}$=1.4 Å]. For a cluster $C_j = \{\pi_1, \ldots, \pi_m\}$, a representative path π^* was chosen to be the one that minimizes the LJ potential in the conformations stored on the path edges and nodes, i.e. $\mathcal{P}(\pi^*) = \min_{1\le i\le m}\{\mathcal{P}(\pi_i)\}$. Different paths may comprise different lengths (number of nodes in the path), still, the above criterion (minimizing $\mathcal{P}$) is more dominant than the path lengths. However, if several paths in a cluster had the same values $\mathcal{P}(\pi^*)$, then the representative path was chosen to be the shortest path among them.

Clusters with a goal conformation that is close to the initial conformation were ignored. A score was assigned to the remaining clusters based on the LJ potential of the goal conformation $\mathcal{G}(\pi^*)$ and the number of paths in the cluster. We integrated the two terms into a form of the colony function (Xiang *et al*., 2002). Thus, the score of a cluster is $\mathcal{F}(C_j) = \sum_{\pi_i \epsilon C_j} e^{-\mathcal{G}(\pi_i)}$. In other words, the score favors clusters comprising many paths leading to a mutual energetically favorable conformation. The representative paths of the highest-score clusters were selected to produce movies that simulate the motion of the TM helices.

4 RESULTS

To explore the utility of the motion-planning algorithm in suggesting possible pathways for conformational changes in proteins, we used it to investigate the TM domain of the RTK ErbB2, overexpression of which has been implicated in many types of cancer [reviewed in Burgess *et al*. (2003)]. The protein, which is a member of the epidermal growth factor-receptor (EGFR) family, includes large extra- and intra-cellular domains that are connected by a single TM helix. It is known to form homo- and heterodimers with other EGFRs. It was proposed that ErbB2 activation involves a rotation in the relative orientation of the cytoplasmic kinase domains within a receptor dimer that is driven by a rotation of the TM helices (Jiang *et al*., 1999). A molecular mechanism for such rotation-coupled activation was suggested based on a computational exploration of conformations of the ErbB2 TM domain (Fleishman *et al*., 2002), yielding two symmetrical, and apparently stable, conformations. The more stable of the two conformations, involved packing of the helices with Gly668 and Gly672 on consecutive helical turns, invoking the Gly-xxx-Gly sequence motif (Curran and Engelman, 2003), at the inter-helix interface. In the less stable conformation, the interface was composed of Ser656 and Gly660 residues on consecutive turns. Based on these calculations it was suggested that activation of the ErbB2 receptor involves rotation of the helices within the TM domain in switching between these two conformations (Fleishman *et al*., 2002), in harmony with the proposition of rotation-coupled activation (Jiang and Hunter, 1999).

The aforementioned computations that served as the basis for suggesting a molecular model for rotation-coupled activation of ErbB2 (Fleishman *et al*., 2002) used a drastically simplified representation of the helices, which comprised solely C_α atoms forming canonical α-helices. To test the feasibility of the suggested molecular mechanism in a more realistic context, we used the method presented in this paper starting from the stable conformation involving the Gly668 and Gly672 residues. Two peptides, each of

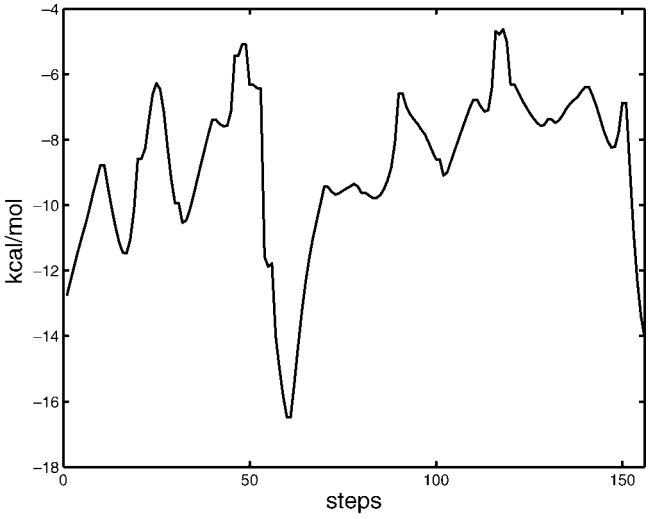


Fig. 3. The LJ potential curve of the conformations along the motion pathway of ErbB2. The curve shows the energy of the preferred pathway according to the colony energy function (Section 3.2). Step 0 corresponds to the initial conformation where the helices were packed via the glycine residues in positions 668 and 672, whereas step 156 corresponds to the goal conformation where the helices interacted through Ser656 and Gly660. The energy minimum in step 60 refers to packing via the Gly668-xxx-Gly672 motif in a conformation that is energetically more favorable than the initial conformation. As expected, it was assigned a lower potential than in step 156, suggesting that packing via Gly668-xxx-Gly672 is more stable than via Ser656-xxx-Gly660 motif as suggested previously (Fleishman *et al*., 2002).

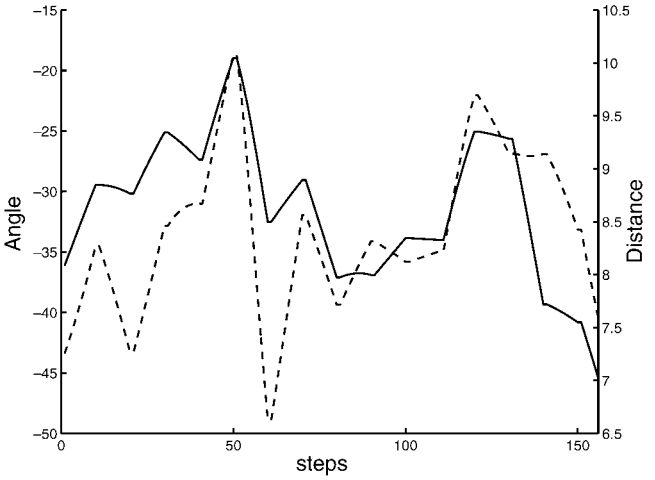


Fig. 4. Crossing angles (°) and interaxial distance (Å) between the helices axes along the most favorable motion pathway simulating the motion in the ErbB2 homodimer. Crossing angles are marked by the continuous curve whereas interaxial distances are marked by the dashed curve. Step 0 corresponds to the initial conformation where the helices were packed via the glycine residues in positions 668 and 672, whereas step 156 corresponds to the goal conformation where the helices interacted through the Ser656-xxx-Gly660 motif.

which corresponds to the TM domain of ErbB2 [LTSIVSAVV-GILLVVVLGVVFGILI], were built as canonical α-helices. They were assembled in a structure that resembled the stable conformation, and side-chains were added to the structure using the SCWRL software (Canutescu *et al*., 2003). Each atom was assigned a vdW radius according to the CHARMM19 forcefield (Neria *et al*., 1996), and the conformational space (external and internal dofs) was explored using the RRT procedure, subjected to two opposing constraints on the distance between the helices. The first was self avoidance: vdW clashes between atoms were not allowed beyond 40% overlap between their radii (i.e. $\mathcal{K} = 60\%$, Section 2.1). An opposing constraint was imposed on the maximal distance between the helices: conformations in which the LJ potential was above a pre-defined cutoff of -5 kcal/mol were excluded. The cutoff value was empirically found to facilitate an efficient exploration of the conformational space. It was the lowest cutoff that yielded motion pathways, i.e. a cutoff value of -6 kcal/mol resulted in paths comprising conformations in the vicinity of the initial state only, and larger values of up to -2 kcal/mol gave similar pathways to those using the -5 kcal/mol cutoff, but also sampled many irrelevant conformations, in which the helices formed little if any contact with one another. We also tried other measures of the helix tightness instead of the LJ potential. For example, each conformation was ranked by the buried-surface area of the helices (calculated with a probe sphere of 1.4 Å) or the number of pairs of atoms that were in contact. The resulting pathways were similar to those obtained by the LJ potential (data not shown), implying that the method is quite robust to the choice of energy function.

A homodimer, such as the ErbB2 TM domain simulated here, is expected to show some degree of symmetry in its conformations. To verify that our implementation retrieves this tendency towards symmetric conformations, we did not impose symmetry on the helices. Nevertheless, the resulting pathways showed that the two helices were symmetry-related throughout all of the simulations. In fact, superimposition of one helix over the other, using a rotation of π radians around the axis of symmetry of the helices' principal axes[2], gave a mean rmsd of 0.57Å (Supplementary Material, Fig. 6). These results encouraged us to impose symmetry on all dofs during the exploration of the conformational space, resulting in a reduction of the number of dofs.

Starting from the initial conformation of the helices, 10 random trees were generated, each of which contained $\sim$320 nodes, i.e. conformations. The conformations were clustered based on the rmsd between the α carbons, and 29 different clusters were found. The next step was to rank the clusters according to their stability. Two different criteria, the total number of conformations in each cluster and the value of the potential of the goal conformation in each cluster, were used to this end. A cluster that contained 79 conformations was ranked first by the colony function (Section 3.2). Encouragingly, the representative conformation of this cluster corresponded to the less stable conformation suggested by Fleishman *et al*. (2002). Each of the pathways was assigned a feasibility score as described in Section 3.2, and the pathway that was assigned the best score was presented in the movie (Supplementary Material, Movie 1). The optimal pathway was composed of a sequence of the most stable conformations. This is in analogy to the path of minimum energy in chemical kinetics. Other

[2]For the two axes ℓ_1 and ℓ_2 of the helices, we choose an axis of symmetry, namely a line ℓ such that rotation of π radians around ℓ will align ℓ_1 with ℓ_2. Further details can be found in the Supplementary Material.

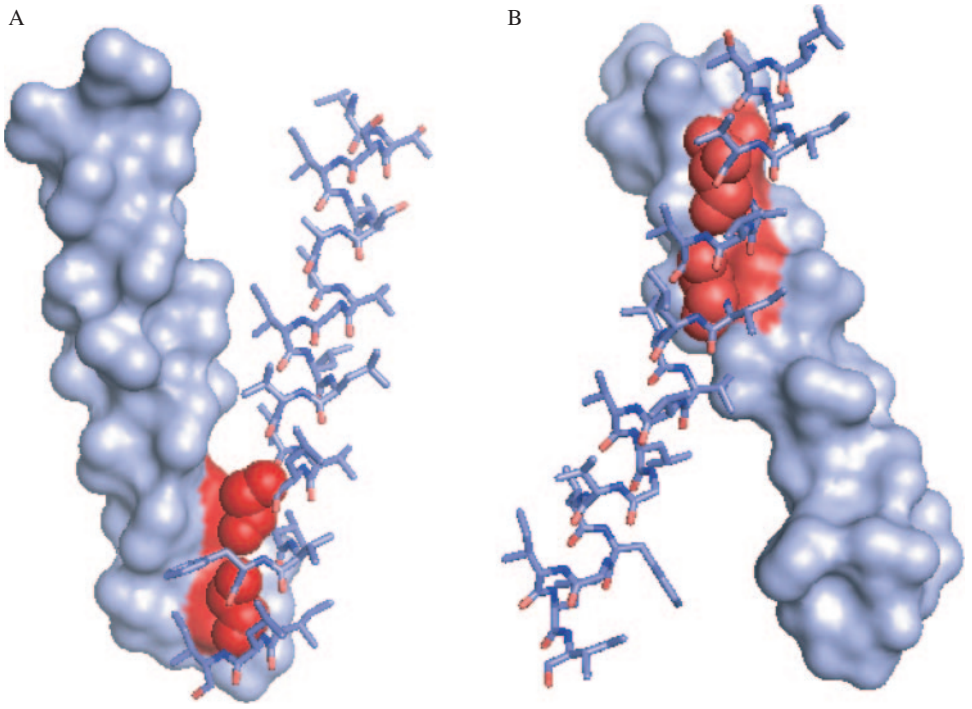


Fig. 5. The initial Gly668-xxx-Gly672 (**A**) and final Ser656-xxx-Gly660 (**B**) conformations of the TM domain of the ErbB2 homodimer. The Gly-xxx-Gly (A) and Ser-xxx-Gly (B) interfaces are marked in dark gray on the molecular surface of the helix at the back. The helix on the front is presented using a balls-and-sticks model, and the glycine and serine residues that comprise the motifs are presented using space-filled model.

characteristics of this pathway are presented in Figures 3 and 4, and representative snapshots from this pathway are provided in Figure 4. It is interesting to note that pathways that were ranked below this one partially overlapped with it.

Figure 3 shows the potential curve of the pathway that was ranked first according to the colony function. The pathway starts from the stable conformation involving the Gly668 and Gly672 residues (Fig. 5A) towards the less stable conformation involving the Ser656 and Gly660 residues (Fig. 5B). The energy is indicative of the stability of the conformation, e.g. in step 60, the pathway leads to the energetically most favorable conformation of packing via the Gly668-xxx-Gly672 motif where the distance between the helices is very small (6.5 Å) and the crossing angle is around $-35°$. The path ends in a conformation where the helices are packed via the Ser656-xxx-Gly660 motif. This conformation is associated with a less pronounced trough in the curve, where the interaxial distance between the helices is 7.5 Å and the angle is around $-45°$. Both this and the initial conformation (Fig. 5A) correspond to ridges-into-groves packing between the helices (Chothia *et al*., 1981) via the Ser-xxx-Gly and Gly-xxx-Gly motifs, respectively. In fact, it is evident from the movie (Supplementary Material, Movie 1) that the helices move subjected to the ridges-into-groves packing and that the stability at each step along the pathway is determined by the steric properties of the residues that mediate the inter-helix contact. For example, the least stable conformation (around Step 120) corresponds to the packing via Val664 residues. As suggested by Fleishman *et al*. (2002), the bulkiness of this residue interferes with the ridges-into-groves packing and this conformation, which determines the height of the energy barrier between the initial and final conformations in our suggested motion pathway. It is encouraging that the search, which started from a conformation that was in the vicinity of the most stable conformation, yielded both the most stable conformation (step 60) and a less favorable, but stable, conformation (step 156).

In addition, we examined the backward motion from a conformation where the helices are packed via the Ser656-xxx-Gly660 motif towards the conformation in which the helices are packed via the Gly668-xxx-Gly672 motif. The results (Supplementary Material, Movie 2) showed that the motion that was ranked first was very similar (in reverse order) to the original path. It ended in a goal conformation with an rmsd of $\sim$1.4Å from the initial conformation of the original path.

Glycophorin A is a bitopic TM protein that forms stable homodimers, and the NMR structure of this protein shows that the two TM helices are packed together via Gly79 and Gly83, similar to the Gly-xxx-Gly motif in one of the conformations suggested for ErbB2 above (MacKenzie *et al*., 1997). We carried out calculations using the NMR structure as the initial conformation. The calculation, which can be thought of as a negative control experiment, resulted in a few redundant pathways, comprising of native-like conformations (Supplementary Material, Movie 3).

5 DISCUSSION

A new RRT algorithm for the detection of stable conformations in TM proteins and putative pathways between them was presented here. In its pure form, the algorithm is based on geometric considerations, and energetic criteria may be added in a flexible

way. The current implementation is based on the LJ potential [Equation (1)].

It should be noted, however, that the calculated energy is unrealistically large in magnitude (e.g. Fig. 3), which is typical for force fields. Thus, the results should be examined only qualitatively. The reason for the apparent success of the potential of Equation (1) to provide reasonable pathways may be indicative of the significance of vdW interactions in stabilizing the conformations. Alternatively, the success of such a rudimentary potential, that excludes all other components of the inter-protein interactions, as well as the effects of the lipids and membrane structure, may be fortuitous. This issue will be clarified as more examples are investigated.

The calculations are very fast. For example, the 10 trees that were used to investigate the ErbB2 dimer (Section 4) were produced within <4 h on a standard desktop PC, which is significantly faster than typical molecular dynamics simulations of a similar system. The short simulation time and the flexible nature of the algorithm enable testing many aspects of the system, including the effects of changes in the energy function. Given a TM protein of interest, one can conduct a few test runs to converge to a reasonable procedure, as we demonstrated here for the TM domain of the ErbB2 and Glycophorin A homodimers.

In this preliminary work, we have focused on simple systems comprising pairs of α-helices, thus circumventing the complexities of modeling loops that connect pairs of helices. Our method can be generalized to TM proteins with an arbitrary number of helices and possibly also to water-soluble proteins of the α-helix bundle class. The addition of more helices will obviously increase the number of dofs. However, it will also reduce C_{feasible} owing to self-avoidance effects. C_{feasible} may be reduced further because many conformations of the helices may be incompatible with the lengths of the loops that connect them (Enosh *et al*., 2004).

ACKNOWLEDGEMENTS

This study was supported by a grant 222/04 from the Israel Science Foundation to N.B.-T. S.J.F was supported by a doctoral fellowship from the Clore Israel Foundation. Work by A.E. and D.H. has been supported in part by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University.

Conflict of Interest: none declared.

REFERENCES

Abramson,J. *et al*. (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**, 610–615.

Amato,N.M. *et al*. (2003) Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, **10**, 239–255.

Apaydin,M.S. *et al*. (2001) Capturing molecular energy landscapes with probabilistic conformational roadmaps. *Proceedings of IEEE International Conference Robotation Automation,* Seoul, pp. 932–939.

Bahar,I. and Rader.,A.J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin Struct. Biol*., **15**, 1–7.

Burgess,A.W. *et al*. (2003) An open-and-shut case? Recent insights into the activation of EGF/ErbB receptors. *Mol. Cell*, **12**, 541–552.

Canutescu,A.A. *et al*. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*., **12**, 2001–2014.

Choset,H. *et al*. (2005) *Principles of Robot Motion: Theory, Algorithms, and Implementations*. The MIT Press, chapter 7.

Chothia,C. *et al*. (1981) Helix to helix packing in proteins. *J. Mol. Biol*., **145**, 215–250.

Cortes,J. *et al*. (2005) A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, **21**, i116–i125.

Curran,A.R and Engelman,D.M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Op. in Struct. Biol*., **13**, 412–417.

Dunbrack,R.L.O. and Karplus,M. Jr (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains. *Nat. Struct. Biol*., **1**, 334–340.

Enosh,A. *et al*. (2004) Assigning transmembrane segments to helics in intermediate-resolution structures. *Bioinformatics*, **20**, i122–i129.

Finn,P.W. *et al*. (1996) Geometric manipulation of flexible ligands. *Proceedings of Workshop on Applied Computational Geometry,* Berlin, pp. 67–78.

Fleishman,S.J. *et al*. (2002) A putative molecular-activation switch in the transmembrane domain of erbb2. *Proc. Natl Acad. Sci*., **99**, 15937–15940.

Hsu,D. *et al*. (1999) Path planning in expansive configuration spaces. *Int. J. Comput. Geometry Appl*., **9**, 495–512.

Jiang,G. and Hunter,T. (1999) Receptor signaling: when dimerization is not enough. *Curr Biol*., **9**, 568–571.

Karplus,M. and McCammon,J.A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol*., **9**, 646–652.

Kavraki,L. *et al*. (1996) Probabilistic roadmaps for path planning in high dimensional configuration spaces. *In Proceedings of IEEE Transactions on Robotics and Automation*, Vol. 12, 566–580.

Latombe,J.-C. (1991) *Robot Motion Planning*. Kluwer Academic Publishers Boston, MA.

Latombe,J.-C. (1999) Motion planning: A journey of robots, molcules, digital actors, and other artifacts. *Int. J. Robotics Res*., **10**, 1119–1128.

LaValle,S.M. (2006) *Planning Algorithms. Cambridge University Press,* chapter 5. http://msl.cs.uiuc.edu/planning/.

LaValle,S.M. and Kuffner,J.J. (2001) Rapidly-exploring random trees: progress and prospects. In Donald,B.R., Lynch,K.M. and Rus,D. (eds), *Algorithmic and Computational Robotics: New Directions.* A.K. Peters, Wellesley, MA, 293–308.

Lotan,I. *et al*. (2004) Algorithm and data structures for efficient energy maintenance during Monte Carlo simulation of proteins. *J. Comput. Biol*., **11**, 902–932.

MacKenzie,K.R. *et al*. (1997) A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.

Neria,E. *et al*. (1996) Simulation of activation free energies in molecular systems. *J. Chem. Phys*., **105**, 1902–1921.

Sharir,M. (2004) Algorithmic motion planning. In Goodman,J.E. and O'Rourke,J. (eds), *Handbook of Discrete and Computational Geometry*. 2nd edn., 1037–1064. Chapman and Hall/CRC Press, Boca Raton.

Singh,A.P. *et al*. (1999) A motion planning approach to flexible ligand binding. *Int. Sys. for Molec. Biol*., 252–261.

Xiang,Z. *et al*. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of protein loop prediction. *Proc. Natl. Acad. Sci*., **99**, 7432–7437.

Structure, Vol. 12, 2265–2275, December, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.str.2004.10.006

A Putative Mechanism for Downregulation of the Catalytic Activity of the EGF Receptor via Direct Contact between Its Kinase and C-Terminal Domains

Meytal Landau, Sarel J. Fleishman, and Nir Ben-Tal*
Department of Biochemistry
George S. Wise Faculty of Life Sciences
Tel-Aviv University
Ramat-Aviv 69978
Israel

Summary

Tyrosine kinase receptors of the EGFR family play a significant role in vital cellular processes and in various cancers. EGFR members are unique among kinases, as the regulatory elements of their kinase domains are constitutively ready for catalysis. Nevertheless, the receptors are not constantly active. This apparent paradox has prompted us to seek mechanisms of regulation in EGFR's cytoplasmic domain that do not involve conformational changes of the kinase domain. Our computational analyses, based on the three-dimensional structure of EGFR's kinase domain suggest that direct contact between the kinase and a segment from the C-terminal regulatory domains inhibits enzymatic activity. EGFR activation would then involve temporal dissociation of this stable complex, for example, via ligand-induced contact formation between the extracellular domains, leading to the reorientation of the transmembrane and intracellular domains. The model provides an explanation at the molecular level for the effects of several cancer-causing EGFR mutations.

Introduction

The epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases (RTKs), also known as ErbB or HER, consists of four members, ErbB1, -2, -3, and -4 (Schlessinger, 2000). The receptors, which are activated by some dozen ligands, including EGF and TGFα, play an important role in the control of many fundamental cellular processes (Schlessinger, 2000). Mutations and overexpression of the ErbBs have been implicated in malignant diseases such as carcinoma and glioblastoma and are linked with aggressive disease, resistance to chemotherapy, and poor survival (Dancey, 2004). Accordingly, the ErbBs are attractive targets for anticancer drugs (Cho et al., 2003; Yarden and Sliwkowski, 2001). Structurally, the ErbBs consist of an N-terminal, extracellular domain that is connected by a short transmembrane span to a tyrosine kinase domain, which is in turn followed by a C-terminal domain.

In all RTKs, including the ErbBs, the active kinase triggers a wide spectrum of crucial intracellular signaling events (Schlessinger, 2000), and their catalytic activity is encapsulated in multiple layers of regulation (Huse and Kuriyan, 2002). A primary means of regulation in RTKs is ligand binding to the extracellular domain, leading to dimerization or formation of higher-order oligomers of the receptors and enzymatic activation (Schlessinger, 2000, 2003). Similarly, activation of ErbB1, -3, and -4 involves ligand-induced contact formation between the extracellular domains of different members of the ErbB family to form homo- and heterodimers (Schlessinger, 2000). Some studies have shown that, without a ligand, EGFR exists mostly in a monomeric form and that ligands induce its dimerization and activation (Yarden and Schlessinger, 1987). On the other hand, recent studies have demonstrated that while required, dimerization is not sufficient for activation and that in the absence of a ligand, stable, inactive dimers exist in a form in which contact between monomers involves the transmembrane and intracellular domains (Biswas et al., 1985; Gadella and Jovin, 1995; Moriki et al., 2001; Yu et al., 2002). Experimental evidence (Cadena et al., 1994), as well as the computational results presented below, demonstrates that the C-terminal domain plays a role in such contact formation.

In most tyrosine kinases (TKs) excluding the ErbBs, an important means of regulation involves profound structural changes along with transautophosphorylation of the kinase domain (Schlessinger, 2000). In contrast, the ErbB family is unique in that activation is independent of its phosphorylation state (Gotoh et al., 1992). The structure of the apo-EGFR kinase domain demonstrated that its unphosphorylated conformation was, in essence, identical to the phosphorylated conformations of other TKs (Stamos et al., 2002).

Recently, a structure of the kinase domain of the EGFR in complex with the inhibitor GW572016 (Lapatinib) was determined (Wood et al., 2004). This structure shows several differences, including different conformations of the substrate and ATP binding sites (Wood et al., 2004), from either the structure of the apo-EGFR or of EGFR bound to the OSI-774 (Tarceva) inhibitor (Stamos et al., 2002). The authors have suggested that these differences are due to the fact that the conformation seen in the GW572016 bound kinase domain reflects an inactive state that is accessible to the kinase domain under physiological conditions. However, GW572016 is very bulky in comparison to OSI-774. Thus, as the authors indicated, another possibility is that the differences in the structures are due to the inhibitor's large size, which forces a conformation that is far from native. That the apo-EGFR kinase domain is seemingly in a constitutively active conformation (Stamos et al., 2002) leads to an apparent paradox, since it is well established that ErbBs are not constitutively active (Schlessinger, 2000). Hence, our working hypothesis, as presented in Figure 1, was that ErbBs are regulated by another mechanism intrinsic to the intracellular domain; one that is phosphorylation independent.

The orphan receptor ErbB2 presents an even more intriguing case than other members of the EGFR family because its activation is not only phosphorylation inde-

*Correspondence: bental@ashtoret.tau.ac.il

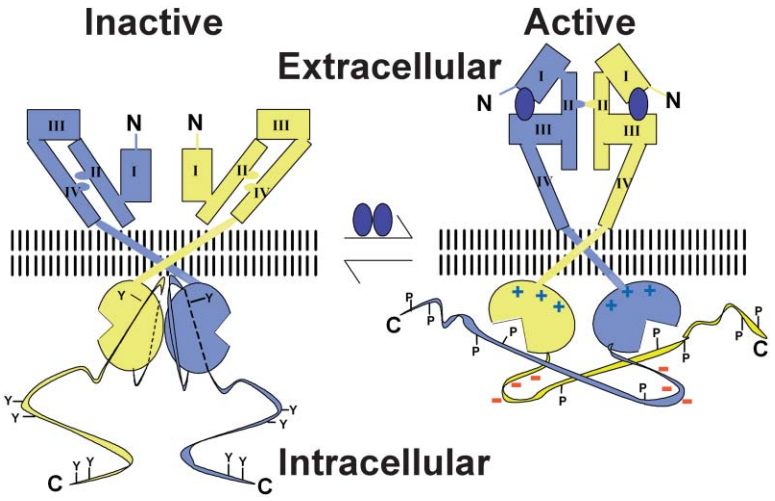


Figure 1. Schematic Diagram Representing the Suggested Model of EGFR Activation

Two EGFR monomers are colored light purple and yellow. The extracellular domain (residues 1–620, labeled I, II, III, and IV according to its subdomains) and the kinase domain (residues 685–957) are connected via a transmembrane helix (residues 621–642) and a short juxtamembrane segment (not shown). The C-terminal domain, comprising 229 amino acids, whose structure has not been determined, follows the kinase domain. Tyrosine residues (Y) known as the autophosphorylation sites in the C-terminal domain are indicated. In the inactive conformation (left), each of the extracellular domains assumes a compact structure, and the intracellular domains contact via the C-terminal fragments, leading to an inactive and stable form. Activation (right) occurs when ligands (purple ovals) bind to the extracellular domains, leading to the formation of a stable extracellular contact, which is followed by the rotation of the transmembrane helices and the subsequent destabilization of the contacts between the C-terminal and kinase domains. The kinase can now transautophosphorylate the tyrosine residues of its own C-terminal domain, as well as tyrosine residues of its protein substrates. The figure displays an illustration of the transmembrane domain; the suggested molecular model for the transmembrane domain in the active and inactive states was presented in (Fleishman et al., 2002). Positive and negative charges are marked in the active conformation on the kinase and C-terminal domains, respectively. In the inactive conformation, they roughly neutralize each other (Figure 2).

pendent, but also ligand independent (Cho et al., 2003). The absence of clear regulation of ErbB2 activation prompted us to propose a molecular mechanism for rotation-coupled activation of this receptor (Fleishman et al., 2002). Specifically, the transmembrane domain of an ErbB2 homodimer may occupy one of two stable conformations, corresponding to the active and inactive states of the receptor. The switch between the two conformations, involving a rotation of the transmembrane domain (Jiang and Hunter, 1999), induces the reorientation of the cytoplasmic domains within receptor dimers, thus leading to transautophosphorylation and stimulation of enzymatic activity. In this paper, we shall analyze the implications of this mode of activation on the conformation of the intracellular kinase domain.

The C-terminal domain plays a role in the internalization and degradation of the EGFR (Chang et al., 1995) and in EGFR's regulation by other molecules (Huse and Kuriyan, 2002). This domain also serves as a docking site for protein modules that bind the phosphotyrosines on the activated receptors (Schlessinger, 2000). In addition to these roles, the importance of the C-terminal domain for proper functioning of the EGFR was previously noted on the basis of studies of viral and other mutant EGFR members (Boerner et al., 2003; Chang et al., 1995; Wedegaertner and Gill, 1992).

Naturally occurring retroviral oncogene variants (v-ErbB) are an extracellular domain-truncated form of the EGFR gene that affects cell growth, motility, and survival (Gamett et al., 1986). These v-ErbB variants share striking homology with mutants of the human EGFR members that have been identified in gliomas and carcinomas (Frederick et al., 2000). Truncation of the extracellular domain is insufficient to manifest the transforming properties of the different v-ErbB variants; these properties are probably related to amino acid replacements, insertions, and deletions in the C-terminal domain (Boerner et al., 2003; Massoglia et al., 1990; Riedel et al., 1987). Variations in the C-terminal domain of ErbB receptors are known to be responsible for the alterations in the transforming potential and type of malignant diseases due to the expression of v-ErbBs in affected cells (Gamett et al., 1986; Pelley et al., 1989; Raines et al., 1988). The increased substrate-phosphorylation capacity of the C-terminally impaired EGFR is not attributed to lesser degradation and internalization, but rather to an enhanced rate of autophosphorylation (Robinson et al., 1992), thus providing direct evidence for a relationship between C-terminal domain impairment and increased catalysis.

Here, we propose a molecular model clarifying some of the ambiguity regarding the role of the C-terminal domain in ErbB regulation. According to the model (Figure 1), contact between the intracellular domains of the EGFR within a dimer leads to receptor inactivation, while ligand-induced contact between the extracellular domains leads to rotation-coupled activation (Fleishman et al., 2002; Jiang and Hunter, 1999; Moriki et al., 2001) by destabilization of the intermonomer contacts in the cytoplasmic domain. According to this scenario, interactions between the intracellular domains regulate activation (Burgess et al., 2003; Chantry, 1995), and the C-terminal domain acts as an inherent negative regulator of the EGFR's activity. This model offers a molecular mechanism that underlies the tumorigenic activity of EGFR mutants.

Results

Geometric Complementarity between the Kinase and C-Terminal Domains

The crystal structure of the EGFR (Stamos et al., 2002) (PDB entries 1m14 and 1m17) includes the kinase domain (residues 685–957) and a segment from the

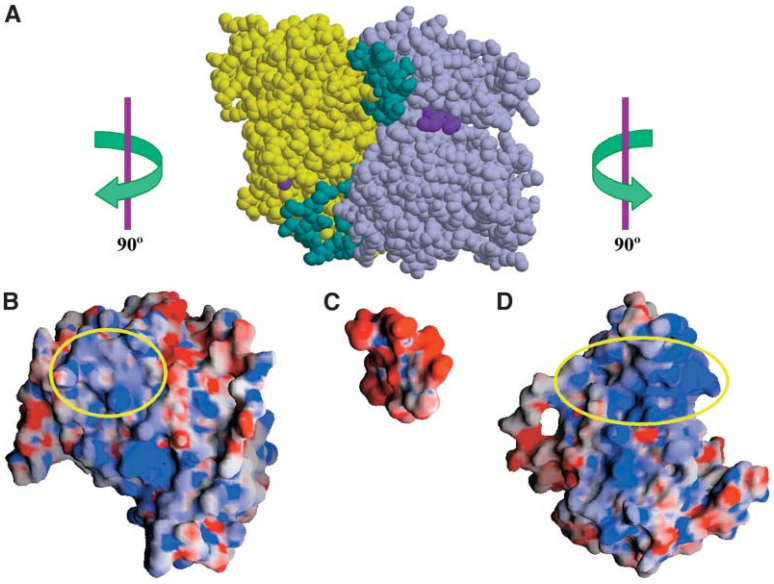


Figure 2. Geometric and Electrostatic Complementarity between the Kinase and C-Terminal Domains

(A) A space-filled model of EGFR's homodimeric complex (Stamos et al., 2002) showing the geometric complementarity within the complex. The kinase domains are colored light purple and yellow, the C-terminal fragments are colored cyan, and the inhibitor is colored purple. The dimer is symmetric, which means that each kinase domain is in contact with both C-terminal fragments, yielding one large and one small interface per monomer. The interactions with the C terminus are identical in both monomers. Figure 2A was made by using MOLSCRIPT (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997).

(B–D) A projection of the electrostatic potential (ϕ) onto the molecular surface of the kinase domain and the C-terminal fragment that comprise the complex in (A); $\phi > 10$ kT/e is dark blue, $\phi = 0$ is white, and $\phi < -10$ kT/e is dark red. Yellow ellipses mark the interfaces between the kinase domains and the C-terminal fragments. The figures were produced by using GRASP (Nicholls et al., 1991). (B) The left-most kinase domain shown in (A) (yellow) was rotated 90° to the left relative to its orientation in (A). (C) The C-terminal fragment is shown in the same orientation as the upper segment in (A). (D) The right-most kinase domain shown in (A) (light purple) was rotated 90° to the right relative to its orientation in (A). The electrostatic complementarity between the negatively charged C-terminal fragment and the positively charged residues of the kinase domain that interact with it is noticeable.

C-terminal domain (residues 977–995). The crystal structure reveals six putative dimer forms (Stamos et al., 2002). We focus here on the one with the largest intersubunit interface. In this complex, the kinase domain was found as a symmetric homodimer (Figure 2A), in which two copies of the fragment of the C-terminal domain mediate contact between the two kinase domains. This dimer is also the only one in which the kinase domains' N termini are facing in the same direction, in accordance with the physiological requirement that the two domains connect to the membrane bilayer.

We calculated the water-accessible surface area of the kinase domain alone and within the homodimeric complex. Each kinase monomer contacts two C-terminal fragments (Figure 2A). The water-accessible surface areas of these interfaces are 1419 Å^2 and 1048 Å^2. Thus, the total interface between each monomer of the kinase domain and the C-terminal fragments is 2467 Å^2, and the interface within the entire complex is twice as large, constituting a very large interface compared to typical interprotein interfaces (Jones and Thornton, 1996).

Charge Complementarity between the Kinase and C-Terminal Domains

Electrostatic calculations show strong positive potential in the kinase domain at its interface with the C-terminal fragments (Figures 2B and 2D). This potential originates from positively charged residues in both subunits, suggesting that the kinase domains would repel one another in the absence of the C-terminal fragments. Kinase domains from other ErbBs, which were constructed by using homology modeling, displayed similar positive electrostatic potentials in the corresponding regions (data not shown). The C-terminal fragment of all of the ErbBs contained 8–10 acidic and no basic residues (Figure 3). These residues produced a highly negative electrostatic potential (Figure 2C). Thus, the kinase domain and the C-terminal fragments form complementary surfaces in terms of their electrostatic potential. The geometric and charge complementarity (Figures 2B–2D), together with the significant size of the interface (Figure 2A), are indicative of the stability of the complex and suggest that it may be biologically meaningful.

Following the experiments of Chang et al. (1995) discussed below, we simultaneously substituted each of the negatively charged residues 979–982 (DEED, Figure 3) in the C-terminal domain with its polar equivalent, i.e., D→N and E→Q. The mutated C-terminal fragment is much less negatively charged than the native fragment (Figure 4B), and this difference in charge obstructs its electrostatic complementarity with the kinase domain and presumably destabilizes the complex. We further mutated the same positions to four positively charged lysine residues (Figure 4C). Electrostatic analysis of the mutated C-terminal fragment displayed a positive potential at the N-terminal region of the fragment, which would lead to its electrostatic repulsion from the kinase domain. To test whether the charge complementarity is

EGFR	979	D	E	E	D	M	D	D	V	V	D	A	D	E	Y	L	I	P	Q	996
ERB2	986	E	D	D	D	M	G	D	L	V	D	A	E	E	Y	L	V	P	Q	1003
ERB3	977	E	E	V	E	L	E	P	E	L	D	L	D	L	D	L	E	A	E	994
ERB4	985	D	E	E	D	L	E	D	M	M	D	A	E	E	Y	L	V	P	Q	1002

Figure 3. Abundance of Acidic Residues in the Fragment of the C-Terminal Domain

The multiple sequence alignment of the C-terminal segments of the four human members of the ErbB family. Each segment contains between 8 and 10 acidic residues (marked in red).

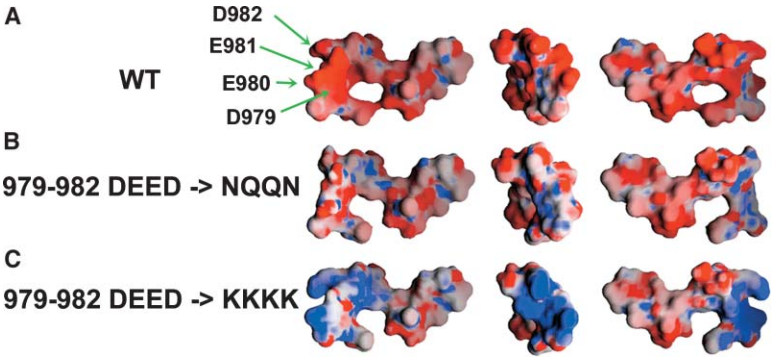


Figure 4. Electrostatic Analysis of Reported and Novel Mutations in the EGFR C-Terminal Fragment

A projection of the electrostatic potential (ϕ) onto the molecular surface of the C-terminal fragment; the color coding is as in Figures 2B–2D.

(A) The C-terminal fragment of the native EGFR, in the same orientation as in Figure 2C (central image), rotated 90° to the right (right image) or left (left image). The location of selected residues is marked.

(B) The C-terminal fragment, in the same orientations as in (A), in which the negatively charged residues in the 979–982 positions (DEED) were mutated to their polar equivalent (NQQN).

(C) The C-terminal fragment, in the same orientations as in (A), in which the same positions were mutated to positively charged lysine residues.

unique to ErbBs among TKs, we examined the electrostatic potentials of a few TKs of known structures as described in the Supplemental Data (available with this article online; Electrostatic calculations). These domains, which were derived from remotely related proteins, display diverse electrostatic characteristics. In particular, they do not share EGFR's strong positive electrostatic potential at the interface with the C-terminal fragments (data not shown), suggesting that such electrostatic interactions between the kinase and the C-terminal domains are specific to the ErbBs.

A Network of Ion Pairs and Hydrogen Bonds at the Interface

Our analysis demonstrated that a network of salt bridges and hydrogen bonds connects the two adjacent kinase domains through the C-terminal fragments (Figure 5B). We identified four charged residues within this network that are involved in several interactions with neighboring residues and are buried at the interface of the EGFR complex. Of these residues, two are positively charged (Lys822 and Lys828 on the kinase domain) and two are negatively charged (Asp988 and Asp990 on the C-terminal fragment) (Figure 5B).

Polar networks, such as the one observed in the EGFR interface (Figure 5B), significantly increase the stability of complexes and contribute to the binding specificity (Sheinerman et al., 2000). Therefore, mutations of charged positions in the network would alter the stability of the complex (Serrano et al., 1990). An even larger effect would be obtained by mutating them in pairs. For example, mutating Lys822 and Lys828 to aspartates or Asp988 and Asp990 to lysines altered the electrostatic surface of the kinase domain and the C-terminal fragment, respectively (Figure 6). Such mutations would impinge on the formation of the EGFR complex and kinase activation.

The importance of the network for the stability of the

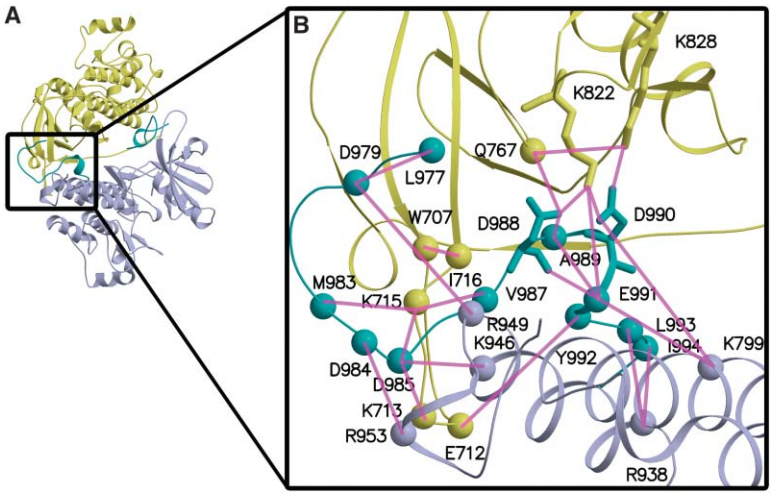


Figure 5. A Network of Ion Pairs and Hydrogen Bonds across the Interface of the EGFR Complex

The kinase domain monomers are displayed as ribbons and colored light purple and yellow. The C-terminal fragment is colored cyan.

(A) The EGFR homodimeric complex (Stamos et al., 2002) as viewed with a clockwise rotation of about 90° compared to Figure 2A.

(B) A close view, in the same orientation as in (A), of the polar network connecting the C-terminal fragment with its two adjacent kinase domains. The Cα atoms of residues comprising the polar network are displayed as spheres. Four selected residues in the network (Lys822, Lys828, Asp988, Asp990; their side chains displayed as sticks) are buried in the core of the kinase/C-terminal fragment interface, suggesting that they play a key role in complex stabilization (Sheinerman et al., 2000). Solid pink lines connect the Cα (or nearest neighbors) atoms of residues that form ion pairs and hydrogen bonds in the network. By symmetry, identical interactions connect residues between the second C-terminal fragment and the kinase domains (not shown). Each residue in the network is involved in a few interactions with neighboring residues. For instance, Asp990, located on the C-terminal domain, interacts with Lys822, located on one kinase domain monomer (yellow), and with Lys799, located on the second kinase domain monomer (light purple), presumably stabilizing the complex.

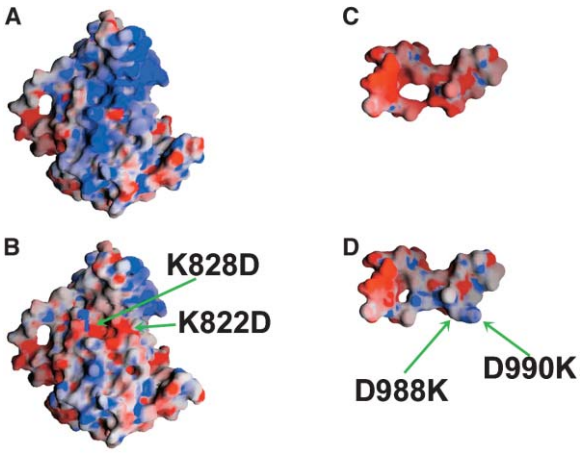


Figure 6. Electrostatic Analysis of Novel Mutations in the EGFR Kinase and C-Terminal Domains

A projection of the electrostatic potential (ϕ) onto the molecular surface of the kinase domain and the C-terminal fragment; the color coding is as in Figures 2B–2D.

(A) The native kinase domain in the same orientation as in Figure 2D.

(B) The R822D/R828D double mutant EGFR kinase domain in the same orientation as in (A).

(C) The native C-terminal fragment of the EGFR rotated 90° to the left relative to its orientation in Figure 2C.

(D) The D988R/D990R double mutant C-terminal domain in the same orientations as in (C).

complex can be tested experimentally by using the double mutant cycle approach (Serrano et al., 1990). Briefly, if the additive effects of mutating two residues separately (e.g., Lys822→Asp and Asp990→Lys) is significantly different from the effect of mutating the same two residues simultaneously, then the two positions are interdependent (Serrano et al., 1990), e.g., are involved in a salt bridge. Based on our analysis of the network, we suggest using a double mutant cycle, in which each step involves mutating a pair of similarly charged residues in the EGFR interface, as specified above.

Model of C-Terminal Domain Regulation of Kinase Activity

One of the phosphorylation sites of the C-terminal domain (Tyr992) is located on the fragment that forms contact with the kinase domain and is therefore inaccessible to phosphorylation in this conformation. The catalytic sites in the kinase domains are facing away from each other in the complex; therefore, transphosphorylation of residues on the kinase domain is improbable. The above two observations suggest that the EGFR crystal structure represents an inactive form of the receptor. The EGFR participates in imperative cell processes and ought to remain inactive under most physiological conditions (Huse and Kuriyan, 2002). Therefore, its inactive state should be very stable. Indeed, the complex in the crystal structure of the EGFR appears to be stable, based on the geometric and charge complementarity, further supporting the notion that this complex is inactive. It has been suggested that ligand-induced contact formation of the extracellular domains would lead to reorientation of the transmembrane domains (Fleishman et al., 2002; Jiang and Hunter, 1999; Moriki et al., 2001) and, subsequently, to rearrangements in the cytoplasmic domains (Figure 1). Any reorganization of the cytoplasmic complex, followed by a change in the position of the negatively charged C-terminal fragment, would lead to electrostatic repulsion between the two positively charged kinase monomers (Figure 2). Hence, this model of conformational changes during receptor activation may constitute a hitherto unknown mode of regulation.

Strong reinforcement of this model of regulation is provided by data on the EGFR analog c-ErbB (Chang et al., 1995). Deletions of a C-terminal fragment of this receptor (corresponding to residues 966–1006 of the EGFR) lead to higher autokinase activity compared to normal c-ErbB and transforming ability in vitro and in vivo. Moreover, a mutant in which the four consecutive acidic residues EEED were replaced by the polar segment QQQN showed higher autokinase activity and a partial transformation phenotype. Since the two mutants and normal receptors have similar rates of degradation, the higher transforming ability of the mutants could not be attributed to a longer half-life of the mutant receptor (Chang et al., 1995). These data are consistent with our results. The four acidic residues, which correspond to the DEED segment (Asp979–Asp982) in the EGFR, are located on the C-terminal fragment (Figure 3) that forms contact with the kinase domain. Our analysis showed that these positions contribute significantly to the negative electrostatic potential of the fragment (Figure 2), and their substitution with polar residues reduces the complementarity between the kinase and C-terminal domains (Figure 4B), presumably destabilizing the inactive complex.

Internal deletions of segments in the C-terminal domain of the EGFR have also been detected in naturally occurring EGFR mutants, which display tumorigenic properties. For example, an internal deletion of residues 959–1030 has been detected in EGFRs sequenced from human glioblastomas (Boerner et al., 2003; Chang et al., 1995; Frederick et al., 2000). Some viral ErbBs contain an in-frame deletion of 139 residues within the intracellular region, immediately following the kinase domain (Boerner et al., 2003; Chang et al., 1995; Frederick et al., 2000). This region contains the C-terminal fragment contacting the kinase domain according to the X-ray structure (Stamos et al., 2002). Our model suggests that the internal deletions in the C-terminal domain yield constitutively active forms of EGFR by means of destabilization of the inactive complex.

Evolutionary Conservation Analysis

The kinase domain of ErbB3 has no catalytic activity, yet it dimerizes with other members of the ErbB family to produce heterodimers with highly efficient catalytic activity (Schlessinger, 2000). These distinct features are manifested in the evolutionary-conservation analysis. ErbB3's kinase domain displays variations in the catalytic site in comparison to other members of the ErbB family, thus rendering it inactive. However, the interface between the kinase domain and the C-terminal fragment

is highly conserved within the ErbBs and their orthologs, including ErbB3. As a reference, an analysis of 121 kinase domains from various TKs showed that the catalytic site, including the ATP and substrate binding loop, was highly conserved, whereas the interface between the kinase domain and the C-terminal fragment was highly variable (data not shown).

Overall, the conservation analysis provides further support for the suggestion that the dimeric complex observed in the crystal structure is not common to all the TKs. However, the contact area between the kinase and C-terminal domains in this complex is common to the ErbBs, which thus maintain the ability to produce homo- and heterodimers through the same interface.

A Network of Correlated Amino Acid Substitutions between Regulatory Elements

By and large, all TKs carry out the same catalytic process. Thus, key residues in the kinase domain, which are responsible for catalysis of phosphotransfer, are under strong evolutionary constraint, as mentioned above. However, in order for the kinases to be involved in numerous and distinct signal transduction pathways, each kinase family exhibits variations in its amino acid sequence that are necessary for the modification of the mode of regulation. Since multiple positions are involved in determining these traits, these sequence variations should occur concomitantly in relevant regulatory elements. In other words, during evolution, substitutions of one residue in regulatory elements may be compensated by a concurrent change in another residue, in order to maintain the structural or functional relationship between these positions (Fleishman et al., 2004b).

In order to look for particular positions that could play a role in regulation, we analyzed the set of 121 multiply aligned TKs of diverse families in search of pairs of amino acid positions that might be evolutionarily correlated (Fleishman et al., 2004b). The analysis revealed 152 pairs of correlated residues, among which we identified a network of 14 highly intercorrelated positions (Figure 7A and Table 1).

The kinase domain includes several regulatory elements, such as the αC helix and activation loop, which play a role in allosteric regulation and are responsible for conformational changes. These elements function together to control activation, i.e., their movements are concurrent and their conformations are mutually dependent (Huse and Kuriyan, 2002). Our analysis showed pairs of evolutionarily correlated positions in these known regulatory elements. For example, Ala743, which is located on the αC helix, is correlated with Gly849 of the activation loop (Figure 7A).

The LVI segment (residues 955–957) of EGFR and its equivalent segments in other ErbBs are necessary for ligand-independent dimerization of the EGFR intracellular domains and for transphosphorylation in ErbB2 and ErbB3 heterodimers through allosteric regulation (Stamos et al., 2002). Leu955 in this LVI segment is correlated with Tyr740, which is located on the αC helix (Figure 7A). The association of the αC helix with a known dimerization motif exemplifies interdomain relationships between regulatory elements in the ErbBs. Both of these

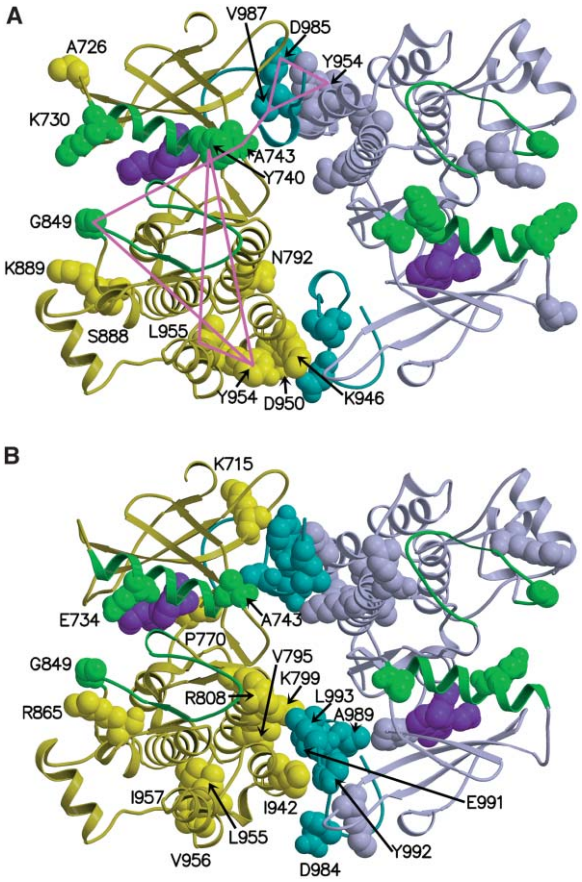


Figure 7. Evolutionarily Correlated and Specificity-Determining Amino Acid Sites

The EGFR homodimeric complex (Stamos et al., 2002) as viewed with an upward rotation of about 180° compared to Figure 2A. The kinase domains, presented by using trace models, are colored light purple and yellow, the C-terminal fragments are colored cyan, and the inhibitor is shown as a purple space-filled model. The αC helix (residues 729–744) and the activation loop (residues 831–852) are colored green.

(A) The residues in the cluster of the most significant pairs of correlated amino acid sites are displayed as space-filled models. Solid pink lines connect a few of the pairs of correlated residues (highlighted in Table 1) in the EGFR homodimer. Correlations within the kinase domain are demonstrated only on the left monomer, and correlations between the kinase and the C-terminal domains are demonstrated only on the upper interface. The correlations between known regulatory elements, such as the αC helix and the activation loop and the interface between the kinase domain and C-terminal fragment, suggest that the latter may also be involved in regulation.

(B) The main specificity-determining residues are located on the αC helix, the activation loop, the C-terminal fragment, and its interfaces on the kinase. This suggests that the regulatory elements in the EGFR had evolved specifically to stabilize the active conformation. Concurrently, an alternative negative regulatory mechanism had evolved in the form of the inactive complex between the kinase and the C-terminal domains. The figures were made by using MOLSCRIPT (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997).

domains are important for regulation (Huse and Kuriyan, 2002; Stamos et al., 2002); for example, mutations in Leu955 or Tyr740 severely impaired the kinase activity of the EGFR (Stamos et al., 2002; Walker et al., 1998).

Based on these results, we concluded that this net-

Table 1. Correlated Pairs in the TK Family

Pairs of Correlated Positions	Correlation Coefficients
Ala726-Lys730	0.65 (0.49, 0.78)
Ala726-Ser888	0.60 (0.34, 0.77)
Ala726-Lys946	0.45 (0.20, 0.64)
Lys730-Ser888	0.59 (0.37, 0.76)
Tyr740-Tyr954	**0.45 (0.19, 0.67)**
Tyr740-Leu955	**0.48 (0.21, 0.69)**
Ala743-Asn792	0.50 (0.19, 0.76)
Ala743-Gly849	**0.47 (0.18, 0.72)**
Ala743-Ser888	0.45 (0.21, 0.69)
Ala743-Lys889	0.51 (0.14, 0.73)
Ala743-Val987	0.54 (0.34, 0.72)
Asn792-Gly849	0.48 (0.26, 0.70)
Asn792-Ser888	0.50 (0.25, 0.72)
Gly849-Tyr954	**0.56 (0.29, 0.76)**
Ser888-Lys946	0.61 (0.40, 0.78)
Lys946-Asp950	0.52 (0.31, 0.69)
Asp950-Tyr954	0.48 (0.31, 0.65)
Asp950-Leu955	0.50 (0.26, 0.67)
Tyr954-Leu955	**0.54 (0.32, 0.72)**
Tyr954-Asp985	**0.50 (0.30, 0.67)**
Tyr954-Val987	**0.52 (0.36, 0.67)**
Asp985-Val987	**0.52 (0.36, 0.67)**

A list of 22 pairwise correlations between positions comprising the most significant cluster of correlated residues. The trimmed means in the 95% confidence interval of correlations (r), which were calculated from 400 bootstrapping samples, are indicated, and the 95% confidence interval is shown in parentheses (see the Supplemental Data). The numbering of the positions is done according to the EGFR sequence. The pairs of positions that are located on known regulatory regions are highlighted in bold and are connected by solid pink lines in Figure 7A.

work of correlations identified amino acids playing a role in regulation. Interestingly, the same cluster also displays correlations between residues mediating contact between the kinase and C-terminal domains. Tyr954 is located on the kinase domain and contacts the C-terminal fragment. This residue is in close proximity to, and is highly correlated with, residues Asp985 and Val987 of the C-terminal fragment (Figure 7A). Taken together, these correlations consolidate our hypothesis that the contact between the kinase and C-terminal domains is biologically meaningful.

The same cluster of 14 highly intercorrelated positions also includes correlations between positions at the interface of the kinase domain and the C-terminal fragment and known regulatory elements. For instance, Tyr954 located on this interface is correlated with Gly849 of the activation loop, with Leu955 of the LVI segment, and with Tyr740 of the αC helix (Figure 7A). This network of correlations suggests that this interface is also involved in regulation.

Val987 of the C-terminal fragment is correlated with Ala743, which is located on the αC helix (Figure 7A). In this context, it is important to note that the C-terminal domain is a vital modulator of TKs' activity (Jorissen et al., 2003; Schlessinger, 2000), as was elaborated above. For example, structure determination and mutagenesis experiments have shown that the kinase domains of the insulin, the Tie2, and the platelet-derived growth factor β receptor (PDGFR) TKs are autoinhibited by their C-terminal domains through direct contacts with the kinase domain (Chiara et al., 2004; Noelle et al., 2000; Shewchuk et al., 2000). Accordingly, evolutionary correlation between the kinase and C-terminal domains is expected to be general. The mechanism by which direct contacts control activation may vary between the kinases and could not be inferred from the evolutionary-correlation analysis. We anticipate that in the ErbBs, the direct contact between the kinase and C-terminal domains regulates catalysis by the formation of the inactive dimer shown in Figure 2A.

Specificity Determinants in Regulatory Regions

Although TKs share an identical catalytic mechanism, each kinase family is regulated by various means, responds to different ligands, and activates diverse substrates. It is anticipated that certain positions would be responsible for these different traits, and would be reflected in their patterns of substitution (Fleishman et al., 2004a). Due to such differences in functions, these positions are not expected to be strictly conserved in evolution. Rather, they should be conserved among kinases of similar functions in different species (orthologs), and would differ in paralogs. Substitutions involving these residues are presumably responsible for certain alterations in the functions of the various families of the TK superfamily.

We have identified some of these specificity-determining amino acid positions in a set of 121 multiply aligned TKs. The main specificity-determining residues are presented in Figure 7B, and their locations are indicated in Table 2. The list includes residues from the known regulatory regions, as well as residues that connect the kinase and the C-terminal domains and participate in the polar network across the interface (Figure 5B).

Discussion

ErbBs are structurally unique among TKs, as all of the catalytic elements in their kinase domains are ready for phosphotransfer at all times (Stamos et al., 2002). Yet, various functional assays show them not to be constitutively active (Schlessinger, 2000). The absence of a central regulatory module raises a fundamental dilemma, namely, what prevents the receptors from being spuriously activated? One possible mechanism is that changes in the relative orientation of the subunits within a dimer control activation, as suggested by the model of rotation-coupled activation (Jiang and Hunter, 1999). According to this view, contact formation between the extracellular domains leads to reorientation in the transmembrane domain, which is propagated into the cytoplasm (Fleishman et al., 2002; Jiang and Hunter, 1999; Moriki et al., 2001). Thus, the reorientation of the kinase domains vis-à-vis each other serves as a molecular switch that turns the kinase domains "on." What might be the mechanism by which this reorientation is translated into kinase activation is not yet clear.

Understanding the molecular details of how the ErbB proteins are regulated will most probably have to await the emergence of a structure of the full-length receptor in oligomeric complexes. The structures of parts of the kinase and the extracellular domains available today only provide a fragmentary view of the regulatory ele-

Table 2. Specificity Determinants in the TK Family

Position	Correlation Coefficients	Location and Putative Functional Role in the EGFR
Lys715	0.23 (0.11, 0.36)	Located on the kinase at the large interface with the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Pro770	0.24 (0.07, 0.39	Located on the kinase at the large interface with the C-terminal fragment.
Val795	0.29 (0.17, 0.40)	Located on the kinase at the small interface with the C-terminal fragment.
Ile942	0.24 (0.13, 0.36)	Located on the kinase at the small interface with the C-terminal fragment.
Lys799	0.24 (0.10, 0.36)	Located on the kinase at the small interface with the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Glu734	0.27 (0.12, 0.42)	Located on the αC helix of the kinase domain; involved in regulation.
Ala743	0.29 (0.14, 0.42)	Located on the αC helix of the kinase domain; involved in regulation.
Arg808	0.23 (0.12, 0.35)	Located on the kinase domain, close to the activation loop. Involved in hydrogen bonds that stabilize the activation loop (Stamos et al., 2002).
Arg865	0.26 (0.10, 0.41)	Located on the kinase domain, close to the activation loop. Involved in hydrogen bonds that stabilize the activation loop (Stamos et al., 2002).
Gly849	0.28 (0.18, 0.37)	Located on the activation loop of the kinase domain; involved in regulation.
Leu955	0.29 (0.13, 0.42)	A part of the "LVI motif". Important for dimerization of the kinases.
Val956	0.29 (0.17, 0.42)	A part of the "LVI motif". Important for dimerization of the kinases.
Ile957	0.25 (0.08, 0.39)	A part of the "LVI motif". Important for dimerization of the kinases.
His964	0.29 (0.15, 0.43)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Leu965	0.24 (0.10, 0.39)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Ser967	0.23 (0.10, 0.35)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Pro968	0.26 (0.13, 0.40)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Ser971	0.27 (0.13, 0.38)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Tyr974	0.30 (0.16, 0.43)	A putative negative regulator of EGFR's activity; located on the C-terminal domain.
Asp984	0.31 (0.19, 0.44)	A putative negative regulator of EGFR's activity; located on the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Ala989	0.25 (0.13, 0.38)	A putative negative regulator of EGFR's activity; located on the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Glu991	0.25 (0.11, 0.41)	A putative negative regulator of EGFR's activity; located on the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Leu993	0.29 (0.17, 0.41)	A putative negative regulator of EGFR's activity; located on the C-terminal fragment; participates in the polar network across the interface (Figure 5).
Tyr992	0.24 (0.11, 0.38)	An autophosphorylation site, located on the C-terminal fragment; participates in the polar network across the interface (Figure 5).

A list of 24 out of 47 residues that were identified as specificity determinants (Fleishman et al., 2004a). The location of each residue in the EGFR sequence and its functional role are indicated. The trimmed means in the 95% confidence interval of correlations (r), which were calculated from 400 bootstrapping samples, are indicated, and the 95% confidence interval is shown in parentheses (see the Supplemental Data). In addition to the residues presented above, the list of specificity determinant includes the following residues: V750, Q763, L775, E780, D783, N792, V821, Q825, T830, S888, K889, I899, S901, I902, P910, K925, S933, D950, Q952, Q958, G959, D960, and E961. Their putative roles remain to be tested experimentally.

ments in the structure. In Figure 1, we suggest a model for such regulation in the ErbB family; this model is based on the available structures and is supported by a large body of biochemical and physiological data.

The role of the C-terminal domain as a modulator of kinase activity has been discussed extensively (Cadena et al., 1994; Jorissen et al., 2003), especially in the v-ErbB products (Boerner et al., 2003). Our results offer a model of the molecular mechanism for this modulation (Figure 1). In the inactive state (Figure 1, left), the EGFR extracellular domains assume a tethered structure (Ferguson et al., 2003) that hinders contact formation between the two subunits (Burgess et al., 2003). In this conformation, the extracellular domains are connected to the transmembrane helices in their inactive state (Fleishman et al., 2002), thereby maintaining the intracellular domains as a stable, inactive dimer (Figure 2A). In this state, the C-terminal domain is in contact with the kinase domain and is inaccessible to downstream substrates (Cadena et al., 1994). Ligand-induced activation of the EGFR (Figure 1, right) leads to conformational changes in the extracellular domains, allowing contact formation between the two subunits (Ogiso et al., 2002), followed by a rotation of the transmembrane helices toward their active state (Fleishman et al., 2002; Jiang and Hunter, 1999; Moriki et al., 2001). This switch in the orientation of the transmembrane helices leads to the destabilization of the inactive intracellular dimer. The C-terminal domain detaches from the kinase domain and may undergo phosphorylation, making the kinase accessible to its substrates (Moriki et al., 2001).

The structure of the GW572016 bound EGFR comprises the kinase domain and part of the C-terminal domain that is packed along the kinase domain. In this structure, the C-terminal domain partly blocks the ATP binding site (Wood et al., 2004), as in the inactive forms of the myosin light chain kinase of the Ser/Thr kinase family (Huse and Kuriyan, 2002) and the Tie2 RTK (Shewchuk et al., 2000). That the GW572016 bound EGFR structure shows an inactive conformation that is not primed for catalysis suggests that activation of the EGFR involves conformational changes within the kinase domain, in contrast to the view that the kinase domain is constitutively ready for phosphotransfer (Stamos et al., 2002). We note, however, that the new structure suggests an important role for the C-terminal domain in stabilizing an inactive conformation of the kinase domain (Wood et al., 2004); this finding is in harmony with the model of activation suggested here.

The proposed molecular model may explain the un-

derlying molecular causes of malignancy mediated by EGFRs that contain mutations in their C-terminal domain. According to the model, the transforming properties of these mutations (Boerner et al., 2003; Chang et al., 1995; Frederick et al., 2000) are due to destabilization of the inactive EGFR.

All TKs catalyze the same reaction, which is the transfer of the γ-phosphate of ATP to the hydroxyl group of tyrosine. Indeed, the active conformation of the kinase domain of most TKs is nearly identical. In contrast to the uniform active conformation, TKs differ from each other in their inactive conformations (Huse and Kuriyan, 2002). In some RTKs, as in the PDGFR family, the juxtamembrane domain serves to block the active conformation. Autophosphorylation of tyrosine residues in highly conserved juxtamembrane motifs, specific to each family, relieves autoinhibition (Griffith et al., 2004). In the case of the EGFR family, inhibition by the juxtamembrane domain is less likely, since there are no tyrosine residues in the juxtamembrane segment that can be phosphorylated.

Various regulatory mechanisms could play an important role in ensuring the signaling specificity in the TK superfamily. Accordingly, we suggest that certain amino acid substitutions in regulatory elements were sustained during evolution, leading to alterations in the regulatory mechanisms. This hypothesis is supported by the analysis of specificity determinants (Figure 7B). In the vast majority of the TKs, kinase activity is regulated through a change in the conformation of the activation loop and αC helix. Nevertheless, these regulatory regions undergo different conformational changes in different isoforms, and their inactive conformations are stabilized by fastidious means specific to each kinase family (Huse and Kuriyan, 2002). The ErbBs are further exceptional among TKs, in that the activation loop and αC helix are constitutively stable in the active conformation (Stamos et al., 2002). Our analysis of correlated mutations (Figure 7A) suggests that in order to complement the role of these known regulatory elements in maintaining an active conformation, other residues in ErbBs have evolved to keep the enzyme dormant, as in the "inactive" complex shown in Figure 2A.

We propose that members of the EGFR family utilize the unique regulatory mechanism that is presented in Figure 1. These receptors contain a long C-terminal domain that is involved in signal transmission inside the cell and is also an inherent regulator of kinase activity (Chang et al., 1995). Our results suggest that the complex between the kinase and C-terminal domains of Figure 2A is stable and biologically significant, as indicated by the large intersubunit interface, the electrostatic and geometric complementarity between the C-terminal segments and the kinases (Figures 2B–2D and 5), as well as the evolutionary correlation between specified amino acid sites (Figure 7A). This complex appears to correspond to the basal, inactive form of the receptor, as delineated above and in accordance with previous experimental data (Boerner et al., 2003; Chang et al., 1995). Although our computational analysis and the experimental data support the presence of an inactive dimer (Yu et al., 2002) and the necessity of contact between the kinase and C-terminal domains (Chang et al., 1995), the biological relevance of the crystal dimer has yet to be determined. The importance of the interface between the kinase domain and the C-terminal fragment for the regulation of EGFR activity can be tested experimentally, as delineated in the section entitled "A Network of Ion Pairs and Hydrogen Bonds at the Interface."

Our model of EGFR's regulation (Figure 1) and its relevance to cancer could be further tested by examining the properties of a short peptide analog to the C-terminal fragment. Such a peptide may have a regulatory effect on EGFR activation. For instance, in tumorigenic cells, the short peptide may associate with the kinase domain instead of the truncated C-terminal domain. This would stabilize the inactive configuration and thereby thwart the constitutive activation of the mutant receptor. Interestingly, a similar approach was applied successfully in a recent study on the PDGFR, which is also selfinhibited by direct contact with its C-terminal domain (Chiara et al., 2004). In this work, the authors showed that a soluble peptide, corresponding to the inhibitory fragment in the PDGFR C-terminal domain, delayed the activation of the receptor and inhibited the enhanced kinase activity of a C-terminal truncated PDGFR. Hence, the small peptide mimicked the role of the C-terminal fragment in regulating kinase activity (Chiara et al., 2004). It will be interesting to examine the therapeutic properties of such a peptide in the case of the EGFR.

Experimental Procedures

Biophysical and Structural Analysis

Electrostatic, solvent-accessible surface area calculations and homology modeling were carried out as described in the Supplemental Data.

Collection of Sequence Homologs and Their Alignment

A multiple-sequence alignment (MSA) of homologous kinase domains was produced by combining multiple-structure and sequence alignments to obtain high-quality alignments as described by Al-Lazikani et al. (2001) and in the Supplemental Data. This resulted in an MSA of 121 homologous sequences comprising the kinase domain and about 50 positions C-terminal to it (corresponding to positions 683–998 of the EGFR). The MSA is shown in Supplemental Figure S1 in the Supplemental Data.

Evolutionary Conservation

Evolutionary conservation scores were calculated by using the MSA and *Rate4Site*'s maximum-likelihood algorithm (Pupko et al., 2002), as implemented in the ConSurf web server (Glaser et al., 2003) (http://consurf.tau.ac.il/).

Correlated Amino Acid Substitutions

Pairs of amino acids that appear to change concomitantly during evolution within the TKs were detected by using the MSA and the *CorrMut* algorithm (Fleishman et al., 2004b). The methodological details are provided as Supplemental Data.

Specificity Determinants

Residues in the TK superfamily, which may be responsible for determining specific characteristics in different kinase families, were detected by using the MSA and the *SpecDet* algorithm (Fleishman et al., 2004a). A description of the algorithm is provided as Supplemental Data.

Supplemental Data

Supplemental Data including analysis of the electrostatic potential of representative TKs of known structure; solvent-accessible surface area calculations and homology modeling of selected TKs;

the MSA of the TK family; methodological details of the correlated mutations analysis; and a description of the algorithm used for detecting the specificity-determining residues are available at http://www.structure.org/cgi/content/full/12/12/2265/DC1/.

Acknowledgments

We thank Tony Hunter, Antony Burgess, Joseph Schlessinger, Idit Kopatz, Amit Kessel, and Saul Yankofsky for their critical comments on the manuscript, Miriam Eisenstein for her help in the identification of the dimeric conformation of the EGFR kinase, and Lisa Shewchuk for sharing the coordinates of the EGFR/GW572016 structure before their release. This study was supported by a grant from the Israel Cancer Association (ICA) and by a Research Career Development Award from the Israel Cancer Research Fund (ICRF) to N.B.-T. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation. M.L. was supported by a Travel Scholarship from the Constantiner Institute for Molecular Genetics.

Received: June 16, 2004
Revised: September 22, 2004
Accepted: October 8, 2004
Published: December 7, 2004

References

Al-Lazikani, B., Sheinerman, F.B., and Honig, B. (2001). Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc. Natl. Acad. Sci. USA *98*, 14796–14801.

Biswas, R., Basu, M., Sen-Majumdar, A., and Das, M. (1985). Intrapeptide autophosphorylation of the epidermal growth factor receptor: regulation of kinase catalytic function by receptor dimerization. Biochemistry *24*, 3795–3802.

Boerner, J.L., Danielsen, A., and Maihle, N.J. (2003). Ligand-independent oncogenic signaling by the epidermal growth factor receptor: v-ErbB as a paradigm. Exp. Cell Res. *284*, 111–121.

Burgess, A.W., Cho, H.S., Eigenbrot, C., Ferguson, K.M., Garrett, T.P., Leahy, D.J., Lemmon, M.A., Sliwkowski, M.X., Ward, C.W., and Yokoyama, S. (2003). An open-and-shut case? Recent insights into the activation of EGF/ErbB receptors. Mol. Cell *12*, 541–552.

Cadena, D.L., Chan, C.L., and Gill, G.N. (1994). The intracellular tyrosine kinase domain of the epidermal growth factor receptor undergoes a conformational change upon autophosphorylation. J. Biol. Chem. *269*, 260–265.

Chang, C.M., Shu, H.K., Ravi, L., Pelley, R.J., Shu, H., and Kung, H.J. (1995). A minor tyrosine phosphorylation site located within the CAIN domain plays a critical role in regulating tissue-specific transformation by erbB kinase. J. Virol. *69*, 1172–1180.

Chantry, A. (1995). The kinase domain and membrane localization determine intracellular interactions between epidermal growth factor receptors. J. Biol. Chem. *270*, 3068–3073.

Chiara, F., Bishayee, S., Heldin, C.H., and Demoulin, J.B. (2004). Autoinhibition of the platelet-derived growth factor beta-receptor tyrosine kinase by its C-terminal tail. J. Biol. Chem. *279*, 19732–19738.

Cho, H.S., Mason, K., Ramyar, K.X., Stanley, A.M., Gabelli, S.B., Denney, D.W., Jr., and Leahy, D.J. (2003). Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. Nature *421*, 756–760.

Dancey, J.E. (2004). Predictive factors for epidermal growth factor receptor inhibitors—The bull's-eye hits the arrow. Cancer Cell *5*, 411–415.

Ferguson, K.M., Berger, M.B., Mendrola, J.M., Cho, H.S., Leahy, D.J., and Lemmon, M.A. (2003). EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. Mol. Cell *11*, 507–517.

Fleishman, S.J., Schlessinger, J., and Ben-Tal, N. (2002). A putative molecular-activation switch in the transmembrane domain of erbB2. Proc. Natl. Acad. Sci. USA *99*, 15937–15940.

Fleishman, S.J., Unger, V.M., Yeager, M., and Ben-Tal, N. (2004a). A C-alpha model for the transmembrane alpha-helices of gap-junction intercellular channels. Mol. Cell *15*, 879–888.

Fleishman, S.J., Yifrach, O., and Ben-Tal, N. (2004b). An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J. Mol. Biol. *340*, 307–318.

Frederick, L., Wang, X.-Y., Eley, G., and James, C.D. (2000). Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. Cancer Res. *60*, 1383–1387.

Gadella, T.W., Jr., and Jovin, T.M. (1995). Oligomerization of epidermal growth factor receptors on A431 cells studied by time-resolved fluorescence imaging microscopy. A stereochemical model for tyrosine kinase receptor activation. J. Cell Biol. *129*, 1543–1558.

Gamett, D.C., Tracy, S.E., and Robinson, H.L. (1986). Differences in sequences encoding the carboxyl-terminal domain of the epidermal growth factor receptor correlate with differences in the disease potential of viral erbB genes. Proc. Natl. Acad. Sci. USA *83*, 6053–6057.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163–164.

Gotoh, N., Tojo, A., Hino, M., Yazaki, Y., and Shibuya, M. (1992). A highly conserved tyrosine residue at codon 845 within the kinase domain is not required for the transforming activity of human epidermal growth factor receptor. Biochem. Biophys. Res. Commun. *186*, 768–774.

Griffith, J., Black, J., Faerman, C., Swenson, L., Wynn, M., Lu, F., Lippke, J., and Saxena, K. (2004). The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. Mol. Cell *13*, 169–178.

Huse, M., and Kuriyan, J. (2002). The conformational plasticity of protein kinases. Cell *109*, 275–282.

Jiang, G., and Hunter, T. (1999). Receptor signaling: when dimerization is not enough. Curr. Biol. *9*, R568–R571.

Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA *93*, 13–20.

Jorissen, R.N., Walker, F., Pouliot, N., Garrett, T.P., Ward, C.W., and Burgess, A.W. (2003). Epidermal growth factor receptor: mechanisms of activation and signalling. Exp. Cell Res. *284*, 31–53.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. *24*, 946–950.

Massoglia, S., Gray, A., Dull, T.J., Munemitsu, S., Kun, H.J., Schlessinger, J., and Ullrich, A. (1990). Epidermal growth factor receptor cytoplasmic domain mutations trigger ligand-independent transformation. Mol. Cell. Biol. *10*, 3048–3055.

Merritt, E.A., and Bacon, D.J. (1997). Raster3D photorealistic molecular graphics. Methods Enzymol. *277*, 505–524.

Moriki, T., Maruyama, H., and Maruyama, I.N. (2001). Activation of preformed EGF receptor dimers by ligand-induced rotation of the transmembrane domain. J. Mol. Biol. *311*, 1011–1026.

Nicholls, A., Sharp, K.A., and Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins *11*, 281–296.

Noelle, V., Tennagels, N., and Klein, H.W. (2000). A single substitution of the insulin receptor kinase inhibits serine autophosphorylation in vitro: evidence for an interaction between the C-terminus and the activation loop. Biochemistry *39*, 7170–7177.

Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J.H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., and Yokoyama, S. (2002). Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell *110*, 775–787.

Pelley, R.J., Maihle, N.J., Boerkoel, C., Shu, H.K., Carter, T.H., Moscovici, C., and Kung, H.J. (1989). Disease tropism of c-erbB: effects of carboxyl-terminal tyrosine and internal mutations on tissue-specific transformation. Proc. Natl. Acad. Sci. USA *86*, 7164–7168.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional

regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18*, S71–S77.

Raines, M.A., Maihle, N.J., Moscovici, C., Moscovici, M.G., and Kung, H.J. (1988). Molecular characterization of three erbB transducing viruses generated during avian leukosis virus-induced erythroleukemia: extensive internal deletion near the kinase domain activates the fibrosarcoma- and hemangioma-inducing potentials of erbB. J. Virol. *62*, 2444–2452.

Riedel, H., Schlessinger, J., and Ullrich, A. (1987). A chimeric, ligand-binding v-erbB/EGF receptor retains transforming potential. Science *236*, 197–200.

Robinson, H.L., Tracy, S.E., Nair, N., Taglienti-Sian, C., and Gamett, D.C. (1992). Characterization of an angiosarcoma-inducing mutation in the erbB oncogene. Oncogene *7*, 2025–2030.

Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. Cell *103*, 211–225.

Schlessinger, J. (2003). SIGNAL TRANSDUCTION: autoinhibition control. Science *300*, 750–752.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M., and Fersht, A.R. (1990). Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. Biochemistry *29*, 9343–9352.

Sheinerman, F.B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. Curr. Opin. Struct. Biol. *10*, 153–159.

Shewchuk, L.M., Hassell, A.M., Ellis, B., Holmes, W.D., Davis, R., Horne, E.L., Kadwell, S.H., McKee, D.D., and Moore, J.T. (2000). Structure of the Tie2 RTK domain: self-inhibition by the nucleotide binding loop, activation loop, and C-terminal tail. Structure *8*, 1105–1113.

Stamos, J., Sliwkowski, M.X., and Eigenbrot, C. (2002). Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. J. Biol. Chem. *277*, 46265–46272.

Walker, F., Kato, A., Gonez, L.J., Hibbs, M.L., Pouliot, N., Levitzki, A., and Burgess, A.W. (1998). Activation of the Ras/mitogen-activated protein kinase pathway by kinase-defective epidermal growth factor receptors results in cell survival but not proliferation. Mol. Cell. Biol. *18*, 7192–7204.

Wedegaertner, P.B., and Gill, G.N. (1992). Effect of carboxyl terminal truncation on the tyrosine kinase activity of the epidermal growth factor receptor. Arch. Biochem. Biophys. *292*, 273–280.

Wood, E.R., Truesdale, A.T., McDonald, O.B., Yuan, D., Hassell, A., Dickerson, S.H., Ellis, B., Pennisi, C., Horne, E., Lackey, K., et al. (2004). A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. Cancer Res. *64*, 6652–6659.

Yarden, Y., and Schlessinger, J. (1987). Epidermal growth factor induces rapid, reversible aggregation of the purified epidermal growth factor receptor. Biochemistry *26*, 1443–1451.

Yarden, Y., and Sliwkowski, M.X. (2001). Untangling the ErbB signalling network. Nat. Rev. Mol. Cell Biol. *2*, 127–137.

Yu, X., Sharma, K.D., Takahashi, T., Iwamoto, R., and Mekada, E. (2002). Ligand-independent dimer formation of epidermal growth factor receptor (EGFR) is a step separable from ligand-induced EGFR signaling. Mol. Biol. Cell *13*, 2547–2557.

An Automatic Method for Predicting Transmembrane Protein Structures Using Cryo-EM and Evolutionary Data

Sarel J. Fleishman,* Susan Harrington,† Richard A. Friesner,† Barry Honig,‡ and Nir Ben-Tal*
*Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat-Aviv 69978, Israel; †Department of Chemistry, Columbia University, New York, New York 10027 USA; and ‡Department of Biochemistry and Molecular Biophysics, Columbia University and Howard Hughes Medical Institute, New York, New York 10032 USA

ABSTRACT The transmembrane (TM) domains of many integral membrane proteins are composed of α-helix bundles. Structure determination at high resolution ($<$4 Å) of TM domains is still exceedingly difficult experimentally. Hence, some TM-protein structures have only been solved at intermediate (5–10 Å) or low ($>$10 Å) resolutions using, for example, cryo-electron microscopy (cryo-EM). These structures reveal the packing arrangement of the TM domain, but cannot be used to determine the positions of individual amino acids. The observation that typically, the lipid-exposed faces of TM proteins are evolutionarily more variable and less charged than their core provides a simple rule for orienting their constituent helices. Based on this rule, we developed score functions and automated methods for orienting TM helices, for which locations and tilt angles have been determined using, e.g., cryo-EM data. The method was parameterized with the aim of retrieving the native structure of bacteriorhodopsin among near- and far-from-native templates. It was then tested on proteins that differ from bacteriorhodopsin in their sequences, architectures, and functions, such as the acetylcholine receptor and rhodopsin. The predicted structures were within 1.5–3.5 Å from the native state in all cases. We conclude that the computational method can be used in conjunction with cryo-EM data to obtain approximate model structures of TM domains of proteins for which a sufficiently heterogeneous set of homologs is available. We also show that in those proteins in which relatively short loops connect neighboring helices, the scoring functions can discriminate between near- and far-from-native conformations even without the constraints imposed on helix locations and tilt angles that are derived from cryo-EM.

INTRODUCTION

TM proteins are crucial mediators of cell-to-cell signaling and transport processes, and constitute some 50% of contemporary drug targets (Fleming, 2000). In recent years the pace of structural determination of TM proteins has increased, but technical problems related to protein purification and crystallization still hamper TM-protein structure determination. Thus, despite their biomedical importance, $<$40 distinct folds of TM proteins have been solved to date by high-resolution methods such as x-ray crystallography. The lack of a large set of solved TM proteins also restricts the usefulness of computational methods based on the statistics of solved protein structures, and in particular, of comparative or homology modeling, which has been a very successful approach in soluble proteins.

In general, computational prediction of soluble-protein structures is difficult, largely because of the variety of possible folds, which implies a vast number of degrees of freedom. In contrast, all TM proteins that inhabit the plasma membrane of eukaryotic cells form α-helix bundles, thus reducing the desolvation penalty of exposing polar main-chain groups. The high propensity to form secondary structures reduces the number of degrees of freedom, which determine the protein's fold, and hence, lowers the complexity of predicting the structures of these proteins.

Structure prediction of TM proteins often relies conceptually on the two-stage model for protein assembly in the membrane (Popot and Engelman, 1990). According to this model, the first step of folding is the insertion of the TM domains into the membrane as α-helices. Only in the second stage do these helices associate to form bundles (reviewed by White and Wimley, 1999 and Popot and Engelman, 2000). One of the implications of the two-stage model is that, overall, the stability of individual TM domains is independent of that of other domains. Hence, prediction of TM-protein structures can begin with experimental determination (or prediction, reviewed by von Heijne, 1996 and Chen et al., 2002) of the locations of the TM helices in the amino-acid sequence of the protein.

Some early attempts were made to predict helix orientations relative to one another by using the concept of the hydrophobic moment (Eisenberg et al., 1984; Rees et al., 1989). However, in view of the low-dielectric character of the membrane, the hydrophobic driving force is probably less dominant in this medium than in soluble proteins, and the hydrophobic moment proved to be of limited use in TM-protein structure prediction (Pilpel et al., 1999; Stevens and Arkin, 1999).

Attempts were also made to predict the structures of specific TM proteins or protein families (Tuffery and Lavery, 1993; Stokes et al., 1994; Taylor et al., 1994;

Submitted May 24, 2004, and accepted for publication August 12, 2004.

Address reprint requests to Nir Ben-Tal, Dept. of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel. Tel.: 972-3-640-6709; Fax: 972-3-640-6834; E-mail: bental@ashtoret.tau.ac.il.

0006-3495/04/11/3448/12 $2.00

doi: 10.1529/biophysj.104.046417

Adams et al., 1995; Baldwin et al., 1997; Heymann and Engel, 2000; Hirokawa et al., 2000; Zhdanov and Kasemo, 2001; Sorgen et al., 2002; Trabanino et al., 2004). For high-resolution structure prediction of pairs of TM α-helices, a method that was based on molecular dynamics was developed, in which data derived from large-scale mutational assays were utilized to derive constraints for the conformation search (Adams et al., 1995). Extensions to this method were suggested, which used phylogenetic instead of mutational data (Briggs et al., 2001) and lowered the computational load associated with the conformation search (Pappu et al., 1999). Recently, a method based on Monte-Carlo sampling of conformations, which selects tightly packed conformations, was shown to reproduce the structures of homooligomers (Kim et al., 2003). Another method that was founded on a knowledge-based potential constructed on the basis of TM proteins of known structures and energy terms that simulate the membrane environment was also shown to retrieve the conformations of small homooligomers (Pellegrini-Calace et al., 2003).

A major limitation of many of the methods in this class is the large computational load. In fact, computational complexity has restricted the applicability of these methods mostly to the cases of homooligomers of single-spanning TM proteins. A more fundamental handicap is the reliance of many of these methods on contemporary force fields. Recent results indicate that the forces specifying and stabilizing TM-helix interactions are still unclear (Bowie, 2000), casting doubt on the ability of methods based on existing force fields to yield accurate predictions.

We recently examined the possibility of reducing the computational burden by using low resolution from the outset (Fleishman and Ben-Tal, 2002), i.e., by considering only the helices' C^{α} traces. We developed a scoring function and a search methodology to seek stable conformations of pairs of closely packed TM helices. The use of a reduced representation of the helices allowed us to conduct an exhaustive search of conformation space, and to test the method systematically on many different examples. This approach proved useful in studying the involvement of the TM domain in the activation of the erbB2 receptor tyrosine kinase (Fleishman et al., 2002). However, it could only be applied reliably to helix pairs that are closely packed (<9 Å separation between the helix axes) (Fleishman and Ben-Tal, 2002). Because many of the helices in TM proteins have greater interhelical separations (Bowie, 1997), in general, this method cannot be used to predict entire protein domains.

Here, we explored whether such an approach can be extended to deal with large TM domains by incorporating the evolutionary-conservation profile of the protein and the hydrophobicity of its constituent amino-acid residues. The underlying idea is that amino-acid positions that mediate interhelical contacts would be more evolutionarily conserved than those that face the lipid (Donnelly et al., 1993; Stevens and Arkin, 2001; Beuming and Weinstein, 2004), because mutation of positions that form contact would most likely destabilize the protein, and render it dysfunctional (Fig. 1). Hydrophobicity can be used to discard potential conformations that expose charged positions (e.g., Arg and Glu) to the membrane environment (Cronet et al., 1993) due to the prohibitive cost in desolvation of their highly polar side chains (Honig and Hubbell, 1984).

To reduce the computational burden associated with conformational searches of large TM domains, the targets for our approach are those proteins for which intermediate-resolution (5–10 Å in-plane) structural data are available, e.g., from cryo-EM (Unger, 2001). At such resolution, cryo-EM maps reveal the organization of TM helices relative to one another including the helices' positions and tilt angles, but do not disclose the locations of the individual amino acids. Based on the cryo-EM data, it is possible to approximate the helices' principal axes either manually (Baldwin et al., 1997; Fleishman et al., 2004) or computationally (Jiang et al., 2001). Then, the conformational search

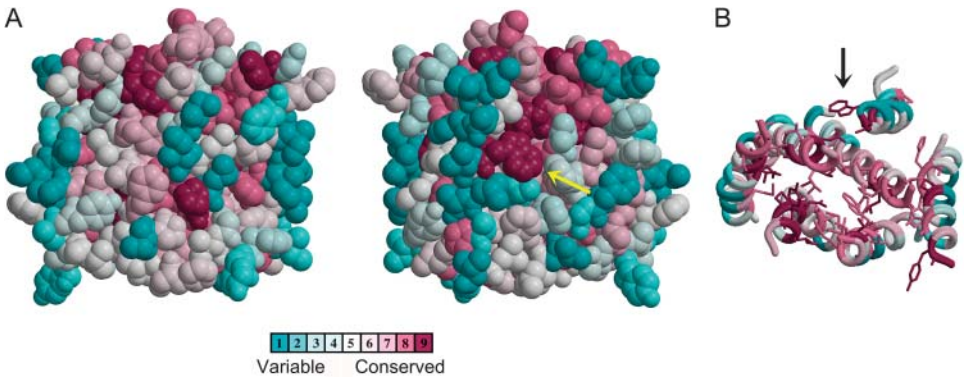


FIGURE 1 The conservation profile of the TM domain of rhodopsin (PDB code 1l9h). Conservation scores were computed using the ConSurf server with the Rate4Site algorithm (Pupko et al., 2002), and are mapped according to the color scale with turquoise through burgundy signifying variable through conserved positions. (*A*) Two side views looking from within the membrane plane. The space-filling models show that the lipid-facing parts of the protein are mostly variable. (*B*) Looking from the cytoplasmic side. Stick models of residues that belong to the two highest categories of the conservation scale (8 and 9) are indicated. The vast majority of these highly conserved residues face the protein interior. The arrows identify the highly conserved Trp-161, which is exceptional in that it is exposed to the membrane despite its high conservation. This and all other molecular representations were generated with MOLSCRIPT (Kraulis, 1991) and rendered with Raster3D (Merritt and Bacon, 1997).

need only explore the orientations of the helices around their principal axes.

Intermediate-resolution cryo-EM maps of TM proteins often provide accurate data on the lateral positions of the helices and their tilt angles within the lipid bilayer, but much poorer data on the positions of the helices along the vertical axis (Unger and Schertler, 1995; Unger et al., 1999). In this study, we limited the methods' validation to the hydrophobic portion of each of the TM helices. As these segments are most likely to align with one another within the hydrophobic core of the lipid bilayer, the inaccuracy due to the low vertical resolution of cryo-EM data does not present a significant problem. In a refinement stage of the conformational search described below, a limited exploration of all degrees of freedom, including the vertical axis, was conducted.

Baldwin et al. (1997) used a similar approach to predict the orientations of helices in rhodopsin based on the receptor's cryo-EM map at 9 Å in-plane resolution (Unger et al., 1997). This prediction was shown (Bourne and Meng, 2000) to compare very well with the high-resolution structure, which was solved a few years later (Palczewski et al., 2000). However, Baldwin et al.'s conservation analysis was highly labor intensive and required substantial subjective intervention at various stages (Baldwin et al., 1997), making it difficult to apply to a large set of proteins. As conservation analyses have grown in rigor and sophistication in recent years, we have employed automatic and more sensitive tools, to construct score functions for ranking conformations of TM proteins. This has allowed us to test various formulations of the prediction rule and search methodology on a variety of TM proteins. The tests were based on perturbations of the native-state structures as they are found in the PDB, except in the case of rhodopsin, in which they were conducted using data extracted (Baldwin et al., 1997) from its cryo-EM map at 9-Å resolution (Unger et al., 1997).

Our analysis leads us to conclude that an approach based on evolutionary conservation, hydrophobicity, and intermediate-resolution structures can retrieve near-native structures subject to two principal requirements. First, the cryo-EM map must show that all helices have a face that is buried in the protein bundle and another that is exposed to the membrane milieu or the pore lumen. This requirement is necessary because it is only the heterogeneity of environments that allows the correct orientation of the helices. Second, evolutionarily conserved and variable residues must be distributed in the TM domain in accordance with a helical pattern (Fig. 1). This distribution ensures that a clearly higher score is assigned to an orientation, in which conserved residues face the interior of the helix bundle, whereas the variable residues are directed toward the lipid. Hence, a typical case in which this approach is expected to yield a near-native structure is a protein or an oligomer, where all helices face the lipid environment or a relatively large internal pore, and a sufficiently heterogeneous set of sequences are available.

Score functions

In developing the conformation-search methodology and the score functions, we initially used the structure of bacteriorhodopsin for parameterization (Luecke et al., 1998). That is, various formulations of the scoring function were attempted with the aim of detecting the native structure among contending templates. For instance, formulations that gave a more dominant effect to hydrophobicity were found to do more poorly than the formulation that is given below, which stresses conservation, in agreement with the notion that the hydrophobic moment is a relatively poor indicator of helix orientations (Pilpel et al., 1999; Stevens and Arkin, 1999).

The so-called burial function, which we first introduced in Fleishman and Ben-Tal (2002), is a major component of the scoring schemes defined here. It is an estimate of the extent of an amino acid's contact with another helix. Because the model describes amino acids merely in terms of C^{α} positions, only an approximate measure of contact can be attained. To achieve this approximation, the function considers the distance between an amino acid's C^{α} position and the other helix's principal axis. It also considers the angle formed between two vectors: one that connects the two helix axes, and a second that connects the C^{α} position to its own axis (Fleishman and Ben-Tal, 2002). If both the angle and the distance are small, the burial function is assigned relatively high values ($\rightarrow 1$). Low values ($\rightarrow 0$) are assigned otherwise.

This burial function takes into account the details of the local interactions of the helices. The alternative use of a moment to account for hydrophobicity or conservation treats all helices as being perpendicular to the membrane plane (e.g., Eisenberg et al., 1982; Pilpel et al., 1999), thus giving a particular helix face the same weight in computing the optimal conformation throughout the TM span. In contrast, the use of the burial function tests the extent of contact for each amino-acid residue, and treats each position according to its actual contact with other helices, thus treating tilted and kinked helices more realistically (Fleishman and Ben-Tal, 2002).

We used three schemes for ranking template conformations. The simplest form is the ''singlewise'' score (Fleishman et al., 2004). This function assigns a high score to conformations that bury conserved faces in the α-helix bundle, and expose the helices' variable faces to the lipid. The function is singlewise in the sense that for any given amino acid, only the locations of the axes of its neighboring helices are taken into account. Because these locations can be derived from the cryo-EM data to a reasonable degree of confidence, the contributions of each amino-acid residue to the overall score is independent of the positions of other residues. The underlying notion in the singlewise score is that positions that are buried in the protein core are typically conserved evolutionarily (Fig. 1). Indeed, some conserved positions may be exposed to the membrane in contradiction to this ''rule'' (see the arrows in Fig. 1). However, summation

across the entire helix span reduces the prediction's sensitivity to such cases.

Another term penalizes the exposure to the lipid (burial values <0.5) of the most polar amino-acid residues that are associated with high (>7 kcals/mol) desolvation energies upon transfer from water to membrane according to the Kessel & Ben-Tal scale (Kessel and Ben-Tal, 2002). The residues for which the penalty applies are Arg, Asn, Asp, Glu, and Lys. In essence, this term associates conformations that expose very polar residues with very unfavorable scores. Polar residues at the terminal turns (four amino-acid residues) of helices were disregarded in computing this penalty, because at these locations, residues may interact favorably with the relatively polar environment at the lipid-water interface (von Heijne, 1996). Proline residues are ignored in calculating the conservation scores because they are often conserved owing to kinks that they induce in the helix secondary structure rather than to the formation of interhelical contacts (Baldwin et al., 1997).

A second scheme, called the ''pairwise'' function, included, in addition to the singlewise score, a term that favors contact formation between highly conserved residues, and penalizes contacts among highly variable residues. Hence, this function takes into account the positions of pairs of residues in contrast to the singlewise score, which considers residues separately. The underlying concept here is that positions that form contact should be highly conserved, because introducing even mild changes in these positions would abrogate interhelical contact.

The singlewise and pairwise score functions do not include terms that penalize the formation of possible steric clashes between the helices. Generally, the positions and tilt angles can be derived from cryo-EM data. However, these data are potentially inaccurate due to limited resolution. A scoring function that contains an approximation of penalties due to steric clashes could be useful for a limited exploration of the conformation space with respect to helix positions and tilt angles. We thus defined a third score function, which included, in addition to the terms in the pairwise score, penalties for conformations, in which a helix is potentially in violation of another's approximate exclusion volume.

METHODS

Conservation analysis

The conservation of amino-acid residues in the TM domains of the proteins were calculated using the ConSurf server (Glaser et al., 2003) with the Rate4Site algorithm (Pupko et al., 2002). Homologs were collected using 5 PSI-BLAST iterations and a BLAST e-value cutoff of 1 (Altschul et al., 1997). We asserted by visual inspection of the alignments that there were no significant gaps in the TM domains of all the proteins under study.

Score functions

To each configuration of the helix bundle produced by the search method, we assign a score. The score is based on four terms, such that:

1. Hydrophobic residues face the lipid environment and hydrophilic residues are directed toward the protein core.
2. Conserved residues face the protein core and variable residues face the lipid environment (Fig. 1).
3. Highly conserved residues on different helices are in close proximity, whereas highly variable residues are distal.
4. A penalty for potential steric clashes.

For each conformation the score is generally defined as follows, where the summation is on every residue pair i,j in the TM domain:

$$Score = \sum_{i}(2(B^{i} - ½)H^{i} + 2(B^{i} - ½)C^{i}) + \sum_{i,j}(P^{i,j} - Q^{i,j}). \quad (1)$$

In Eq. 1, C^{i} are the normalized evolutionary-conservation scores assigned by Rate4Site (Pupko et al., 2002) (Fig. 1) and H^{i} the desolvation free energies of transfer from water to membrane (Kessel and Ben-Tal, 2002); B^{i} is the burial score associated with each residue, i.e., the extent of that residue's contact with other helices (Fleishman and Ben-Tal, 2002); $P^{i,j}$ is a pairwise term that promotes contact between highly conserved residues and penalizes contact between highly variable residues; and $Q^{i,j}$ is a penalty for formation of severe van der Waals clashes.

High C^{i} and H^{i} values indicate that a residue is conserved and hydrophilic, respectively. Hydrophobicity is taken into account only for residue types that are associated with free energies of transfer >7 kcal/mol (Kessel and Ben-Tal, 2002), and are counted only for residues i, for which the burial scores B^{i} are <0.5. Thus the hydrophobicity scale serves as a significant penalty on the exposure of the most polar residues to the membrane environment. The terminal turns (four amino-acid residues) from each side of the TM segments are ignored in computing this penalty, because residues in these regions may be accommodated by the polar environment at the lipid-water interface (von Heijne, 1989). The contributions of proline residues to the score is also ignored because they are often conserved due to kinks they form in secondary structure rather than due to the promotion of interhelical contacts (Baldwin et al., 1997).

C^{i} and H^{i} are singlewise terms that depend on the amino-acid site itself, regardless of the protein conformation. In contrast, B^{i} is the burial score associated with each residue i, and depends on the maximal contact formed by each residue with other helices in the bundle (elaborated below). It assumes values in the range 0−1, where zero indicates complete exposure to the membrane environment and 1 indicates complete burial in another helix.

Maximization of the score defined in Eq. 1 favors the burial of hydrophilic residues in the α-helix bundle and penalizes their exposure to the membrane (the first term in Eq. 1). Similarly, the second term in Eq. 1 favors the burial of conserved amino acids in the bundle interior and penalizes their exposure to the lipid. The third is a pairwise-contact term favoring contact between well-conserved residues and penalizing contact between highly variable residues.

$$P^{i,j} = B^{i}B^{j}(C^{i} + C^{j}), \quad (2)$$

where residues j and i are not >7 Å apart, and their respective burial scores (B) are >0.2.

The fourth term in Eq. 1, $Q^{i,j}$, produces a severe penalty on steric-clash formation, and is summed on all pairs of residues i, j in the TM domain:

$$Q^{i,j} = \begin{cases} \infty & d^{i,j} \le \Theta \\ \frac{1}{d^{i,j} - \Theta} + \frac{d^{i,j} + \Theta - 2\mu}{(\mu - \Theta)^{2}} & \Theta < d^{i,j} < \mu \\ 0 & d^{i,j} \ge \mu \end{cases}, \quad (3)$$

where $d^{i,j}$ is the distance between residues i and j, Θ is the threshold below which the penalty assumes infinite magnitude, and μ is the threshold above

which the penalty cancels out. We chose this formulation for the penalty because it produces a function that is continuous for $d^{i,j} > \Theta$, as is its first derivative. A value of 2 Å was chosen for Θ, to approximate the C^{α} van der Waals radius (1.88 Å) (Tsai et al., 1999), and 2.5 Å was chosen for μ. The penalty is very large for distances close to 2 Å, but drops off quickly toward zero at 2.5 Å. Thus, conformations are penalized only for severe steric clashes.

We tested different formulations of the score function presented in Eqs. 1–3 by assigning different weights to the various terms, and by using different hydrophobicity scales. This formulation was found to work well in identifying bacteriorhodopsin's native-state structure from decoys. Hydrophobicity appears to be a poor indicator on its own for TM-helix orientations, whereas contact between highly conserved residues is a good indicator.

The singlewise score function is defined as in Eq. 1 (Fleishman et al., 2004), except that the pairwise contact terms P and the penalties on steric clashes Q are neglected. Essentially this score function favors the burial of conserved and hydrophilic residues in the protein core, but does not favor contact between conserved residues. The pairwise score is similarly defined as in Eq. 1 with the penalties for steric clashes being neglected.

Assessing the extent of interresidue contact

The score function defined in Eq. 1 is based on a quantification of the burial of amino acids that mediate interhelical contact. In measuring the extent of burial B^i of amino acid i we consider two criteria, as elaborated by Fleishman and Ben-Tal (2002). The first is the distance between the amino acid and the principal axis of the other helix; the smaller the distance, the more deeply buried the amino acid. The second is the orientation of the amino acid with respect to the principal axis of the other helix; the more the amino acid is directed toward the other helix, the better its burial.

Formally, we consider two parameters: the distance D^i between amino acid i and the axis of the other helix, and the angular orientation A^i of amino acid i with respect to the axis of the other helix. We define the burial of an amino acid as the intersection of these two criteria:

$$B^i = S(D^i)\,S(A^i), \tag{4}$$

where $S(D^i)$ and $S(A^i)$ are transformations of the distance and angular criteria as defined in Eqs. 5 and 6 below.

The parameters used by Fleishman and Ben-Tal (2002) for the burial function B were tailored specifically to TM-helix pairs with short interaxial separations. In the more general case treated here, it was necessary to reparameterize the function. By manually modulating these parameters with regard to the structure of bacteriorhodopsin, we found the parameter values $t = 60°$ and $p = 4$ to be suitable for transformation of the angle A^i. For transformation of the distance, we first subtract 4.3 Å from the value of D^i calculated for the distance between the amino acid and the axis of the other helix. This value approximates the smallest possible distance between an amino acid and another helix (the radius of an α-helix to its C^{α} atoms is 2.3 Å plus 2 Å for two exclusion radii), and approaches a value of 1 for $S(D^i)$ if the amino acid is as close as possible to the axis of the other helix. The parameter values chosen for transformation of the distance are $t = 10$ Å and $p = 6$. Thus the two transformations for amino acid i are:

$$S(D^i) = \frac{1}{\left(\dfrac{D^i - 4.3}{10}\right)^6 + 1} \tag{5}$$

$$S(A^i) = \frac{1}{\left(\dfrac{A^i}{60}\right)^4 + 1}, \tag{6}$$

where A^i and D^i are expressed in degrees and Ångstroms, respectively.

Conformation search in TM proteins with short loops

In those cases, where the TM helices are connected via short loops, e.g., rhodopsin, it is possible to sample the constrained conformation space available to the α-helix bundle by using a modification of the method of Monge et al. (1994), in which α-helices are treated as rigid bodies, and their exclusion volumes and the lengths of the interconnecting loops are taken into account. The software and low-resolution potential used were developed by Eyrich et al. (1999) (J. Gunn, private communication).

We began with the native-state structure, and systematically perturbed the helix positions as follows. One helix was selected and moved around its close-contact interfaces with other helices by shifting up and down, twisting, and rotating; all of these changes were made by adding appropriate quadratic bonus functions to the low-resolution potential and minimizing. The resulting structures were then used as starting points for another round of minimization of the low-resolution potential. In both cases, another bonus function was added to the potential to help reward the TM orientations of the helices. (Because in this software the conformational space is given in Φ-Ψ coordinates and no consistent embedding into Euclidean space is done by the program, it was not possible to impose the membrane constraints in the straightforward way.) This membrane function was based on the distances between the termini of all of the helices besides the one designated to move. It rewarded those intertermini distances (excluding those of the selected perturbed helix) that remained within 4.5 Å of their original values. Thus steric clashes resulting from the helix perturbations would tend to be resolved inside the membrane, and conformations that did not respect the TM orientations were penalized.

Several rounds of this procedure were completed using the best-scoring structures as the initial structures to perturb. The resulting structures were then screened for steric clashes and inappropriate TM orientations using the energy functions, and finally clustered at 0.8 Å to produce our test set.

RESULTS

Rhodopsin and the bacterial rhodopsins

We used rhodopsin as our main test case because it represents the typical case for which the method is intended. That is, it is a medium-size protein (7 TM segments), which has been solved at intermediate in-plane resolution (9 Å) (Unger et al., 1997), and shares sequence homology with a large set of other G-protein-coupled receptors. Moreover, its high-resolution structure (2.8 Å) (Palczewski et al., 2000) allows us to test the prediction's quality.

Baldwin et al. (1997) used the intermediate-resolution cryo-EM maps of rhodopsin (Unger et al., 1997), as well as conservation data, to manually infer a template structure, which included the coordinates of C^{α} atoms. We did not use their model structure of rhodopsin, but did employ the helix-tilt angles and positions that they extracted from the cryo-EM maps (Baldwin et al., 1997). The assignment of individual TM segments to the helices seen in the cryo-EM maps was also taken from Baldwin et al.'s analysis. In addition, we used their data on the positions, directions, and extents of kinks in the TM domain. In summary, the C^{α} positions of each helix were generated according to the helix parameters of canonical α-helices as observed in the intermediate-resolution data (Unger et al., 1997).

To test the singlewise function's performance, each helix was rotated in 5° increments around its principal axis (range: 0–360°), and its best-scoring orientation was selected. Because the contribution to the singlewise score of each of the helices is essentially independent of that of the others, we superimposed the best-scoring orientations of each of the seven helices to obtain an optimal template structure. The root-mean-square deviation (RMSd) of this template from the native-state structure of rhodopsin was 3.7 Å.

The search in orientation space is confined within a seven-dimensional hypercube, where each degree of freedom sets the orientation of one of the seven helices. To calculate the distribution of RMSd values of conformations within this hypercube to the native-state structure of rhodopsin, we generated 2000 template conformations. In each of these templates, every helix's orientation was randomly selected from a distribution with uniform probability in the range 0–360°. The RMSds of each of these templates from the native-state structure of rhodopsin (Palczewski et al., 2000) was then computed (Fig. 2). The optimal structure was found within the lowest 3.5 percentiles of RMSd values, demonstrating that even the relatively simple singlewise score function is capable of retrieving a near-native structure from a set of decoys (Table 1).

We also tested the singlewise score on the three homologous bacterial rhodopsins, bacterio-, halo-, and sensory rhodopsin II (PDB codes are 1c3w, 1e12, and 1jgj, respectively). These three proteins share ~30% sequence identity and their structures are quite similar (1–1.7 Å RMSd; Fischer et al., 1992), but show some local structural differences and no homology with rhodopsin. We extracted the helix-axes parameters (tilt angles and positions) (Fleishman and Ben-Tal, 2002) from the proteins' high-resolution structures, and constructed canonical α-helices accordingly, without modeling explicitly any deviations from helicity, such as kinks and bulges. We then employed the singlewise score and searched the conformation space (seven-dimensional hypercube) exhaustively in the same manner as explained above for rhodopsin. Table 1 summarizes the results of the conformation searches. In all cases, as in rhodopsin, the singlewise score detected templates that were much closer to native than expected by chance.

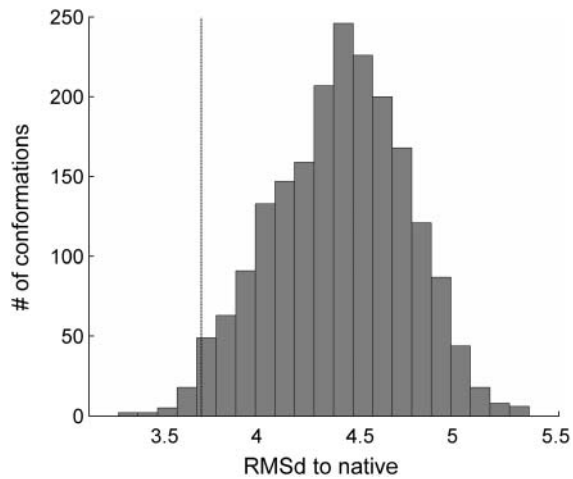


FIGURE 2 A histogram of RMSd values to the native-state structure of 2000 randomly generated templates of rhodopsin. The templates were constructed according to the helical axes parameters obtained (Baldwin et al., 1997) from the cryo-EM data of rhodopsin at 9 Å in-plane resolution (Unger et al., 1997). The RMSd of the conformation with the best singlewise score (3.7 Å from native) is marked by a dashed line, a value that is at the lowest 3.5 percentiles of the random conformations.

Using the result of the exhaustive singlewise search as a starting template structure of rhodopsin, we conducted a conformation search employing the pairwise score function that avoids steric clashes, and the Simplex optimization method, which is a line-search algorithm for finding a local optimum (Nelder and Mead, 1965). The RMSd of the predicted structure from the native state (PDB code 1l9h) was 3.1 Å, which is an improvement over the result obtained by using the singlewise score function alone (3.7 Å). This result is comparable with that obtained by Baldwin et al. (3.2 Å) (Baldwin et al., 1997). We tested whether subsequent use of the two scores constitutes a viable search strategy on the three homologous bacterial rhodopsin structures. However, in these cases the pairwise score improved the RMSd of the predicted conformations only marginally (data not shown).

The acetylcholine receptor

The nicotinic acetylcholine receptor (AchR) transfers the electrical signal at the nerve-muscle synapse by the gating of its TM pore (Hille, 2001). The channel is composed of five homologous subunits (β, γ, δ, or ε, and two α-subunits), where each monomer consists of four TM domains (M1–M4). The five M2 segments from each of the subunits line the pore. The recently solved structure of the closed AchR at 4-Å resolution revealed an unexpected architecture, in which the M2 helices appear to be embedded in water and surrounded by an outer ring of the other TM helices (Miyazawa et al., 2003), to which they form only a very loose attachment. These loose contacts are thought to facilitate the substantial changes in the orientations of the M2 helices (Unwin, 1995).

We constructed a model of the AchR TM domain by deriving the helix-tilt angles and positions (Fleishman and Ben-Tal, 2002) from its native-state structure (PDB code 1oed). Canonical α-helices that fit the parameters of these helix axes were then constructed. To predict the optimal structure based on the pairwise score, we sampled 20,000 different combinations of orientations of the four helices comprising a subunit. Fivefold symmetry across the AchR subunits was enforced, and the best-scoring conformation according to the pairwise score was selected. In contrast to the cases of the rhodopsins, the relatively small number of helices in each monomer of the AchR ensures that this number of

TABLE 1 Summary of the results of using the singlewise score function to calculate a near-native conformation of rhodopsin and the three bacterial rhodopsins, bacterio-, halo- and sensory rhodopsin II

Protein	RMSd of randomly generated conformations (± SD) Å	RMSd of the highest-score conformation from the native-state structure (Å)	Percentile of highest-scoring conformation
Bacteriorhodopsin	3.9 ± 0.4	3.2	5.6
Halorhodopsin	3.3 ± 0.4	2.5	4.2
Sensory rhodopsin II	3.5 ± 0.4	1.8	0.01
Rhodopsin	4.5 ± 0.4	3.7	3.5

The three bacterial proteins are related to one another in terms of sequences and structures, but show some local structural differences. Rhodopsin is different in terms of architecture and sequence. Templates for the three bacterial rhodopsins were constructed on the basis of their high-resolution PDB structures. Rhodopsin's templates were constructed on the basis of helix-axes parameters (Baldwin et al., 1997) taken from its 9-Å in-plane resolution structure (Unger et al., 1997). Percentiles were computed on the basis of a distribution of expected RMSd values for each protein (see Results). In all cases, the best-scoring conformation is significantly closer to the native state than predicted by chance.

orientations will adequately cover the conformation space. This search yielded a structure that was 2.5 Å RMSd from the native-state structure (Miyazawa et al., 2003) (Fig. 3 *A*).

In this predicted conformation (Fig. 3 *A*) the orientations of helices M1 and M3 match the native state quite closely, except for deviations from helical ideality in M3. Helix M4 is largely exposed to the lipid (Fig. 3 *B*), a feature not typical of other solved TM protein structures, which usually show tighter interhelical interactions. Owing to this exposure, there is a larger degree of uncertainty concerning the prediction of this helix's orientation, and indeed the optimal orientation is skewed by ~100° relative to the native state. The predicted orientation of M2 is offset to a slightly lesser extent. The reason for the deviation of M2 from the native state is that this helix is conserved quite homogeneously throughout the segment (Fig. 3 *B*). The lack of a clear conservation versus variability pattern precludes this helix's orientation with confidence.

Constraints imposed by short interconnecting loops instead of by cryo-EM data

Many of the extramembrane loops that connect TM helices are relatively short (<10 amino-acid residues) (Tusnady and Simon, 1998). In principle, such short loops can impose severe constraints on the conformation space that a pair of helices is free to sample. Here, we were interested in testing whether considering the constraints imposed by loop lengths improves the prediction's quality.

For conformation sampling, we adapted a technique that was developed by Monge et al. (1994) for sampling the conformations of secondary-structural elements in soluble proteins. The method starts from the native-state structure of the protein, and perturbs the secondary-structural elements' positions and tilt angles while treating them as rigid bodies. In contrast, the regions of the interconnecting loops that are devoid of defined secondary structure are allowed to sample conformations freely.

To construct a complete native-state structure, we added the positions of the loop residues that are missing from the PDB structure (1l9h). These missing loop residues were built into our native state via minimization of our low-resolution energy function of these loop residues, whereas the rest were constrained to their positions as observed in the PDB structure. The native state was then systematically perturbed, and the resultant conformations were assessed with a low-resolution energy function to penalize the formation of steric clashes and covalent-bond strains. Nonphysical conformations were thus penalized (Monge et al., 1994). Hence, the constraint on the helices' positions and tilt angles is that the lengths of the interconnecting loops are respected.

Another penalty was imposed on TM helices that assumed a nontransmembrane orientation, i.e., for helices whose termini were not located on opposite sides of the presumed

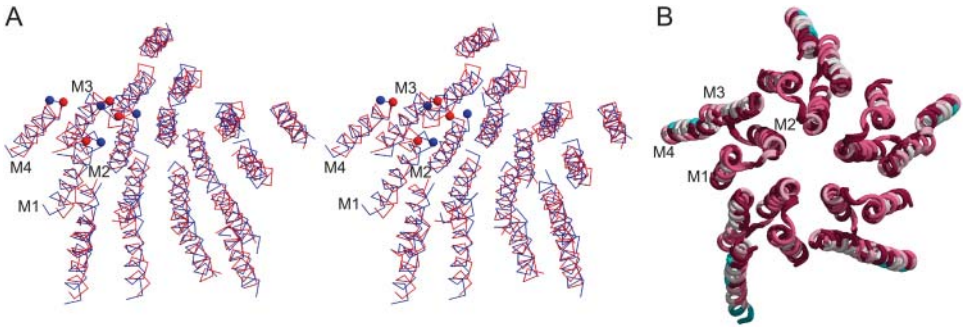


FIGURE 3 (*A*) A stereo view of the TM domain of AchR (*blue*) superimposed on the predicted template (*red*). Spheres mark the positions of the cytoplasmic ends of the helices for clarity. The RMSd between the native-state and the calculated structures is 2.5 Å. Helices M1 and M3 were predicted quite accurately, but helices M2 and M4 were skewed by 90 and 100°, respectively. (*B*) A view of the AchR structure from the cytoplasmic side. The residues are colored according to the evolutionary-conservation scale shown in Fig. 1. M2 is homogeneously conserved explaining the inaccurate prediction. M4 is highly exposed to the membrane. Hence, despite the clear conservation signal, there is a large degree of uncertainty in its orientation.

membrane. Hence, the search method samples conformation space that is available to the helix bundle, but penalizes nonphysical orientations. Structures with high penalties were then discarded to eliminate those that were clearly nonphysical.

Based on the high-resolution structure of rhodopsin (PDB code 1l9h) as the template structure, we generated 108 modified templates, each differing from all the others by at least 0.8 Å RMSd (Table 2). The structures were quite evenly distributed in conformation space; sampled conformations were up to 6.2 Å RMSd from rhodopsin's native-state structure.

Because the conformation-sampling method usually does not generate conformations that form steric clashes (Monge et al., 1994), we used the pairwise score without the terms that penalize the formation of clashes. We note that in ranking the resultant conformations, the score did not incorporate any terms from the Monge et al. (1994) conformational sampling technique. Strikingly, the native-state structure of rhodopsin ranked second according to the pairwise function (Table 2), demonstrating that short interconnecting loops may indeed be used for identifying near-native conformations, even without the constraints on helix positions and tilts derived from cryo-EM data.

A more stringent criterion, testing the Pearson correlation coefficient between the conformations' scores and their RMSds from the native-state structure, resulted in $r = -0.78$ (Fig. 4). This high anticorrelation demonstrates that the pairwise score is capable not only of detecting the native-state conformation, but also of discriminating near-native and far-from-native conformations. We also analyzed the performance of this combination of pairwise score and search method on the structures of bacteriorhodopsin and aquaporin-1 (PDB codes 1c3w and 1j4n, respectively). The results are summarized in Table 2. Despite the sequence, structural, and functional heterogeneity of the three proteins, the results for all are encouraging.

Deviations from α-helicity have only a local effect on the prediction's quality

Many TM helices exhibit deviations from α-helicity, including π-bulges and kinks. These deviations were shown to have functional importance in some cases (Ubarretxena-Belandia and Engelman, 2001). Kinks are sometimes discernible in cryo-EM maps, e.g., in rhodopsin's 9-Å map (Unger et al., 1997). When observed, the kinks can be incorporated into the conformational search methodology in a straightforward manner, as we have done for rhodopsin above. Recently, it was shown that the positions and directions (though not the magnitudes) of the majority of the kinks observed in high-resolution structures could also be inferred from sequence data alone (Yohannan et al., 2004). However, no computational method is yet available to identify π-bulges.

Fig. 5 shows the consequences of modeling as α-helices domains that contain π-bulges and bent regions in the case of sensory rhodopsin II. As mentioned above, to generate the calculated template (Fig. 5 *B*), the tilt angles and positions of the helix axes were inferred from the high-resolution structure (PDB code 1jgj), and canonical α-helices were constructed. The singlewise score was then used to rank all the possible orientations of each of the helices, and the best-scoring conformation was selected (Fig. 5 *B*). Obviously, the prediction's accuracy in the region surrounding the deviations from helicity is relatively low, but is quite high in other regions of the same helices, and in other helices (RMSd of the prediction from the native-state structure is 1.8 Å). Hence, we conclude that the adverse effects of helical deviations on the prediction quality are mostly local.

Uncertainties in the TM helix boundaries have a negligible effect on the prediction's accuracy

Even when helix positions and tilts are derived reliably from cryo-EM measurements, different TM boundaries can be fitted into the intermediate-resolution images. Qualitatively, changes at the TM-domain termini are not expected to have very large effects on the prediction's quality according to the scoring schemes suggested here, because the calculations are based on the average properties of relatively long helical stretches (5–6 helical turns).

To examine the implications of erroneous choices of the boundaries, we changed the boundaries of the TM spans in the construction of templates of rhodopsin and reevaluated the prediction. Juxtamembrane regions are often spotted by charged residues. Because the score functions penalize

TABLE 2 Summary of results using a modified version of the conformation-sampling method of Monge et al. (1994) in conjunction with the pairwise score function

Protein	Number of structures sampled	Maximal RMSd from native of sampled structures (Å)	RMSd of the highest-score conformation from the native-state structure (Å)	Score rank of the native structure	Correlation coefficient (r) of RMSd values versus pairwise scores
Rhodopsin	109	6.2	1.5	2	−0.78
Bacteriorhodopsin	96	4.0	1.9	30	−0.54
Aquaporin-1	26	3.7	0.9	6	−0.63

The three TM proteins that were tested are heterogeneous in terms of functions, structures, and sequences. The anticorrelations obtained in all three cases demonstrate that the pairwise score is capable of ranking conformations according to their similarity to the native-state structure in a variety of cases.

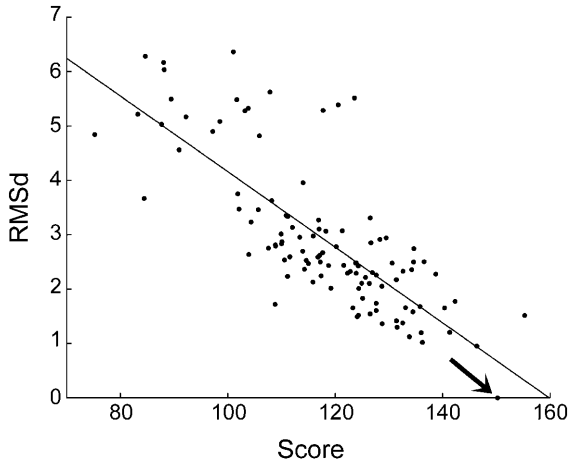


FIGURE 4 A scatter plot showing the RMSd values from the native state (PDB code 1l9h) versus the pairwise score for 109 different template structures of rhodopsin. The two measures are anticorrelated ($r = -0.78$). The solid line marks the linear regression of the data points. The arrow marks the point of the native state structure.

conformations that expose very polar residues to the lipid environment, we tested only helix stretches that are shorter than the TM-domain definitions. Thus, in each iteration, every helix was shortened by variable amounts according to a uniform-probability distribution (0−4 positions). We drew 200 such domain definitions, and used the singlewise score to identify a near-native conformation for each of these definitions according to the method outlined above.

The RMSd values of the highest-scoring conformations to the native-state structure of rhodopsin for this sample were very dense around 3.7 Å, which is the value obtained for the original TM-boundary definition, with a standard deviation of 0.1 Å. This result demonstrates that the score function is indeed minimally sensitive to moderate changes in the hydrophobic boundaries.

DISCUSSION

Structure determination of TM proteins at high resolution remains an intricate task despite recent advances. On the other hand, several TM proteins have been solved at intermediate resolution (5–10 Å). These data have mostly been employed to gain a general understanding of the proteins' architectures, but the positions of individual amino-acid residues could not be inferred (e.g., Holm et al., 2002; Ubarretxena-Belandia et al., 2003)). Hence, it has been impossible to gain a clear view of the molecular determinants affecting protein stability and function from these data. Here, we have explored how TM helices' conservation profiles and hydrophobicity can be used in conjunction with data on helix tilts and positions for structure prediction.

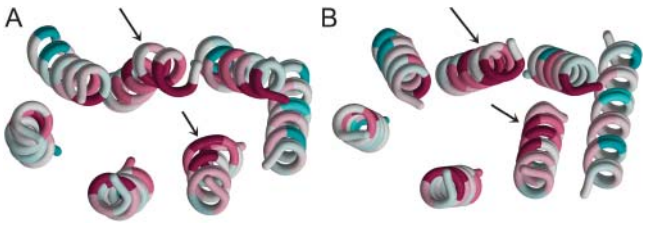


FIGURE 5 (*A*) A view from the extracellular side of the TM domain of sensory rhodopsin II (PDB code 1jgj). The locations of a π-bulge and a kink are marked with arrows. (*B*) The template of sensory rhodopsin II that was assigned the highest singlewise score. Even though the calculated template shown in panel *B* is based on canonical α-helices, the deviations from α-helicity have a minor effect on the calculated conformation. Panels *A* and *B* are colored according to the evolutionary-conservation scale shown in Fig. 1.

We employed accurate measures of conservation (Pupko et al., 2002) and hydrophobicity (Kessel and Ben-Tal, 2002) in a fully automated method. Such measures have been used previously to predict structures from cryo-EM maps (e.g., Baldwin et al., 1997), but these methods were mostly manual, and often required an alignment of a large number of homologous sequences. Here, we showed that even a relatively small set of sequences (36 in the case of the bacterial rhodopsins) may be sufficient to engender accurate predictions thanks to the more sensitive measures of conservation that are currently available (Pupko et al., 2002).

Importantly, the fact that the methods are automatic provides a more objective and reproducible way of modeling TM domains. In particular, in many cryo-EM maps of TM proteins, the connectivity between helices is not discernible, leading to an ambiguity with regard to the assignment of hydrophobic sequences to the helices seen in the map (e.g., Ubarretxena-Belandia et al., 2003). In principle, there may be up to $n!$ different assignments, where n is the number of helices in the bundle. In practice, many of the assignments may be eliminated at the outset if they imply the connection of distant helices by short loops (Enosh et al., 2004). In some cases, biochemical data may provide sufficient constraints for assignment, e.g., regarding the positions of pore-lining helices (Fleishman et al., 2004). Still, it may be that several contending assignments would need to be carefully considered in view of experimental data (Enosh et al., 2004). The methods we have suggested can be helpful in automatically generating and comparing models for different assignments, in which the combinatorial complexity would preclude manual model building.

Thus, after parameterization using bacteriorhodopsin, we tested and challenged this approach with a variety of different TM-protein structures, including rhodopsin, bacterial rhodopsins, aquaporin 1, and the AchR. We have used several different search methodologies for structure prediction, and all produced relatively promising results. This is encouraging, because it demonstrates that the score functions are robust, in the sense that their outcomes are sound independently of the search method used.

Our study has yielded a number of rules that must be met for the protein under study, if this approach is to succeed. First, the cryo-EM map must show that each helix is neither overly buried in the protein core nor overly exposed to the

membrane (or the pore lumen in the case of large channels). Accordingly, it is due to the uncharacteristic exposure of the M4 helix in AchR that its calculated orientation is far from the native state (Fig. 5). Second, the conservation profile of each helix must be sufficiently variable. Helices that are highly conserved throughout (such as M2 of AchR) do not contain a clear enough signal to reveal their orientations. A threshold of sequence variability necessary for accurate predictions is difficult to set a-priori. However, a rule of thumb is that the TM domain should show a helical pattern of variability versus conservation, as seen in most of the cases studied here (e.g., Fig. 4).

Reassuringly, our results on AchR demonstrate, that even in those cases in which a number of helices in the structure cannot be oriented reliably (M2 and M4), the others can still be accurately retrieved (M1 and M3). In the setting of a structure-prediction exercise, it would be possible to determine which helices cannot be oriented reliably on the basis of their conservation profiles and their exposures to the membrane according to intermediate-resolution data.

Our results show that in other cases, the score functions can identify near-native conformations (Figs. 2 and 4; Tables 1 and 2). The fact that the parameterization, which was conducted to reproduce the native structure of bacteriorhodopsin, also retrieved quite closely the native structures of two homologous proteins (sensory rhodopsin II and halorhodopsin) and three very different TM proteins (rhodopsin, aquaporin-1, and AchR) is an indication of the method's predictive ability. The results show that this scoring scheme, though simple, is capable of reliably ranking decoy structures according to their RMSds from the native state (Table 2, Figs. 2–4).

The main focus of this study has been the development of score functions for structure prediction in conjunction with intermediate-resolution cryo-EM maps. However, the results using a conformational search method that takes into account interconnecting loop lengths (Monge et al., 1994) have been encouraging for proteins with small extra-membrane domains. Further research should be devoted to the possibility of predicting the structures of TM domains with short loops even without the constraints imposed by cryo-EM data on helix positions and tilt angles. Furthermore, the results based on rhodopsin's intermediate-resolution structure (Table 2) indicate that a limited exploration of the conformational space defined by the helix positions, tilt, and azimuthal angles may improve structure prediction in cases, in which these parameters cannot be approximated with high confidence from the cryo-EM data. The inclusion of atomistic detail may improve these results further by capturing the subtleties of helix-packing interactions.

It was demonstrated that short sequence motifs could drive the dimerization of TM domains (Lemmon et al., 1992; Javadpour et al., 1999; Russ and Engelman, 1999, 2000; Dawson et al., 2002). For instance, the GxxxG motif, in which two Gly residues are separated by three other residues was shown to induce the close association of two TM helices (MacKenzie et al., 1997). It was also shown that Ala and small polar residues (Ser and Thr) could replace the Gly residues in the motif and induce contact formation (Dawson et al., 2002). We previously used such sequence rules for predicting likely conformations of pairs of TM helices (Fleishman and Ben-Tal, 2002; Fleishman et al., 2002). Here, we did not explicitly utilize information regarding amino-acid packing propensities, because the importance of these residues for packing is reflected in their evolutionary conservation (Sternberg and Gullick, 1989).

We note that the results presented here show that the methods are quite robust in terms of sensitivity to structural or sequence differences. Changes in TM boundaries, for example, did not have a significant effect on the predicted templates of rhodopsin. Some recently solved TM protein structures show helices that are not straight (e.g., Jiang et al., 2002; Miyazawa et al., 2003). In the case of the AchR we used canonical α-helices, even though there are some marked deviations from α-helicity in M2 and M3, yet the predictions did not suffer to any great extent due to these deviations (Fig. 3). Nor have π-bulges and kinks affected the prediction's quality extensively (Fig. 5). Furthermore, although retinal was not modeled in the rhodopsins, the helices' orientations in all cases were reproduced quite accurately. Indeed, explicitly modeling these deviations from α-helicity and the addition of prosthetic groups should improve prediction accuracy. However, from the cases we have examined, we conclude that the strong conservation signal in many TM proteins (exemplified in Fig. 1) ensures that various structural deformations, that might not be accounted for in the cryo-EM data, have mostly a local effect on the accuracy of the prediction, and that this effect is much diminished in unaffected helices.

We acknowledge the Bioinformatics Unit at Tel Aviv University for providing us with infrastructure for the computations.

This study was supported by grant 222/04 from the Israel Science Foundation to N.B.T. and in part by a grant to B.H. from the National Science Foundation (MCB-9808902). S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation and by the Constantiner Institute of Molecular Biology at Tel Aviv University. S.H. was supported by a National Science Foundation postdoctoral fellowship

REFERENCES

Adams, P. D., I. T. Arkin, D. M. Engelman, and A. T. Brunger. 1995. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat. Struct. Biol.* 2:154–162.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Baldwin, J. M., G. F. Schertler, and V. M. Unger. 1997. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* 272:144–164.

Beuming, T., and H. Weinstein. 2004. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*. 20:1822–1835.

Bourne, H. R., and E. C. Meng. 2000. Structure. Rhodopsin sees the light. *Science*. 289:733–734.

Bowie, J. U. 1997. Helix packing in membrane proteins. *J. Mol. Biol.* 272:780–789.

Bowie, J. U. 2000. Understanding membrane protein structure by design. *Nat. Struct. Biol.* 7:91–94.

Briggs, J. A., J. Torres, and I. T. Arkin. 2001. A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins*. 44:370–375.

Chen, C. P., A. Kernytsky, and B. Rost. 2002. Transmembrane helix predictions revisited. *Protein Sci.* 11:2774–2791.

Cronet, P., C. Sander, and G. Vriend. 1993. Modeling of transmembrane seven helix bundles. *Protein Eng*. 6:59–64.

Dawson, J. P., J. S. Weinger, and D. M. Engelman. 2002. Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.* 316:799–805.

Donnelly, D., J. P. Overington, S. V. Ruffle, J. H. Nugent, and T. L. Blundell. 1993. Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci.* 2:55–70.

Eisenberg, D., E. Schwarz, M. Komaromy, and R. Wall. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 179:125–142.

Eisenberg, D., R. M. Weiss, and T. C. Terwilliger. 1982. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*. 299:371–374.

Enosh, A., S. J. Fleishman, N. Ben-Tal, and D. Halperin. 2004. Assigning transmembrane segments to helices in intermediate-resolution structures. *Bioinformatics*. 20:I122–I129.

Eyrich, V. A., D. M. Standley, and R. A. Friesner. 1999. Prediction of protein structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.* 288:725–742.

Fischer, D., O. Bachar, R. Nussinov, and H. Wolfson. 1992. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn.* 9: 769–789.

Fleishman, S. J., and N. Ben-Tal. 2002. A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.* 321:363–378.

Fleishman, S. J., J. Schlessinger, and N. Ben-Tal. 2002. A putative activation switch in the transmembrane domain of erbB2. *Proc. Natl. Acad. Sci. USA*. 99:15937–15940.

Fleishman, S. J., V. M. Unger, M. Yeager, and N. Ben-Tal. 2004. A C-alpha model for the transmembrane alpha-helices of gap-junction intercellular channels. *Mol. Cell*. In press.

Fleming, K. G. 2000. Riding the wave: structural and energetic principles of helical membrane proteins. *Curr. Opin. Biotechnol.* 11:67–71.

Glaser, F., T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*. 19:163–164.

Heymann, J. B., and A. Engel. 2000. Structural clues in the sequences of the aquaporins. *J. Mol. Biol.* 295:1039–1053.

Hille, B. 2001. Ion Channels of Excitable Membranes. Sinauer Associates, Sunderland, MA.

Hirokawa, T., J. Uechi, H. Sasamoto, M. Suwa, and S. Mitaku. 2000. A triangle lattice model that predicts transmembrane helix configuration using a polar jigsaw puzzle. *Protein Eng*. 13:771–778.

Holm, P. J., R. Morgenstern, and H. Hebert. 2002. The 3-D structure of microsomal glutathione transferase 1 at 6 A resolution as determined by electron crystallography of p22(1)2(1) crystals. *Biochim. Biophys. Acta*. 1594:276–285.

Honig, B. H., and W. L. Hubbell. 1984. Stability of ''salt bridges'' in membrane proteins. *Proc. Natl. Acad. Sci. USA*. 81:5412–5416.

Javadpour, M. M., M. Eilers, M. Groesbeek, and S. O. Smith. 1999. Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys. J.* 77:1609–1618.

Jiang, W., M. L. Baker, S. J. Ludtke, and W. Chiu. 2001. Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* 308:1033–1044.

Jiang, Y., A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon. 2002. The open pore conformation of potassium channels. *Nature*. 417:523–526.

Kessel, A., and N. Ben-Tal. 2002. Free energy determinants of peptide association with lipid bilayers. *In* Current Topics in Membranes. S. Simon and T. McIntosh, editors. Academic Press, San Diego, CA. 205–253.

Kim, S., A. K. Chamberlain, and J. U. Bowie. 2003. A simple method for modeling transmembrane helix oligomers. *J. Mol. Biol.* 329:831–840.

Kraulis, P. J. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24: 946–950.

Lemmon, M. A., J. M. Flanagan, J. F. Hunt, B. D. Adair, B. J. Bormann, C. E. Dempsey, and D. M. Engelman. 1992. Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J. Biol. Chem.* 267:7683–7689.

Luecke, H., H. T. Richter, and J. K. Lanyi. 1998. Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science*. 280:1934–1937.

MacKenzie, K. R., J. H. Prestegard, and D. M. Engelman. 1997. A transmembrane helix dimer: structure and implications. *Science*. 276: 131–133.

Merritt, E. A., and D. J. Bacon. 1997. Raster3D photorealistic molecular graphics. *Methods Enzymol.* 277:505–524.

Miyazawa, A., Y. Fujiyoshi, and N. Unwin. 2003. Structure and gating mechanism of the acetylcholine receptor pore. *Nature*. 424:949–955.

Monge, A., R. A. Friesner, and B. Honig. 1994. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA*. 91:5027–5029.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. *Comput. J.* 7:308–313.

Palczewski, K., T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano. 2000. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*. 289:739–745.

Pappu, R. V., G. R. Marshall, and J. W. Ponder. 1999. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.* 6:50–55.

Pellegrini-Calace, M., A. Carotti, and D. T. Jones. 2003. Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins*. 50:537–545.

Pilpel, Y., N. Ben-Tal, and D. Lancet. 1999. kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* 294:921–935.

Popot, J. L., and D. M. Engelman. 1990. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*. 29:4031–4037.

Popot, J. L., and D. M. Engelman. 2000. Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* 69:881–922.

Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*. 18:S71–S77.

Rees, D. C., L. DeAntonio, and D. Eisenberg. 1989. Hydrophobic organization of membrane proteins. *Science*. 245:510–513.

Russ, W. P., and D. M. Engelman. 1999. TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc. Natl. Acad. Sci. USA*. 96:863–868.

Russ, W. P., and D. M. Engelman. 2000. The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.* 296:911–919.

Sorgen, P. L., Y. Hu, L. Guan, H. R. Kaback, and M. E. Girvin. 2002. An approach to membrane protein structure without crystals. *Proc. Natl. Acad. Sci. USA.* 99:14037–14040.

Sternberg, M. J., and W. J. Gullick. 1989. Neu receptor dimerization. *Nature.* 339:587.

Stevens, T. J., and I. T. Arkin. 1999. Are membrane proteins ''inside-out'' proteins? *Proteins.* 36:135–143.

Stevens, T. J., and I. T. Arkin. 2001. Substitution rates in alpha-helical transmembrane proteins. *Protein Sci.* 10:2507–2517.

Stokes, D. L., W. R. Taylor, and N. M. Green. 1994. Structure, transmembrane topology and helix packing of P-type ion pumps. *FEBS Lett.* 346:32–38.

Taylor, W. R., D. T. Jones, and N. M. Green. 1994. A method for alpha-helical integral membrane protein fold prediction. *Proteins.* 18:281–294.

Trabanino, R. J., S. E. Hall, N. Vaidehi, W. B. Floriano, V. W. Kam, and W. A. Goddard 3rd. 2004. First principles predictions of the structure and function of g-protein-coupled receptors: validation for bovine rhodopsin. *Biophys. J.* 86:1904–1921.

Tsai, J., R. Taylor, C. Chothia, and M. Gerstein. 1999. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* 290:253–266.

Tuffery, P., and R. Lavery. 1993. Packing and recognition of protein structural elements: a new approach applied to the 4-helix bundle of myohemerythrin. *Proteins.* 15:413–425.

Tusnady, G. E., and I. Simon. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* 283:489–506.

Ubarretxena-Belandia, I., J. M. Baldwin, S. Schuldiner, and C. G. Tate. 2003. Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer. *EMBO J.* 22:6175–6181.

Ubarretxena-Belandia, I., and D. M. Engelman. 2001. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* 11:370–376.

Unger, V. M. 2001. Electron cryomicroscopy methods. *Curr. Opin. Struct. Biol.* 11:548–554.

Unger, V. M., P. A. Hargrave, J. M. Baldwin, and G. F. Schertler. 1997. Arrangement of rhodopsin transmembrane alpha-helices. *Nature.* 389: 203–206.

Unger, V. M., N. M. Kumar, N. B. Gilula, and M. Yeager. 1999. Three-dimensional structure of a recombinant gap junction membrane channel. *Science.* 283:1176–1180.

Unger, V. M., and G. F. Schertler. 1995. Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy. *Biophys. J.* 68: 1776–1786.

Unwin, N. 1995. Acetylcholine receptor channel imaged in the open state. *Nature.* 373:37–43.

von Heijne, G. 1989. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature.* 341:456–458.

von Heijne, G. 1996. Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* 66:113–139.

White, S. H., and W. C. Wimley. 1999. Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* 28:319–365.

Yohannan, S., S. Faham, D. Yang, J. P. Whitelegge, and J. U. Bowie. 2004. The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA.* 101:959–963.

Zhdanov, V. P., and B. Kasemo. 2001. Folding of bundles of alpha-helices in solution, membranes, and adsorbed overlayers. *Proteins.* 42: 481–494.

doi:10.1016/j.jmb.2004.04.064 J. Mol. Biol. (2004) 340, 307–318

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

An Evolutionarily Conserved Network of Amino Acids Mediates Gating in Voltage-dependent Potassium Channels

Sarel J. Fleishman[1], Ofer Yifrach[2] and Nir Ben-Tal[1]*

[1]*Department of Biochemistry George S. Wise Faculty of Life Sciences, Tel-Aviv University Ramat Aviv 69978, Israel*

[2]*Department of Life Sciences and the Zlotowski Center for Neurosciences, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel*

**Corresponding author*

A novel sequence-analysis technique for detecting correlated amino acid positions in intermediate-size protein families (50–100 sequences) was developed, and applied to study voltage-dependent gating of potassium channels. Most contemporary methods for detecting amino acid correlations within proteins use very large sets of data, typically comprising hundreds or thousands of evolutionarily related sequences, to overcome the relatively low signal-to-noise ratio in the analysis of co-variations between pairs of amino acid positions. Such methods are impractical for voltage-gated potassium (Kv) channels and for many other protein families that have not yet been sequenced to that extent. Here, we used a phylogenetic reconstruction of paralogous Kv channels to follow the evolutionary history of every pair of amino acid positions within this family, thus increasing detection accuracy of correlated amino acids relative to contemporary methods. In addition, we used a bootstrapping procedure to eliminate correlations that were statistically insignificant. These and other measures allowed us to increase the method's sensitivity, and opened the way to reliable identification of correlated positions even in intermediate-size protein families. Principal-component analysis applied to the set of correlated amino acid positions in Kv channels detected a network of inter-correlated residues, a large fraction of which were identified as gating-sensitive upon mutation. Mapping the network of correlated residues onto the 3D structure of the Kv channel from *Aeropyrum pernix* disclosed correlations between residues in the voltage-sensor paddle and the pore region, including regions that are involved in the gating transition. We discuss these findings with respect to the evolutionary constraints acting on the channel's various domains. The software is available on our website http://ashtoret.tau.ac.il/~sarel/CorrMut.html

Keywords: correlated mutations; phylogenetic analysis; maximum likelihood; voltage-gated potassium channel; structural biology

Introduction

Many potassium channels are gated in response to changes in transmembrane voltage.[1–3] This form of gating underlies the production of action potentials: electrical impulses that run across the cell membrane, allowing neurons, for example, to transmit signals over their lengths.[4] Voltage-gated potassium (Kv) channels are tetramers,[5] where each monomer consists of six hydrophobic stretches (S1–S6).[6,7] The S1–S4 region comprises a voltage-sensing domain, in which the S4 segment is thought to be the voltage-sensor element,[8,9] whereas the S5–S6 regions from the four channel subunits form a central pore. This pore domain contains, in addition to the outer (S5) and inner (S6) helices, the pore helix, and the selectivity filter, which are responsible for the channel's high potassium selectivity and throughput[10] (Figure 1(a)).

Comparison of the three-dimensional pore structures of K^+ channels in the closed[10] and open[11] states revealed significant structural rearrangement

Abbreviation used: Kv, voltage-gated potassium.

E-mail address of the corresponding author: bental@ashtoret.tau.ac.il

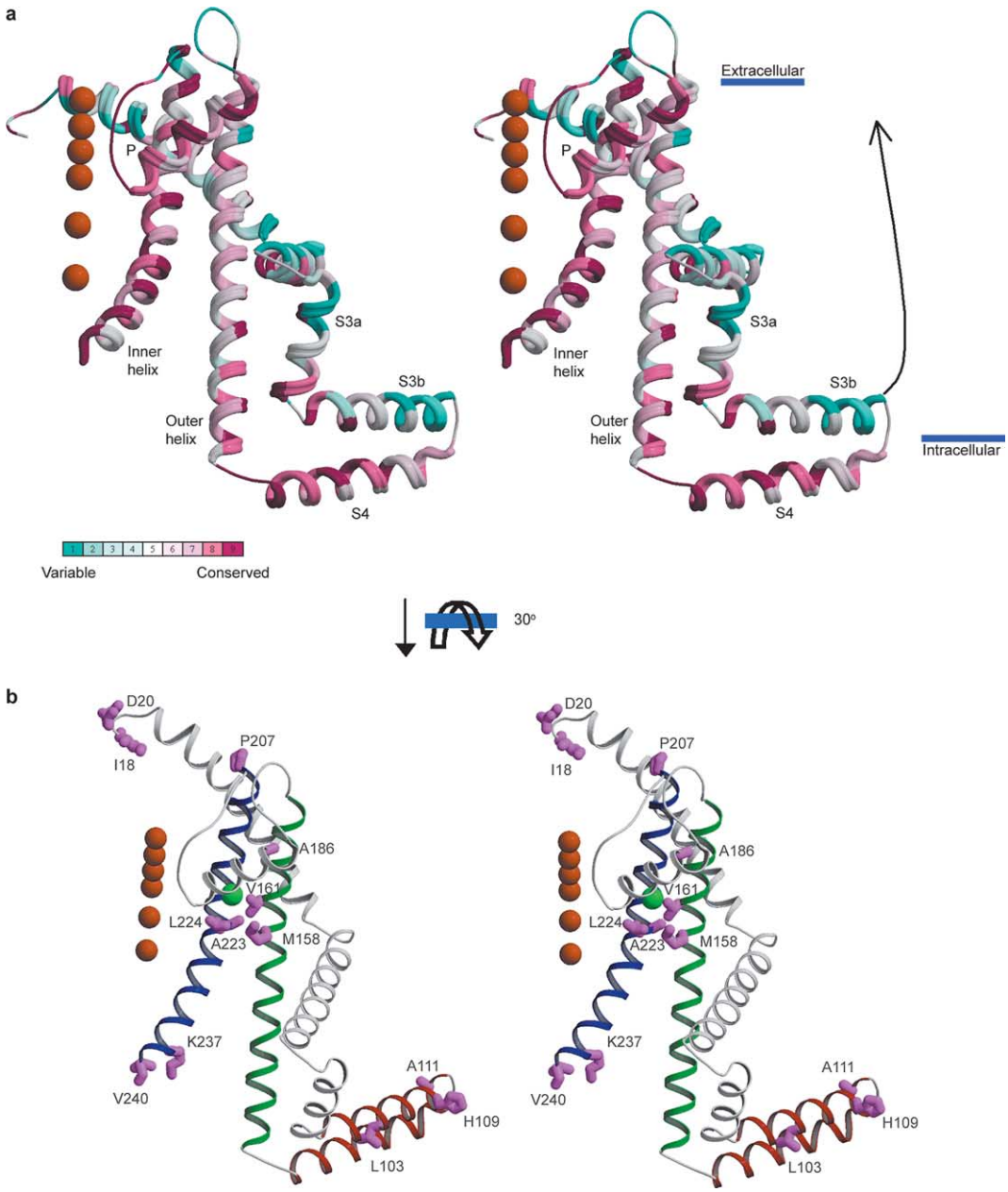


Figure 1. Stereo images of a monomer of the voltage-dependent potassium channel from *Aeropyrum pernix* (KvAP).[13] a, The trace model is color-coded according to evolutionary conservation,[61] with burgundy through turquoise, indicating conserved through variable residues (see color bar). Potassium ions are shown as orange spheres. The selectivity filter, S4 and parts of the inner helix are highly conserved, whereas the outer helix and S3b are more variable. The arrow indicates the direction of motion of the voltage sensor during the opening transition from a membrane-exposed to an extracellular-exposed orientation according to MacKinnon and co-workers' model.[14] b, The S3b–S4 segment is colored red, the outer helix green, and the inner helix blue. The position of the Gly220 gating hinge[11] is marked with a green sphere. A cluster of highly inter-correlated positions is indicated by magenta stick models. The cluster includes the following residues, where the pairwise alignment of the positions with the sequence of the *Shaker* channel[13] is indicated in parentheses: Ile18 (Tyr219) and Asp20 (Glu221) on S1; Leu103 (Thr326), His109 (Ala332) and Ala111 (Glu334) on S3b; Met158 (Ile405), Val161 (Val408) on the outer helix; Ala186 (Ala432) on the pore helix; Pro207 (Val453), Ala223 (Thr469), Leu224 (Ile470), Lys237 (Tyr483), Val240 (His486) and Glu242 (Glu488) on the inner helix. Glu242 (not shown) is missing from the KvAP structure. Table 1 lists some of the correlations connecting this cluster. The Figure was generated using MOLSCRIPT[62] and rendered with Raster3D.[63]

at the pore upon opening. The opening transition involves a large kink in the inner helix around a highly conserved Gly residue, which serves as a gating hinge. This bending allows the movement of the inner helices that leads to the disassembly of the activation gate (Figure 1b).[11] These conformational changes are mostly restricted to the intracellular portion of the channel (Figure 1a). The

region spanning the highly conserved selectivity filter remains, for the most part, rigid during gating.[11,12]

Studies by MacKinnon and co-workers of the voltage-dependent potassium channel from *Aeropyrum pernix* (KvAP) revealed several unexpected findings.[13,14] In contrast to earlier models that identified S4 as the major voltage-sensing element, they found that the S3 segment contains two helices (S3a and S3b), where the S3b helix and the N-terminal portion of the S4 helix form a tight helix hairpin, which they termed the voltage-sensor paddle.[13] Secondly, their results indicated that the paddle moves approximately 20 Å across much of the membrane span in response to the changes in transmembrane voltage[14] (Figure 1a). On the basis of these findings, MacKinnon and co-workers proposed that channel opening occurs *via* coupling of the voltage-sensor paddle's movement to that of the outer helix.[14] According to this proposition (Figure 1a), this movement, in turn, induces conformational changes in the inner helix[11] that open the energetically more stable closed structure.[15]

During the gating transition, at least three charged arginine residues per subunit in the tetramer cross the membrane[16–18] (Figure 1a). Contrary to previous models of activation (summarized by Bezanilla,[19]) the KvAP model argues that these charges move mostly in an unshielded manner through the hydrophobic membrane environment.[7,14] This conclusion[14] is astonishing from a thermodynamic point of view because of the prohibitive cost in desolvation free energy[20] associated with the transfer from water to lipid of at least a dozen charged arginine residues per channel.[14]

Following MacKinnon and co-workers' new view of voltage dependence, several studies have been devoted to test its validity.[21–23] The model has been criticized[24] for its reliance on a structure that may not be physiologically relevant owing to possible artefacts originating from co-crystallization with Fab fragments that may have distorted its conformation.[13,23] Studies on the *Shaker* homologue of KvAP provided evidence that residues within S4 are in close proximity to residues at the extracellular part of the outer helix, in apparent contradiction to the KvAP model.[21,25] In addition, it was shown that S3b does not move significantly in response to changes in the transmembrane voltage,[24] and based on accessibility studies using the homologous Kv2.1, it was suggested[22] that the motion of the voltage sensor is not as large as that implied by MacKinnon and co-workers.[14] On the basis of these results, an alternative model of the gating transition has been proposed[24] that is coherent with the previous view, in which the voltage sensor, which is comprised solely of S4, is encapsulated within a proteinaceous environment. This model further argues that the channel's conformational changes upon gating are of smaller magnitude, when compared to that suggested on the basis of experiments on KvAP.[14]

Nevertheless, relatively large conformational changes are anticipated in both gating models.[14,24] Such large changes make it exceedingly difficult to plan and interpret mutation and accessibility studies aimed at uncovering conformational substates.[14,21,22,24,26] For instance, it is difficult to control whether the modifications introduced in these studies trap the molecule in physiologically relevant states. The fact that some of these recent studies were performed on the *Shaker* homologue of KvAP,[21,22,24,25] which contains a long segment between S3 and S4 that is missing in the KvAP structure[13] (Figure 2), adds another layer of complexity.

Here, we study the inter-domain relationships in Kv channels from an evolutionary perspective. We found a network of inter-correlated amino acid positions, which cluster in functionally important regions when mapped on the KvAP structure. Specifically we show that residues on S3b, which forms part of the voltage sensor according to the KvAP structure[13,14] (Figure 1), but not according to alternative models,[24] are coupled to pore residues distributed in the vicinity of the activation gate and the gating-hinge position. These regions experience major structural rearrangements upon pore opening.[11,15]

Phylogeny-based Detection of Correlations

In silico analysis of correlated mutations has been used to identify positions that are implicated in contact formation or allosteric regulation and conformational changes in large protein families.[27–33] The underlying assumption in these studies was that functional or structural associations between a pair of positions force a coherent change in their amino acid identities during evolution. In other words, substitution of one position would induce

Figure 2. A multiple-sequence alignment showing the S3b–S4 segment of a few divergent sequences of Kv channels. Residues that were identified as part of the cluster of inter-correlated positions are shaded. The S4 segment is relatively conserved, whereas S3b, which contains the three correlated positions, is highly variable. The two helices are connected by a linker of variable length.

the other to undergo a compensatory change in order to maintain the structural or functional relationships between the two positions.

Detection of co-variation in amino acid positions within proteins, when combined with experimental data, may indicate what differences are necessary for modifying function. For instance, all isoforms of Kv channels are known to have the same ion selectivity and permeation characteristics, yet they show differences in terms of voltage sensitivity and closing and opening kinetics. Such changes might be reflected in variations in the amino acid sequences of the family. Since multiple positions are involved in determining these traits, such sequence variations should occur concomitantly in the relevant locations.

A key problem in identifying correlations in amino acid positions along multiply aligned sequences of a protein family is the difficulty in distinguishing co-variation (signal) from noise. Therefore, contemporary methods for identifying correlations often rely on very large multiple-sequence alignments of homologous proteins (typically hundreds or thousands of sequences) in order to obtain good signal-to-noise ratios.[30–32] In the case of the Kv channel family, however, only a few tens of protein sequences have been discovered. Such paucity of homologous protein sequences is typical for many protein families. Nevertheless, a collection of sequences that is sufficiently heterogeneous in terms of functions and sequences can be constructed by the inclusion of various Kv paralogues (see Materials and Methods). We present a novel method for detecting co-varying amino acid positions that is applicable for the analysis of intermediate-size protein families (50–100 sequences) that are sufficiently heterogeneous. The method is similar to that of Shindyalov *et al*.[28] in that it is based on phylogenetic reconstruction rather than on multiple-sequence alignment alone.

Generally speaking, by tracing the evolutionary pathway for every pair of amino acid positions within the protein, it is possible to substantially increase detection accuracy. As a first step in the analysis, we reconstruct the evolutionary history of the protein family by inferring the sequences of hypothetical (now-extinct) ancestral proteins of the family.[34] The phylogenetic tree (Figure 3) together with the set of reconstructed and contemporary sequences specify the evolutionary pathway that has generated the protein family as we observe it today, where each branch connects evolutionarily close sequences. By following the reconstructed pathway, we trace the changes that occurred at each evolutionary step for every position, thus reducing the errors that arise when comparing sequences that are phylogenetically distant.

Many contemporary methods for detecting correlations employ a simplistic amino acid substitution scheme, whereby all changes are treated equally.[28–30,32] Since we consider the changes that occurred at each position in subsequent evolutionary steps, we can employ a substitution matrix that reflects the subtleties of amino acid replacements in proteins more realistically, e.g. a Val for Ile change would be considered of smaller magnitude than a Gly for Trp substitution. That is, in each evolutionary step, represented by a branch on the phylogenetic tree, the changes in amino acid identities are measured. The correlations between changes in different positions of the alignment can then be calculated in a straightforward manner. We note that the method does not consider back mutations or multiple mutations in a single branch.

Here, we used the Miyata matrix,[35] which provides a measure for the physicochemical differences between amino acids. The advantages of using a phylogenetic tree are hence twofold: first, only changes that occurred at the same evolutionary interval are compared; and second, we may discriminate between small and large amino acid substitutions. Thus, the method not only detects the positions that change concurrently, but also identifies those that undergo changes of similar magnitude during evolution. We note that the Miyata[35] substitution matrix may be replaced by other substitution schemes, such as the Dayhoff matrix that was derived from the observed substitution frequencies in homologous proteins.[36]

The difficulty in detecting correlations in intermediate-size protein families is compounded by the uneven sampling or bias of homologues in sequence space. In many cases there is an over-representation of particular families of sequences, while others are under-represented. Thus, high correlations might simply be the result of a lack of variability in the given collection of sequences. To decrease bias in the set of sequences, we manually removed those that shared high homology in the S1–S6 segments with others.

The phylogenetic tree of the Kv family demonstrates that in the current selection of sequences, bias resulting from lack of variability is rather low (Figure 3). The majority of the sequences are from mammals; however, by including many paralogous sequences we were able to gain sequence variability. Following the computation of the phylogenetic tree and the reconstruction of ancestral sequences,[34] we eliminated from the alignment all positions showing relatively low entropy or information content,[37] which is a measure of the heterogeneity of amino acid identities in a given position of the alignment. This step is applied to avoid the detection of pairs of positions that changed a small number of times in the family's evolutionary history. We also eliminated positions exhibiting at least one gap in the multiple-sequence alignment because of the unreliability of ancestral-sequence reconstruction at such sites.[34]

To further reduce the possibility of errors due to bias, we derived confidence intervals for the correlation coefficients using bootstrap sampling.[38] Briefly, bootstrapping randomly generated samples

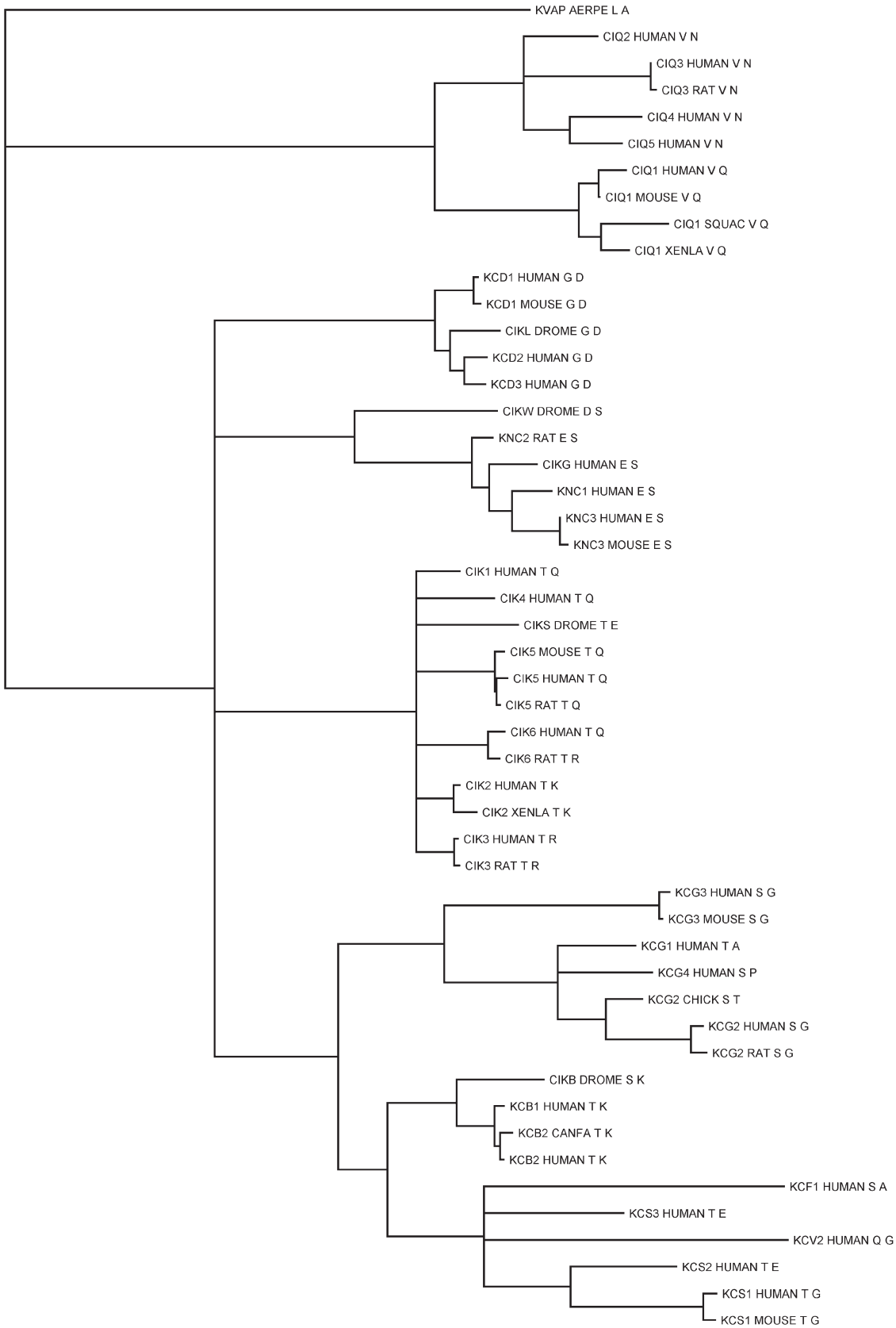


Figure 3. The phylogenetic tree used in this study, displaying at the terminal nodes one-letter codes for residues aligned with positions Leu103 (left) and Ala111 (right) on the KvAP sequence. The two correlated positions are located on the S3b helix, which forms part of the voltage sensor according to the KvAP structures[11] (Figure 1). The phylogenetic tree was computed[50] on the basis of the multiple-sequence alignment of 50 voltage-gated potassium channel sequences, and was used throughout the analysis (see Materials and Methods).

of phylogenetic branches. On each sample, the correlation coefficients of the changes occurring in each pair of positions were computed. Thus, we obtained a set of correlation coefficients for each pair of positions, on which we computed the average value as well as the confidence interval. We then eliminated correlations that were statistically insignificant. By applying the bootstrapping procedure on our data set we also rejected those correlations that were highly dependent on a particular subset of phylogenetic branches, thus reducing the possibility that the evolutionary pattern in certain parts of the tree would have a dominant effect.

We detected a large set of correlated positions, which we then subjected to principal-components analysis[39] in order to identify amino acid networks, in which all residues are highly inter-correlated. This filtering step also provided a means for reducing the effects of spurious correlations.

Results

A network of inter-correlated residues mediates channel opening

Overall, 158 correlations between pairs of amino acids were identified that met the requirements of high mean Pearson correlation coefficients ($r > 0.5$), and for which the lower confidence boundary, measured by bootstrapping, was judged to be statistically significant ($r_{low} > 0.15$). The list of correlations shows the amino acid positions to be heterogeneous, with some positions being linked to many, and others to just a few. To identify networks of highly inter-correlated positions within this list, we used principal-components analysis.[39] Several distinct sets of highly inter-correlated positions were identified. Figure 1b shows a mapping of one such set that was identified as the most significant cluster of correlated positions, on the KvAP structure. This cluster of 14 positions is linked by 50 significant correlations according to the above criteria; some of the positions, which were identified in this cluster, were associated with all of the others. Representative correlations are listed in Table 1. It may be seen that most of the inter-correlated residues are in the pore domain. However, several others were identified in the voltage-sensing paddle.

Ten out of 14 positions in the cluster of highly inter-correlated positions were previously tested in scanning-mutagenesis studies for their effects on voltage-dependent gating of the *Shaker* homologue of KvAP. Tryptophan-scanning mutagenesis showed that mutations of *Shaker* positions aligned with Leu103 and His109 on the S3b helix,[40] and Met158, Val161, Ala223 and Leu224 of the pore region[41] caused high-impact changes in gating transitions (Figure 1b). For these mutants, the voltage-activation relations were dramatically different compared to that of the wild-type channel (effects on the stability of the closed *versus* the open states). An alanine scan showed that Lys237 is another gating-sensitive residue, but mutations of three positions, including Ala186, Pro207 and Val240, that are part of the cluster on the inner helix did not alter the gating equilibrium.[15] Thus, seven out of the ten positions tested experimentally have large effects on channel-gating transitions, implying that this network of correlated amino acids may have a functional role in the voltage-induced conformational changes that lead to channel opening.

From a structural point of view, if indeed the cluster of inter-correlated residues comprises mostly gating-sensitive positions, we would expect these residues to occupy pore regions that are involved in the conformational changes during channel gating. The inter-correlated residues distributed at the intracellular end of the inner helix lie roughly two helical turns below the channel's activation gate, which opens to grant potassium ions entry into the channel during the gating transition[42,43] (Table 1; Figure 1b). The gate itself, which consists of a relatively conserved Pro-X-Pro sequence motif (*Shaker* positions 473–475), likely

Table 1. A list of representative pairs of correlated positions involving the S3b and S4 segment,[13] activation gate (C terminus of the inner helix[43]) and the gating-hinge region[11] (surrounding Gly220, shown as a green sphere in Figure 1b)

S3b-S4 inter-correlations		
Leu103	His109	0.53 (0.27, 0.74)
Leu103	Ala111	0.59 (0.34, 0.77)
Glu108	Gly114	0.50 (0.17, 0.74)
Leu113	Leu118	0.58 (0.23, 0.87)
S3b-outer helix		
Leu103	Val161	0.50 (0.24, 0.75)
His109	Val161	0.55 (0.20, 0.77)
Leu110	Ala140	0.58 (0.15, 0.83)
Leu110	Asp143	0.51 (0.19, 0.73)
Ala111	Val161	0.66 (0.34, 0.86)
S3b-gate		
Leu103	Glu242	0.69 (0.46, 0.84)
His109	Val240	0.57 (0.28, 0.79)
Ala111	Val240	0.56 (0.22, 0.83)
Ala111	Glu242	0.71 (0.42, 0.90)
Gate-gating hinge region		
Met158	Lys237	0.66 (0.31, 0.91)
Val161	Val240	0.58 (0.19, 0.83)
Ala186	Val240	0.53 (0.20, 0.79)
Ala223	Lys237	0.64 (0.35, 0.91)
Ala223	Val240	0.55 (0.18, 0.81)
Inter-correlations in the gating hinge region		
Met158	Ala223	0.59 (0.27, 0.90)
Met158	Leu224	0.71 (0.36, 0.91)
Val161	Ala223	0.58 (0.16, 0.89)
Val161	Leu224	0.79 (0.51, 0.91)
Ala186	Leu224	0.74 (0.49, 0.94)
Pro207	Ala223	0.56 (0.30, 0.80)
Pro207	Leu224	0.55 (0.26, 0.73)

The trimmed means in the 95% confidence interval of correlations (r), which were calculated from 400 bootstrapping samples, are indicated and the 95% confidence interval is parenthesized (see Materials and Methods).

forms a bend and adds flexibility to the intracellular part of the S6 segment, which is important for the opening transition.[44] The five positions in the extracellular region are all within 10 Å of a mostly conserved glycine (Gly220 in KvAP (green sphere in Figure 1b)).[10,13] This position serves as the gating hinge during pore opening, where a bend occurs in the inner helix.[11] This co-variance can be rationalized by assuming that substitutions in one region of the pore domain can be compensated for by mutations in the other. An alternative interpretation is that in order to modify existing function or to gain a new one, both regions have changed in evolution concomitantly. Since the set of sequences used in the current study consisted of many different paralogous sequences (Figure 3), where each may have slightly different characteristics of voltage sensing and gating kinetics, it is tempting to adopt the latter explanation.

The distribution pattern of inter-correlated residues on the pore domain, determined by the correlated-mutations analysis is in very good agreement with an energetics analysis of pore opening performed for the *Shaker* Kv channel.[15] In that study as well, gating-sensitive positions at the pore were found to cluster at the activation gate of the channel and at the region just extracellular to the glycine gating-hinge residue. The particular pairs that were identified in that study[15] were not highlighted in our analysis because of the fact that residues at some of the positions that were analyzed experimentally are highly conserved.

From a structural perspective, it appears unexpected that many of the positions identified in this cluster are distant in 3D space, and yet are inter-correlated. However, this result is in line with experimental data on the *Shaker* channel pore that demonstrated, by using double-mutant cycle analysis,[15] that gating-sensitive positions at the pore are energetically coupled to each other even at distances as large as 15 Å.[10,15] Such long-range energetic couplings betweens residue pairs are indicative of large tertiary or quaternary conformational changes[45] as was indeed verified upon comparison of the closed (KcsA[10]) and open (MthK[11]) pore channel structures.

The amino acid correlations detected here and their distribution pattern on the pore imply that during the evolutionary process, the activation gate and the gating hinge regions of the channel have accumulated substitutions in order to assume slightly different gating characteristics. Since the correlated positions occur mostly in regions that mediate inter-helical contacts, where extensive packing interactions occur, their substitution from one channel to another may increase or decrease the thermodynamic stability of the closed *versus* the open states of the channel. Such changes in the packing interfaces of regions that experience conformational changes during gating are expected to alter the gating kinetics.

Another interesting result in our analysis is the finding that several of the highly correlated residues occupy positions on the S3b helix, and co-vary with the pore domain residues that affect channel gating (Figure 1b; Table 1). This co-variation implies that S3b affects the opening transition, along with the activation gate and gating hinge regions. An important role for S3b in affecting the gating transition makes sense in the light of the KvAP structure, which shows that the helices S3b and S4 form one structural unit (a "paddle").[13] Moreover, electrophysiological assays demonstrated that the two helices move together between membrane and extracellular exposures in response to transmembrane voltage changes.[14] This result is also in agreement with alanine and tryptophan-scanning mutagenesis analyses, which showed that some positions on S3b are gating-sensitive.[40,46] On the other hand, a recent accessibility study showed that the *Shaker* channel's S3b segment is externally exposed in both open and closed channel states, thereby contradicting the notion that S3b and S4 move as one structural unit.[24]

Other residues in this cluster of correlated positions are less readily explained within the context of a network of positions that are involved in the gating transition. One of these positions is Pro207 at the N-terminal end of the inner helix. Mutation of the *Shaker* position that is aligned with Pro207 to either alanine[15] or tryptophan[41] did not alter channel gating significantly, thus making it unlikely that this position is involved in the gating transition. Two other positions in the cluster, Ile18 and Asp20, which are located N-terminal to S1, have not been tested experimentally. Indeed, it is difficult to imagine a role for these residues in gating according to the KvAP structure (Figure 1), but it is very likely that the structure does not represent the physiological position of S1, whose N-terminal part is intracellular.[13,47,48] It would be interesting to experimentally test whether and how this segment is coupled to the opening transition.

Sensitivity of the analysis to the phylogenetic inference method and to the amino acid substitution matrix

Using neighbor-joining

To gauge the results' sensitivity to the particular phylogenetic inference methods used here, we also computed the phylogenetic tree using the neighbor-joining algorithm.[49] This method, when compared to the maximum-likelihood program Tree-Puzzle[50] is computationally less intensive, but is less robust, in particular when analyzing a very divergent sequence set. Aside from this step, all others used to compute correlations and to derive the cluster of highly inter-correlated positions were the same as elaborated in Materials and Methods.

The list of pairs of positions showing high ($r \geq 0.5$) and significant ($r_{low} \geq 0.15$) correlations detected using the neighbor-joining tree was

approximately half the one obtained by using the phylogenetic tree of maximum-likelihood. The fact that a smaller number of correlations were found to be statistically significant indicates that the analysis based on the neighbor-joining tree was noisier than that of the maximum-likelihood tree. Based on the neighbor-joining tree, the most significant network of correlated positions included the six amino acid residues on the voltage sensor and the activation-gate region that were detected also using the tree of maximum-likelihood. However, the cluster of five positions surrounding the gating-hinge Gly220 was missing. This result indicates that while an analysis based on a neighbor-joining tree retrieves some of the most significant correlations, the results based on the maximum-likelihood tree are more sensitive, as is indeed expected.

The amino acid replacement matrix

Many existing approaches for the detection of correlated positions within protein families treat all amino acid substitutions in the same way, i.e. without differentiating among small and large changes.[28–30,32] In contrast to these methods, we have employed the Miyata substitution matrix,[35] which assigns small values to physicochemically moderate substitutions such as Val for Ile, and large values to drastic substitutions such as Gly for Trp. To test whether the use of the Miyata matrix increases the method's sensitivity, we replaced the Miyata matrix with a binary substitution matrix, in which every change in amino acid identity is given a value of 1, whereas no change is assigned a value of 0. Based on the phylogenetic tree and ancestral-sequence reconstruction of maximum likelihood, we computed the correlations among amino acids using this binary matrix.

The list of high ($r \geq 0.5$) and significant ($r_{low} \geq 0.15$) correlations was larger by more than 60% in the case of the binary matrix than when using the Miyata matrix.[35] Many of the pairs of positions showed very similar correlation coefficients and smaller confidence intervals, reflecting the lower discriminating strength of the binary substitution matrix. Importantly, whereas in the case of the Miyata matrix more than 50% of the correlations were deemed statistically insignificant using the bootstrap criterion $r_{low} \geq 0.15$, none of the correlations using the binary matrix was rejected on this basis. We conclude that, at least in cases in which relatively small sets of sequences are analyzed, a physicochemical substitution matrix is preferable.

Discussion

Recently, an alternative model for channel gating was suggested for voltage-dependent potassium channels based on biochemical, electrophysiological, and structural studies.[13,14] In addition, it has been shown that the pore opening transition involves coupled interactions in different regions of the pore domain.[15,51] In view of the alternative models of channel opening, we set out to investigate whether residues outside the pore are also coupled to regions that are important for the gating transition. We did so by examining inter-domain relationships from an evolutionary perspective.

We developed a novel method that identifies evolutionarily co-varying amino acid positions in intermediate-size protein families, and applied it to study voltage-dependent potassium channels. One of the method's strengths is its use of phylogenetic inference, allowing the algorithm to trace the evolutionary pathway for every pair of positions in the protein family. Various measures have been employed to limit the effects of bias in sequence space and of errors in ancestral-sequence reconstruction.

Despite the method's enhanced sensitivity, it cannot be used reliably to detect correlations within a family represented by a very small sequence set. The actual boundary, below which the method's dependability is too low, cannot be determined *a priori* as it is contingent on a variety of factors. However, a critical element to a reliable analysis is that the collection of homologues spans as much as possible of the function and sequence space of the protein family. This may be achieved by the inclusion of a variety of paralogues. The family of Kv channels comprises many paralogues (see Materials and Methods), providing the necessary functional and sequence variability.

The Kv channel family provided this study with a wealth of experimental data for validation and for advancing hypotheses that may not be available in families that are less well characterized. To interpret an analysis of correlations in such cases, it is possible to employ the standards used in the study of double-mutant cycles, where a coupling between positions that are proximal is deemed a consequence of physical contact, whereas the coupling of distant positions is a result of allostery (e.g. Yifrach & Mackinnon[15]).

Of the cluster of highly inter-correlated positions detected here, a large fraction (seven of ten) were also identified in mutagenesis studies as being gating-sensitive,[15,40,41] providing support for the method's capabilities in identifying functionally related positions. For comparison, Miller and co-workers found roughly 50% of the positions in the S1, S2, and S3 segments to be sensitive to substitution by tryptophan.[40,52] The fraction of highly inter-correlated positions that were found to be gating-sensitive should be considered an understatement, since two out of the three positions that were presumably identified erroneously as correlated were only tested using an alanine scan, which is relatively stringent.[15] Notably, as has been observed before, gating-sensitive positions do not necessarily map to evolutionarily conserved regions of the protein, e.g. the S3b segment[40]

(Figure 1b). It has therefore been difficult to identify such positions without the use of large-scale scanning mutagenesis experiments. Thus, an analysis of evolutionarily correlated mutations may provide a means to focus experimental efforts.

Analysis of double-mutant cycles provides a direct means to test the functional implications of evolutionarily detected couplings between positions.[32,53] Our results agree well with experimental findings on gating of Kv channels that show that the regions encompassing the glycine gating hinge on the inner helix and the activation gate in the intracellular part of the channel are energetically coupled in the context of the gating transition.[15] The correlations indicate that the energetic coupling between positions in the pore domain is also reflected by the positions' co-variation in the Kv family's evolution.

The correlations that we have identified extend this coupling, and include the S3b segment as well (Table 1, Figure 1b). The results imply that these three functional elements of the channel are evolutionarily coupled, suggesting that substitutions in S3b would have an effect on channel gating. It is interesting to note that the three positions identified on S3b are all evolutionarily variable (Figure 1). In fact, the segment's hypervariability contrasts with the relatively high conservation of S4. Moreover, the two segments are connected *via* a linker of variable length in different paralogues (Figure 2).

It has been suggested that the low conservation of the S3b helix implies lack of a functional constraint and is, therefore, an indication that its structural association with S4 is not universal.[23,24] Our analysis suggests that the functional constraint is manifested in this case through the pattern of substitutions of pairs of positions (Table 1), i.e. through the inter-correlated amino acids detected. This argument is strengthened by the observation that many positions at the C-terminal part of the S3b helix, where two of the highly inter-correlated positions are located (Figure 1b), are gating-sensitive in two different channel subtypes, despite the segment's sequence variability.[40,46,54]

The evolutionary advantage of residue substitutions in important functional regions such as the S3b part of the voltage-sensor paddle, the activation gate, and the region encompassing the conserved gating hinge is clear. Modifications in these regions would have significant effects on the gating characteristics of the channel. For instance, if S3b indeed forms part of the voltage sensor,[13] its modification might alter the sensitivity of the channel to changes in transmembrane voltage, whereas substitutions in the gate and the region surrounding the gating hinge might alter the thermodynamic stability of the open or closed states. Given that these effects are intertwined in the sense that they all modulate the gating transition, the changes that are observed in amino acid identities should be coupled as is evident through correlated-mutation analysis.

In view of these conclusions, an interesting experiment might be to identify what constitutes a minimal set of mutations that alter the function of a given Kv paralogue to obtain a channel with the characteristics of another. That the domains of Kv channels are at least grossly modular was exemplified by an experiment, in which the voltage-sensor segment (S1–S4) of the *Shaker* Kv channel was connected to the voltage-insensitive pore domain from KcSA to produce a voltage-sensitive chimera.[55] Also, substitutions of just three hydrophobic amino acid positions of the S4 segment of the *Shaw* channel to the corresponding ones of the *Shaker* member of the Kv family were enough to switch the gating characteristics of the former channel into those of the latter.[56] The fact that the S3b segment, the activation gate, and the region encompassing the gating hinge are evolutionarily coupled (Figure 1(b)) suggests that chimera that include subsequences of these regions and parts of the S4 segment from different channels might indeed switch channel characteristics in a manner that can be rationalized.

In summary, our results demonstrate that along with conservation analysis, a study of correlated mutations is an important part of phylogenetic investigation. Often in such analyses, positions that are not evolutionarily conserved are assumed to have little functional role. Here, we show, however, that certain positions that are evolutionarily variable or only show intermediate conservation (e.g. residues His109 and Lys237 of the S3b and activation-gate regions, respectively) are coupled in a functionally meaningful way (Table 1; Figure 1b). Hence, whereas the functional importance of the S4 helix as the principal carrier of the gating charge[17,18] is immediately apparent from its conservation profile (Figure 1a), the implied importance of S3b in modulating voltage sensitivity is evident only through its co-variation with other domains that are relevant to gating.

On the basis of our results, we argue that the architectural design of Kv channels, in terms of evolutionary conservation, is two-tiered. The selectivity filter and parts of the pore and inner helices are all evolutionarily constrained to maintain the channel's hallmark features of selectivity and high throughput[11,12] (Figure 1a). S4 is also highly conserved to maintain the nominal gating charge. In contrast, the activation gate and parts of the outer helix, the gating-hinge region, and the S3b helix, all of which do not directly control ion selectivity or carry the gating charge, are freer to accumulate substitutions in order to change certain gating characteristics. These substitutions are correlated, reflecting the concerted effect of these domains on the gating transition.

Kv channels show a large degree of modularity, with the activation gate, gating hinge, selectivity filter, and voltage sensor occupying different regions of the protein. This modularity contrasts with the apparently more "parsimonious" architecture of the ClC chloride channel, in which the gate

and selectivity filter inhabit the same region.[57] The evolutionary advantage of a modular architecture is that it is possible to introduce modifications to particular functions of the protein without undermining others. For instance, in Kv channels, changes in voltage sensitivity or gating kinetics upon the evolutionary pathway need not interfere with the channel's selectivity for potassium. With respect to the Kv channel family, whose members respond differently to a large spectrum of voltages, this separation of functionally important regions may have been a highly important evolutionary force shaping the voltage-gated channel structure.

Materials and Methods

Data

We constructed an initial multiple-sequence alignment of a few tens of sequences of Kv channels derived from the SWISS-PROT database.[58] On the basis of this alignment, a hidden Markov model was then constructed,[59] calibrated, and used to search for more Kv channel sequences. From the final list of homologues we removed sequences that showed very high homology to others in the data set, and retained 50 mostly mammalian sequences with a few fly, frog, fish, and chicken representatives. These consisted of Kv1.1-1.6, Kv2.1-2.2, Kv3.1-3.4, Kv4.1-4.3, Kv5.1, Kv6.1-6.3, Kv9.1-9.3, Kv11.1, the KQT members 1–5, the *Drosophila melanogaster* sequences *Shab*, *Shal*, *Shaker*, and *Shaw*, and the KvAP sequence.

All sequences except for KvAP were aligned using the CLUSTAL W algorithm using default parameters,[60] and KvAP was then added manually, based on its pairwise alignment with the *Shaker* Kv channel.[13] Because the N and C termini of the multiply aligned sequences contained many gaps, thus reducing the alignment's reliability, the sequences were trimmed to produce a core alignment consisting mostly of the transmembrane S1–S6 segments. The final alignment contained the positions corresponding to 138–505 of the *Shaker* Kv channel.

Phylogenetic reconstruction

An unrooted phylogenetic tree was computed using the maximum-likelihood method Tree-Puzzle,[50] using eight Gamma rate categories, the Muller–Vingron model of amino acid substitution, and default parameters (Figure 3). An alternative tree was constructed using the neighbor-joining method[49] based on Jukes–Cantor distances. The ancestral (now-extinct) sequences were reconstructed on the basis of both trees with the maximum-likelihood program PAML[34] using marginal reconstruction, eight Gamma rate categories, the JTT substitution matrix, and default parameters. Positions in the multiple-sequence alignment that exhibited one or more gaps were discarded owing to the uncertainty in sequence reconstruction at sites with insertion or deletion.

To gauge the extent of change at each position during the evolutionary process, we followed the phylogenetic tree, and for each amino acid position and branch, the differences in amino acid identities were converted to physicochemical distances according to the Miyata substitution matrix.[35] Alternatively, we used a binary matrix, where every change in amino acid identity was given an equal weight of 1 and no change a value of 0.

Calculating entropy

We calculated the entropy (or information content),[37] which is a measure of the heterogeneity of amino acid identities, at each position in the alignment of extinct and extant sequences according to $\Sigma_{i=1..20} - f_{p,i} \ln(f_{p,i})$, where $f_{p,i}$ is the frequency of amino acid i at position p. Positions showing entropy lower than 1.1 were eliminated. Such positions were judged to be too conserved, and therefore unlikely to contain enough information for computing correlations.

Calculating correlations among residues

Pearson correlation coefficients (r) between the physicochemical distances[35] of each pair of amino acid positions were calculated by taking into account the changes that occurred along all of the branches of the phylogenetic tree. Hence, high correlations are expected for pairs of positions, whose identities change at similar physicochemical magnitudes and at the same evolutionary time. Pairs of positions with $r < 0.5$ were assumed to be poorly correlated and were not further analyzed.

We used the bootstrap method[38] to obtain confidence intervals for the correlation coefficient of every pair of amino acid positions that were not rejected in previous steps. Bootstrapping redrew 400 samples of branches with replacements from the phylogenetic tree, i.e. pairs of evolutionarily related proteins. For each such sample, we calculated the Pearson correlation coefficient for all the pairs of positions. For each pair of positions, these 400 correlation coefficients were then sorted numerically, and the correlation coefficients at the lower (r_{low}) and upper (r_{high}) 2.5 percentiles were considered to be the lower and upper 95% confidence boundaries. For each pair of positions i, j the trimmed mean $r_{i,j}$ of the correlation coefficients in the 95% confidence interval was also computed. Pairs of positions, whose lower 2.5% confidence boundary was $r_{\mathrm{low}} < 0.15$ were rejected, as were positions with trimmed means of correlation coefficients $r_{i,j} < 0.5$.

As a further test of significance, we discarded pairs of positions if one or both of the residues i in each pair was found to have $\mathrm{cov}(i, i) = 0$ in at least 2.5% of the bootstrap samples. This eliminated correlations that were high simply because of homogeneous evolutionary conservation in large parts of the phylogenetic tree.

Principal components analysis

In searching for networks of highly inter-correlated amino acid positions, we subjected the complete set of correlated pairs of positions to principal-components analysis.[39] We constructed a symmetric matrix of all of the correlations identified in the study, where each element i, j of the matrix corresponded to the trimmed mean of correlations between positions i and j, $r_{i,j}$. Pairs showing low ($r_{i,j} < 0.5$) or insignificant ($r_{\mathrm{low}} < 0.15$) correlations were assigned a value of 0, and the diagonal elements a value of 1. The matrix was decomposed into eigenvalues and eigenvectors. The eigenvector associated with the eigenvalue of highest magnitude was regarded as the most significant correlated network. Position i was considered to be part of this cluster if

$|e_i| \geq 0.15$, where e_i corresponds to element i in the eigenvector.

Availability

The programs used in this analysis are available†.

Acknowledgements

The authors thank Christopher Miller, Dan Graur, and Meytal Landau for helpful discussions. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund to N.B.T. and by a doctoral fellowship from the Clore Israel Foundation to S.J.F. Some computations were conducted using the facilities of the Bioinformatics Service Unit at Tel-Aviv University.

References

1. Sigworth, F. J. (1994). Voltage gating of ion channels. *Quart. Rev. Biophys.* **27**, 1–40.
2. Yellen, G. (1998). The moving parts of voltage-gated ion channels. *Quart. Rev. Biophys.* **31**, 239–295.
3. Bezanilla, F. (2000). The voltage sensor in voltage-dependent ion channels. *Physiol. Rev.* **80**, 555–592.
4. Hille, B. (2001). *Ion Channels of Excitable Membranes*, 3rd edit., Sinauer Associates, Sunderland, MA.
5. MacKinnon, R. (1991). Determination of the subunit stoichiometry of a voltage-activated potassium channel. *Nature*, **350**, 232–235.
6. Tempel, B. L., Papazian, D. M., Schwarz, T. L., Jan, Y. N. & Jan, L. Y. (1987). Sequence of a probable potassium channel component encoded at Shaker locus of Drosophila. *Science*, **237**, 770–775.
7. Miller, C. (2003). A charged view of voltage-gated ion channels. *Nature Struct. Biol.* **10**, 422–424.
8. Papazian, D. M., Timpe, L. C., Jan, Y. N. & Jan, L. Y. (1991). Alteration of voltage-dependence of Shaker potassium channel by mutations in the S4 sequence. *Nature*, **349**, 305–310.
9. Liman, E. R., Hess, P., Weaver, F. & Koren, G. (1991). Voltage-sensing residues in the S4 region of a mammalian K^+ channel. *Nature*, **353**, 752–756.
10. Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*, **280**, 69–77.
11. Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. T. & MacKinnon, R. (2002). The open pore conformation of potassium channels. *Nature*, **417**, 523–526.
12. Kelly, B. L. & Gross, A. (2003). Potassium channel gating observed with site-directed mass tagging. *Nature Struct. Biol.* **10**, 280–284.
13. Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T. & MacKinnon, R. (2003). X-ray structure of a voltage-dependent K^+ channel. *Nature*, **423**, 33–41.
14. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. (2003). The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature*, **423**, 42–48.
15. Yifrach, O. & MacKinnon, R. (2002). Energetics of pore opening in a voltage-gated K(+) channel. *Cell*, **111**, 231–239.
16. Schoppa, N. E., McCormack, K., Tanouye, M. A. & Sigworth, F. J. (1992). The size of gating charge in wild-type and mutant Shaker potassium channels. *Science*, **255**, 1712–1715.
17. Seoh, S. A., Sigg, D., Papazian, D. M. & Bezanilla, F. (1996). Voltage-sensing residues in the S2 and S4 segments of the Shaker K^+ channel. *Neuron*, **16**, 1159–1167.
18. Aggarwal, S. K. & MacKinnon, R. (1996). Contribution of the S4 segment to gating charge in the Shaker K^+ channel. *Neuron*, **16**, 1169–1177.
19. Bezanilla, F. (2002). Voltage sensor movements. *J. Gen. Physiol.* **120**, 465–473.
20. Honig, B. H. & Hubbell, W. L. (1984). Stability of "salt bridges" in membrane proteins. *Proc. Natl Acad. Sci. USA*, **81**, 5412–5416.
21. Laine, M., Lin, M. C., Bannister, J. P., Silverman, W. R., Mock, A. F., Roux, B. & Papazain, D. M. (2003). Atomic proximity between S4 segment and pore domain in Shaker potassium channels. *Neuron*, **39**, 467–481.
22. Lee, H. C., Wang, J. M. & Swartz, K. J. (2003). Interaction between extracellular Hanatoxin and the resting conformation of the voltage-sensor paddle in Kv channels. *Neuron*, **40**, 527–536.
23. Cohen, B. E., Grabe, M. & Jan, L. Y. (2003). Answers and questions from the KvAP structures. *Neuron*, **39**, 395–400.
24. Gandhi, C. S., Clark, E., Loots, E., Pralle, A. & Isacoff, E. Y. (2003). The orientation and molecular movement of a K(+) channel voltage-sensing domain. *Neuron*, **40**, 515–525.
25. Broomand, A., Mannikko, R., Larsson, H. P. & Elinder, F. (2003). Molecular movement of the voltage sensor in a k channel. *J. Gen. Physiol.* **122**, 741–748.
26. Sigworth, F. J. (2001). Potassium channel mechanics. *Neuron*, **32**, 555–556.
27. Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.
28. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358.
29. Kass, I. & Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins: Struct. Funct. Genet.* **48**, 611–617.
30. Oliveira, L., Paiva, A. C. & Vriend, G. (2002). Correlated mutation analyses on very large sequence families. *Chembiochem*, **3**, 1010–1017.
31. Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12**, 368–373.
32. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
33. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10**, 59–69.
34. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.

† http://ashtoret.tau.ac.il/~sarel/CorrMut.html

35. Miyata, T., Miyazawa, S. & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236.
36. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1979). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), pp. 353–358, National Biomedical Research Foundation, Washington, DC.
37. Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656.
38. Bradley, E. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
39. Lebart, L., Morineau, A. & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*, Wiley, New York.
40. Hong, K. H. & Miller, C. (2000). The lipid–protein interface of a Shaker K(+) channel. *J. Gen. Physiol.* **115**, 51–58.
41. Li-Smerin, Y., Hackos, D. H. & Swartz, K. J. (2000). A localized interaction surface for voltage-sensing domains on the pore domain of a K^+ channel. *Neuron*, **25**, 411–423.
42. Holmgren, M., Shin, K. S. & Yellen, G. (1998). The activation gate of a voltage-gated K^+ channel can be trapped in the open state by an intersubunit metal bridge. *Neuron*, **21**, 617–621.
43. del Camino, D. & Yellen, G. (2001). Tight steric closure at the intracellular activation gate of a voltage-gated K(+) channel. *Neuron*, **32**, 649–656.
44. Labro, A. J., Raes, A. L., Bellens, I., Ottschytsch, N. & Snyders, D. J. (2003). Gating of shaker-type channels requires the flexibility of S6 caused by prolines. *J. Biol. Chem.* **278**, 50724–50731.
45. Ranganathan, R., Lewis, J. H. & MacKinnon, R. (1996). Spatial localization of the K^+ channel selectivity filter by mutant cycle-based structure analysis. *Neuron*, **16**, 131–139.
46. Li-Smerin, Y., Hackos, D. H. & Swartz, K. J. (2000). alpha-helical structural elements within the voltage-sensing domains of a K(+) channel. *J. Gen. Physiol.* **115**, 33–50.
47. Blaustein, R. O., Cole, P. A., Williams, C. & Miller, C. (2000). Tethered blockers as molecular "tape measures" for a voltage-gated K^+ channel. *Nature Struct. Biol.* **7**, 309–311.
48. Santacruz-Toloza, L., Huang, Y., John, S. A. & Papazian, D. M. (1994). Glycosylation of shaker potassium channel protein in insect cell culture and in Xenopus oocytes. *Biochemistry*, **33**, 5607–5613.
49. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
50. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
51. Lu, Z., Klem, A. M. & Ramu, Y. (2002). Coupling between voltage sensors and activation gate in voltage-gated K^+ channels. *J. Gen. Physiol.* **120**, 663–676.
52. Monks, S. A., Needleman, D. J. & Miller, C. (1999). Helical structure and packing orientation of the S2 segment in the Shaker K^+ channel. *J. Gen. Physiol.* **113**, 415–423.
53. Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. (1984). The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus*). *Cell*, **38**, 835–840.
54. Perozo, E. (2000). Structure and packing orientation of transmembrane segments in voltage-dependent channels. Lessons from perturbation analysis. *J. Gen. Physiol.* **115**, 29–32.
55. Lu, Z., Klem, A. M. & Ramu, Y. (2001). Ion conduction pore is conserved among potassium channels. *Nature*, **413**, 809–813.
56. Smith-Maxwell, C. J., Ledwell, J. L. & Aldrich, R. W. (1998). Uncharged S4 residues and cooperativity in voltage-dependent potassium channel activation. *J. Gen. Physiol.* **111**, 421–439.
57. Dutzler, R., Campbell, E. B. & MacKinnon, R. (2003). Gating the selectivity filter in ClC chloride channels. *Science*, **300**, 108–112.
58. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
59. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
60. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
61. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
62. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
63. Merritt, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524.

Edited by G. von Heijne

(Received 21 January 2004; received in revised form 26 April 2004; accepted 30 April 2004)

BIOINFORMATICS

Vol. 00 no. 00 2003
Pages 1–7

Assigning Transmembrane Segments to Helices in Intermediate-Resolution Structures

Angela Enosh [a,*], *Sarel J. Fleishman* [b], *Nir Ben-Tal* [b] *and Dan Halperin* [a]

[a]*School of Computer Science,* [b]*Department of Biochemistry, Tel Aviv University, Ramat Aviv, 69978, Israel*

ABSTRACT

Motivation: Transmembrane (TM) proteins that form α-helix bundles constitute approximately 50% of contemporary drug targets. Yet, it is difficult to determine their high-resolution ($<$ 4Å) structures. Some TM proteins yield more easily to structure determination using cryo electron microscopy (cryo-EM), though this technique most often results in lower resolution structures, precluding an unambiguous assignment of TM amino-acid sequences to the helices seen in the structure. We present computational tools for assigning the TM segments in the protein's sequence to the helices seen in cryo-EM structures.

Results: The method examines all feasible TM helix assignments and ranks each one based on a score function that was derived from loops in the structures of soluble α-helix bundles. A set of the most likely assignments is then suggested. We tested the method on eight TM chains of known structures such as bacteriorhodopsin and the lactose permease. Our results indicate that many assignments can be rejected at the outset, since they involve the connection of pairs of remotely placed TM helices. The correct assignment received a high score, and was ranked highly among the remaining assignments. For example, in the lactose permease, which contains 12 TM helices, most of which are connected by short loops, only 12 out of 479 million assignments were found to be feasible, and the native one was ranked first.

Availability: The program and the non-redundant set of protein structures used here are available at:
`http://www.cs.tau.ac.il/~angela`

Contact: angela@post.tau.ac.il

1 INTRODUCTION

In recent years, the pace of structure determination of TM proteins has increased, but technical problems related to protein purification and crystallization still hamper TM protein structure determination. Thus, notwithstanding their biomedical importance, less than 30 distinct folds of TM proteins have been solved to date by high-resolution methods such as X-ray crystallography.

Eukaryotic TM proteins form predominantly α-helix bundles in the membrane. These proteins are composed of TM helices and loops, which are typically located on the internal or external sides of the membrane, and connect pairs of consecutive helices. Structure prediction in this class of proteins often relies conceptually on the two-stage model for their assembly in the membrane (Popot and Engelman, 1990). According to this model, TM protein folding begins with the insertion of the TM segments into the membrane as α-helices. In the second stage these helices assemble to form a bundle (reviewed in Popot and Engelman, 2000; White and Wimley, 1999).

Some of the factors stabilizing TM protein structures have been elucidated in recent years on the basis of solved structures and biochemical experiments (e.g., Choma *et al*., 2000; Eilers *et al*., 2000; MacKenzie and Engelman, 1998; Russ and Engelman, 2000). A number of computational methods have been suggested for positioning and orienting the helices comprising the TM domain with respect to one another (e.g., Adams *et al*., 1995; Fleishman and Ben-Tal, 2002; Kim *et al*., 2003; Pellegrini-Calace *et al*., 2003).

Here, we consider a situation in which the locations of the TM helices in $3D$-space can be deduced experimentally. The challenge is then to assign the TM segments in the protein sequence into the corresponding helices in $3D$-space. For concreteness, let us focus on proteins that were solved at intermediate in-plane resolution ($5-10$Å) (Unger, 2001). From these data, one can derive helix positions, as well as their tilt and azimuthal angles with respect to the membrane. However, the individual amino acids cannot be identified, so that the correspondence between the TM segments and the cryo-EM helices cannot be decided unambiguously. So far, no method has tackled this problem.

Providing a solution to the helix-assignment problem is a first step toward modeling of TM proteins. That is, by assigning the TM segments to the helices in the cryo-EM data, conformation space in a modeling exercise can be limited substantially. In addition, helix assignment is directly useful for

*To whom correspondence should be addressed.

structural studies of membrane proteins, as it reveals which helices are in contact with each other, and outlines helices that are located in critical positions, such as around a pore in channels and pumps.

We show here that many putative helix assignments can be eliminated based on the (estimated) maximal lengths of each of the loops in the protein. In addition, we present a novel score function, that was derived on the basis of conformations of loops in α-helix bundles (of soluble proteins), in order to rate the capability of loops to connect each pair of helices. Based on this score function, we ranked assignments of 8 TM-protein chains of known structures taken from the Protein Data Bank (`http://www.rcsb.org/pdb/`), and our results show that the native-state assignment ranks high in many cases.

Terminology and Formal Statement of the Problem. The sequence of a TM protein of the α-helix bundle type, denoted by S, is composed of TM and extra-membrane segments, which connect TM segments that are consecutive in the sequence (Figure 1(a)). The locations of TM segments in protein sequences can be predicted fairly precisely on the basis of sequence data alone (Chen *et al.*, 2002). We denote a TM segment, $T_i \in S$, by $T_i = \{t_{i1}, t_{i2} \dots t_{ik_i}\}$, as an ordered sequence of amino acids from the N- to the C-terminus. Similarly, we denote an extra-membrane segment, $X_i \in S$, by $X_i = \{x_{i1}, x_{i2} \dots x_{ik_i}\}$, as an ordered sequence of amino acids from the N- to the C-terminus. The length of an extra-membrane segment X_i, denoted by $\mathrm{length}(X_i)$, is the number of amino acids in the segment. The maximal distance between two points that can be connected by X_i is denoted by $\mathrm{max\_dist}(X_i) = (\mathrm{length}(X_i) + 1) \times \mathrm{dist}(C\alpha, C\alpha)$, where $\mathrm{dist}(C\alpha, C\alpha)$ is the distance between two consecutive $C\alpha$ atoms, which is typically taken as 3.8Å (Creighton, 1993).

A Helix, $C_i \in C$. Positions, tilt and azimuthal angles of each helix can be extracted from intermediate-resolution cryo-EM maps (Unger, 2001). Canonical α-helices are constructed, and made to fit the cryo-EM map. We represent each such helix by a sequence of coordinates of its $C\alpha$ atoms, $C_i = \{c_{i1}, c_{i2} \dots c_{ik_i}\}$. The membrane can be regarded as a region in $3D$ bounded by two planes, to which we refer as the inner and the outer planes of the membrane. We define an order on a helix C_i in the sense that c_{i1} is the closest atom to the inner plane of the membrane, and c_{ik_i} is the closest atom to the outer plane of the membrane. We denote the internal $C\alpha$ atom by $\mathrm{internal}(C_i) = c_{i1}$, and the external $C\alpha$ atom by $\mathrm{external}(C_i) = c_{ik_i}$.

It should be noted that the positions of helices deduced from cryo-EM in this manner suffer from imprecision. First, the orientation of the helices around their principal axes cannot be derived from the cryo-EM map due to the limited in-plane resolution (typically, $5 - 10$Å (Unger, 2001)). Moreover, the low resolution along the axis normal to the membrane plane ($12 - 30$Å) entails a large distortion in the positions of helices along this axis. For simplicity we avoid dealing with these inaccuracies in the description of our algorithm. However, as described in Appendix A, our program takes the noisiness that results from the limited resolution into account by also testing helix positions that are in the vicinity of those seen in the cryo-EM data.

Formal Definition of Our Goals. Given the secondary structure classification of a TM protein sequence $S = \{T_1, X_1, T_2, \dots X_{n-1}, T_n\}$ and a set of helix locations in $3D$-space $C = \{C_1, C_2, \dots C_n\}$, derived from the cryo-EM map, *(i)* find all the *feasible* assignments between the $T_i's$ and the $C_i's$, namely find a permutation σ such that for each $1 \le i \le n$, T_i is assigned to $C_{\sigma(i)}$, and *(ii)* attribute a score to each assignment based on its compatibility with the locations of the helices in $3D$-space.

In principle, a TM segment can be assigned to a helix in $3D$-space with its N- and C-termini on the inner and outer sides of the membrane, respectively, or vice versa. However, it is possible to resolve this ambiguity experimentally. Hence, the number of all the assignments is $n!$. A brute-force approach would require the generation of all these assignments. To reduce this immense computational burden, at the outset we exploit the maximal lengths of the extra-membrane segments to filter out impossible assignments. Suppose we want to match two consecutive segments T_i and T_{i+1} to the helices, C_k and C_m, correspondingly, such that the extra-membrane segment X_i lies on the external side of the membrane. A necessary condition for this assignment to be valid is that the maximal length of the extra-membrane segment ($\mathrm{max\_dist}(X_i)$) is longer than the distance between $\mathrm{external}(C_k)$ and $\mathrm{external}(C_m)$). In the same manner, if X_i should connect the helices on the internal side of the membrane, its maximal length should be larger than the distance between $\mathrm{internal}(C_k)$ and $\mathrm{internal}(C_m)$. Consequently, if this condition does not hold, the assignment should be ignored from the outset.

2 THE ALGORITHM

Our algorithm proceeds in two stages: *Pruning by Distance Constraints* — construction of an assignment graph that contains *only* the set of feasible assignments, i.e., assignments in which the maximal lengths of the extra-membrane segments are longer than the distances between the helices that they connect (Figure 1). This stage is followed by *Loop Conformation Scoring* — attributing scores to the feasible assignments based on their compatibility with the locations of the helices in $3D$-space.

2.1 Pruning by Distance Constraints

We wish to filter out as many assignments as possible, without eliminating the right one. For this purpose we construct a directed acyclic graph $G(V, E_{int} \bigcup E_{ext})$, such as the one in Figure 1(c), where:

$$V = \{(T_i, C_j) \mid 1 \le i, j \le n\},$$
$$E_{int} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \mathrm{dist}(\mathrm{internal}(C_j), \mathrm{internal}(C_m)) \le \mathrm{max\_dist}(X_i)\},$$
$$E_{ext} = \{(T_i, C_j) \rightarrow (T_{i+1}, C_m) \mid \mathrm{dist}(\mathrm{external}(C_j), \mathrm{external}(C_m)) \le \mathrm{max\_dist}(X_i)\}$$

V stands for the vertices and E stands for the edges in G. There are two kinds of edges in G: external (E_{ext}) and internal (E_{int}). There is an edge $e \in E_{ext}$, if and only if the two consecutive TM segments T_i and T_{i+1} can be matched congruently to C_j and C_m. Namely, the extra-membrane segment X_i between T_i and T_{i+1} is sufficiently long to connect the two points $\mathrm{external}(C_j)$ and $\mathrm{external}(C_m)$ on the external side of the membrane. The same applies to the E_{int} edges where X_i is sufficiently long to connect $\mathrm{internal}(C_j)$ and $\mathrm{internal}(C_m)$ on the internal side of the membrane.

We construct G in a bottom-up fashion, i.e., the levels in G are constructed from the nth to the 1st level (where n is the number of TM segments in the protein). The kth level in the graph consists of vertices composed of T_k, namely $\{(T_k, C_j) \mid 1 \le j \le n\}$. Given the set of nodes $\{(T_k, C_j) \mid 1 \le j \le n\}$ in the kth level, we construct the $(k-1)$st level as follows. For each vertex (T_k, C_j) we go over all the helices $C_t \in C \setminus \{C_j\}$ and if X_{k-1} can connect the two helices C_t and C_j on the external or internal side of the membrane, we add the vertex (T_{k-1}, C_t) (if it is still missing) to the $(k-1)$st level, and a directed edge $e = ((T_{k-1}, C_t), (T_k, C_j))$, where $e \in E_{ext}$ or $e \in E_{int}$. Thus, a directed edge $e \in \{E_{ext} \bigcup E_{int}\}$ can appear only between two consecutive levels. At the beginning, all of the vertices (T_n, C_j) in the nth level are examined against the pairs (T_{n-1}, C_t) where $C_t \in C \setminus \{C_j\}$, and created if and only if the above condition holds. After construction of the graph G we can eliminate all of the nodes between the second to the nth level that do not have at least one entering edge.

A path $\pi = \{v_1, e_1, v_2, e_2, v_3 \ldots e_{n-1}, v_n\}$ in the graph G is considered valid if it starts at the first level of G, ends at the nth level of G, and it is comprised of an alternating sequence of external and internal edges (either $\{e_k \mid k \text{ even}\}$ are external and $\{e_k \mid k \text{ odd}\}$ are internal, or vice versa). In addition, we require that π does not contain two vertices with the same helix (the C_k's in all the vertices $v_i = (T_i, C_k)$ are distinct). Each valid path π defines a feasible assignment between the TM segments of S and the helices in C. It will be shown that this pruning phase eliminates many infeasible assignments when the protein contains short loops (namely, loops whose lengths are less than 6).

2.2 Ranking the Feasible Assignments

In the following stage, a score is assigned to the feasible assignments that are stored in G based on the suitability of the loops to connect helices in the structure. Each feasible assignment is a permutation σ^k which assigns the TM segments $T_1 \ldots T_n$ to the helices $C_{\sigma^k(1)} \ldots C_{\sigma^k(n)}$, where $1 \le k \le n!$. We define the score function F of a permutation σ^k as follows:

$$F(\sigma^k) = \sum_{i=1}^{n-1} f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)})$$

where f scores the suitability of assigning the consecutive TM segments T_i and T_{i+1} to helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$. Namely, f defines the feasibility of connecting the two helices in $3D$-space by X_i.

The problem of adjusting an extra-membrane segment to connect two fixed secondary structures is related to the well-known kinematics problem of loop-closure (Canutescu and Dunbrack, 2003; Manocha and Zhu, 1994; Wedemeyer and Scheraga, 1999; Wojcik *et al.*, 1999; Xiang *et al.*, 2002). However, our problem is slightly different. We wish to rank the assignments instead of predicting the conformation of the extra-membrane loops as in the classic loop-closure problem, since the native extra-membrane segment, which connects the two helices, is unknown. Hence, we seek to define a *score* for matching the extra-membrane segment to connect the two helices in a way that the native match is assigned the highest score.

The evaluation of f is based on the length of the extra-membrane segment X_i and on a statistical analysis we have conducted on solved structures of soluble

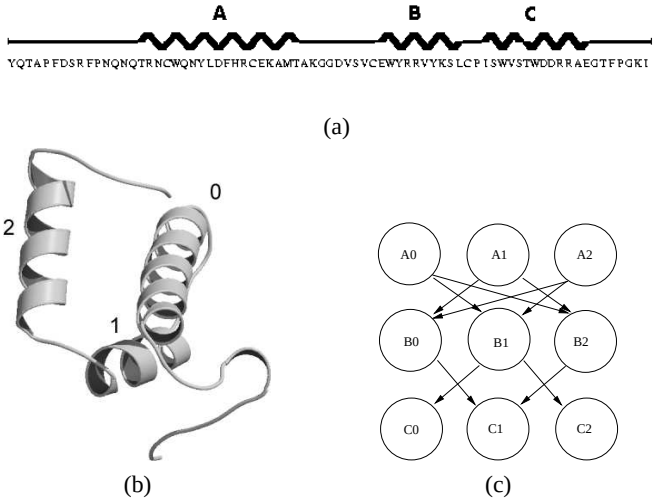


Fig. 1. (a) The locations of the three TM segments in the sequence of chain H of the cytochrome c oxidase. (b) The corresponding $3D$ structure. (c) The assignment graph of this chain. The numbers represent the helices and the letters represent the TM segments. There are four valid paths (feasible assignments) in the graph which are: $(A0, B1, C2)$, $(A0, B2, C1)$, $(A2, B0, C1)$ and $(A2, B1, C0)$. Notice that there is no edge between $(B0)$ and $(C2)$, for example, since the loop between the TM segments B and C is too short to connect helices 0 and 2.

proteins taken from the Protein Sequence Culling Server (http://www.fccc.edu/research/labs/dunbrack/pisces/) in a preprocessing phase. We restricted our survey to protein sections comprised of two consecutive helices with a loop region between them, namely to *helix-loop-helix* motifs, where secondary-structure elements are assigned according to DSSP (Kabsch and Sander, 1983).

The preprocessing phase. We denote the two consecutive helices in a helix-loop-helix motif, by A and B, and the loop region which connects them by L, and set $l = \text{length}(L)$. Let us examine the helix-loop-helix motifs with the same loop length l ($2 \leq l \leq 7$). All of these motifs (A, L, B) were placed in a common orthogonal reference frame, so that the helices A of all of the motifs overlap. Transforming these motifs to the common reference frame yields a set of points in $3D$-space that represents the starting points of the second helices (i.e., B's) relative to the common first helix (i.e., the overlapping A's).

All of these starting points, denoted by p_i ($1 \leq i \leq N$, where N is the number of helix-loop-helix motifs), were stored in a KD-tree data structure [1]. Since the lengths of the loops in these motifs have a great impact on the locations of the points, p_i's, in $3D$-space, these points were stored in 6 distinct KD-trees which we denote by KD_l, $2 \leq l \leq 7$, one tree per length l. Our results indicate that these points are distributed non-uniformly in $3D$-space. For an illustration, Figure 2 shows the starting points in the common reference frame for $l = 3$ and $l = 4$.

The scoring phase. We compute $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)})$ as follows. We place the two helices $C_{\sigma^k(i)}$ and $C_{\sigma^k(i+1)}$ in the common orthogonal reference frame in the same manner as we have done in the preprocessing phase, and obtain the new starting point q of the helix $C_{\sigma^k(i+1)}$. Given q and the starting points of helix-loop-helix motifs with loop length $x = \text{length}(X_i)$ from the preprocessing phase, the score depends on two criteria: the number of neighboring points in the vicinity of q and the distances between these neighboring points and q.

Let Q be a cube centered at q with side size $(10 \cdot x)$Å. We query KD_x to find the points that were stored in the preprocessing phase which occur in Q. Q represents the region in $3D$-space for the clusters of points in the appropriate KD-tree we wish to examine. The score for this assignment is based on the sum of the distances between q and the derived points that were found inside Q. The score was constructed with the aim of favoring loops that have been observed many times in the protein database we have used. It is, therefore, defined in the form of a colony function (Xiang *et al.*, 2002), whereby loops in the database that are similar to the query make a more significant contribution to the loop's score. Formally, $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)}) = \sum_{r \in Q} e^{-\text{dist}(q,r)}$.

When $x \geq 8$, we do not obtain significant information about the quality of the assignment due to the low frequency of occurrence of long loops in the helix-loop-helix motif in the specified protein database. Thus, for $\text{length}(X_i) \geq 8$, we have set $f(X_i, C_{\sigma^k(i)}, C_{\sigma^k(i+1)}) = 0$.

Given the assignment graph G that was generated in the pruning phase, we assign a weight, $\text{weight}(e) = f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)})$ to each edge in the graph, namely to each $e = (u, v)$ where $u = (T_i, C_{\sigma(i)})$ and $v = (T_{i+1}, C_{\sigma(i+1)})$. G is an acyclic directed weighted graph. Each valid path in G defines a feasible assignment, and its score is the sum of the weights of the edges in the path, i.e., $F(\pi) = \sum_{e \in \pi} \text{weight}(e)$.

3 THE DISTRIBUTION OF END POINTS OF SHORT LOOPS IS HIGHLY NONUNIFORM

Structures of helix-loop-helix motifs (resolution of 2Å or less, and R-factor of 0.3 or less) of soluble proteins were selected from the Protein Sequence Culling Server (http://www.fccc.edu/research/labs/dunbrack/pisces/). To reduce the bias inherent in the Protein Data Bank, only proteins whose sequences were less than 20% identical were selected. The secondary structures were assigned by DSSP (Kabsch and Sander, 1983). We looked only at helix-loop-helix motifs containing two helical regions of at least 8 amino acids each, which are connected by loops of lengths 2 to 7 amino acids (Table 1). The order of the two helices was specified from the N- to the C-terminus. Entries were classified by the loop lengths. Each loop of length l (where $2 \leq l \leq 7$) contributed to our analysis a point in $3D$-space corresponding to the beginning of helix B. The distribution of the examined points in the common reference frame for short loops (i.e., lengths three and four) is shown in Figure 2. Loops longer than seven were not considered, due to their low frequency of occurrence in our dataset.

Table 1. Helix-loop-helix motifs classified by loop length

Loop Length	2	3	4	5	6	7
Number of motifs	456	260	171	167	98	36

Helix-loop-helix motifs derived from the Protein Sequence Culling Server and classified by their loop lengths.

The scoring function is greatly dependent on this protein database analysis. To understand why our scoring function performs well (as indicated by the results reported below), consider for example the case where $l = 4$ (Figures 2(d-f)), i.e., the loop L has four $C\alpha$'s. In this case L has 8 degrees of freedom (each $C\alpha$ contributes two degrees of freedom ϕ and

[1] KD-trees are orthogonal range-search structures. They are used to store a set P of points in R^d so that the subset of P inside a query axis-aligned hyperbox can be reported efficiently. See, e.g., (de Berg *et al.*, 2000) for details.

ψ). By sheer kinematics considerations, if we fix one end of the loop, the reachable space by the other end (we refer to it as the *free end*) is large, practically limited only by the stretch of the loop (the conformation that has the largest diameter). However, Figures 2(d-f) show that the locus of the free end in length-four loops connecting two helices is limited to a few clusters of points in $3D$-space. Our scoring takes advantage of this phenomenon, which is highly significant in loops of lengths two through five, but is still substantially noticeable in loops of lengths up to seven.

4 IMPLEMENTATION

We verified the scoring function by applying it to 8 TM protein chains, whose structures were solved using X-ray crystallography (Table 2). We restricted our study to those chains, whose TM segments did not contain half-helices or loops (except for the glycerol facilitator, as discussed below). Moreover, we did not consider proteins that contain long extra-membrane segments that could form large domains. It should be noted that the results reported below were derived solely from the solved structures of TM proteins.

The algorithm has been implemented for two distinct cases: (i) using accurate data of the locations of helices as derived from the Protein Data Bank; and (ii) using noisy data, i.e., uncertainty with regard to the positions of the helices. In case (i) the algorithm assumes that the helices are located and oriented in their native conformations. In case (ii), the algorithm assumes that the orientations and locations of the helices are known only approximately. However, in real cases, thanks to the cryo-EM data, we will know that the native helices are located in bounded regions. Therefore, we examine all of the possible orientations and locations of the helices in these bounded regions. The exact definition of these regions is provided in Appendix A.

The two implemented cases (using accurate and noisy data) are examined in Table 2 by the number of feasible assignments that remain after the pruning phase and by the rank of the score of the native assignment with respect to other assignments. In most of the examined TM proteins, the table shows that the native assignment ranks very highly, which implies that the combination of the pruning and scoring phases yields a reliable tool for assigning TM segments to helices.

For example, bacteriorhodopsin (1c3w) is composed of 7 helices, and thus has $7! = 5040$ possible assignments. The number of feasible assignments that remained after the pruning phase is 44. Applying the score function and sorting all of the 44 assignments by their scores, the native assignment was ranked third. When using the noisy data, the list of feasible assignments expanded, but the rank of the native state (13) did not change dramatically, which implies that our score function deals well with this level of noise.

The strength of the pruning phase is clearly shown for the lactose permease (1pv6), where out of 479 million possible assignments, the number of feasible assignments in both cases (i, ii) was below 13 and the running time was relatively short since the assignment graph ruled out many assignments which were not examined. Our method yielded poor results for the glycerol facilitator (1fx8) due to a 24 residue loop which contains a half TM helix. It is rather encouraging that even in this pathological case the algorithm removed approximately half of the potential assignments (352 out of 720) and ranked the native state to be 119.

5 DISCUSSION

A novel method for assigning TM spans in the sequence of an integral membrane protein to the approximate locations of the helices in $3D$-space was presented here. Each of the possible assignments is evaluated based on the compatibility of the extra-membrane segments with the suggested relative locations of the helices. Our results show that in TM proteins with extra-membrane segments of 7 residues or less, the vast majority of the putative assignments can be rejected from the outset, since they involve the connection by short loops of pairs of TM helices that are spatially distant from each other. In the lactose permease, for example, only 12 out of 479 million putative assignments were found to be feasible based on this criterion. The significant reduction in the number of assignments is due to the short lengths of the extra-membrane segments. It demonstrates that, in practice, the complexity of the TM helix assignment problem scales with the lengths of these segments rather than with the number of TM helices.

The feasible assignments are then screened based on the suitability of each of the extra-membrane segments to adopt a conformation that could connect the adjacent TM helices. This is done using a novel knowledge-based score function that was derived from the conformations of loops in helix-loop-helix motifs. Our results show that this function ranks the TM helix assignment of the native structure high among the other feasible assignments. This is best demonstrated with chain H of the cytochrome c oxidase, where the native structure ranks first among the feasible assignments.

In the typical case, the locations of the TM helices in $3D$-space will be determined using medium-resolution data, e.g., from cryo-EM studies at in-plane resolutions of $5 - 10$Å. At such resolution, one can only derive the approximate locations of the TM helices in $3D$-space. The results demonstrate that the method is robust to changes in the locations of the TM helices; the native-state assignment ranks high among the feasible assignments, even when using noisy data (Table 2).

Our results are very encouraging in that the problem of TM helix-assignment is significantly reduced, and yet in the typical case, the analysis is likely to result in several putative assignments rather than only one. We anticipate that the set of potential assignments may be further reduced based on available empirical data, e.g., from biochemical, molecular and genetic studies. Finally, forward-looking experiments

Table 2. The performance of the two-stage (pruning and scoring) algorithm using accurate and noisy data

Name	PDB	Loop Lengths	n_h	$n_{\rm pos}$	(i) Accurate		(ii) Noisy	
					$n_{\rm feas}$	rank	$n_{\rm feas}$	rank
Bacteriorhodopsin	1c3w	3,14,2,3,10,4	7	5040	44	3	948	13
Sensory rhodopsin	1h68	7,12,2,3,3,4	7	5040	84	2	512	48
Cytochrome c oxidase	1occC	3,5,19,2,7,7	7	5040	74	7	335	62
Cytochrome c oxidase	1occE	5,6,1,1	5	120	2	2	2	1
Cytochrome c oxidase	1occH	7,2	3	6	4	1	6	1
Glycerol facilitator	1fx8	6,19,24,8,4	6	720	236	8	352	119
Halorhodopsin	1e12	2,20,2,4,1,5	7	5040	34	5	73	22
Lactose permease	1pv6	3,2,1,3,1,24,3,1,3,1,1	12	$> 10^8$	7	3	12	1

Classification and comparison of the results using (i) accurate helix positions derived from the PDB and (ii) noisy data. The set of TM proteins of known $3D$ structures that were studied are indicated by their names and pdb entries. The subunit is indicated by the last letter. The proteins are classified by the number of TM helices (n_h), their loop lengths, and the number of possible assignments $n_{\rm pos} = n_h!$. The results are categorized by the number of feasible assignments ($n_{\rm feas}$) that remained following the pruning phase and by the position of the native assignment (rank) with respect to other feasible assignments. We ran the program on PC Intel Pentium IV, CPU 2.4GHz, 256 MB RAM, and the running time using the accurate data was below 2 seconds for each of the proteins. When using the noisy data, the running time varied between 8 seconds for 1occ subunit H and 6.5 minutes for 1pv6. Currently we are working on additional TM proteins. The results obtained from processing these cases will be available soon at `http://www.cs.tau.ac.il/~angela`.

may be designed to select the native assignment out of a few possibilities.

The application of the method to oligomeric TM proteins, such as cytochrome c oxidase may complicate the analysis. In the present study, the subunit boundaries were taken as a given, but, if these are unknown, it may be necessary to examine various molecular boundaries, which would entail an increase in the dimensionality of the problem.

To demonstrate the method's usefulness we are applying it to the assignment of the TM helices in the microsomal glutathione transferase 1 (MGST1). This protein is a member of the MAPEG (membrane-associated proteins in eicosanoid and glutathione metabolism) superfamily of TM enzymes (Jakobsson *et al*., 1999).

MGST1 is a homotrimer, in which each monomer is composed of 4 TM segments. The $3D$ structure of MGST1 was determined at an in-plane resolution of 6Å using cryo-EM (Holm *et al*., 2002; Schmidt-Krey *et al*., 2000). The electron-density map shows three repeats of 4 rod-like densities, which presumably correspond to the 12 TM helices of the homotrimer. Our preliminary results show that only a few assignments are consistent with the structure (data not shown).

ACKNOWLEDGEMENT

Work reported in this paper has been supported in part by the IST Programmes of the EU as Shared-cost RTD (FET Open) Projects under Contract No IST-2000-26473 (ECG - Effective Computational Geometry for Curves and Surfaces) and No IST-2001-39250 (MOVIE - Motion Planning in Virtual Environments), by The Israel Science Foundation founded by the Israel Academy of Sciences and Humanities (Center for Geometric Computing and its Applications), by the Hermann Minkowski – Minerva Center for Geometry at Tel Aviv University and by Nofar grant from the Israel Ministry of Trade and Industry. SJF was supported by a doctoral fellowship from the Clore Israel Foundation.

REFERENCES

Adams,P.D., Arkin,I.T., Engelman,D.M. and Brunger,A.T. (1995) Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat. Struct. Biol.*, **2**, 154-162.

de Berg,M., van Kreveld,M., Overmars,M. and Schwarzkopf,O. (2000) Computational Geometry: Algorithms and Applications, 2^{nd} Edition. Springer-Verlag, Berlin.

Canutescu,A.A. and Dunbrack,R.L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.*, **12**, 963-972.

Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774-2791.

Choma,C., Gratkowski,H., Lear,J.D. and DeGrado,W.F. (2000) Asparagine-mediated self-association of a model transmembrane helix. *Nat. Struct. Biol.*, **7**, 161-166.

Creighton,T.E. (1993) Poteins, Structures and Molecular Properties. Freeman, New York.

Eilers,M., Shekar,S.C., Shieh,T., Smith,S.O. and Fleming,P.J. (2000) Internal packing of helical membrane proteins. *Proc. Natl. Acad. Sci. USA.*, **97**, 5796-5801.

Fleishman,S.J. and Ben-Tal,N. (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.*, **321**, 363-378.

Holm,P.J., Morgenstern,R. and Hebert,H. (2002) The 3-D structure of microsomal glutathione transferase 1 at 6 A resolution as determined by electron crystallography of p22(1)2(1) crystals. *Biochim. Biophys. Acta*, **1594**, 276-285.

Jakobsson,P.J., Morgenstern,R., Mancini,J., Ford-Hutchinson,A. and Persson,B. (1999) Common structural features of MAPEG – a widespread superfamily of membrane associated proteins

with highly divergent functions in eicosanoid and glutathione metabolism. *Protein Sci.*, **8**, 689-692.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonding and geometrical features, `http://www.cmbi.kun.nl/gv/dssp/`. *Biopolymers*, **22**, 2577-2637.

Kim,S., Chamberlain,A.K. and Bowie,J.U. (2003) A simple method for modeling transmembrane helix oligomers. *J. Mol. Biol.*, **329**, 831-840.

MacKenzie,K.R. and Engelman,D.M. (1998) Structure-based prediction of the stability of transmembrane helix- helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl. Acad. Sci. USA.*, **95**, 3583-3590.

Manocha,D. and Zhu,Y. (1994) Kinematic manipulation of molecular chains subject to rigid constraints. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 285-293.

Pellegrini-Calace,M., Carotti,A. and Jones,D.T. (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins*, **50**, 537-545.

Popot,J.L. and Engelman,D.M. (1990) Membrane protein folding and oligomerization: The two-stage model. *Biochemistry*, **29**, 4031-4037.

Popot,J.L. and Engelman,D.M. (2000) Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.*, **69**, 881-922.

Russ,W.P. and Engelman,D.M. (2000) The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.*, **296**, 911-919.

Schmidt-Krey,I., Mitsuoka,K., Hirai,T., Murata,K., Cheng,Y., Fujiyoshi,Y., Morgenstern,R. and Hebert,H. (2000) The three-dimensional map of microsomal glutathione transferase 1 at 6 A resolution. *EMBO*, **19**, 6311-6316.

Unger,V.M. (2001) Electron Cryomicroscopy Methods. *Curr. Opin. Struct. Biol.*, **11**, 548-554.

White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**, 319-365.

Wedemeyer,W.J. and Scheraga,H.A. (1999) Exact analytical loop closure in proteins using polynomial equations. *J. Comp. Chem.*, **20**, 819-844.

Wojcik,J., Mornon,J.P. and Chomilier,J. (1999) New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol.*, **289**, 1469-1490.

Xiang,Z., Soto,C.S. and Honig,B. (2002) Evaluating conformational free energies: The colony energy and its application to the problem of protein loop prediction. *Proc. Natl. Acad. Sci.*, **99**, 7432-7437.

APPENDIX A: DEALING WITH THE UNCERTAINTY IN CRYO-EM DATA

Cryo-EM studies at $5-10$Å in-plane resolution provide only the approximate locations of the helix-axes positions and orientations. The uncertainty in $3D$-space is mainly due to two reasons (Figure 3): (i) the unknown orientation of the helix with respect to its axis; (ii) the unknown translation of the helix along its axis.

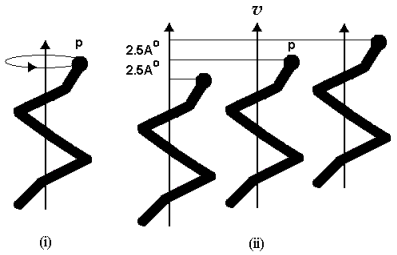


Fig. 3. The detection of helix location in cryo-EM data involves two types of uncertainties: (i) the exact orientation of the helix with respect to its axis is unknown; (ii) the helix derived from the cryo-EM (middle) may shift toward the outer plane ((ii) right) or toward the inner plane ((ii) left) of the membrane.

We now redefine the score function $f(X_i, C_{\sigma(i)}, C_{\sigma(i+1)})$ that was introduced in Section 2.2 to suit the noisiness of the data. For simplicity, we assume that X_i should connect the two helices in the external side of the membrane. We denote by p' and q' the native positions of the external $C\alpha$ atoms of helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$ respectively. The above uncertainties may affect f dramatically, since it strongly depends on the points $p =$ external $(C_{\sigma(i)})$ and $q =$ external $(C_{\sigma(i+1)})$, whose locations are known only approximately. However, the locations of p' and q' are restricted to bounded regions as shown below.

Let us examine the surface where p' can possibly be located accounting for the imprecision in the model. We call this surface the envelope of p and denote it by $E(p)$ (the same discussion applies to q'). $E(p)$ is defined as follows (the numbering corresponds to the numbers of the reasons for imprecision): (i) p' can be located on a circle in $3D$-space centered at the helix axis (Figure $3(i)$); (ii) p' can be located in the range $[p - v \cdot 2.5, p + v \cdot 2.5]$ where v is the unit vector that coincides with the helix axis toward the external side of the membrane (Figure $3(ii)$). It follows that $E(p)$ has a cylindrical envelope shape with radius 2.5Å (typically, radius of a helix) and its height is set to 5Å.

Given p and q as specified above, each pair of points $p_k \in E(p)$ and $q_j \in E(q)$, can be regarded as the external $C\alpha$ atoms of the native helices. We pick uniformly distributed random points $p_k \in E(p)$ and $q_j \in E(q)$ and transform the helices $C_{\sigma(i)}$ and $C_{\sigma(i+1)}$ so that p and q will coincide with p_k and q_j, respectively (without changing their axes' directions). The transformed helices are denoted by $T_k(C_{\sigma(i)})$ and $T_j(C_{\sigma(i+1)})$. To account for this imprecision, we modify the score function f to be: $\max_{k \in E(p), j \in E(q)} f(X_i, T_k(C_{\sigma(i)}), T_j(C_{\sigma(i+1)}))$.

It can be shown that in order to cover the envelope $E(p)$ adequately, we need to sample $n = 135$ points on $E(p)$. By adequately we mean that with high probability (> 0.98), the native point p' will be at distance $\epsilon = 1$Å at most from at least one of the samples points in $E(p)$.

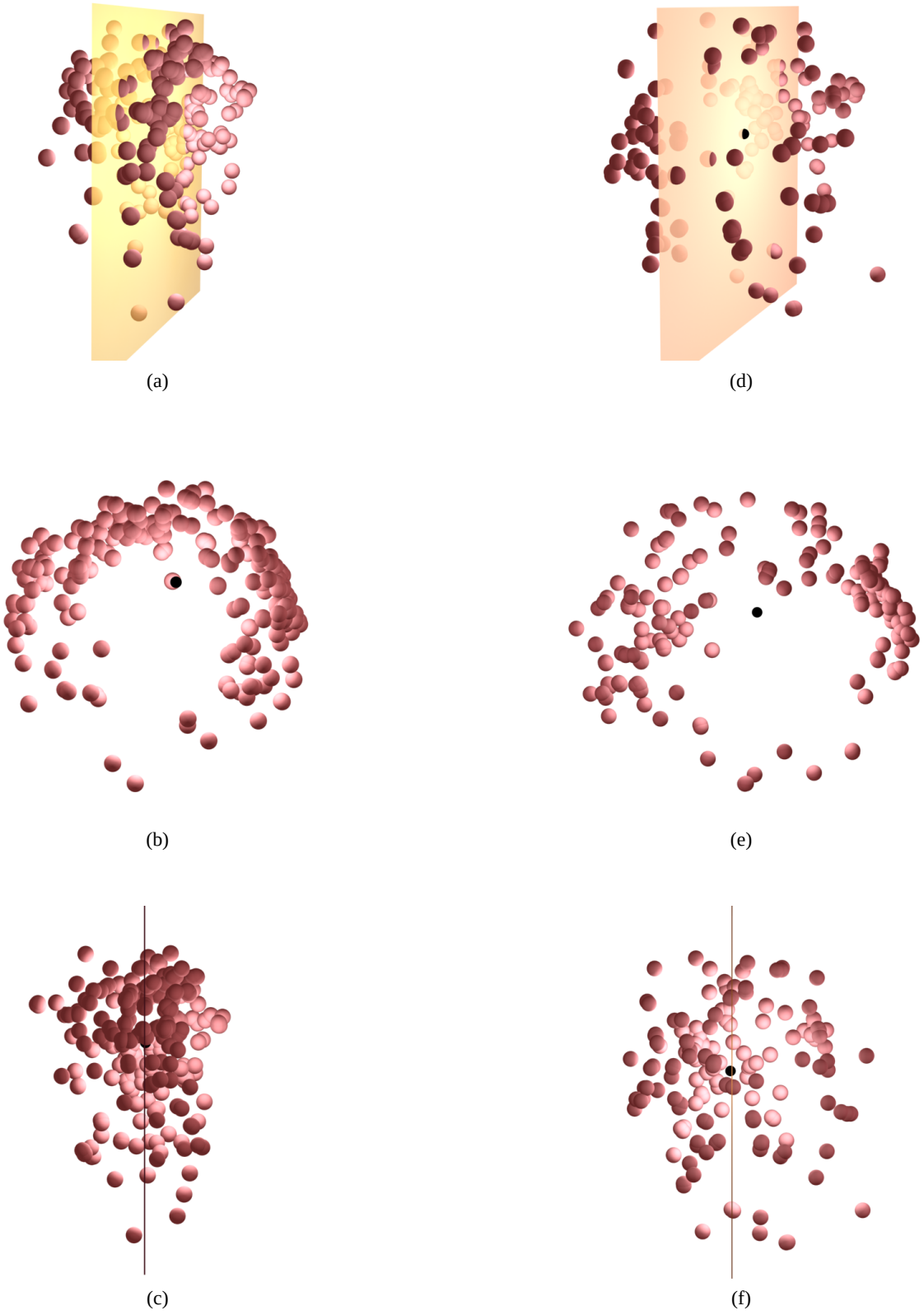


Fig. 2. The distribution of the starting points of helices B's in $3D$-space derived from the helix-loop-helix motifs (A, L, B) with loop lengths 3(a-c) on the left and 4(d-f) on the right. The black spot marks the origin of the common reference frame. Figures (a,d) display the points together with their least-mean-square (LMS) plane. The view point of figures (b,e) is the normal to the LMS plane. Figures (c,f) present a view from the side on the LMS plane. It can be seen that the starting points of 3 amino acids loops create a torus-like shape in $3D$-space.

Molecular Cell, Vol. 15, 879–888, September 24, 2004, Copyright ©2004 by Cell Press

A C^{α} Model for the Transmembrane α Helices of Gap Junction Intercellular Channels

Sarel J. Fleishman,[1] Vinzenz M. Unger,[2] Mark Yeager,[3,4] and Nir Ben-Tal[1,*]
[1]Department of Biochemistry
George S. Wise Faculty of Life Sciences
Tel-Aviv University
Ramat Aviv, 69978
Israel
[2]Department of Molecular Biophysics and Biochemistry
Yale University
P.O. Box 208024
New Haven, Connecticut 06520
[3]Department of Cell Biology
The Scripps Research Institute
10550 North Torrey Pines Road
La Jolla, California 92037
[4]Division of Cardiovascular Diseases
Scripps Clinic
10666 North Torrey Pines Road
La Jolla, California 92037

Summary

Gap junction channels connect the cytoplasms of apposed cells via an intercellular conduit formed by the end-to-end docking of two hexameric hemichannels called connexons. We used electron cryomicroscopy to derive a three-dimensional density map at 5.7 Å in-plane and 19.8 Å vertical resolution, allowing us to identify the positions and tilt angles for the 24 α helices within each hemichannel. The four hydrophobic segments in connexin sequences were assigned to the α helices in the map based on biochemical and phylogenetic data. Analyses of evolutionary conservation and compensatory mutations in connexin evolution identified the packing interfaces between the helices. The final model, which specifies the coordinates of C^{α} atoms in the transmembrane domain, provides a structural basis for understanding the different physiological effects of almost 30 mutations and polymorphisms in terms of structural deformations at the interfaces between helices, revealing an intimate connection between molecular structure and disease.

Introduction

A gap junction channel is formed by the end-to-end docking of two hexameric hemichannels or connexons (Kumar and Gilula, 1996). Each hexamer is formed by six connexin subunits (Cascio et al., 1995) that are composed of four hydrophobic transmembrane (TM) segments designated M1–M4 from the N- to the C terminus (Milks et al., 1988). The intercellular pore of gap junction channels is roughly 15 Å in diameter and allows transport of cytoplasmic secondary messengers, thereby mediating signaling and ion current flow between neighboring cells. Over the past several decades, the important role that gap junctions play in coordinating tissue and organ physiology, e.g., in the heart, ear, skin, and pancreas, has been increasingly recognized (Harris, 2001). A number of genetic conditions in humans and mouse models involving the skin, neurodegenerative and developmental diseases, and most cases of nonsyndromic hereditary deafness have been attributed to mutations in connexins (reviewed by Kelsell et al., 2001).

We previously used electron cryomicroscopy (cryo-EM) and image analysis to solve the structure of a recombinant gap junction channel formed by a C-terminal truncation mutant of Cx43. The three-dimensional (3D) density map at 7.5 Å in-plane resolution revealed the close packing of 24 α helices within each connexon (Unger et al., 1999). Since publication of the original map, improvements in the data analysis have allowed calculation of a map with 5.7 Å in-plane and 19.8 Å vertical resolution. Each of the helices is clearly resolved from its neighbors in the TM domain, and the helices' centers of gravity are also discernible, allowing accurate determination of the helix positions, tilt, and azimuthal angles. However, even in this improved map, connecting loops remained largely undefined either because of limitations in the vertical resolution (in the nonhelical structure of extracellular loops) or disorder (in the cytoplasmic domains). This precluded direct assignment of the helices in the map to the TM domains in the connexin sequence. Consequently, the molecular basis for ionic conduction, channel permeability, and gating properties among the various connexin isoforms could not be inferred directly from the cryo-EM map (Harris, 2001).

However, there is a large body of biochemical and biophysical evidence (reviewed by Harris, 2001) that provides insight into the TM boundaries for M1–M4 and subunit topology (Bennett et al., 1994) and the identities of the pore-lining helices (Kronengold et al., 2003; Skerrett et al., 2002; Zhou et al., 1997). We used these data to assign the TM segments M1–M4 to the helices observed in the cryo-EM map (Unger et al., 1999). We then combined the helix positions, tilt, and azimuthal angles from the improved cryo-EM map with computational methods for the analysis of evolutionary conservation and hydrophobicity of amino acid residues (Fleishman et al., 2004b) to generate a C^{α} trace model of the 24 helices in the connexon. Even though the cryo-EM map corresponds to Cx43, our analysis was based on the human Cx32 sequence since there is a wealth of biochemical, mutational, and genetic data for this isoform. Modeling Cx32 on the basis of the Cx43 structure is justified because the two proteins exhibit 50% sequence identity in the predicted TM residues of M1–M4 (Yeager and Gilula, 1992). Moreover, various connexins assemble to form heteromeric connexons (Harris, 2001). It is therefore very likely that connexins share a common architecture, at least in the TM domain. Consequently, the model we describe should serve as a template for other connexins.

Our approach followed that used by Baldwin et al. (1997) to predict the structure of the TM domain of vertebrate rhodopsin based on a cryo-EM map at 9 Å in-

*Correspondence: nirb@tauex.tau.ac.il

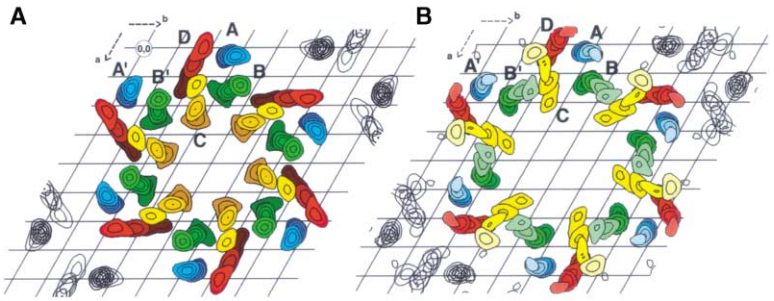


Figure 1. Overlay of Cross-Sections of the 3D Density Map of One Connexon Derived by Electron Cryocrystallography

Counting from the middle of the extracellular gap and toward the observer, sections +14, +18, and +24 (A) and +20, +24, +29, and +34 (B) were used. The approximate boundary between the membrane and the extracellular gap corresponds to section +8 (not shown). The vertical distance between consecutive sections is 2 Å. Densities belonging to the same helices are represented by the same base color, with the darkest and lightest shades corresponding to densities in sections +14 and +34, respectively. Helices were arbitrarily marked A–D and A′ and B′ (which are symmetry related to A and B) to provide a reference for discussion. The position marked (0,0) was used to generate grid coordinates for the locations of helices A–D given in Table 1. The spacing between grid lines is 10 Å, and the map was contoured starting at 1.5σ above the mean.

plane and 16.5 Å vertical resolution (Unger et al., 1997). The model of rhodopsin was shown to be quite accurate (Bourne and Meng, 2000) when compared to the subsequent high-resolution X-ray structure (3.2 Å rmsd) (Palczewski et al., 2000). We have used a similar approach (Fleishman et al., 2004b), which relies on the assumption that conserved amino acid residues preferentially pack at helix-helix interfaces, whereas the positions that face the lipid or the pore lumen are variable (Baldwin et al., 1997). In addition, it is unfavorable for charged residues to face the lipid, except for the terminal helical turns, where charged positions may interact favorably with the polar-headgroup region. Where conservation and hydrophobicity did not suffice to produce an unambiguous conformation, we applied a computational tool for identifying pairs of positions that exhibit correlated evolution, which is often associated with contact formation in the protein's tertiary structure (Fleishman et al., 2004a; Gobel et al., 1994). We thus computed a structure for the entire TM domain of the gap junction hemichannel.

Results

Helix Assignment

Analysis of superimposed cross-sections from the TM density of one connexon (Figure 1) revealed the following helix tilts (Table 1): 9.1° (A), 15.6° (B), 27.5° (C), and 29.2° (D). The contoured sections identified section 29 (second from the top in Figure 1B) as being the last section of helix A that exhibited significant density. Based on the necessity that the aqueous pore be shielded from the membrane lipids, we concluded that section 29 was located close to the cytoplasmic boundary of the membrane. Densities past section 29 for helices B, C, and D likely represented parts of the cytoplasmic domains (N-tail, C-tail, and the M2-M3 cytoplasmic loop).

There are 24 different assignments of the hydrophobic domains M1–M4 to the four helices in the cryo-EM map (Nunn et al., 2001). At the outset, we stress that no single helix assignment can be reconciled with all of the experimental data on connexins (Harris, 2001). This is in part due to the channel's plasticity and the heterogeneity of methods and connexin isoforms on which relevant studies were based. Our approach has therefore been to use primarily the cryo-EM map together with data on hydrophobicity and evolutionary conservation. We relied on other experimental evidence to provide support only in cases where there was substantial agreement between different studies. With Figure 1 as a reference, the following describes clues from different sources that were used to derive an assignment of helices A–D to the TM segments M1–M4 in the connexin sequences.

The substituted-cysteine accessibility method (SCAM) (Karlin and Akabas, 1998) demonstrated that specific residues on M1 as well as in the N-terminal part of E1 are accessible for labeling by water-soluble sulfhydryl reagents (Kronengold et al., 2003; Skerrett et al., 2002; Zhou et al., 1997). A detailed analysis showed that M3 was the major pore-lining helix (Skerrett et al., 2002). Notably, both M1 and M3 contain several evolutionarily

Table 1. Estimated Axes of the TM α Helices

	Tilt and Azimuthal Angles		Positions			
Helix	Θ (°)	φ (°)	a14 (Å)	b14 (Å)	a24 (Å)	b24 (Å)
A (blue)	9.1	0.0	2.4	24.8	2.4	28.8
B (green)	15.6	28.0	15.6	32.4	12.4	34.8
C (yellow)	27.5	90.0	23.2	27.6	11.6	21.6
D (red)	29.2	60.0	10.4	18.0	−0.8	18.0

Colors refer to Figure 1. Positions a14, b14, a24, and b24 were derived from the grid shown in Figure 1 using (0,0) as common origin. With +z pointing towards the observer, sections 14 and 24 are located +28 Å and +48 Å from the center of the extracellular gap. The values for the azimuthal angles (φ) were derived by centering orthogonal x,y-coordinate systems at the a14, b14 positions for each of the helices and measuring the angles between the x-axis, oriented parallel to b, and the projected paths of the helices connecting the points (a14, b14) and (a24, b24). Positive φ angles were measured counterclockwise from x in the direction of y. The values for the tilt angles (Θ) were measured as the angle between the projected path of the helices and the z axis. The estimated axes assume that the α helices are straight.

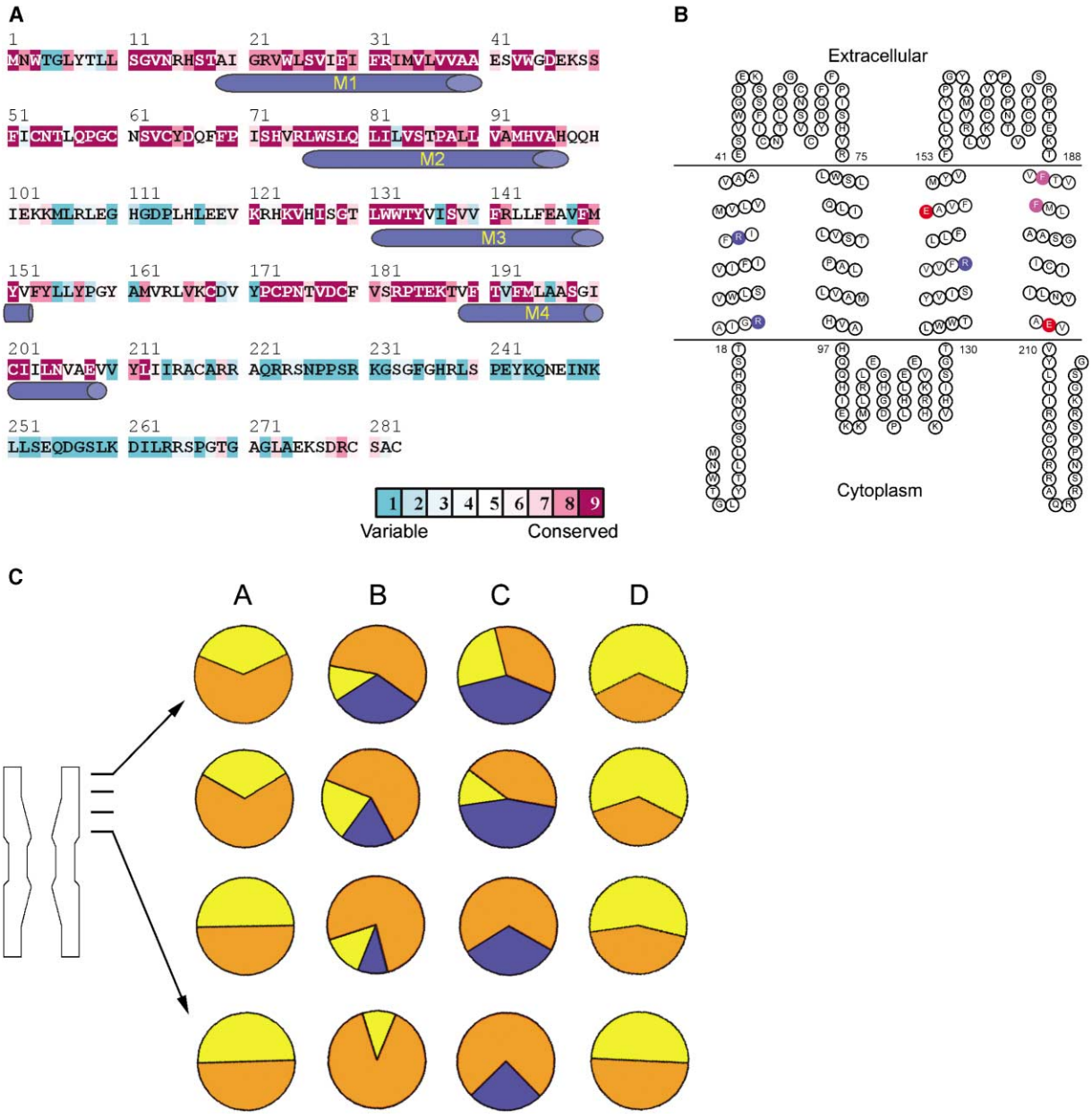


Figure 2. Connexin Architecture and Amino Acid Conservation

(A) The sequence of human Cx32 color-coded according to evolutionary conservation using the *ConSeq* server (Berezin et al., 2004), with turquoise-through-maroon corresponding to variable-through-conserved positions (see color bar). The hydrophobic segments M1–M4 are marked on the sequence.

(B) Membrane topology of Cx32. Acidic and basic amino acids in the TM domain are marked red and blue, respectively. The transmembrane segments M1–M4 and the two extracellular loops E1 and E2 are indicated. Two aromatic residues are colored magenta. Numbers indicate the positions of the extramembrane domain boundaries. Part of the C terminus was truncated.

(C) As indicated in the schematic model (left), four cross-sections evenly distributed within the membrane region of one connexon were evaluated. The approximate total areas facing the aqueous pore (blue), the membrane lipid (yellow), and neighboring α helices (orange) were estimated in each section for each of the four helices A–D. The orientations of the pie charts are arbitrary. As suggested in Figure 1, this representation clearly reveals that only helices B and C have access to the aqueous pore. Furthermore, each of the helices has a characteristic accessibility pattern that was used in combination with the conservation profile of (A) to assign each helix to a specific TM sequence (see text).

conserved charged residues (Figure 2A). The important role of M3 in lining the pore is also suggested by the amphipathic pattern of its conserved polar and charged amino acids (Milks et al., 1988). In contrast, M2 is devoid of any charges, and M4 contains a single Glu residue at position 208 toward the cytoplasmic domain, which is likely to be outside the hydrophobic core of the bilayer (Figure 2B).

Hence, without committing to the specific identities of the pore-lining helices, a generalized assignment for

M1 and M3 could be made by assessing which of the helices in the structure had access to the aqueous pore. From Figure 1, it was clear that only helices B and C lined the aqueous pore, which suggested that these segments corresponded to M1 and M3. However, this first assignment step did not allow us to distinguish between the two possible alternatives. Nevertheless, if M1 and M3 corresponded to helices B and C, then it followed that helices A and D corresponded to TM segments M2/M4 in the connexin sequence.

After division of the four TM segments into two groups (i.e., B/C = M1/M3 and A/D = M2/M4), the number of options for a specific assignment could be limited by a comparison of connexin amino acid sequences using an approach similar to Baldwin's analysis of the G protein coupled receptor family (Baldwin, 1993). Specifically, residues in the lipid-facing positions of TM helices were the least conserved among the receptors. A similar analysis based on 60 connexin sequences (Berezin et al., 2004) showed that the relative conservation of the TM segments was M2 > M4 and M1 > M3 (Figure 2A). We reasoned that evolutionary variability within the TM segments indicated that amino acid residues in these positions were not very important for helix packing and were therefore more likely to face the membrane lipid or the large pore lumen.

A specific helix assignment could then be made by assessing the extent to which the α helices in the structure had access to the lipid and the aqueous pore. Cross-sections similar to those shown in Figure 1 were chosen throughout the membrane-spanning part of one connexon (Figure 2C). In each cross-section, we estimated what part of each of the helices faced the aqueous pore, packed against neighboring helices, or was exposed to the lipid. Helix C was found to be more exposed to the aqueous pore than B. Hence, of the M1/M3 pore-lining pair, the highly conserved M1 most likely corresponded to B, and M3 to the major pore-lining helix C. A similar analysis showed that helix D was more exposed to the lipid environment than was helix A (Figure 2C). Therefore, of the M2/M4 lipid-exposed pair, the conserved M2 most likely corresponded to the more buried helix A, and M4 to the lipid-exposed helix D. Interestingly, the evolutionary conservation of M3 showed a decrease in the central part of the bilayer (Figure 2A), which coincided with an increase in the exposure of helix C to the pore lumen (Figure 2C). Similarly, the conservation of M4 decreased toward the cytoplasmic side, correlating with an increase in its exposure to the membrane.

Helix Orientations

Canonical α helices were constructed based on the parameters defined in Table 1 (Figure 3). A starting C^{α} model for the 24 α helices in the hexameric connexon was built using the assignment M1 = B, M2 = A, M3 = C, and M4 = D. We used an exhaustive search and scoring function to sample the rotational orientation of each of the helices around their principal axes, while maintaining 6-fold symmetry around the channel axis (Fleishman et al., 2004b). This search yielded the optimal conformation shown in Figure 4A. It is evident that helices M3 and M4 show a very clear evolutionary variability versus conservation signal, with the variable residues mapping to one helical face. Indeed, the optimal conformation placed all of the evolutionarily variable positions of M3 and M4 in lumen- or lipid-exposed positions, respectively, whereas conserved faces were packed inside the protein core.

In contrast to M3 and M4, the residues in M1 and M2 are homogeneously conserved (Figure 2A), so that the orientations around their principal axes cannot be determined reliably on the basis of conservation alone. Correlated amino acid evolution has been used previously to identify interresidue contact (e.g., Gobel et al., 1994). The underlying assumption was that pairs of residues that form contact undergo dependent evolution, i.e., a substitution in one position would induce the other to change in order to maintain the protein fold.

To detect correlations, we applied a method that was especially designed for treating intermediate-sized protein families (50–100 sequences) (Fleishman et al., 2004a) such as connexins. We identified five pairs of correlations in the TM and juxtamembrane domains that are connected by solid lines in Figure 4B. Positions in the juxtamembrane domain (3 positions from the end of the hydrophobic stretch at most) were assumed to conform to α-helical ideality. Based on these correlations we manually oriented helices M1 and M2 to obtain a conformation in which each of the two positions of a correlated pair would be in proximity (Figure 4B). The correlations that pertained to helix M3 were in accordance with the helix's orientation around its principal axis as determined above by the evolutionary conservation analysis. Moreover, the five pairs of correlations were accommodated by the model, thus providing additional support for the model at various levels, including the TM-domain boundaries, helix assignment, and the orientations of the helices around their principal axes.

Structural Features

It is difficult to provide a detailed structural interpretation of the model at this resolution since the computed structure does not contain information regarding side chain conformations. Moreover, we estimate that the orientations of the helices around their principal axes may vary by up to 40°. Nevertheless, even at this level of uncertainty, it is possible to provide a rough description of the factors that stabilize the structure.

The lipid-exposed residues of M2 and M4 are mostly hydrophobic. In fact, these helices are devoid of charged amino acids, except for Glu208 on M4 (Figures 2B and 5A). This residue is just two amino acid positions from the C-terminal end of the hydrophobic segment and is located in the protein core, toward the cytoplasmic side of the protein. Hence, it is not exposed to the membrane environment and, due to the tilt of helix M4, might be surrounded by water from the cytoplasm. Position Arg22 on M1 faces the protein core on the cytoplasmic side of the protein (Figure 5B). Likely, this position "snorkels" (von Heijne, 1996) to the cytoplasmic side of the lipid bilayer according to the positive-inside rule (von Heijne, 1989). Another possibility is that Glu208 and Arg22, which are oriented toward one another, form a salt bridge.

Most of the charged residues in M1 and M3 are posi-

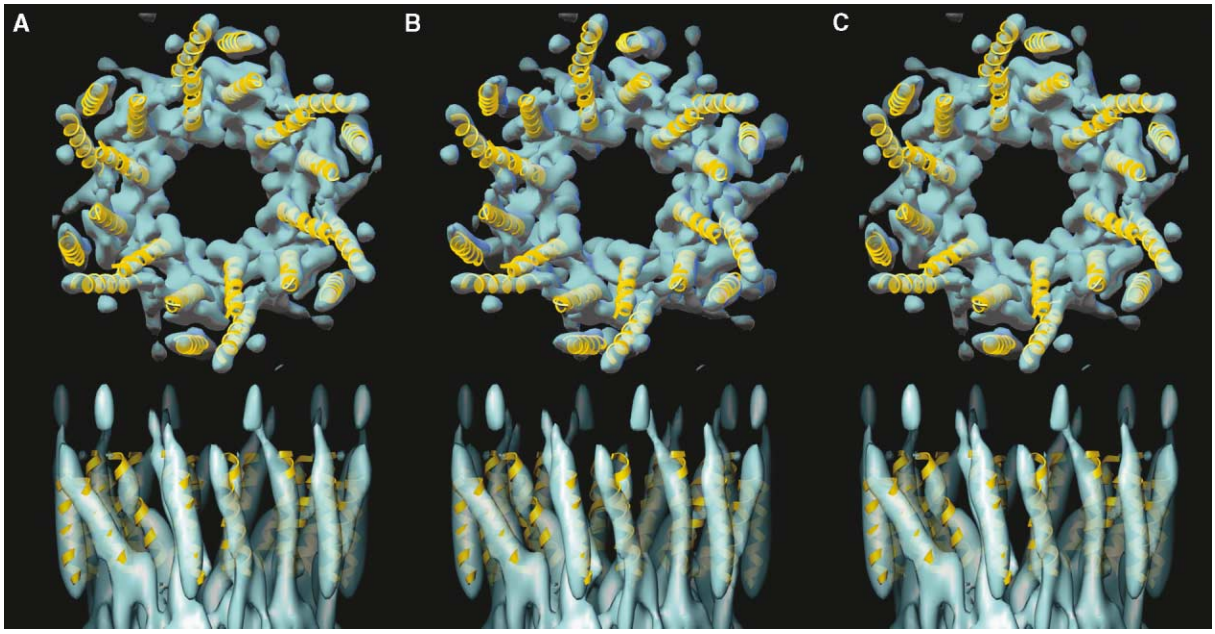


Figure 3. Fit of Canonical α Helices to the Cryo-EM Density Map of the Gap Junction Channel

Top and side views of one connexon showing the fit of canonical α helices (gold) to the cryo-EM density map of Cx43 (blue), according to the helix-axis parameters provided in Table 1. The left and right pairs are wall-eyed and cross-eyed stereo views, respectively.

tioned where they could extend their side chains into the pore lumen (Figure 5B). Arg142 and Glu146 on M3 are only partly pore lining, and interact in part with helix M1, in register with Arg32 of M1. Possibly, the two charged positions of M3, which are one helical turn above each other, form a salt bridge. Being roughly in register with one another, the three charged positions form a thin (4–5 Å) polar belt around the pore lumen about two-thirds of the way from the cytoplasmic to the extracellular ends of the TM domain (Figure 5B). Charged residues in the extracellular loops have been shown to be determinants of charge selectivity in gap junctions (Trexler et al., 2000). It is possible that this polar belt plays a secondary role in charge selectivity.

Roughly in register with one another, a number of conserved polar residues are found throughout the pro-

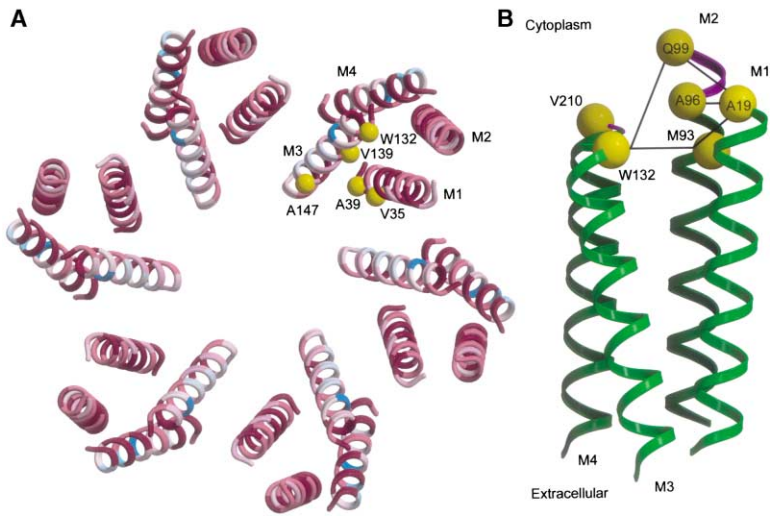


Figure 4. A Model for the Structure of the Gap Junction Connexon

(A) Conservation is color-coded as in Figure 2A. Helices were rotated around their principal axes and evaluated according to a scoring function that (1) favors the burial of conserved and charged amino acids in the protein interior and (2) the exposure of variable positions to the pore lumen or the lipid. Hydrophobic segments M3 and M4 show a clear conservation signal, with a well-defined variable face. Yellow spheres indicate putative specificity determinants, all of which map to pore-lining positions, where they may modulate permeability and conductance. Significantly, specificity determinants span five helical turns on the M3 segment in support of its role as the major pore-lining helix.

(B) M1 and M2 are almost homogeneously conserved (Figures 2A and 4A) and were oriented using a method for the detection of correlated positions (Fleishman et al., 2004a). Positions in the juxtamembrane domain (three positions from the end of the hydrophobic stretch at most) were assumed to extend the α helix (colored magenta). Correlated positions are connected by solid lines. The three correlated pairs of positions on M1 and M2 were assumed to interact, so the helices were rotated manually for these positions to be roughly in proximity. The orientation of M3 around its principal axis was determined solely on the basis of evolutionary conservation (Figure 4A), but the two pairs of correlations between positions on M3 and M2 are congruent with the orientation of M3, serving as partial verification of this helix's orientation around its principal axis. A sixth correlation between Gln99 (M2) and Val210 (M4) could not be reconciled with the model.

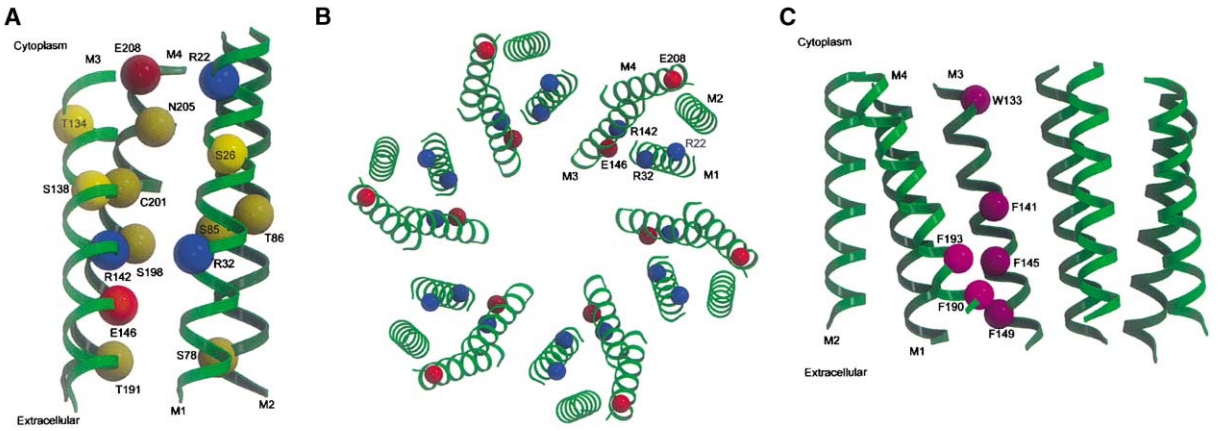


Figure 5. Structural Features of the TM Domain of the Gap Junction Connexon

(A) Polar and charged amino acid residues in the protein interior. The polar residues (yellow spheres) are roughly in register and could be involved in the formation of a network of hydrogen bonds that would stabilize interhelical contacts.

(B) Acidic and basic residues in the protein interior and facing the pore lumen are indicated by red and blue spheres, respectively. Arg22 is near the boundary of the hydrophobic domain and could be accessible to the cytoplasmic side of the membrane (von Heijne, 1989). Glu208 also resides at this boundary and is likely to be exposed to the cytoplasm. The pore-lining charged residues form a slender (4–5 Å) belt of charge around the pore lumen. None of the charged residues is exposed to the membrane.

(C) Aromatic residues on M3 and M4 are shown as purple spheres. The two Phe positions on M4 coincide with the position of a protrusion of density on helix D in the cryo-EM map (Unger et al., 1999). Stacked aromatic residues have been shown to generate such protrusions of density (Henderson et al., 1990). The clustering of aromatic residues from M3 and M4 could stabilize interhelical contacts. Furthermore, the ridge of aromatic residues on M3 could serve to shield the water-filled pore from the lipids in this region of the protein structure, in which helices are not tightly packed.

tein core (yellow spheres in Figure 5A). An attractive hypothesis is that these residues form a hydrogen bonding network to stabilize interhelical contacts. This could explain why many of these positions are intolerant to substitution; even fairly conservative mutations at these positions have been implicated in disease. We note, however, that no terms in the scoring function used to orient the helices around their principal axes favored a particular hydrogen bonding pattern among amino acid residues (Fleishman et al., 2004b).

Significantly, the criteria used for orienting M3 and M4, i.e., evolutionary conservation and hydrophobicity (Fleishman et al., 2004b), did not take into account interactions among aromatic residues. Nevertheless, a prominent structural feature of the model is the clustering of five conserved Phe residues near the extracellular side of the bilayer between helices M3 and M4 (Figure 5C), which may stabilize interhelical contacts. There is also a ridge of aromatic residues on M3 that extends almost without interruption between the extracellular and the intracellular ends of the channel, from Trp133 on the cytoplasmic side to Phe149 on the extracellular side of the bilayer (Figure 5C). Notably, the density map shows that helices C (M3) and B′ (M1) are separated by a relatively large distance (Figure 1). This ridge of aromatic residues could shield the water-filled pore from the lipid.

It is also notable that the previous (Unger et al., 1999) and current cryo-EM maps show a relatively large "shoulder" of density on helix D toward the extracellular side of the gap junction channel. Such protrusions of density can arise from stacked aromatic residues in intermediate-resolution maps (Henderson et al., 1990). A map of bacteriorhodopsin that we computed at the resolution of the gap junction map (5.7 Å in-plane and 19.8 Å vertical) showed a thickening of density corresponding to Phe153 and Phe156 in helix E (data not shown). Although aromatic residues are present in all four TM segments of connexins, only M4 contains two conserved Phe residues near the extracellular side of the bilayer (positions 190 and 193) that occupy the same helical face (magenta circles in Figure 2B). In contrast, helix M2 contains only one aromatic residue (Trp77) in its extracellular part. While it is not an ultimate proof, the interpretation of the shoulder of density on helix D provides independent support for the assignment of helix D to M4 and the orientation of this helix around its principal axis.

Specificity Determinants

Gap junction channels manifest very little ionic selectivity and yet do show differences in ionic preferences between different connexin isoforms. Based on this behavior, one would expect that pore-lining residues would vary among different types of connexins (paralogs) but be conserved for identical connexins in different species (orthologs) (Harris, 2001). Such positions are termed specificity determinants, as their identities determine the specific functional behavior of the given channel.

We analyzed the connexin sequences to identify putative specificity determinants. Connexins of similar functions in different species (orthologs) are the products of speciation events, whereas those with different functions (paralogs) arise from gene duplication (Graur and Li, 1999). It is therefore expected that orthologous sequences would cluster in the termini of the phylogenetic tree, whereas the events leading to paralogy would be reflected in deeper nodes. Thus, using a phylogenetic tree (Yang, 1997) and reconstructed ancestral sequences (Schmidt et al., 2002), we automatically traced the evolutionary history of each amino acid position in

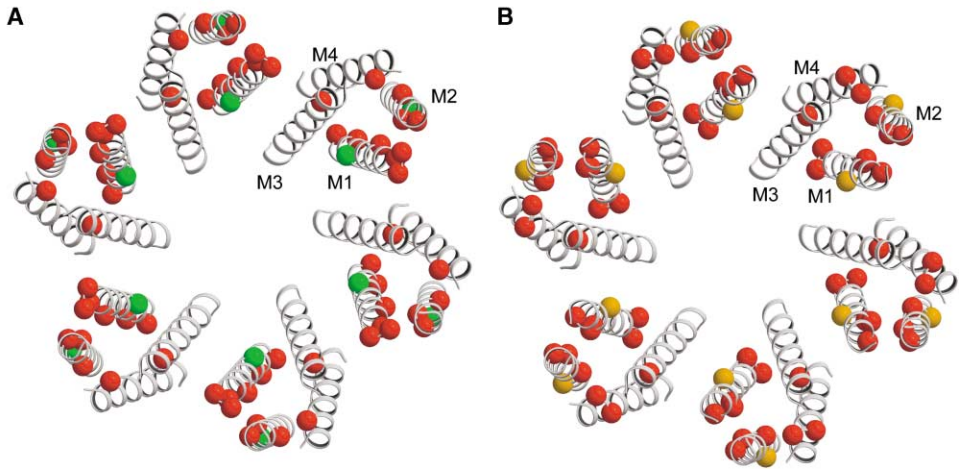


Figure 6. The Distribution of Disease-Causing and Benign Polymorphisms in the Gap Junction Model

(A) The model provides an explanation for the differential effects of mutations that cause nonsyndromic hereditary deafness, erythrokeratodermia variabilis (EKV), and polymorphisms in the TM domain. Physicochemically conservative disease-causing mutations (e.g., Val for Ile) were colored red, and radical polymorphisms (e.g., Ser for Tyr) were colored green. As expected, conservative disease-causing mutations all map to structurally dense regions of the protein, whereas the radical polymorphisms map to more spacious regions.

(B) Similarly, 11 of 13 conservative Charcot-Marie-Tooth (Fischbeck et al., 1999) causing mutations (red spheres) map to structurally packed regions, whereas only two such mutations (orange spheres) map to pore-lining or lipid-exposed helix faces.

search of those that exhibited relatively minor evolutionary differences in the branches separating terminal nodes and large differences in the inner branches (see Experimental Procedures).

We identified five putative specificity determinants on M1 and M3, all of which are pore lining as expected (yellow spheres in Figure 4A). Notably, the putative specificity determinants on M3, the major pore-lining helix, span five helical turns from the cytoplasmic end of the channel, up to roughly two-thirds of the way toward the extracellular side of the bilayer. Since pore-lining positions are expected to specify the different conductance and permeability traits of connexins (Harris, 2001), these results serve as independent verification of our model and predict which residues have important effects on channel properties.

The Locations of Mutations and Polymorphisms

To see whether the model can provide insight on the molecular basis for the effects of mutations that have been reported clinically, we analyzed mutations related to skin, deafness, and developmental diseases that are documented in the Connexin-Deafness Homepage (http://www.crg.es/deafness). The logic underlying our analysis is that mild substitutions such as Val for Ile will cause disease only if they occur in regions of the protein that are structurally well packed. Similarly, radical substitutions such as Ser for Tyr will only be tolerated if they occur in structurally spacious regions.

Figure 6A displays the structure of the gap junction hemichannel with all 11 physicochemically conservative substitutions of Cx26 causing nonsyndromic deafness and erythrokeratodermia variabilis (EKV) in red, and the two radical but benign substitutions (polymorphisms) in green. Strikingly, all mutations indeed map to structurally packed regions, whereas both polymorphisms map to either the pore region or the lipid-exposed face. Analysis of all 13 mild substitutions of Cx32 causing Charcot-Marie-Tooth (CMT) neuropathy (Fischbeck et al., 1999) (Figure 6B) reveals a similar pattern, with only two disease-causing mutations (orange) mapping to spacious regions of the protein structure. A list of all of the mutations shown in Figure 6 is available upon request and at http://ashtoret.tau.ac.il/~sarel/GJ.html.

Discussion

Determining the positions of amino acid residues in the gap junction channel has defied experimental approaches for many years. In part, this is due to the complicated organization of gap junctions when compared to other membrane channels whose structures have been solved. That is, gap junction channels are composed of two connexons in separate membranes and can form different channel varieties depending on the types of connexins that are associated within a connexon.

We used a cryo-EM map of the gap junction channel (Figure 1) to guide the positioning of model α helices in the membrane (Figure 3). The four hydrophobic segments M1–M4 in the connexin sequence were assigned to the helices according to biochemical and evolutionary evidence. The orientation of each of the helices around its principal axis was then computed by analyses of evolutionarily conserved (Figure 4A) and correlated amino acid substitutions (Figure 4B). The resultant conformation placed positions that we identified as specificity determinants in pore-lining locations, as expected (Figure 4A).

We note that the validity of the model is entirely contingent on the assignment of the hydrophobic segments, M1–M4, to the helices A–D in the cryo-EM map (that is, A = M2, B = M1, C = M3, and D = M4). While no single assignment is completely in harmony with all available

biochemical evidence (Harris, 2001), the assignment we have used is compatible with a large body of data.

Several different lines of evidence have converged in the computation and verification of the model. The agreement between these methods is encouraging, but the model should be treated only as an approximation. In fact, there are some inherent inaccuracies in the modeling. For instance, the effective resolution of the cryo-EM map perpendicular to the membrane plane is only 19.8 Å, thereby precluding accurate vertical positioning of the helices. However, the helices are all relatively short, and their tilt angles are fairly small (Table 1). Hence, it is reasonable to position the geometric centers of the helices in the middle of the membrane-spanning part of the cryo-EM map. We note that the correlated pairs of positions are roughly in register (Figure 4B), as are the polar amino acids in the protein core (Figure 5A), serving as support for the positions of the helices' geometric centers.

Another complication is that the limited resolution of the cryo-EM map does not allow us to detect deviations from α-helicity. Nevertheless, the fit of canonical α helices to the cryo-EM map is energetically reasonable (Nunn et al., 2001), and the map does not show any kinks in the TM domain. For comparison, large kinks have been observed in the cryo-EM map of vertebrate rhodopsin at 9 Å in-plane resolution, which still yielded a correct assignment for the positions and orientations of the helices (Baldwin et al., 1997). We cannot rule out the existence of small kinks and bulges at this resolution (Ri et al., 1999), but these would likely have only a local effect on the resultant model (Fleishman et al., 2004b).

The limited vertical resolution of the cryo-EM map also does not reveal the connecting loops between the TM helices, thus precluding the unambiguous assignment of the molecular boundary of each connexin subunit. There are two reasonable subunit boundaries, encompassing either the helices marked as ABCD or A′B′CD in Figure 1. Certainly, more experiments are needed to distinguish these alternatives, and the model provides a detailed structural template for testing these possibilities biochemically. Nevertheless, it is important to note that this ambiguity regarding the connexin subunit boundary is independent of and does not adversely affect the assignment of TM sequences to the helices in the cryo-EM map (i.e., A = M2, B = M1, C = M3, and D = M4).

We are encouraged that the model provides an explanation why substitutions at certain positions can lead to disease (mutations), whereas in other positions, substitutions result in no apparent phenotype (polymorphism). Helices M1 and M2 are considerably more sensitive to mutations than M3 and M4, consistent with the tighter packing of M1 and M2 according to the model (Figure 6). The somewhat higher incidence of mild disease-causing mutations toward the cytoplasmic ends of M1 and M2 coincides with a closer approach of these two helices in this region. We note that sequence conservation alone is not as informative as the model in identifying the portions of the sequences in which substitutions would have deleterious consequences (Figure 6). That is, residues on M1 and M2 are all highly conserved (Figure 2A), but only substitutions in relatively narrow segments on these helices, which are packed in the protein interior, result in disease. Given the striking compatibility of data on mutations and polymorphisms with the model, it appears that the effects of a significant fraction of disease-causing mutations in the TM domain may be explained quite simply in terms of deformations of local structure at the interfaces between helices.

Without a model that explicitly defined amino acid positions, it has been difficult previously to plan rational biochemical experiments. Many studies tested connexin chimeras by swapping large segments from various isoforms (e.g., Hu and Dahl, 1999; Oh et al., 2000; Trexler et al., 2000). While such approaches have provided important insight into broad characteristics, such as charge selectivity and channel permeability, they do not provide an understanding of fine structural and functional details. In recent years, scanning mutagenesis and SCAM provided more detailed information (e.g., Kronengold et al., 2003; Skerrett et al., 2002; Zhou et al., 1997). Nevertheless, without a detailed model, it has not been possible to assess their reliability within one consistent structural framework. Another difficulty in interpreting results from SCAM analyses is that negative results at particular positions (i.e., no labeling) cannot be reliably associated with inaccessibility of these residues. As the labeling reaction depends very strongly on the local environment, neighboring side chains might obstruct accessibility to an otherwise pore-lining position.

The model we describe provides the first integration of a large body of biochemical, mutational, structural, and computational data on the structure of gap junction channels. The model should prove valuable for deriving testable hypotheses related to structure and function. For instance, the model provides certain clues regarding the factors that stabilize interhelical contacts and the determinants of connexin oligomerization. Studies on the roles of the pore-lining positions in affecting channel permeability and selectivity may also be focused with the help of the model, in particular to the residues that we identified as putative specificity determinants (Figure 4A). Moreover, the model can guide studies on the folding of individual connexins and their association to form connexons. A fascinating prediction of the model is that the phenotypic effects of a disease-causing mutation on one helix can be rescued by a substitution on a neighboring helix.

Experimental Procedures

Electron Cryomicroscopy and Image Analysis

Preparation of two-dimensional crystals, cryo-EM, and lattice straightening were performed as described before (Unger et al., 1999). A list that contained the data from 69 crystalline areas was edited to exclude measurements where the sampling of reciprocal space was too sparse to allow a meaningful fit of lattice lines. The final fit was limited to a maximum z* value of 0.065 $Å^{-1}$ generating 1734 unique structure factors compared to 1022 that were included in the previous reconstruction (Unger et al., 1999). Using image data with signal-to-noise ratios $\geq$1.8, the overall merging phase residual for each crystal was $<25^{\circ}$ compared with the entire data set. The 3D map was computed using an inverse B factor of -350. Analysis of the point-spread function indicated a maximum in-plane resolution of 5.7 Å and a vertical resolution of 19.8 Å.

Sequence Data

60 connexin sequences were obtained from SWISS-PROT (Bairoch and Apweiler, 2000) and aligned using CLUSTAL W (Thompson et

al., 1994) with default parameters. For each position in the alignment, evolutionary conservation was computed using the *ConSeq* server (Figure 2A) (Berezin et al., 2004), and hydrophobicity using the Kessel and Ben-Tal scale (Kessel and Ben-Tal, 2002).

The topology of Cx32 was determined experimentally (Milks et al., 1988). Definition of the N- and C termini of the four TM segments (Bennett et al., 1994) was adjusted slightly to include hydrophobic stretches that were as long as possible. That is, we eliminated positions from the hydrophobic segments' termini that were occupied by polar or charged amino acids in any of the sequences in the multiple-sequence alignment of 60 homologs. The resulting topology and boundaries of the hydrophobic stretches are shown in Figures 2A and 2B.

Scoring Function

The conformational search was performed using the scoring function described by Fleishman et al. (2004b). In brief, this scoring function favors the burial of evolutionarily conserved amino acid positions in the protein core and the exposure of variable positions to the lipid or the pore. Conformations that expose charged amino acids to the lipid milieu are penalized. Since the gap junction pore is relatively large, pore-lining and lipid-exposed residues were treated equally as unburied positions, with no need for introducing modifications to the functions. However, since charged residues can be exposed to the lumen of the pore with no consequence on desolvation energy, we abolished the penalty for exposure of charged positions on the pore-lining helices M1 and M3 (Figure 5B). Each conformation was scored according to the following equation:

$$Score = \sum_i (2(B^i - \tfrac{1}{2})(H^i - C^i)), \quad (1)$$

where B^i quantifies the extent of burial of amino acid i in the protein core (Fleishman and Ben-Tal, 2002). It assumes values of 0 to 1; 1 signifying complete burial against another helix, and 0 complete exposure to the lipid or the pore lumen. The function is computed by iterating over all of the helices in the structure other than the one on which i is located, and taking into account i's distance from, and orientation with respect to, each of these helices. B^i is then taken as the maximum of the values calculated for each of the helices (Fleishman and Ben-Tal, 2002; Fleishman et al., 2004b). Thus, high values of B^i imply that i is in close contact with another helix, whereas low values indicate that it is not interacting with any of the helices.

The C^i values are the normalized evolutionary-rate scores assigned by *Rate4Site* (Figure 2A) (Berezin et al., 2004; Pupko et al., 2002). High-through-low values of C^i are assigned to variable-through-conserved positions, respectively. Proline residues are ignored in calculating the conservation scores, as they are often conserved due to kinks they induce in the helix secondary structure rather than due to the formation of interhelical contacts (Baldwin et al., 1997).

H^i is the free energy of transfer from water to lipid of amino acid i according to the Kessel and Ben-Tal scale (Kessel and Ben-Tal, 2002). H^i values are taken into account only if they are greater than 7 kcal/mole, and only for residues i that are exposed to the membrane, i.e., for which the burial scores B^i are less than 0.5. Thus, the hydrophobicity scale serves as a significant penalty on the exposure of the most polar residues to the membrane environment. The terminal turns (4 amino acid residues) from each side of the TM segments were ignored in computing this penalty, since the polar environment at the lipid-water interface could accommodate these residues (von Heijne, 1989).

Conformational Search

Canonical C^{α}-trace models of four α helices were constructed according to the helix-axes parameters derived from the cryo-EM map (Table 1), and their geometric centers were placed at the hypothetical membrane midplane. The amino acid identities of positions in the hydrophobic segments M1–M4 were assigned to the relevant positions on these helices. The channel's axis of symmetry was inferred from the map (Figure 1), and 6-fold symmetry around this axis was strictly maintained throughout all conformational searches. Hence, only the rotation angles around the principal axes of each of four helices comprising a single connexin were explored, and applied to all 24 helices.

Each helix was rotated around its principal axis independently, in 5° steps, and its optimal orientation was derived. Then, the optimal orientations of all helices were superimposed to yield the optimal conformation of the entire complex.

Correlated Mutations

The multiple-sequence alignment of 60 connexin homologs was used to compute the phylogenetic tree of maximum-likelihood (Schmidt et al., 2002). Subsequently, the most likely ancestral (now-extinct) sequences were inferred (Yang, 1997). We then identified correlated positions in the TM domain of connexins (Fleishman et al., 2004a) (Figure 4B). The informational-entropy threshold (Shannon, 1948), which is a measure of the heterogeneity of amino acid identities in a particular position in the alignment, was set to 1.1 in order to remove highly conserved positions. To obtain confidence intervals for each of the computed correlations, 400 bootstrap iterations (Bradley and Tibshirani, 1993) with replacement were conducted. The lower (r_{low}) and upper (r_{high}) boundaries of the 95% confidence interval were determined as the correlation coefficient at the 2.5 and the 97.5 percentiles, respectively, and the trimmed mean (r) of correlation coefficients was calculated. Pairs of positions showing lower confidence boundaries of $r_{low} < 0.1$ were eliminated as were pairs with trimmed means of $r < 0.5$.

Specificity Determinants

The phylogenetic tree and ancestral-sequence reconstruction (see Correlated Mutations, above) were used to detect putative specificity determinants in the connexin family. Conserved positions in the sequence alignment exhibiting information entropy (Shannon, 1948) of less than 1.1 were eliminated. For each position in the alignment, and in each phylogenetic branch, we measured the physicochemical distance between the amino acid identities occupying those positions using the Miyata substitution matrix (Miyata et al., 1979). Multiple and back substitutions in a single branch were not considered. Each node in the phylogenetic tree was assigned a "depth" value, which was an integer calculated as the minimal distance between that node and any terminus, counting intervening nodes. Thus, the termini were assigned depth values of 0, neighboring nodes values of 1, etc.

For each amino acid position, we then computed the Pearson correlation coefficient between physicochemical distances traversed in each phylogenetic branch and the average depths of each of the nodes that were connected by that particular branch. Hence, high correlation coefficients were associated with positions that exhibited relatively low variability among terminal nodes (orthologs) and relatively high variability in deeper nodes (separating paralogs). We conducted 400 bootstrap iterations with replacement (Bradley and Tibshirani, 1993), and calculated the trimmed mean of the 95% confidence interval of these correlation values (r). The lower (r_{low}) and upper (r_{high}) bounds of the 95% confidence interval were determined as the correlation coefficient at the 2.5 and the 97.5 percentiles, respectively. Positions showing lower confidence bounds of $r_{low} < 0$ were eliminated as were positions with trimmed means $r < 0.1$.

Acknowledgments

The authors thank Michael Bennett, Ted Bargiello, and Vyto Verselis for helpful discussions. This study was supported by a Research Career Development Award from the Israel Cancer Research Fund to N.B.-T. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation. During part of this work, V.M.U. was supported by a postdoctoral fellowship from the American Heart Association. M.Y. was supported by NIH grant RO1HL48908 and a Clinical Scientist Award in Translational Research from the Burroughs-Wellcome Fund. Additional information is available on request and at http://ashtoret.tau.ac.il/~sarel/GJ.html.

Received: March 26, 2004
Revised: July 6, 2004
Accepted: July 22, 2004
Published: September 23, 2004

References

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48.

Baldwin, J.M. (1993). The probable arrangement of the helices in G protein-coupled receptors. EMBO J. *12*, 1693–1703.

Baldwin, J.M., Schertler, G.F., and Unger, V.M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J. Mol. Biol. *272*, 144–164.

Bennett, M.V., Zheng, X., and Sogin, M.L. (1994). The connexins and their family tree. Soc. Gen. Physiol. Ser. *49*, 223–233.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, R., Cassadio, R., and Ben-Tal, N. (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics, in press.

Bourne, H.R., and Meng, E.C. (2000). Structure. Rhodopsin sees the light. Science *289*, 733–734.

Bradley, E., and Tibshirani, R. (1993). An Introduction to the Bootstrap (New York: Chapman and Hall).

Cascio, M., Kumar, N.M., Safarik, R., and Gilula, N.B. (1995). Physical characterization of gap junction membrane connexons (hemi-channels) isolated from rat liver. J. Biol. Chem. *270*, 18643–18648.

Fischbeck, K.H., Abel, A., Lin, G.S., and Scherer, S.S. (1999). X-linked Charcot-Marie-Tooth disease and connexin32. Ann. N Y Acad. Sci. *883*, 36–41.

Fleishman, S.J., and Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. J. Mol. Biol. *321*, 363–378.

Fleishman, S.J., Yifrach, O., and Ben-Tal, N. (2004a). An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J. Mol. Biol. *340*, 307–318.

Fleishman, S.J., Harrington, S., Friesner, R.A., Honig, B., and Ben-Tal, N. (2004b). An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data. Biophys. J., in press.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Graur, D., and Li, W.H. (1999). Fundamentals of Molecular Evolution, Second Edition (Sunderland, MA: Sinauer Associates).

Harris, A.L. (2001). Emerging issues of connexin channels: biophysics fills the gap. Q. Rev. Biophys. *34*, 325–472.

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., and Downing, K.H. (1990). Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. J. Mol. Biol. *213*, 899–929.

Hu, X., and Dahl, G. (1999). Exchange of conductance and gating properties between gap junction hemichannels. FEBS Lett. *451*, 113–117.

Karlin, A., and Akabas, M.H. (1998). Substituted-cysteine accessibility method. Methods Enzymol. *293*, 123–145.

Kelsell, D.P., Dunlop, J., and Hodgins, M.B. (2001). Human diseases: clues to cracking the connexin code? Trends Cell Biol. *11*, 2–6.

Kessel, A., and Ben-Tal, N. (2002). Free energy determinants of peptide association with lipid bilayers. In Current Topics in Membranes, S. Simon and T. McIntosh, eds. (San Diego: Academic Press), pp. 205–253.

Kronengold, J., Trexler, E.B., Bukauskas, F.F., Bargiello, T.A., and Verselis, V.K. (2003). Single-channel SCAM identifies pore-lining residues in the first extracellular loop and first transmembrane domains of Cx46 hemichannels. J. Gen. Physiol. *15*, 389–405.

Kumar, N.M., and Gilula, N.B. (1996). The gap junction communication channel. Cell *84*, 381–388.

Milks, L.C., Kumar, N.M., Houghten, R., Unwin, N., and Gilula, N.B. (1988). Topology of the 32-kd liver gap junction protein determined by site-directed antibody localizations. EMBO J. *7*, 2967–2975.

Miyata, T., Miyazawa, S., and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. J. Mol. Evol. *12*, 219–236.

Nunn, R.S., Macke, T.J., Olson, A.J., and Yeager, M. (2001). Transmembrane α-helices in the gap junction membrane channel: systematic search of packing models based on the pair potential function. Microsc. Res. Tech. *52*, 344–351.

Oh, S., Abrams, C.K., Verselis, V.K., and Bargiello, T.A. (2000). Stoichiometry of transjunctional voltage-gating polarity reversal by a negative charge substitution in the amino terminus of a connexin32 chimera. J. Gen. Physiol. *116*, 13–31.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., et al. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. Science *289*, 739–745.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18*, S71–S77.

Ri, Y., Ballesteros, J.A., Abrams, C.K., Oh, S., Verselis, V.K., Weinstein, H., and Bargiello, T.A. (1999). The role of a conserved proline residue in mediating conformational changes associated with voltage gating of Cx32 gap junctions. Biophys. J. *76*, 2887–2898.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Sys. Tech. J. *27*, 379–423, 623–656.

Skerrett, I.M., Aronowitz, J., Shin, J.H., Cymes, G., Kasperek, E., Cao, F.L., and Nicholson, B.J. (2002). Identification of amino acid residues lining the pore of a gap junction channel. J. Cell Biol. *159*, 349–360.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Trexler, E.B., Bukauskas, F.F., Kronengold, J., Bargiello, T.A., and Verselis, V.K. (2000). The first extracellular loop domain is a major determinant of charge selectivity in connexin46 channels. Biophys. J. *79*, 3036–3051.

Unger, V.M., Hargrave, P.A., Baldwin, J.M., and Schertler, G.F. (1997). Arrangement of rhodopsin transmembrane α-helices. Nature *389*, 203–206.

Unger, V.M., Kumar, N.M., Gilula, N.B., and Yeager, M. (1999). Three-dimensional structure of a recombinant gap junction membrane channel. Science *283*, 1176–1180.

von Heijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. Nature *341*, 456–458.

von Heijne, G. (1996). Principles of membrane protein assembly and structure. Prog. Biophys. Mol. Biol. *66*, 113–139.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.

Yeager, M., and Gilula, N.B. (1992). Membrane topology and quaternary structure of cardiac gap junction ion channels. J. Mol. Biol. *223*, 929–948.

Zhou, X.W., Pfahnl, A., Werner, R., Hudder, A., Llanes, A., Luebke, A., and Dahl, G. (1997). Identification of a pore lining segment in gap junction hemichannels. Biophys. J. *72*, 1946–1953.

Accession Numbers

The coordinates of the C^{α} model have been deposited in the Protein Data Bank (accession code 1TXH).

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 281, NO. 39, pp. 28958–28963, September 29, 2006

The Structural Context of Disease-causing Mutations in Gap Junctions*[S]

Received for publication, June 15, 2006 Published, JBC Papers in Press, July 24, 2006, DOI 10.1074/jbc.M605764200

Sarel J. Fleishman[‡1,2], Adi D. Sabag[§1], Eran Ophir[§], Karen B. Avraham[§], and Nir Ben-Tal[‡3]

From the [‡]Department of Biochemistry, George S. Wise Faculty of Life Sciences, and [§]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel-Aviv University, 69978 Ramat Aviv, Israel

Gap junctions form intercellular channels that mediate metabolic and electrical signaling between neighboring cells in a tissue. Lack of an atomic resolution structure of the gap junction has made it difficult to identify interactions that stabilize its transmembrane domain. Using a recently computed model of this domain, which specifies the locations of each amino acid, we postulated the existence of several interactions and tested them experimentally. We introduced mutations within the transmembrane domain of the gap junction-forming protein connexin that were previously implicated in genetic diseases and that apparently destabilized the gap junction, as evidenced here by the absence of the protein from the sites of cell-cell apposition. The model structure helped identify positions on adjacent helices where second-site mutations restored membrane localization, revealing possible interactions between residue pairs. We thus identified two putative salt bridges and one pair involved in packing interactions in which one disease-causing mutation suppressed the effects of another. These results seem to reveal some of the physical forces that underlie the structural stability of the gap junction transmembrane domain and suggest that abrogation of such interactions bring about some of the effects of disease-causing mutations.

Gap junction channels are formed by the docking of two hemichannels or connexons from adjacent membranes (1). Each connexon comprises six connexin subunits (2), proteins which are encoded by ~20 isoforms in the human genome (3). All connexins contain four transmembrane (TM)[4] segments (M1–M4), whose N and C termini are located in the cytoplasm (4). The channels are ~15 Å in diameter at their narrowest point (5), allowing the transport of ions and secondary messengers. They are expressed in nearly all vertebrate tissues and perform critical functions in mediating cell-to-cell signaling and metabolic coupling between apposed cells (6). Connexins have been implicated in several diseases. For example, mutations in the gene encoding connexin 32 (Cx32; gene symbol *GJB1*) cause X-linked Charcot-Marie-Tooth disease, a common form of inherited motor and sensory neuropathy (7), and mutations in the gene encoding connexin 26 (Cx26; gene symbol *GJB2*) are responsible for a large proportion of cases of severe to profound non-syndromic hearing loss (8).

The structure of the gap junction has been solved only at intermediate resolution, revealing the approximate locations of each of the α-helices comprising the TM domain (5) but not the locations of its constituent amino acids. As with many other human membrane proteins that lack bacterial homologs, structural analyses of connexins have been impeded by the absence of an atomic resolution structure (9). In particular, the effects of disease-causing mutations on gap junction structural stability have not been probed experimentally, and it has been difficult to design and interpret biochemical experiments on the structural aspects of connexins relating to individual amino acid residues. Instead, studies have focused on connexin domains (10–13), and normally, pairwise relationships among residues have not been detected other than by serendipity (14). However, based on the intermediate resolution structure (5) and computational inference methods (15, 16), a model of canonical α-helices corresponding to the M1–M4 segments, which specifies the approximate positions of α-carbons in the TM domain, was recently proposed (see Fig. 1) (17). Because the TM domain of connexins has been well conserved through evolution (17, 18), this model structure may serve as a template for all connexin isoforms, although the various isoforms are likely to exhibit slightly different helix-packing interactions. Hence, by specifying which residues are located in proximity to one another, the model can serve as a basis on which to formulate explicit hypotheses on interactions between amino acid positions.

We have studied several hypotheses of interactions between residue pairs by probing the localization of mutated human Cx26 and Cx32 that were C-terminally fused to green fluorescent protein (GFP) and expressed in HeLa cells that do not express endogenous connexins (19). Connexin trafficking to and insertion into the plasma membrane is dependent on several factors and processes, among them is the proper folding and oligomerization of the protein (20); substantial disruption of the protein stability could therefore result in mislocalized protein. We introduced single and double mutations into the connexin TM domain. A destabilizing mutation would show aberrant localization outside the plasma membrane. However, a carefully chosen second-site mutation could stabilize the mutated protein and retrieve the wild-type localization at the

* This work was supported by the European Commission FP6 Integrated Project EUROHEAR, LSHG-CT-20054-512063, Israel Science Foundation Grant 222/04, and National Institutes of Health Grant R01 DC005641. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental data.

[1] These authors contributed equally to this work.

[2] Supported by a doctoral fellowship from the Clore Israel Foundation.

[3] To whom correspondence should be addressed. Tel.: 972-3-640-6624; Fax: 972-3-640-6834; E-mail: nirb@tauex.tau.ac.il.

[4] The abbreviations used are: TM, transmembrane; Cx, connexin.

sites of cell-cell apposition. This procedure is similar in spirit to the double-mutant cycle (21) and to second-site suppression assays (22). In all of these analyses, if the phenotypic effects of one mutation are found to depend on whether or not the other is mutated as well, this will indicate that the two amino acid sites interact (23). It should be noted that these experimental assays cannot, on their own, distinguish between direct physical interactions of a pair of positions and indirect interactions mediated via other residues (24). However, the model structure, although approximate, helped us to constrain the possible explanations. Based on the results of the mutation analyses, we identified the most likely justification for the observed phenotypes by computing approximate models of the side chains of interacting residues. We thus detected, for the first time, interactions that apparently stabilize contacts between TM helices in connexins; abrogation of these interactions leads to aberrant phenotype and disease.

EXPERIMENTAL PROCEDURES

In general, we followed the experimental procedures presented in Ref. 37.

Cloning—Genomic DNA of each of the genes, *GJB2* (Cx26) and *GJB1* (Cx32), was double-digested with HindIII and KpnI and cloned into a pEGFP-N1 expression vector (Clontech, Palo Alto, CA).

Mutant Connexin Expression Constructs—Mutations were introduced into the open reading frame of human *GJB1* and *GJB2* genomic DNA (subcloned into a pEGFP plasmid) by PCR site-directed mutagenesis using the QuikChange kit (Stratagene, La Jolla, CA). DNA extracted from single colonies was sequenced at the Tel-Aviv University Sequencing Unit (Faculty of Life Sciences) using the ABI 377 DNA sequencer (PE Biosystems, Foster City, CA). A DNA template of one single mutation of each set of mutations was used with the mutagenic primers of the second mutation in the same set to generate the double mutants.

Cell Cultures and Transfections—Communication-deficient HeLa cells, which do not express endogenous connexins, were kindly provided by Prof. David Kelsell (University of London). The cells were grown in low-glucose Dulbecco's modified Eagle's medium supplemented with 10% fetal-calf serum, antibiotics (100 μl/ml penicillin/streptomycin), and glutamine (290 μl/ml) in a humidified atmosphere containing 5% CO_2 at 37 °C. The cells were plated onto six-well plates on coverslips and incubated for 24 h to 60–70% confluence.

HeLa cells were transiently transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions with modifications. The amount of reagent was reduced by half and incubated mixed with an equal volume of Neowater (Do-Coop Technologies, Or Yehuda, Israel) for 5 min at room temperature. This mixture and plasmid DNA were incubated separately in OptiMEM for 5 min and combined for another 20 min at room temperature. HeLa cells (60–70% confluence) were washed with OptiMEM and incubated with the combined Lipofectamine/plasmid DNA solution at 37 °C. After five hours, the transfection medium was removed from the cells to prevent toxicity, and cells were incubated in medium without antibiotics overnight.

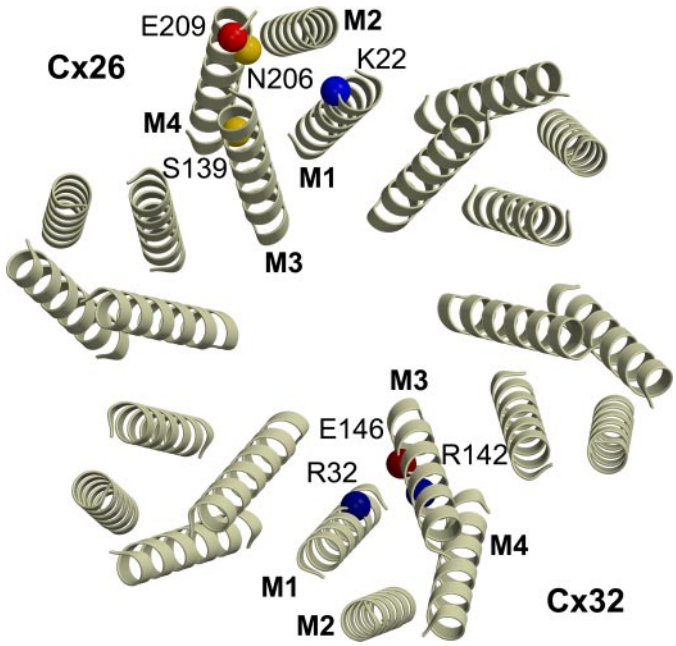


FIGURE 1. **Overall organization of the gap junction TM domain in one of two apposed membranes viewed from the cytoplasm of one cell looking toward the gap.** Six connexin subunits are organized around a central pore. The model structure (Protein Data Bank code 1txh) (17) reports only the positions of α-carbons, assuming that each TM segment forms a canonical α-helix. The model guided the mutation analyses by suggesting which residues form physical interactions. Amino acid positions that were mutated in Cx26 (*top*) and Cx32 (*bottom*) are indicated by *spheres*. *Blue* and *red spheres* represent the positions of positively and negatively charged amino acids, respectively; *yellow spheres* represent polar residues. One-letter codes for the amino acids are shown: *E*, Glu; *K*, Lys; *N*, Asn; *R*, Arg; and *S*, Ser. This and all other molecular representations were generated using MOLSCRIPT (38) and rendered with Raster3D (39).

Cellular Localization—Cells were fixed with either 4% paraformaldehyde for 20 min or with 100% ethanol (−20 °C) for 5 min and mounted on slides using Gel Mount (Biomeda, Foster City, CA). A comparison of the phenotypes using either fixation agent revealed no observable differences (data not shown). The data presented in the paper are based on paraformaldehyde as the fixation agent, as was done, *e.g.* in Refs. 14 and 37. Slides were observed through a Leica TCS SP2 AOBS confocal microscope. For each of the fluorescence images shown here, we examined the phase-contrast image to identify the regions of apposition between cells. The phase-contrast and fluorescence images are available as supplemental data.

Structural Modeling—The template structure of the TM domain of the gap junction, comprising Cx32 monomers (Protein Data Bank code 1txh) (17), was used as a starting point. Backbone atoms were added to this model using the Biopolymer module of the InsightII program (Accelrys, San Diego, CA). The sequences of Cx26 and Cx32 were aligned (17) to generate a model of the gap junction formed by Cx26 monomers. Side chains were then added to both structures using default parameters. Steric clashes were ignored at this stage. The rotameric states of Lys-22 (Cx26), Arg-32, and Arg-142 (Cx32) were examined manually, and rotamers were selected that minimized distances from the putative salt bridge partners, while also minimizing steric clashes with other parts of the protein.

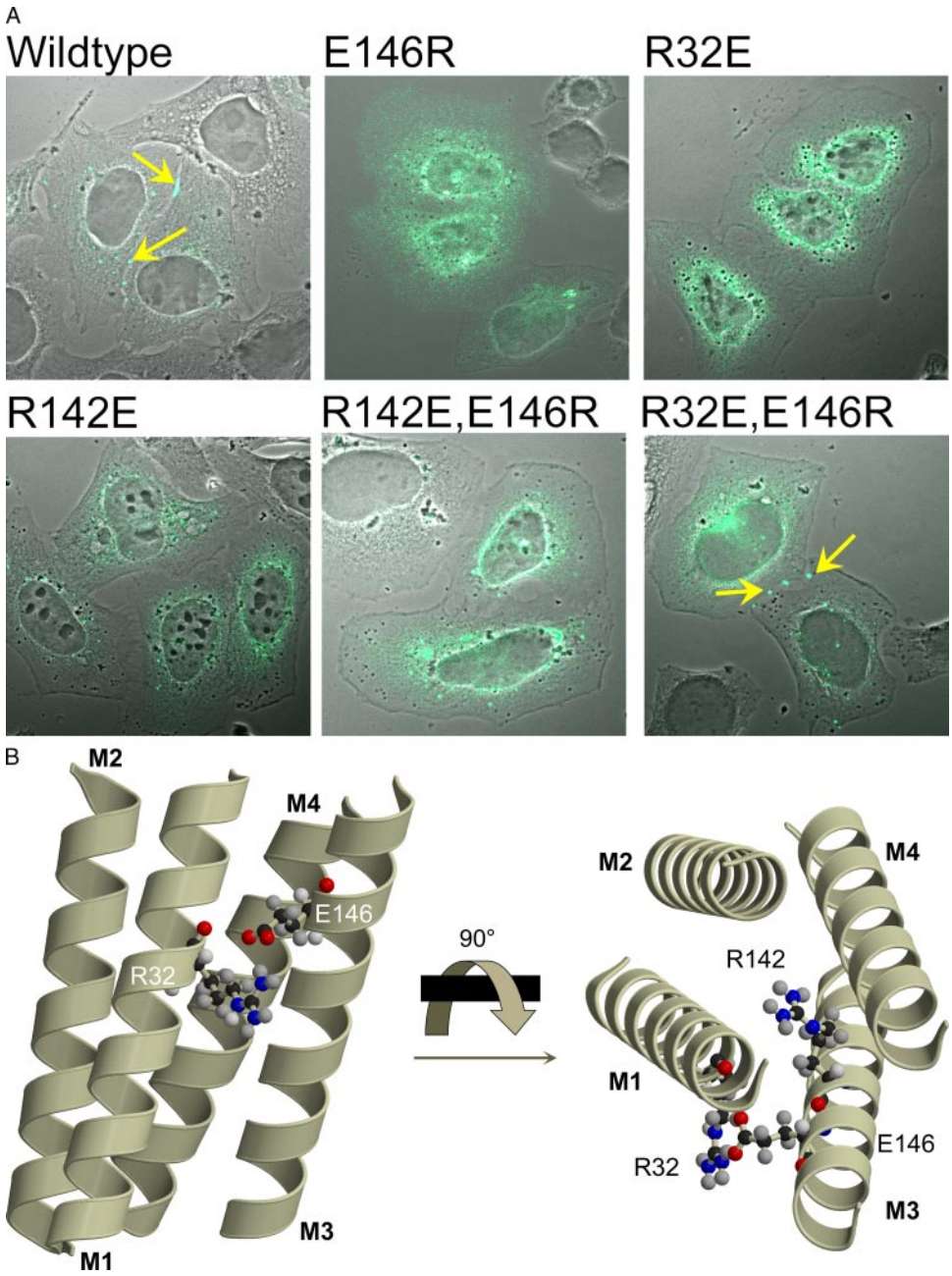


FIGURE 2. *A*, charged amino acids at the interface between the pore-lining helices M1 and M3 of Cx32 were mutated singly and doubly. Localization assays show that the two single charge reversals are mislocalized (*top*). The R32E/E146R double mutant restores membrane localization (*yellow arrows*), indicating that these positions interact. Pictures were taken by confocal microscopy. For each field, three images were taken: phase contrast, to show the boundaries of the cells, green fluorescence, to show the expression of the connexin-GFP chimera protein and a merge of the two, to verify the localization of gap junction plaques in the plasma membrane in points of cell-cell apposition. The phase-contrast and fluorescence images for all figures presented in the manuscript are available as supplemental data. *B*, the modeled side chains suggest that only the paired residues Arg-32/Glu-146 could form a salt bridge without invoking severe steric clashes, as shown by the localization assays. Because the localization assays (*A*) did not detect interactions between Arg-142 and other residues, this modeled side chain (shown on the *right*) should be considered as speculative.

RESULTS

A Salt Bridge between the M1 and M3 Pore-lining Helices—A region of the model structure that showed particular promise for this approach was the interface between the M1 and M3 pore-lining helices, where a triad of charged positions (Arg-32, Arg-142, and Glu-146 of Cx32) is located (Fig. 1). These three positions (the first two occupied by basic residues and the last by an acidic residue) are highly conserved throughout all connexins. In theory, the Arg and Glu residues, which are reciprocally charged and are near the water-filled pore lumen, could be involved in stabilizing electrostatic interactions; it has been estimated that salt bridges embedded in water can add nearly 1 kcal/mol to protein stability (25, 26).

We investigated the possible existence of an interaction between Arg-32 and Glu-146 by reversing the charges of these positions singly and doubly in Cx32 (Fig. 2*A*) (Table 1). Both single mutants, R32E and E146R, were localized outside the plasma membrane, which might be indicative of protein misfolding. Interestingly, a similar charge-reversing mutation, E146K in Cx32, was implicated in Charcot-Marie-Tooth disease (27). Remarkably, the double mutant R32E/E146R, which re-establishes charge complementarity between M1 and M3, restored membrane localization. The compensation suggests that the two residues interact. Given the model structure, it is likely that this interaction involves a salt bridge that stabilizes the interface between M1 and M3 (Fig. 1).

We also investigated the existence of an alternative interaction between Arg-142 and Glu-146. Here too, both single mutants were mislocalized. Despite the proximity between positions 142 and 146, however, the R142E/E146R double mutant did not elicit wild-type localization.

To explain this pattern of phenotypes, we examined the connexin model structure and explored the possible rotameric states (*i.e.* conformations) of the side chains of Arg-32, Arg-142, and Glu-146 (Fig. 2*B*). One of the rotamers of Arg-32 complied with the formation of an interhelical salt bridge with Glu-146. By contrast, none of the potential rotameric states of the side chain of Arg-142 could interact with Glu-146 without generating steric clashes with other parts of the protein. Hence, guided by the mutation assays, modeling of the side chains provided putative mechanistic explanations of the observed localization phenotypes. However, it should be borne in mind that this explanation is based on the model that the localization assay is attempting to support and that correct modeling of side chains is highly sensitive to the accuracy of the C-α model structure (28).

TABLE 1

Frequency of coupled cells transfected with wild-type and mutated connexins

Means and S.D. for each connexin were computed on the basis of three separate experiments and normalized to the levels in the wildtype. For each entry at least 250 transfected cells were counted. Transfection efficiencies in each case were at least 40%. All mutants that showed any coupling (>0%) were statistically indistinguishable on a two-sided *t*-test from the coupling levels in their respective wildtype protein ($\alpha = 5\%$). NA, not applicable.

	Frequency of coupled pairs	S.D.
Cx32 constructs		
Wild type	1	0.06
R32E	0	NA
R142E	0	NA
E146R	0	NA
R142E/E146R	0	NA
R32E/E146R	0.97	0.11
Cx26 construct		
Wild type	1	0.21
K22E	0.75	0.27
E209K	0	NA
K22E/E209K	0.55	0.39
N206S	1.13	0.11
S139N	0	NA
N206S/S139N	1.15	0.09

Although wild-type connexins are membrane-localized, our images show fluorescence also outside the membrane, even in wild-type connexin (*e.g.* Fig. 1). Localization of wild-type connexins outside the membrane, in addition to the existence of gap junction plaques, have been observed in other studies involving overexpressed connexins, and it has been suggested that the cytoplasmic fraction of the protein is at least in part localized in aggresomes (29). The important point to notice from the perspective of the current study is that, along with the localization of some of the protein in the cytoplasm, wild-type and doubly mutated connexins are localized in the plasma membrane, whereas the single mutants are not.

A Salt Bridge at the Intercellular Part of the TM Domain—The charged residues of another pair, Lys-22 (M1) and Glu-209 (M4) of Cx26, face one another in the model structure, potentially forming a salt bridge (Fig. 1). These positions are conserved throughout the connexin family to basic and acidic identities, respectively. Mutations in both positions of the homologous Cx32 were implicated in Charcot-Marie-Tooth disease (30, 31). Our results show that, in Cx26, the single mutant E209R is mislocalized (Fig. 3*A*), similar to the disease-causing mutation to Lys in the homologous Cx32 (30). By contrast, K22E was properly localized in the plasma membrane (Table 1). The double mutant K22E/E209R rescued the aberrant localization phenotype of E209R and restored wild-type localization in the plasma membrane, suggesting that the two positions interact probably through an interhelical salt bridge. However, the normal localization of K22E, where no salt bridge can be formed, suggests that additional forces contribute to stabilization of this region. We did not detect similar compensation when testing this double mutant in Cx32 (data not shown), possibly because of subtle sequence and structure differences between these isoforms (Lys-22 in Cx26 is aligned with Arg-22 in Cx32). It is nevertheless very likely that the overall structure and salt bridge interactions present in Cx26 are also present in Cx32 (17, 18). It should be noted that, although compensation is a sign of physical interaction, lack of compensation

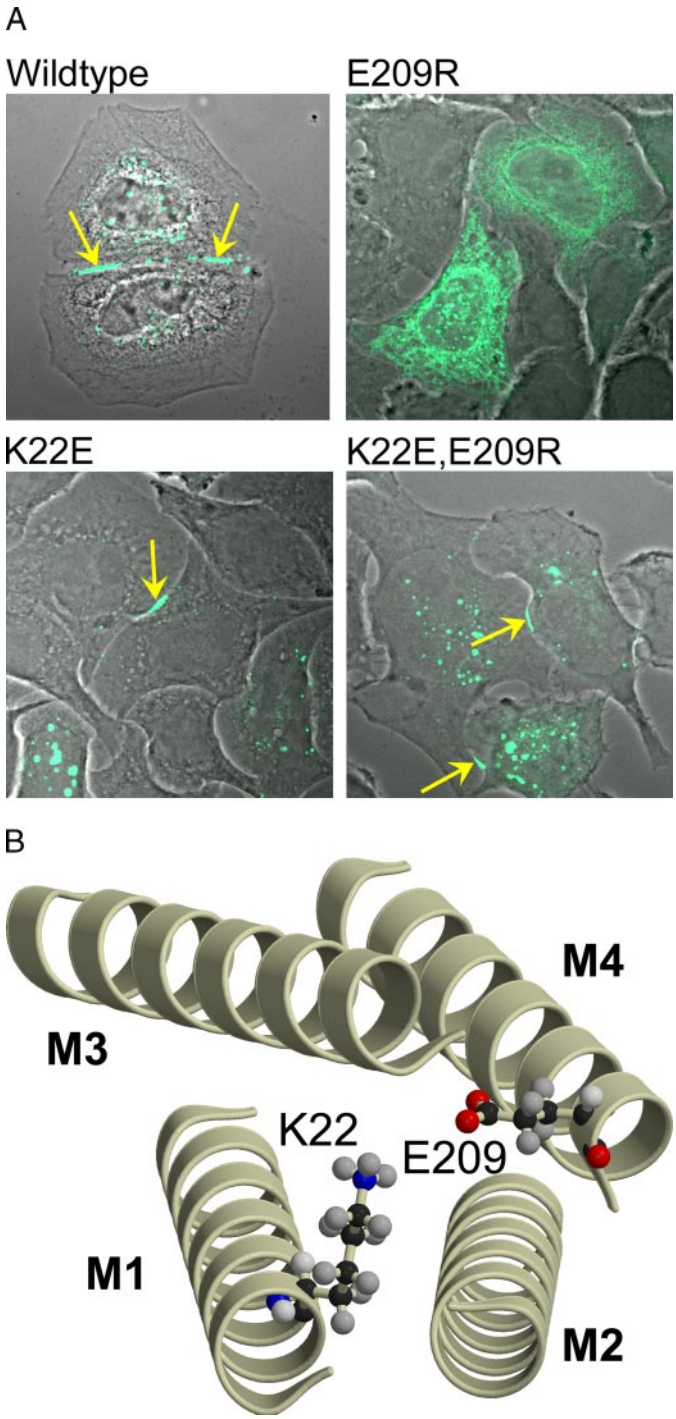


FIGURE 3. *A*, localization assays of wild-type, singly mutated, and doubly mutated Cx26 fused to GFP and expressed in HeLa cells. The single mutants E209R and K22E are cytoplasm- and membrane-localized, respectively. The mutation K22E suppresses the aberrant phenotype of E209R, indicating that the two positions interact. *B*, Lys-22 (M1) and Glu-209 (M4) of Cx26 may form a salt bridge. The two positions are located on the cytoplasmic boundary of the presumed hydrophobic core of the membrane and may thus be embedded in water or in the vicinity of the polar headgroups. The distance between the modeled amino and carboxyl moieties is <5 Å.

(as in Cx32) does not necessarily indicate that the residues are remote (discussed in Refs. 32 and 33).

Further examination of the model structure in light of the results of the localization assay revealed one choice of rotamers for the modeled side chains of Lys-22 and Glu-209, where the car-

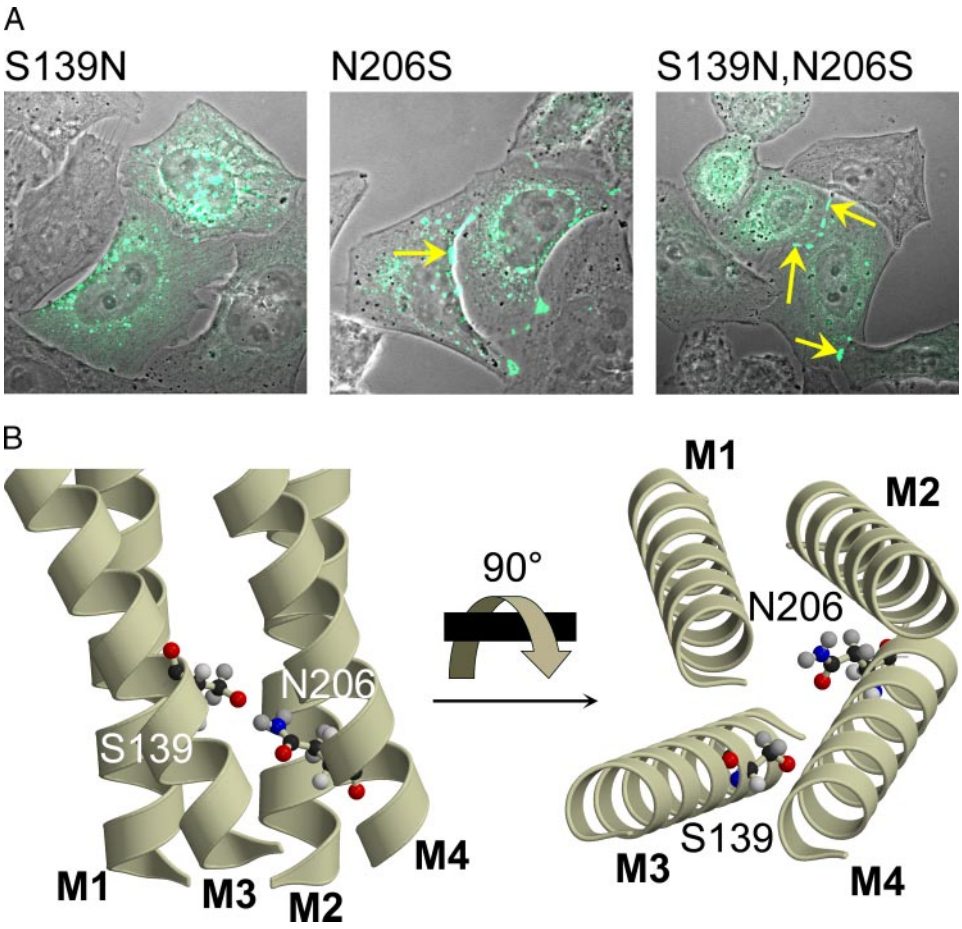


FIGURE 4. *A*, of the two mutations that cause deafness, S139N and N206S in Cx26, only the first is mislocalized. The second mutation compensates for the effects of the first, restoring the wild-type localization. The wild-type localization phenotype of Cx26 is shown in Fig. 3*A*. *B*, Ser-139 and Asn-206 of Cx26 are too far apart (roughly 6–7 Å) to interact directly. They can, however, interact through other residues in their vicinity.

boxyl and amino moieties of the side chains could point toward one another and are roughly 4–5 Å apart (Fig. 3*B*). The two positions are located at the end of the presumed hydrophobic core of the membrane on the cytoplasmic side (Fig. 1) and thus are likely to be exposed to the polar headgroups and the water in this environment. Embedded in water, this salt bridge is also expected to have stabilizing energetic effects on the folding of the protein (25, 26). It is instructive that the model structure provides a framework for understanding why one of the single mutants has deleterious effects on stability, whereas the other does not (Fig. 3*A*). The E209R mutation not only replaces the charge in this position but also adds to the side chain length, destabilizing the structure by bringing a positive charge from position 209 into the vicinity of the endogenous positively charged Lys-22 as well as by adding the potential for forming steric clashes. The compensating K22E mutation places two negative charges in the same region. However, because of the significant length difference between the side chains of Lys and Glu, the charges are more distant from one another in comparison to the distance between these residues in the wild type, reducing the effects of same-charge repulsion. By contrast, the above-mentioned Arg→Glu mutations in positions 32 and 142 did not show wild-type localization (Fig. 2*A*). The model offers an explanation for this difference in phenotypes as well. Whereas the interface between Arg-32 and Glu-146 on the pore-lining helices M1 and M3 is quite tight (Fig. 2*B*), increasing the effects of same-charge repulsion, the interface between M1 and M4, where positions Lys-22 and Glu-209 are located, is less tight (Fig. 3*B*).

One Disease-causing Mutation Compensates for the Aberrant Localization Phenotype of Another—Another pair of residues, Ser-139 (M3) and Asn-206 (M4) of Cx26, is particularly relevant because of its involvement in a genetic disease (34, 35). Mutations in these two amino acids also demonstrate a pattern of localization phenotypes suggestive of an interaction. Theoretically, interactions between these polar positions could result from hydrogen bonding or steric packing (Fig. 1). The localization assays showed that the S139N mutant was mislocalized, whereas N206S exhibited wild-type localization (Fig. 4*A*) (Table 1), although altered voltage-gating properties in this latter mutant have been reported (36). Interestingly, both of these mutations were implicated in nonsyndromic hearing loss (34, 35). When this pair is doubly mutated, however, one disease-causing mutation compensates for the effects of the other, restoring wild-type localization (Fig. 4*A*). Although the model suggests that the two residues are oriented toward one another (Figs. 1 and 4*B*), the modeled side chains do not appear to be in direct contact (~6–7 Å apart). This is mainly because of a difference in the register along the axis vertical to the plane of the membrane, precluding the formation of a hydrogen bond. Hence, interpretation of the localization assay in light of the model suggests that the interaction between these residues results from packing with intermediate amino acid positions (24). Whereas the S139N mutation adds to the volume at the interface between helices M3 and M4, resulting in mislocalization, the N206S mutation compensates for this increased volume and restores membrane localization in the double mutant.

DISCUSSION

Lack of an atomic resolution structure of the gap junction has made it extremely difficult to conduct biochemical investigations within a consistent framework (9). Here, we used a model based on an intermediate resolution structure (17) to formulate testable hypotheses to uncover some of the physical forces underlying the stability of the connexin TM domain at the molecular level. When use of this model is coupled with carefully planned mutagenesis, competing structural explanations may be resolved and mechanistic understanding improved.

Second-site suppression assays need to strike a fine balance

between the dual goals of introducing a mutation radical enough to elicit an aberrant phenotype, on the one hand, but one that does not cripple the protein so completely as to preclude its rescue by a second-site mutation, on the other hand. Thus, two of the three sets of mutants targeted positions that are situated in spacious regions of the model structure (Arg-32 and Glu-146 of Cx32; and Arg-22 and Glu-209 of Cx26), where they are not expected to induce extensive changes to the packing of the helices. In these sets, we used radical charge-reversal substitutions. In another set of mutants, involving positions that are packed inside the core of the helix bundle (Ser-139 and Asn-206 of Cx26), we tested substitutions that were physicochemically mild to reduce the chances that they would bring forth global changes to the protein structure. In all of these sets, clinical and some biochemical data suggested that the effects of the mutations would be observable in our assays.

Based on compensation assays, we suggested atomic models for several amino acid residues, thus refining the structural model of the gap junction TM domain, which specified only the locations of α-carbons (17). It is notable that, since the publication of the intermediate resolution structure of the gap junction in 1999 (5), it has not been supplanted by an experimental atomic resolution structure. Systematic compensation studies could thus provide constraints that specify the nature of the interactions between amino acid residues on apposed helices. In particular, using experimental assays (13) such as electrophysiology and dye-transfer, it may be possible to test mutations that involve more subtle physicochemical changes than those attempted here, including mutations that only change the steric properties of the side chain, *e.g.* Val→Ile. Based on such studies, it might be possible to provide an atomic resolution description of much of the connexin TM domain, circumventing, in part, the impediments to obtaining an experimental atomic resolution structure (9).

Here, we have focused on connexin localization, and the capacity of doubly mutated connexins to form functional gap junctions that conduct ions, secondary messengers, metabolites, etc., should also be examined. Intriguingly, the three sets of compensatory mutations reported here indicate that elimination of specific interhelical contacts might be the cause of a number of connexin-related genetic diseases. The fact that the localization phenotype can be restored by second-site mutations suggests that it might be possible to rescue the aberrant localization of some mutated connexins by applying small molecules that stabilize the structure of the TM domain.

Acknowledgments—We thank Prof. A. Horovitz, Prof. F. Mammano, Dr. V. H. Hernandez, Prof. B. J. Nicholson, Dr. Y. Ofran, Prof. G. Schreiber, T. Sobe, Prof. V. M. Unger, and Dr. O. Yifrach for comments and suggestions, Prof. D. Kelsell for HeLa cells, and Drs. L. Mittelman and A. Barbul for confocal microscopy.

REFERENCES

1. Kumar, N. M., and Gilula, N. B. (1996) *Cell* **84,** 381–388
2. Cascio, M., Kumar, N. M., Safarik, R., and Gilula, N. B. (1995) *J. Biol. Chem.* **270,** 18643–18648
3. Willecke, K., Eiberger, J., Degen, J., Eckardt, D., Romualdi, A., Guldenagel, M., Deutsch, U., and Sohl, G. (2002) *Biol. Chem.* **383,** 725–737
4. Milks, L. C., Kumar, N. M., Houghten, R., Unwin, N., and Gilula, N. B. (1988) *EMBO J.* **7,** 2967–2975
5. Unger, V. M., Kumar, N. M., Gilula, N. B., and Yeager, M. (1999) *Science* **283,** 1176–1180
6. Pitts, J. D. (1998) *BioEssays* **20,** 1047–1051
7. Bergoffen, J., Scherer, S. S., Wang, S., Scott, M. O., Bone, L. J., Paul, D. L., Chen, K., Lensch, M. W., Chance, P. F., and Fischbeck, K. H. (1993) *Science* **262,** 2039–2042
8. Kelsell, D. P., Dunlop, J., Stevens, H. P., Lench, N. J., Liang, J. N., Parry, G., Mueller, R. F., and Leigh, I. M. (1997) *Nature* **387,** 80–83
9. Fleishman, S. J., Unger, V. M., and Ben-Tal, N. (2006) *Trends Biochem. Sci* **31,** 106–113
10. Kronengold, J., Trexler, E. B., Bukauskas, F. F., Bargiello, T. A., and Verselis, V. K. (2003) *J. Gen. Physiol.* **15,** 389–405
11. Hu, X., and Dahl, G. (1999) *FEBS Lett.* **451,** 113–117
12. Trexler, E. B., Bukauskas, F. F., Kronengold, J., Bargiello, T. A., and Verselis, V. K. (2000) *Biophys. J.* **79,** 3036–3051
13. Harris, A. L. (2001) *Q. Rev. Biophys.* **34,** 325–472
14. Skerrett, I. M., Aronowitz, J., Shin, J. H., Cymes, G., Kasperek, E., Cao, F. L., and Nicholson, B. J. (2002) *J. Cell Biol.* **159,** 349–360
15. Fleishman, S. J., Harrington, S., Friesner, R. A., Honig, B., and Ben-Tal, N. (2004) *Biophys. J.* **87,** 3448–3459
16. Fleishman, S. J., Yifrach, O., and Ben-Tal, N. (2004) *J. Mol. Biol.* **340,** 307–318
17. Fleishman, S. J., Unger, V. M., Yeager, M., and Ben-Tal, N. (2004) *Mol. Cell* **15,** 879–888
18. Yeager, M., and Gilula, N. B. (1992) *J. Mol. Biol.* **223,** 929–948
19. Aasen, T., Hodgins, M. B., Edward, M., and Graham, S. V. (2003) *Oncogene* **22,** 7969–7980
20. Evans, W. H., and Martin, P. E. (2002) *Mol. Membr. Biol.* **19,** 121–136
21. Carter, P. J., Winter, G., Wilkinson, A. J., and Fersht, A. R. (1984) *Cell* **38,** 835–840
22. Nikolova, P. V., Wong, K. B., DeDecker, B., Henckel, J., and Fersht, A. R. (2000) *EMBO J.* **19,** 370–378
23. Horovitz, A., Bochkareva, E. S., Yifrach, O., and Girshovich, A. S. (1994) *J. Mol. Biol.* **238,** 133–138
24. Horovitz, A., Bochkareva, E. S., Kovalenko, O., and Girshovich, A. S. (1993) *J. Mol. Biol.* **231,** 58–64
25. Horovitz, A., Serrano, L., Avron, B., Bycroft, M., and Fersht, A. R. (1990) *J. Mol. Biol.* **216,** 1031–1044
26. Hendsch, Z. S., and Tidor, B. (1994) *Protein Sci.* **3,** 211–226
27. Numakura, C., Lin, C., Ikegami, T., Guldberg, P., and Hayasaka, K. (2002) *Hum. Mutat.* **20,** 392–398
28. Fleishman, S. J., and Ben-Tal, N. (2006) *Curr. Opin. Struct. Biol.* **16,** 496–504
29. Das Sarma, J., Meyer, R. A., Wang, F., Abraham, V., Lo, C. W., and Koval, M. (2001) *J. Cell Sci.* **114,** 4013–4024
30. VanSlyke, J. K., Deschenes, S. M., and Musil, L. S. (2000) *Mol. Biol. Cell* **11,** 1933–1946
31. Ressot, C., Latour, P., Blanquet-Grossard, F., Sturtz, F., Duthel, S., Battin, J., Corbillon, E., Ollagnon, E., Serville, F., Vandenberghe, A., Dautigny, A., and Pham-Dinh, D. (1996) *Hum. Genet.* **98,** 172–175
32. Roisman, L. C., Piehler, J., Trosset, J. Y., Scheraga, H. A., and Schreiber, G. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 13231–13236
33. Dall'Acqua, W., Goldman, E. R., Lin, W., Teng, C., Tsuchiya, D., Li, H., Ysern, X., Braden, B. C., Li, Y., Smith-Gill, S. J., and Mariuzza, R. A. (1998) *Biochemistry* **37,** 7981–7991
34. Marlin, S., Garabedian, E. N., Roger, G., Moatti, L., Matha, N., Lewin, P., Petit, C., and Denoyelle, F. (2001) *Arch. Otolaryngol. Head Neck Surg.* **127,** 927–933
35. Kenna, M. A., Wu, B. L., Cotanche, D. A., Korf, B. R., and Rehm, H. L. (2001) *Arch. Otolaryngol. Head Neck Surg.* **127,** 1037–1042
36. Mese, G., Londin, E., Mui, R., Brink, P. R., and White, T. W. (2004) *Hum. Genet.* **115,** 191–199
37. Gottfried, I., Landau, M., Glaser, F., Di, W. L., Ophir, J., Mevorah, B., Ben-Tal, N., Kelsell, D. P., and Avraham, K. B. (2002) *Hum. Mol. Genet.* **11,** 1311–1316
38. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950
39. Merritt, E. A., and Bacon, D. J. (1997) *Methods Enzymol.* **277,** 505–524

doi:10.1016/j.jmb.2006.08.072 J. Mol. Biol. (2006) 364, 54–67

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Quasi-symmetry in the Cryo-EM Structure of EmrE Provides the Key to Modeling its Transmembrane Domain

Sarel J. Fleishman[1], Susan E. Harrington[1], Angela Enosh[2] Dan Halperin[2], Christopher G. Tate[3] and Nir Ben-Tal[1]*

[1]*Department of Biochemistry George S. Wise Faculty of Life Sciences, Tel-Aviv University Ramat Aviv 69978, Israel*

[2]*School of Computer Sciences Tel-Aviv University, Ramat Aviv 69978, Israel*

[3]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

**Corresponding author*

Small multidrug resistance (SMR) transporters contribute to bacterial resistance by coupling the efflux of a wide range of toxic aromatic cations, some of which are commonly used as antibiotics and antiseptics, to proton influx. EmrE is a prototypical small multidrug resistance transporter comprising four transmembrane segments (M1–M4) that forms dimers. It was suggested recently that EmrE molecules in the dimer have different topologies, i.e. monomers have opposite orientations with respect to the membrane plane. A 3-D structure of EmrE acquired by electron cryo-microscopy (cryo-EM) at 7.5 Å resolution in the membrane plane showed that parts of the structure are related by quasi-symmetry. We used this symmetry relationship, combined with sequence conservation data, to assign the transmembrane segments in EmrE to the densities seen in the cryo-EM structure. A C^{α} model of the transmembrane region was constructed by considering the evolutionary conservation pattern of each helix. The model is validated by much of the biochemical data on EmrE with most of the positions that were identified as affecting substrate translocation being located around the substrate-binding cavity. A suggested mechanism for proton-coupled substrate translocation in small multidrug resistance antiporters provides a mechanistic rationale to the experimentally observed inverted topology.

Keywords: dual topology; protein structure prediction; structural bioinformatics; cryo-EM; mechanism of action

Introduction

Bacterial multidrug resistance is a growing challenge to medical treatment, with previously harmless bacteria inducing life-threatening infections.[1] One of the mechanisms for the acquirement of multidrug resistance is the active extrusion of toxic compounds from the bacterial cell through membrane transporters. Efflux of toxic compounds is driven either by ATP hydrolysis, as in the ABC transporter superfamily,[2] or by coupling the extrusion of toxic compounds to the inward movement of protons down their electrochemical gradient, as in the small multidrug resistance (SMR) family of antiporters. Of the SMRs, EmrE is a representative from *Escherichia coli*, which has been extensively characterized structurally, phylogenetically, and biochemically[3,4] These analyses have provided evidence that EmrE contains four transmembrane (TM) segments that form α-helices.[5,6]

A recent electron cryo-electron microscopy (cryo-EM) analysis of 2D crystals of EmrE bound to one of its substrates, tetraphenylphosphonium (TPP^+), clearly resolved the eight α-helices comprising the EmrE dimer at an in-plane resolution of 7.5 Å and 16 Å perpendicular to the membrane plane.[7] However, at this resolution, the individual amino acid residues were not observed, and the TM segments could not be assigned unambiguously to

Abbreviations used: SMR, small multidrug resistance; TM, transmembrane; cryo-EM, cryo-electron microscopy; TPP^+, tetraphenylphosphonium.

E-mail address of the corresponding author: nirb@tauex.tau.ac.il

the densities representing the α-helices. The 2D crystals of EmrE bind TPP^+ with the same high affinity as detergent-solubilized EmrE, and EmrE in the native *E. coli* membrane,[4,8] so it is thought that the cryo-EM structure of EmrE is a faithful representation of the protein's physiological conformation. Quasi-symmetry between six helices was detected around an axis lying within the plane of the membrane, suggesting that the EmrE monomers might assume dual topology in the membrane, with the monomers arranged in an inverted or upside-down manner with respect to one another.[7] In contrast, no obvious symmetry relationship was observed around axes perpendicular to the membrane plane in either the 3D structure or a previous 2D projection map.[9]

Two atomic-resolution X-ray structures of EmrE have been solved in recent years. The first structure at 3.8 Å resolution appears to have trapped the molecule in an unphysiological state,[10] and is incompatible with much of the biochemical data on this protein.[11] Recently, another X-ray structure of EmrE was solved at 3.7 Å resolution,[12] which included one molecule of bound substrate TPP^+ per dimer. However, it has been argued that this structure too may not be physiologically relevant,[13] for three main reasons. (i) The X-ray structure is very different from the cryo-EM structure of EmrE.[12] (ii) Several key residues that were shown to be critical for substrate binding are not in a position to bind substrate in the structure. For instance, it was demonstrated by different experimental approaches that Glu14 residues from both monomers are crucial for translocation, participate in substrate and proton binding,[14–19] and are in proximity to one another.[20] By contrast, the X-ray structure shows that Glu14 from only one monomer forms partial contact with substrate and the two glutamate residues are over 20 Å apart. (iii) Evolutionary conservation has been shown to be a powerful predictor of helix orientations in integral membrane proteins, with conserved amino acid positions usually occupying locations that are buried in the protein core, whereas lipid-facing positions are evolutionarily variable;[21–28] the X-ray structure of EmrE orients many conserved positions (Figure 1(a)) towards lipid, and conversely, variable amino acids are placed at helix–helix interfaces.[29]

The difficulties that have arisen in determining a high-resolution structure of EmrE that accounts for the body of experimental evidence and recent data supporting the dual topology of EmrE and other members of the SMR family[30,31] gave us the impetus to try to understand the cryo-EM structure through modeling strategies. The proposal of the dual topology architecture of EmrE contradicts previous experimental data that suggested EmrE had one unambiguous topology,[32] but could obviously have crucial implications for structural modeling. Here, we show that the most straightforward structural interpretation of dual topology, i.e. that EmrE is arranged as an anti-parallel homodimer, provides the key for determining a model of EmrE based on the cryo-EM structure.

Results

Quasi-symmetry and helix assignment

The assignment of the two sets of four hydrophobic segments seen in the sequence of EmrE to the eight helices observed in the cryo-EM structure is potentially the most significant hurdle in the structural modeling (theoretically having $4\times8! = 161{,}280$ different permutations).[7] However, if a symmetry relationship existed between two parts of the structure, this problem could be greatly simplified (to $2\times4! = 48$ permutations). The previous analysis of the cryo-EM structure of EmrE identified symmetry between two parts of the structure around an axis of symmetry within the plane of the membrane, but there were no symmetry relationships around axes perpendicular to the membrane plane.[7] Recent data suggesting dual topology in EmrE molecules provide additional support for the in-plane 2-fold symmetry axis.[30,31] Indeed, several integral membrane proteins contain two structurally related domains that are related by a rotational axis of quasi-symmetry within the membrane plane (e.g. GlpF,[33] ClC,[34] and SecYEβ[35]).

To derive the most likely helix arrangement for EmrE, four pieces of experimental data were used. (1) Positions of α-helices were based on the cryo-EM structure.[7] (2) The continuous density between the ends of helices F and H suggested that they were adjacent in the amino acid sequence (Figure 1(b)).[7] (3) The two monomers in the EmrE dimer are represented by A-D and E-H, based upon the symmetrical relationship between A-B-C and H-G-F, correspondingly (Figure 1(b)). (4) Densities A-B-C-F-G-H that form the substrate-binding chamber are composed of helices M1, M2, and M3, because amino acid residues that are involved in substrate binding and translocation are found only in these three helices (Table 1). These data alone were insufficient to give a conclusive model, so evolutionary conservation was used to guide the assignment of sequence segments to helices. The rationale behind the use of evolutionary conservation for helix assignment is that residues that are packed against other helices are conserved during evolution, since even minor substitutions in such positions often weaken interhelix contacts and adversely affect protein function.[21,22,25–28,36] Conversely, lipid-exposed positions are expected to be generally accommodating to sequence variability. Hence, we correlated the conservation of sequences with the extent of burial of each of the helices observed in the cryo-EM structure against other helices to constrain the possible assignments.

We found that the most informative helices in the cryo-EM structure were C and F, which are related to one another by the in-plane symmetry axis

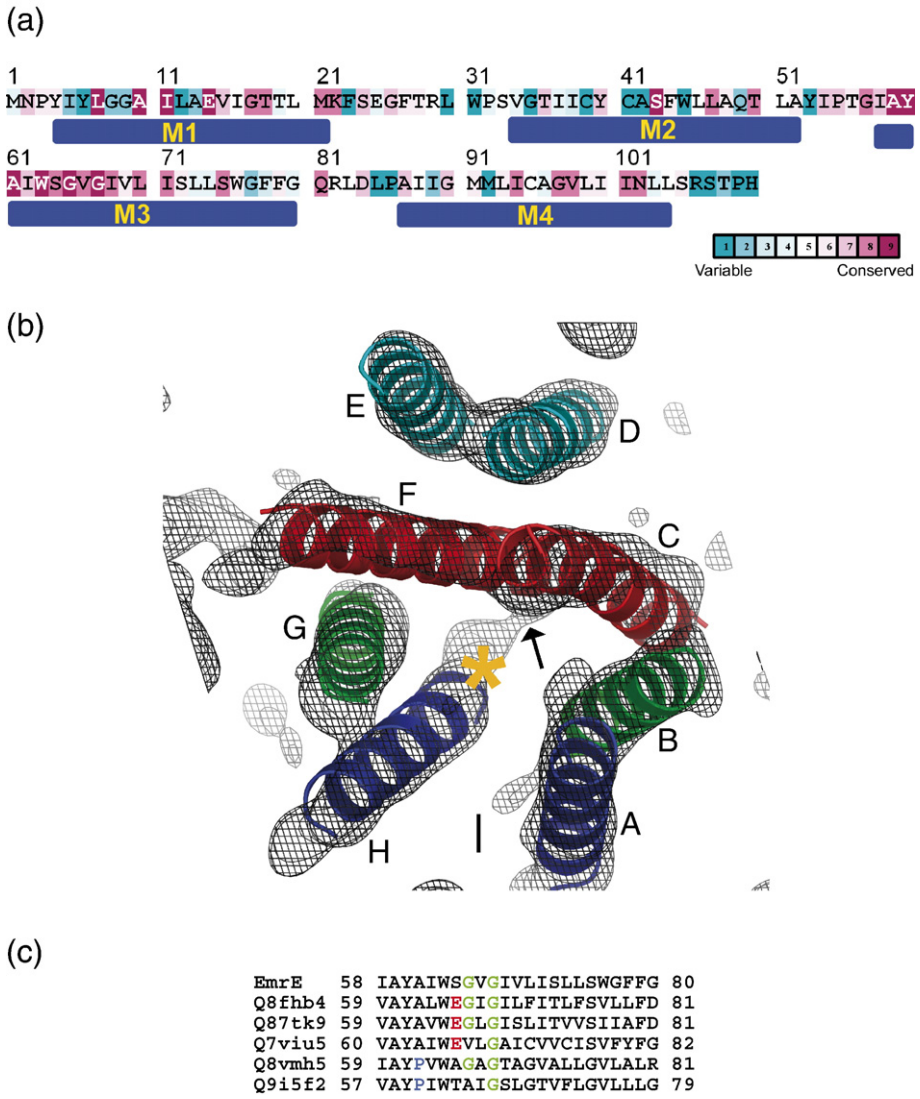


Figure 1. (a) Evolutionary conservation of amino acid residues in EmrE. Sequence conservation was color-coded using the *ConSeq* webserver,[55] and the predicted hydrophobic segments are marked M1–M4. Note that segment M3 is completely conserved in its N terminus with helically periodic variability emerging only in its C terminus. (b) Positions and tilt angles of α-helices inferred from the cryo-EM structure of EmrE[7] viewed perpendicular to the membrane plane. The helices are marked A–H following the notation used by Ubarretxena-Belandia *et al.*[7] The gray mesh indicates electron density at 1.1σ. The arrow marks the position where helices F and H are connected *via* what could be a rigid loop. The orange star marks the approximate in-plane position of the center of the TPP^+ molecule. Notice that helices A-B-C are related to helices H-G-F, respectively, by an approximate 2-fold rotation around the in-plane axis marked by a continuous line. Symmetry-related helices are denoted by the use of the same color. (c) Multiple-sequence alignment of selected SMR sequences in the M3 region. The N terminus of M3 contains the sequence signatures of backbone flexibility, such as fully and highly conserved glycine residues in positions 67 and 64, respectively (green). Some sequences have proline (blue) in positions aligned with Ala61 from EmrE, and others have glutamate aligned with Ser64 of EmrE (red). These polar, small, and helix-deforming residues could elicit flexibility in the M3 segment, correlating with a kink observed in helix C (Figure 1(a)). The complete alignment of SMR homologues is available at http://www.ashtoret.tau.ac.il/~sarel/EmrE.html. Figures 1(b), 3, and 4(b) were generated with PyMol [http://pymol.sourceforge.net/].

(Figure 1(b)). These helices are unique in EmrE, because one half of each helix is buried on all sides by other helices and the other half is exposed to lipid on only one of its faces; of the four hydrophobic segments, only M3 contains conserved amino acid residues in this identical pattern. The N terminus of M3 is highly conserved, implying it is packed on all sides by other helices, but its C terminus shows a helical periodic pattern of variable residues, suggesting that one face is lipid-exposed (Figure 1(a)). In Figure 1(b), the most lipid-exposed, C-terminal portion of M3 is represented by the right-hand (distant) end of C and the left-hand (near) end of F.

The assignment of M3 to C and F is supported partly by the observation that M3 is predicted by our analysis to be the longest hydrophobic stretch (23 residues compared to 18 or 19 for the other TM segments, Figure 1(a)) paralleling its assignment to the most tilted helices in the structure. Further

Table 1. Summary of experimental data gathered on residues of helices M1 to M4 of EmrE, and locations of those residues in the model structure

Residue	Environment predicted from model	Activity data ++=wt	Ref	Environment of label	Ref
M1					
Tyr4	Substrate chamber	−	2,4,5	Accessible	4
Ile5	Lipid facing	++	2,4	Accessible	4
				Lipid-facing	2
Tyr6	Lipid facing	++	2,3,4,5	Partly accessible	4
Leu7	Chamber	−	2,3,4	Accessible	4
				Water-exposed	2
Gly8	Lipid facing	++	2,4	Inaccessible	4
				Lipid-facing	2
Gly9	Lipid facing	++	2,4	Inaccessible	4
Ala10	Interhelix contact	−	2,3,4	Accessible	4
				Water-exposed	2
Ile11	Chamber	++	2	Partly accessible	4
		−	1,3,4	Water-exposed	2
Leu12	Lipid facing	++	1,2,3,4	Partly accessible	4
				lipid-facing	2
Ala13	Lipid facing	++	1,2,3,4	Inaccessible	4
Glu14	Chamber	−	2,3,4,6	Accessible*	4
				Proximal to E14	2
Val15	Interhelix contact	++	2,3,4	Inaccessible	4
Ile16	Lipid facing	++	2,3,4	Inaccessible	4
				Lipid-facing	2
Gly17	Interhelix contact	+	1,2	Partly accessible	4
		−	3,4	Water-exposed	2
Thr18	Chamber	−	1,2,3,4	Accessible	4
				Proximal to T18	2
Thr19	Interhelix contact	++	2,4	Inaccessible	4
				Constrained	2
Leu20	Interhelix contact	++	2,4	Inaccessible	4
				Constrained	2
Met21	Interhelix contact	++	2,3,4	Inaccessible	4
				Constrained	2
M2					
Val34	Substrate chamber	+	1		
Gly35	Lipid facing	++	1		
Thr36	Interhelix contact	+	1		
Ile37	Chamber	++	1		
Ile38	Lipid facing	++	1		
Cys39	Interhelix contact	++	1		
Tyr40	Chamber	−	1,5	Proximal to substrate	4
Cys41	Chamber	+	1		
Ala42	Lipid facing	+	1		
Ser43	Interhelix contact	++	1		
Phe44	Chamber	−	1		
Trp45	Chamber	++	1,7		
Leu46	Interhelix contact	++	1		
Leu47	Chamber	+	1		
Ala48	Chamber	−	1		
Gln49	Lipid facing	+	1		
Thr50	Interhelix contact	++	1		
Leu51	Chamber	++	1		
Ala52	Chamber	−	1		
M3					
Ile58	Interhelix contact				
Ala59	Interhelix contact				
Tyr60	Binding chamber	+	5	Proximal to substrate	5
Ala61	Interhelix contact				
Ile62	Lipid facing				
Trp63	Chamber	−	7	Proximal to substrate	7
Ser64	Interhelix contact				
Gly65	Interhelix contact				
Val66	Interhelix contact				
Gly67	Interhelix contact				
Ile68	Interhelix contact				
Val69	Lipid facing				
Leu70	Interhelix contact				
Ile71	Chamber				
Ser72	Interhelix contact	+	1		
Leu73	Lipid facing	++	1		
Leu74	Interhelix contact	+	1		
Ser75	Interhelix contact	+	1		
Trp76	Lipid facing	++	7		
Gly77	Lipid facing				
Phe78	Chamber				
Phe79	Lipid facing				
Gly80	Lipid facing				
M4					
Ala87	Interhelix contact				
Ile88	Lipid facing				
Ile89	Lipid facing				
Gly90	Interhelix contact				
Met91	Interhelix contact				
Met92	Lipid facing				
Leu93	Interhelix contact	−	1		
Ile94	Interhelix contact	++	1		
Cys95	Lipid facing	++	1		
Ala96	Lipid facing	++	1		
Gly97	Interhelix contact	++	1		
Val98	Interhelix contact				
Leu99	Lipid facing				
Ile100	Interhelix contact				

(continued on next page)

Table 1 (*continued*)

Residue	Environment predicted from model	Activity data ++=wt	Ref	Environment of label	Ref
M4					
Ile101	Interhelix contact				
Asn102	Interhelix contact				
Leu103	Lipid facing				
Leu104	Interhelix contact				

Notice that 22 positions have been probed, but so far, have not been implicated in protein function (++ in the Activity data column); 11 of these are lipid-facing in the model structure. Moreover, 23 positions have been implicated in protein function (− and + in the Activity data column); 21 of these have straightforward structural explanations, with the positions either lining the translocation chamber or situated at helix interaction sites.

Reference 1 (Mordoch *et al.*[45]): Cys-scanning mutagenesis was performed on an active Cys-less mutant. Activity was assessed by performing transport assays for three different substrates. Reference 2 (Koteiche *et al.*[20]): The environment of a spin-label on the Cys mutant is described as either water-exposed, lipid facing or proximal to the corresponding residue in the dimer. Although all residues in M1 were tested, assignments are given only to residues that are unambiguous, with other residues presumably at environmental boundaries. Reference 3 (Gutman *et al.*[16]): Cys mutants were assayed for binding of TPP^+. Reference 4 (Sharoni *et al.*[18]): Accessibility refers to the ability of alkylating agents to react with a Cys mutation at the position indicated and an asterisk (*) shows that the experiment was performed on a heterodimer to ensure proper folding of EmrE. Reference 5: Rotem *et al*. tested the effects of mutations of tyrosine residues to cysteine on function and changes in fluorescence quenching in response to ligand binding. Reference 6: Several studies showed that Glu14 is a critical residue for substrate binding and translocation.[14–19] Reference 7: Elbaz *et al*. tested the effects of mutations of tryptophan residues to cysteine on function and changes in fluorescence quenching in response to ligand binding.[59]

support is provided by the observation that the M3 N terminus contains several sequence signatures that would favor flexibility of the helix backbone, correlating with a kink in helix C observed in the cryo-EM structure (Figure 1(b)). These sequence signatures include (Figure 1(c)): (a) the presence of two highly conserved glycine residues in positions 65 and 67; (b) the observed substitution of position Ser64 with glutamate residues in other SMR members; and (c) the fact that position Ala61 is substituted by proline in several homologues. Notably, proline residues in multiple-sequence alignments of TM domains have been shown to be indicators for kinks, even in cases where the sequence of the protein, for which a structure is available, does not exhibit a proline.[37] Although these sequence features would favor flexibility of the helix backbone, the segment does not necessarily exhibit a kink and, thus, helix F is seen to be mostly straight in the cryo-EM structure (Figure 1(b)).[7]

Given the assignment of M3 to helices C and F, and the experimental constraints listed above, there is only one solution for the assignment of the remaining helices. As the termini of helices F and H are apparently connected by density on the side of the structure away from the viewer in Figure 1(b) (indicated by an arrow), and based on the assignment of the portion of F near the connection to H to be the N terminus of M3, then helix H must be M2. Since the M2–M3 interconnecting loop is predicted to contain only five amino acid residues (Figure 1(a)), it might be rigid and could, therefore, be visible in the cryo-EM structure. If M2 is helix H, then, by symmetry, helix A is also M2. Multiple sources of biochemical data have implicated residues on M1 as crucial for substrate binding and translocation;[14–19] given the assignment of M2 and M3 above, M1 must occupy the symmetry-related B and G helices around the translocation chamber. In contrast, amino acid residues in M4 are not involved directly in substrate binding or translocation (Table 1). The lack of data implicating residues on M4 in substrate binding is in agreement with the location of helices D and E, separated from the substrate-binding chamber by helices C and F. Finally, the helix assignment suggested here (M1 = B,G; M2 = A,H, M3 = C,F, and M4 = D,E) is consistent with constraints imposed by the short interconnecting loops observed in the EmrE sequence (Figure 1(a)) on the distances between the helix ends seen in the cryo-EM structure.[38] In addition to this most likely helix assignment, we tested each of the 47 other permutations against the known functional data on EmrE, the interconnecting loop lengths, and SMR evolutionary conservation. None of these other permutations fit the aggregate data on EmrE nearly as well as the suggested assignment (data not shown).

We note that domain swapping, where helices from one monomer interpenetrate between helices in the other monomer, could confound the proposed helix assignment. However, this possibility would connect helices that are distant from one another, and is therefore made unlikely by the very short lengths of the interconnecting loops (Figure 1(a)).[7] The only loop that would allow domain swapping is between M1 and M2. However, the swapping of these domains involves conformations in which the loop blocks substrate entry to the binding chamber.

Structural modeling

Canonical α-helices were constructed to fit the helix axes[39] extracted from the cryo-EM structure.[7] For each helix, all the rotations around its principal axis were sampled in 5° increments; each conformation was scored according to a rule that favors situations in which evolutionarily conserved amino acid positions were packed inside the protein core, with variable positions facing the lipid.[23] Following the orientation of each of the helices, we introduced a kink into helix C to account for the deviation from α-helical regularity observed for this helix in the cryo-EM structure (Figure 1(b)).[7] At the vertical resolution of the cryo-EM structure (16 Å), the position of the kink cannot be determined unambiguously. We therefore estimated this position on the basis of the direction of the kink observed in the cryo-EM structure and features observed in SMR

sequences of M3 that imply backbone flexibility at the N terminus of this helix (Figure 1(c)), and placed the kink so that it affects mainly the backbone hydrogen bond between positions Ser64 and Ile68. We note that this approximate location for the kink within helix C does not affect the conclusions we draw below on the support that biochemical and biophysical data provide to the model structure.

The computed conformation fits the conservation profile of each helix quite closely (Figure 2) with all of the variable residues facing the lipid, and the conserved residues facing the protein core. It is important to note that no experimentally derived information was used to constrain the orientations of the helices. As shown below, these orientations nevertheless provide a structural framework for understanding most of the biochemical data on EmrE.

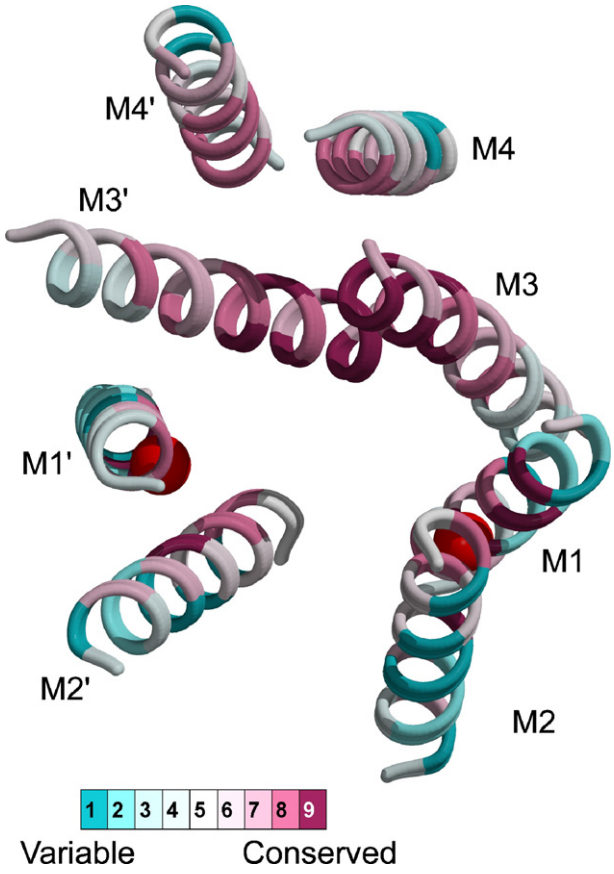


Figure 2. A view of the EmrE model perpendicular to the membrane bilayer color-coded according to evolutionary conservation. On all helices, the conservation signal closely matches the pattern of exposure of residues to lipid, with conserved residues buried at interhelix contact regions, and variable residues placed in membrane-exposed positions. The Glu14 residues on both monomers are shown as red spheres. The two monomers are distinguished by the presence or the absence of an apostrophe. The Figure was generated with MOLSCRIPT[60] and rendered with Raster3D.[61]

Comparison of the EmrE model with data from biochemical and biophysical experiments

It is difficult to interpret the pertinent biochemical and biophysical data on EmrE on the basis of the model structure, because the model does not contain side-chains. Furthermore, we estimate that the orientations of the individual α-helices around their principal axes might vary by up to 20°, and that the positions of C^{α} atoms on the terminal turns of each of the helices might diverge from the positions specified in our model.[23] Even with this level of uncertainty, however, it is possible to provide a rough account of the majority of the experimental data on the basis of the model structure.

The structure of EmrE has been probed using a number of biophysical and biochemical techniques. The model presented here does not seriously conflict with any of these data and, in fact, can be used to rationalize and simplify a number of observations. The experiments discussed in this section used spin labels to probe the environment of helix M1,[20] and site-directed mutagenesis to define amino acid residues important for folding and transport activity.

The TM region M1 of EmrE contains Glu14, which is essential for substrate binding and translocation.[14–16,18,19] Therefore, this region has been studied intensively using a number of biophysical and biochemical approaches. Site-directed spin-labeling experiments were applied to all the residues of M1 to infer which of them are packed against other helices, exposed to lipid, or are in the vicinity of M1 residues of the neighboring monomer.[20] All of the residues that were identified as lipid-exposed by the spin-labeling experiments are predicted to be lipid facing in our EmrE model (dark blue spheres in Figure 3(a) and Table 1). Interestingly, lipid-exposed positions were identified by spin-labeling to be restricted mainly to the N-terminal part of M1; side-chains in its C terminus were found to be motionally more restricted (light blue spheres in Figure 3(a)). These results are in close agreement with the model structure, in which the N-terminal part is more exposed to lipid and the C terminus is packed against other helices from almost all directions. Thus, the spin-labeling data[20] verify the assignment of M1 to helices B and G as well as the helices' orientations around their principal axes. The spin-labeling experiments also identified only two residues on M1 (Glu14 and Thr18) that are vicinal to their counterparts on the other monomer. Indeed, these two residues face one another according to the model structure, and these two pairs have the closest C^{α}–C^{α} distances of all residues on M1 in the model (~16 Å). The model proposed by Koteiche *et al.* for the relative orientations of the two M1s considered only parallel helix packing;[20] however, their results fit equally well with our antiparallel model shown here.

Hsmr is a homologue of EmrE from the archaeon *Halobacterium salinarium*, which is unique among SMR members, in that approximately 40% of its sequence is comprised of Ala and Val residues.[40]

Presumably, this composition reflects an evolutionary pressure to increase the G+C content of the genome, while maintaining the relatively high hydrophobicity necessary for a TM protein. Positions that are not Ala or Val in Hsmr are, therefore, considered important for structure or function; conversely, positions that are Ala or Val in Hsmr, but not Ala or Val in EmrE, can be presumed to be unimportant.[40,41] As expected, the vast majority of these positions are lipid exposed according to the model structure (Figure 3(b)). This observation suggests that, despite the low level of sequence identity between SMR proteins, the overall fold of the homologous proteins is conserved.

The cryo-EM structure of EmrE was derived from crystals of the transporter bound to TPP^+. The position of TPP^+ in the plane of the membrane is clear from the 3D structure and from comparisons of projection maps of EmrE with and without TPP^+.[8] However, the position of TPP^+ along the axis perpendicular to the membrane plane is less certain due to the low resolution along this axis.[7] To provide

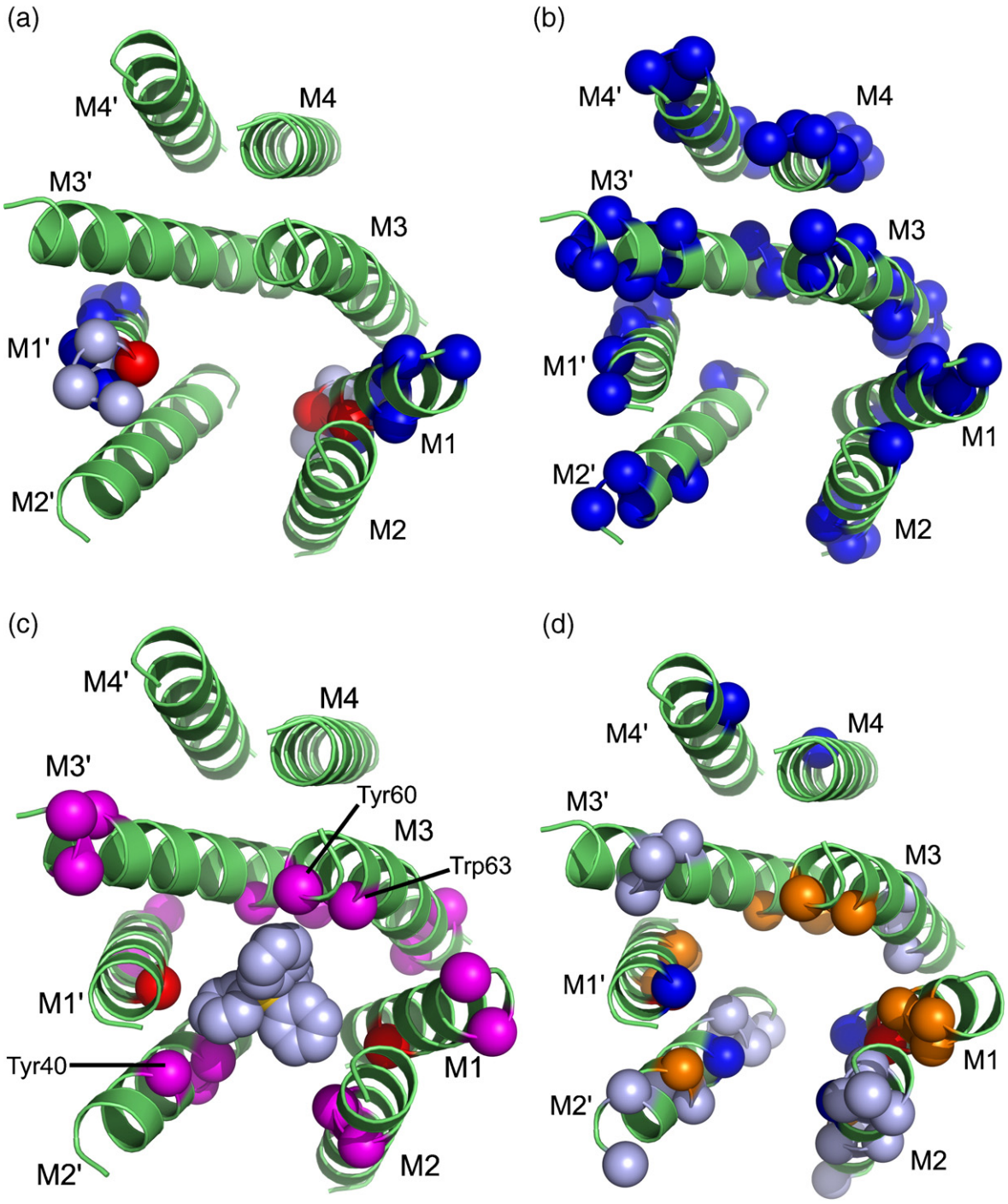


Figure 3 (*legend on opposite page*)

rough constraints for which residues are located around the substrate, we docked TPP^+ manually (Figure 3(c)) based on the constraint that the position of TPP^+ with respect to Glu14 should roughly match that seen in the atomic-resolution structure of the water-soluble multidrug receptor BmrR,[42] which was also crystallized in a TPP^+-bound form. In harmony with various experimental assays, the Glu14 residues from both monomers are in position to form contact with the substrate.[17,19] It is also notable that several aromatic residues are found in the vicinity of the modeled TPP^+, providing partners for aromatic interactions with substrate. In particular, positions Tyr40, Tyr60, and Trp63 were shown experimentally to bind substrate,[18] and their C^α atoms are located within 6 Å from carbon atoms of the modeled TPP^+ (Figure 3(c)). Other aromatic residues are located between α-helices, where they might increase structural stability. Table 1 lists 52 amino acid residues that line the translocation chamber or mediate interhelix contacts in our model; 21 of these residues have been mutated and implicated experimentally in substrate binding and translocation. It should be noted, however, that it is likely that the specific residues mediating EmrE binding to substrates other than TPP^+ might vary from those specified here, in analogy to the differences observed in the binding modes of different substrates to bacterial multidrug gene regulators,[43] and bacterial multidrug resistance transporters.[44]

Mordoch *et al.* conducted an extensive substituted cysteine-accessibility method analysis of Cys-less EmrE[45] (Table 1). They replaced 48 positions throughout the protein with cysteine, and tested the mutant transport properties. Only five positions were absolutely sensitive to replacement, in that the mutants were incapable of conferring resistance to known EmrE substrates. Of these five, two mutations (Ile11Cys and Thr18Cys) led to good expression of EmrE, but transport was reduced severely; these positions are on the same face of M1 as the essential Glu14 and are probably involved in substrate transport (Figure 1(d)). The other three mutations (Tyr40Cys, Phe44Cys, and Leu93Cys) resulted in no expression of EmrE, so it may be that these residues are important in the folding or stability of EmrE in the membrane. Subsequent mutational analysis of Tyr40 showed that this residue is also important in substrate recognition.[18] Both Phe44 and Leu93 are predicted to be at the interfaces between helices (Figure 3(d)), which may explain their effect on protein folding and/or stability. This substituted cysteine-accessibility method study also identified ten residues in the TM regions that showed decreased resistance to only one of the antiporter cognate substrates.[45] Eight out of these ten residues are positioned at or around the substrate-binding chamber (Table 1 and light blue spheres in Figure 3(d)). Presumably, substituting these positions alters the properties of the protein surface that lines the chamber, hence reducing substrate affinity. As Mordoch *et al.* note, the residues on M2 are distributed on two helical faces, and indeed the two sensitive positions (Ala42 and Gln49), which the model does not place at or around the substrate-binding chamber, are located on this helix. Notably, mutants of these two residues are sensitive only to acrylflavine among the cognate substrates that were tested,[45] implying that these residues might be involved in the binding of only certain substrates.[43,44]

There are two reports of experimental data that conflict with the model we present here, because they both suggest that the monomers within EmrE have an identical orientation in the membrane with the N and C termini in the cytoplasm.[32,46] A third study reported single topology for the QacC homologue of EmrE, although the data were inconclusive regarding the localization of the C terminus.[47] These results are, however, contrary to

Figure 3. Structural interpretation of biochemical and phylogenetic data on EmrE and its homologues. (a) Spin-labeling experiments identified lipid-exposed residues (dark blue), and motionally restricted (light blue) positions.[20] Red spheres identify Glu14 and Thr18, which were shown to be close to their counterparts on the other monomer, as indeed they are in the model. Notice that in both monomers, the motionally restricted residues on the N-terminal turn of M1 are surrounded by other helices from almost all sides, whereas positions that were identified experimentally as lipid-exposed are indeed located in lipid-facing parts of the protein. (b) Green spheres mark positions on EmrE that are aligned with Ala or Val in the Hsmr homologue from *H. salinarium*, but are not Ala and Val in EmrE. Such positions are thought to have little structural or functional importance,[40] and indeed the majority face the lipid environment. (c) Docking of a molecule of TPP^+ in the EmrE model structure. TPP^+ was docked manually such that it approximately fits the orientation seen in a crystal structure of the cytoplasmic receptor for TPP^+ BmrR.[42] The two Glu14 residues (red spheres) are in proximity to aromatic rings from the substrate TPP+. Aromatic residues in the TM domain of EmrE are marked by purple spheres. Some of these residues surround TPP^+, thus providing possible interaction partners for the substrate. Positions Tyr40, Tyr60, and Trp63, which are marked on the Figure, have been implicated directly in substrate binding.[18] Others are placed in spacious regions of the structure, where they might serve to enhance the interactions between helices (e.g. the M2/M2′, M1/M3, and M1′/M3′ interfaces). The substrate TPP+ molecule is shown in space-filling spheres, with light blue corresponding to carbon and yellow to phosphate atoms. (d) Blue spheres indicate four positions where mutations to Cys abolish functionality, and green spheres indicate positions that change resistance to only some of the transporter's cognate substrates.[45] Orange spheres mark positions that are involved in substrate binding.[18,19] All of the blue spheres map to positions at the interfaces between the helices, where mutations might disrupt protein folding or oligomerization, or around the binding chamber. Most of the light blue spheres map to positions around the translocation chamber at least in one of the monomers, where changes to the surface of the protein might modify substrate recognition. The orange spheres are all located around the Glu14 residue. A listing of all residues in the TM domain and their experimentally determined structural or functional roles is provided in Table 1.

the topology analysis reported by Daley *et al.*,[30] who showed that the predominant orientation of EmrE has the N and C termini in the periplasm. The cross-linking data identifying helix–helix interactions[46] are difficult to reconcile with our model, and will require an atomic-resolution structure to be determined before the conflict can be resolved, as was the case for the cross-linking data for the lactose permease.[48] The most internally consistent cross-linking data showing that helix M4 lies parallel with and adjacent to M4 from the neighboring monomer could be explained easily by suggesting that EmrE is a tetramer in the membrane, related by a 2-fold perpendicular to the membrane,[4] a proposal that is supported by recent data from studying the interaction between peptides representing individual TM regions of Hsmr, the archaeal homologue of EmrE.[49]

An alternate-access mechanism for substrate translocation

Transport of drug substrates from the cytoplasm or cytoplasmic leaflet of the lipid bilayer out of the bacterium is thought to occur in essentially two steps[3,4] (Figure 4(a)). First, the drug substrate binds

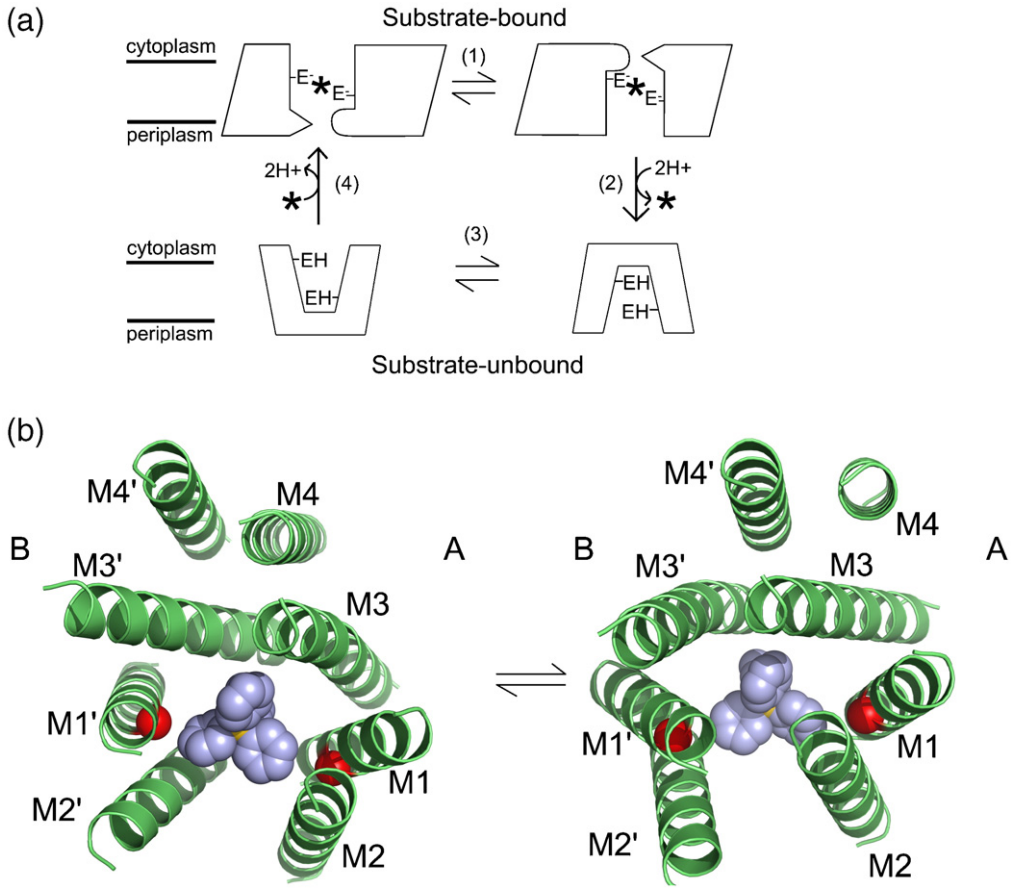


Figure 4. (A) A mechanism for proton-coupled translocation of substrates by the SMR family of proteins. (1) Two substrate-bound forms of the protein interconvert between conformations, in which the substrate, marked by an asterisk (*) faces the cytoplasm or the periplasm due to conformational changes. (2) In the periplasmic-facing conformation, the substrate is supplanted by the binding of two protons to the Glu14 positions (marked by E^-) on both monomers, thus driving the equilibrium towards substrate translocation. (3) A conformational change reorients the binding site towards the cytoplasm. (4) Substrate binding on the cytoplasmic side forces the protons out of the translocation chamber into the cytoplasm. (b) A suggestion for the conformational change represented by Step 1 in Figure 4(a). Periplasmic-facing and cytoplasmic-facing conformations of the EmrE dimer based on the cryo-EM structure. The transition between the two conformations involves a reorientation of the M1–M3 helices in both monomers by approximately 20° with respect to the in-plane axis of symmetry; a kinking and straightening of M3; and a small translation of M1–M3 in both monomers with respect to the M4 helices. As these changes occur in the protein dimer, the TPP^+ substrate, which is accessible from the near end in the conformation on the left, moves downwards and becomes accessible from the far end of the EmrE dimer in the right-hand conformation. Thus, interconversion between these two conformations could alter the accessibility of substrate from cytoplasmic facing to periplasmic facing. The conformation on the right was obtained by rotating the conformation on the left by 180° with respect to the in-plane axis of quasi-symmetry, so that the two conformations are completely superimposable. Thus, inverted topology would reproduce a single substrate-binding mode as two conformations, one of which is accessible to the periplasm and the other to the cytoplasm. The two monomers are arbitrarily marked A and B.

to EmrE, which induces a conformational change, so that the inward-facing binding pocket is opened to the periplasm and closed to the cytoplasm. The high concentration of protons in the periplasm competes directly with the drug for binding at the two Glu14 residues, so protonation results in release of the drug. A further conformation change then re-orients the protein to face the cytoplasm, where it may bind another drug molecule.

The nature of these conformation changes is uncertain, but the cryo-EM structure and our model suggest its basic features. To explain the substrate-translocation process, we propose that M1-M4 from monomer A adopt the conformation of M1-M4 observed in monomer B in our model and *vice versa* (Figure 4(b)) during the step marked as (1) in Figure 4(a). Due to the in-plane symmetry, this transition results in a structure identical with the original model rotated by 180° with respect to the in-plane axis of symmetry; the two symmetry-related structures are shown on both sides of the chemical equilibrium in Figure 4(b). To analyze the details of this transition, it is useful to divide the model structure into three subunits: (1) M1, M2 and M3 (monomer A); (2) M1′, M2′ and M3′ (monomer B); and (3) M4 and M4′.

Helices M1–M3 from monomer A are virtually superimposable on M1′–M3′ of monomer B, except for the kink in M3, suggesting that M1–M3 move as one unit during the transition described in Figure 4(b). The two M4 helices are seen to make minimal movements with respect to one another during the transition, suggesting that they are stable as a helix pair. Indeed, the cryo-EM and model structures show this pair to be closely packed with glycine residues (positions 90 and 97) lining the interhelix interface, which can stabilize helix packing.[50] By contrast, the interfaces between helices M3 and M4 are small in both monomers in comparison to any of the other pairs of interacting helices in the structure (Figure 4(b)). It therefore comes as no surprise that the most significant conformational change that occurs during the transition can be localized to the contact region between M3 and M4, with the crossing angles between these helices changing by approximately 20° around the in-plane axis of symmetry in order to switch the M3-M4 packing from that observed in monomer A to that observed in monomer B, and *vice versa*. The kinking and straightening of the two M3 helices and a small translation of the M1-M3 helices in both monomers with respect to the M4 helices, coupled to the movement of the TPP^+ molecule perpendicular to the membrane plane would then complete the transition. Thus, although residues on M4 have so far not been recognized as important for substrate binding and translocation (Table 1), this putative mechanism suggests a crucial role for M4 in stabilizing the dimer interface during the translocation process. The short M3-M4 loop, consisting of six residues in the SMR family, would hold the two parts of the structure together in the face of these relative motions. It is interesting to note in this connection that recent results have suggested a role for the M4 helix in mediating the formation of SMR tetramers.[49]

The sum of these conformational changes would alternately open the substrate-translocation chamber to the cytoplasmic and periplasmic media, allowing substrate to bind in the cytoplasmic-facing conformation, and then to be replaced by protons when the protein faces the periplasm (Figure 4(a)). Interestingly, this mechanism suggests that the periplasmic- and cytoplasmic-facing conformations of substrate-bound EmrE are essentially identical, and would thus require a single substrate-binding mode to be optimized structurally, which would then by symmetry be reproduced in both cytoplasmic-facing and periplasmic-facing conformations. Currently, there are only two conformations of EmrE (Figure 4(a), the upper panels) for which we have structural information (Figure 4(b)),[7,8] and the structure of further transport intermediates (i.e. Figure 4(a), lower panel) will be essential to identify conformational changes that occur during the transport cycle. However, the availability of our model will now allow the design of specific experiments, such as using site-specific spin labels to monitor movements in EmrE during the transport cycle.

Discussion

The suggestion of dual topology of EmrE,[7] and the recent support for this from global-topology analyses,[30,31] were the key for the successful modeling of EmrE presented here. The presence of 2-fold quasi-symmetry between the monomers of the EmrE dimer within the plane of the membrane (Figure 1(b)) implied an antiparallel orientation of the EmrE monomers. Our previous modeling attempts (not described), which were not guided by the in-plane pseudo 2-fold axis, were unsuccessful in providing explanations for the biochemical and biophysical observations on EmrE. In contrast, our model with the monomers in an antiparallel orientation explains virtually all the biochemical and biophysical data. The model makes many predictions about the structure of EmrE that will provide a platform for further experimental work, such as the identification of other residues in the translocation pathway that have not yet been studied (Table 1), and residues that may be important in mediating helix packing, and therefore could be involved in the conformational changes.

The suggestion of oppositely oriented monomers in EmrE has been made only recently,[7] and is reinforced by global topology analysis of bacterial proteins,[30,31] but, so far, mechanistic advantages of dual topology have not been proposed. The mechanism of translocation that we invoked above suggests a potential advantage. That is, if the cytoplasmic-facing and periplasmic-facing conformations of substrate-bound EmrE are essentially identical, then only one mode of substrate binding should be

devised by evolution, which would be replicated as two conformations, one facing the cytoplasm and another facing the periplasm. This might also provide partial solution to a long-standing puzzle in SMR research; namely, how these small proteins consisting of roughly 100 amino acid residues can catalyze the coupled translocation of substrate and protons,[3] a feat that is accomplished in other antiporter families, such as the major-facilitator family, by much larger proteins.[51] Thus, inverted topology might be a parsimonious evolutionary solution to the problem of vectorial transport. In this connection, it is interesting to note that two of the five proteins identified as having the dual-topology architecture are from the SMR family (the others have not been fully characterized mechanistically), and a sixth case of dual topology was identified involving two homologous proteins (*YdgE* and *YdgF*), which are also SMR members that are likely to have arisen from a gene-duplication event.[30]

Although much of the biochemical and biophysical data gathered on EmrE are in harmony with the model structure, there are one or two pieces of data that are not in agreement. The topology of the protein is clearly the most important point of disagreement, because dual topology provided the basis for the model structure reported here and for the suggested mechanism of substrate translocation; ultimately, if inverted topology for EmrE is incorrect, then so is the model. We have found that dual topology provides the most satisfactory model for EmrE, but Ninio *et al*.[32] predict, on the basis of labeling data, that the monomers have identical topology, conflicting with other lines of experimental data that suggest inverted topology for EmrE;[30,31] the reasons for this discrepancy among different lines of experimental data are unclear. The possible conflicts of our EmrE model with the cross-linking data[46] have been discussed above. The difficulties inherent in the structural interpretations of cross-linking data on dynamic structures are well known, because even rarely sampled conformations might elicit crosslinks, as was underscored recently in the case of lactose permease.[48]

Despite many years of structural studies of the SMR transporter EmrE, an atomic-resolution structure of this representative protein that can explain much of the biochemical and structural data has not emerged. Here, we have used phylogenetic analysis combined with constraints obtained from a cryo-EM structure of EmrE and some biochemical experiments in order to produce a model structure specifying the approximate positions of individual amino acid residues for EmrE and its homologues. Although this model was constrained only by some biophysical data on EmrE, it is encouraging that the model is capable of accounting for so much of the biochemistry. By revealing the locations of individual amino acid residues in the membrane-spanning regions, the model can be used in order to plan and interpret experiments aimed at deciphering the molecular details of the substrate-translocation mechanism in EmrE and its homologues.

Methods

Sequence data

An initial alignment of a few tens of EmrE homologues was constructed using CLUSTAL W.[52] On the basis of this alignment, we then constructed a hidden Markov model (HMM),[53] which was then calibrated and used to search SWISSPROT and TrEMBL[54] for additional sequence homologues. Sequences showing over 90% identity with other sequences in the set were removed to obtain 98 sequences, which were then aligned†.[52] Conservation scores were then computed for each amino acid position using the ConSeq server and the Rate4Site algorithm.[55,56] The sequence alignment was inspected to identify hydrophobic stretches that correspond to the hydrophobic cores of the helices in forming the TM domain. Starting from the secondary structure assignment derived from NMR,[6] we manually modified the N and C termini of each hydrophobic domain so that the longest stretches of hydrophobic residues would be aligned. The following segments of EmrE were used as the hydrophobic stretches: TM1, 4–21; TM2, 34–52; TM3, 58–80; TM4, 87–104. The conservation scores and the hydrophobic segments are shown in Figure 1(b).

Conformation scoring function

The method for conformational search was as described.[23] In brief, this scoring function favors the burial of evolutionarily conserved amino acid positions in the protein core and the exposure of variable positions to the lipid, without biasing helix orientations according to experimentally derived data. Conformations that expose charged amino acids to the lipid milieu are penalized (in EmrE, this applies only to M1 due to position Glu14). The following scoring function is used to score each conformation:

$$Score = \sum_i (2(B^i - 1/2)(H^i - C^i)) \qquad (1)$$

where B^i quantifies the extent of burial of amino acid residue i in the protein core.[39,57] It assumes values of 0 to 1, with 1 signifying complete burial against another helix, and 0 signifying complete exposure to the lipid or the pore lumen. The function is computed by iterating over all of the helices in the structure other than the one on which i is located, and taking into account distance from, and orientation of i with respect to each of these helices. B^i is then taken as the maximum of the values calculated for each of the helices.[23,39] Thus, high values of B^i imply that i is in close contact with another helix, whereas low values indicate that it is not interacting with any of the helices.

The C^i values are the normalized evolutionary-rate scores assigned by *Rate4Site*.[55,56] High-through-low values of C^i are assigned to variable-through-conserved positions, respectively. H^i is the free energy of transfer from water to lipid of amino acid i according to the Kessel and Ben-Tal scale.[58] H^i values are taken into account only if they are greater than 7 kcal/mol, and only for residues i that are exposed to the membrane, i.e. for which the burial scores B^i are less than 0.5. Thus, the hydrophobicity scale

† The multiple-sequence alignment of SMR proteins can be downloaded from http://ashtoret.tau.ac.il/~sarel/EmrE.html

serves as a significant penalty on the exposure of the most polar residues to the membrane environment.

Conformational search

Canonical C^{α}-trace models of eight α-helices were constructed according to the helix axes parameters derived from helical models that were made to fit the cryo-EM structure, and their geometric centers were placed at the hypothetical membrane midplane. The amino acid identities of positions in the hydrophobic segments M1–M4 were assigned to the relevant positions on these helices.

Each helix was rotated around its principal axis independently, in 5° steps, and its optimal orientation was derived. Then, the optimal orientations of all helices were superimposed to yield the optimal conformation of the entire complex.

Data Base accession number

The cryo-EM structure is available from the EM data bank with accession code 1087‡. The coordinates of the model structure of a dimer of EmrE containing backbone atoms has been deposited in the PDB with accession number 2i68.

Acknowledgements

The authors gratefully acknowledge many comments and discussions with S. Schuldiner regarding biochemical and biophysical data, and their implications, for providing unpublished data. This study was supported by a grant 222/04 from the Israel Science Foundation to N.B.-T. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation. Work by A.E. and D.H. was supported in part by the IST Programme of the EU as shared-cost RTD (FET Open) Project under Contract No IST-006413 (ACS - Algorithms for Complex Shapes), and by the Hermann Minkowski-Minerva Center for Geometry at Tel-Aviv University.

References

1. Nikaido, H. (1994). Prevention of drug access to bacterial targets: permeability barriers and active efflux. *Science*, **264**, 382–388.
2. van Veen, H. W., Higgins, C. F. & Konings, W. N. (2001). Multidrug transport by ATP binding cassette transporters: a proposed two-cylinder engine mechanism. *Res. Microbiol.* **152**, 365–374.
3. Schuldiner, S., Granot, D., Mordoch, S. S., Ninio, S., Rotem, D., Soskin, M. *et al.* (2001). Small is mighty: EmrE, a multidrug transporter as an experimental paradigm. *News Physiol. Sci.* **16**, 13013–13014.
4. Ubarretxena-Belandia, I. & Tate, C. G. (2004). New insights into the structure and oligomeric state of the bacterial multidrug transporter EmrE: an unusual asymmetric homo-dimer. *FEBS Letters*, **564**, 234–238.
5. Arkin, I. T., Russ, W. P., Lebendiker, M. & Schuldiner, S. (1996). Determining the secondary structure and orientation of EmrE, a multi-drug transporter, indicates a transmembrane four-helix bundle. *Biochemistry*, **35**, 7233–7238.
6. Schwaiger, M., Lebendiker, M., Yerushalmi, H., Coles, M., Groger, A., Schwarz, C. *et al.* (1998). NMR investigation of the multidrug transporter EmrE, an integral membrane protein. *Eur. J. Biochem.* **254**, 610–619.
7. Ubarretxena-Belandia, I., Baldwin, J. M., Schuldiner, S. & Tate, C. G. (2003). Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer. *EMBO J.* **22**, 6175–6181.
8. Tate, C. G., Ubarretxena-Belandia, I. & Baldwin, J. M. (2003). Conformational changes in the multidrug transporter EmrE associated with substrate binding. *J. Mol. Biol.* **332**, 229–242.
9. Tate, C. G., Kunji, E. R., Lebendiker, M., Schuldiner, S. & Yerushalmi, H. (2001). The projection structure of EmrE, a proton-linked multidrug transporter from *Escherichia coli*, at 7 Å resolution. *EMBO J.* **20**, 77–81.
10. Ma, C. & Chang, G. (2004). Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **101**, 2852–2857.
11. Butler, P. J., Ubarretxena-Belandia, I., Warne, T. & Tate, C. G. (2004). The *Escherichia coli* multidrug transporter EmrE is a dimer in the detergent-solubilised state. *J. Mol. Biol.* **340**, 797–808.
12. Pornillos, O., Chen, Y. J., Chen, A. P. & Chang, G. (2005). X-ray structure of the EmrE multidrug transporter in complex with a substrate. *Science*, **310**, 1950–1953.
13. Tate, C. G. (2006). Comparison of three structures of the multidrug transporter EmrE. *Curr. Opin. Struct. Biol.* **16**, 457–464.
14. Muth, T. R. & Schuldiner, S. (2000). A membrane-embedded glutamate is required for ligand binding to the multidrug transporter EmrE. *EMBO J.* **19**, 234–240.
15. Grinius, L. L. & Goldberg, E. B. (1994). Bacterial multidrug resistance is due to a single membrane protein which functions as a drug pump. *J. Biol. Chem.* **269**, 29998–30004.
16. Gutman, N., Steiner-Mordoch, S. & Schuldiner, S. (2003). An amino acid cluster around the essential Glu-14 is part of the substrate- and proton-binding domain of EmrE, a multidrug transporter from *Escherichia coli*. *J. Biol. Chem.* **278**, 16082–16087.
17. Soskine, M., Adam, Y. & Schuldiner, S. (2004). Direct evidence for substrate-induced proton release in detergent-solubilized EmrE, a multidrug transporter. *J. Biol. Chem.* **279**, 9951–9955.
18. Sharoni, M., Steiner-Mordoch, S. & Schuldiner, S. (2005). Exploring the binding domain of EmrE, the smallest multidrug transporter. *J. Biol. Chem.* **280**, 32849–32855.
19. Weinglass, A. B., Soskine, M., Vazquez-Ibar, J. L., Whitelegge, J. P., Faull, K. F., Kaback, H. R. & Schuldiner, S. (2005). Exploring the role of a unique carboxyl residue in EmrE by mass spectrometry. *J. Biol. Chem.* **280**, 7487–7492.
20. Koteiche, H. A., Reeves, M. D. & McHaourab, H. S. (2003). Structure of the substrate binding pocket of the multidrug transporter EmrE: site-directed spin

‡ http://www.ebi.ac.uk/msd/index.html

labeling of transmembrane segment 1. *Biochemistry*, **42**, 6099–6105.
21. Fleishman, S. J., Unger, V. M. & Ben-Tal, N. (2006). Transmembrane protein structures without X-rays. *Trends Biochem. Sci.* **31**, 106–113.
22. Fleishman, S. J., Unger, V. M., Yeager, M. & Ben-Tal, N. (2004). A C-alpha model for the transmembrane alpha-helices of gap-junction intercellular channels. *Mol. Cell*, **15**, 879–888.
23. Fleishman, S. J., Harrington, S., Friesner, R. A., Honig, B. & Ben-Tal, N. (2004). An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data. *Biophys. J.* **87**, 3448–3459.
24. Baldwin, J. M., Schertler, G. F. & Unger, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* **272**, 144–164.
25. Beuming, T. & Weinstein, H. (2005). Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT. *Protein Eng. Des. Sel.* **18**, 119–125.
26. Briggs, J. A., Torres, J. & Arkin, I. T. (2001). A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins: Struct. Funct. Genet.* **44**, 370–375.
27. Adamian, L. & Liang, J. (2006). Prediction of buried helices in multispan alpha helical membrane proteins. *Proteins: Struct. Funct. Genet.* **63**, 1–5.
28. Hurwitz, N., Pellegrini-Calace, M. & Jones, D. T. (2006). Towards genome-scale structure prediction for transmembrane proteins. *Phil. Trans. Roy. Soc. ser. B*, **361**, 465–475.
29. Fleishman, S. J. & Ben-Tal, N. (2006). Progress in structure prediction of alpha-helical membrane proteins. *Curr. Opin. Struct. Biol.* **16**, 496–504.
30. Daley, D. O., Rapp, M., Granseth, E., Melen, K., Drew, D. & von Heijne, G. (2005). Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science*, **308**, 1321–1323.
31. Rapp, M., Granseth, E., Seppala, S., von Heijne, G., Daley, D. O., Melen, K. & Drew, D. (2006). Identification and evolution of dual-topology membrane proteins. *Nature Struct. Mol. Biol.* **13**, 112–116.
32. Ninio, S., Elbaz, Y. & Schuldiner, S. (2004). The membrane topology of EmrE—a small multidrug transporter from *Escherichia coli*. *FEBS Letters*, **562**, 193–196.
33. Fu, D., Libson, A., Miercke, L. J., Weitzman, C., Nollert, P., Krucinski, J. & Stroud, R. M. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **290**, 481–486.
34. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. (2002). X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–294.
35. Van den Berg, B., Clemons, W. M., Jr, Collinson, I., Modis, Y., Hartmann, E., Harrison, S. C. & Rapoport, T. A. (2004). X-ray structure of a protein-conducting channel. *Nature*, **427**, 36–44.
36. Baldwin, J. M. (1993). The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* **12**, 1693–1703.
37. Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P. & Bowie, J. U. (2004). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl Acad. Sci. USA*, **101**, 959–963.
38. Creighton, T. E. (1993). *Proteins, Structures and Molecular Properties.* Freeman, New York.
39. Fleishman, S. J. & Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.* **321**, 363–378.
40. Ninio, S. & Schuldiner, S. (2003). Characterization of an archaeal multidrug transporter with a unique amino acid composition. *J. Biol. Chem.* **278**, 12000–12005.
41. Gottschalk, K. E., Soskine, M., Schuldiner, S. & Kessler, H. (2004). A structural model of EmrE, a multi-drug transporter from *Escherichia coli*. *Biophys. J.* **86**, 3335–3348.
42. Zheleznova, E. E., Markham, P. N., Neyfakh, A. A. & Brennan, R. G. (1999). Structural basis of multidrug recognition by BmrR, a transcription activator of a multidrug transporter. *Cell*, **96**, 353–362.
43. Godsey, M. H., Zheleznova Heldwein, E. E. & Brennan, R. G. (2002). Structural biology of bacterial multidrug resistance gene regulators. *J. Biol. Chem.* **277**, 40169–40172.
44. Lewinson, O. & Bibi, E. (2001). Evidence for simultaneous binding of dissimilar substrates by the *Escherichia coli* multidrug transporter MdfA. *Biochemistry*, **40**, 12612–12618.
45. Mordoch, S. S., Granot, D., Lebendiker, M. & Schuldiner, S. (1999). Scanning cysteine accessibility of EmrE, an H+-coupled multidrug transporter from *Escherichia coli*, reveals a hydrophobic pathway for solutes. *J. Biol. Chem.* **274**, 19480–19486.
46. Soskine, M., Steiner-Mordoch, S. & Schuldiner, S. (2002). Crosslinking of membrane-embedded cysteines reveals contact points in the EmrE oligomer. *Proc. Natl Acad. Sci. USA*, **99**, 12043–12048.
47. Paulsen, I. T., Brown, M. H., Dunstan, S. J. & Skurray, R. A. (1995). Molecular characterization of the staphylococcal multidrug resistance export protein QacC. *J. Bacteriol.* **177**, 2827–2833.
48. Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R. & Iwata, S. (2003). Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**, 610–615.
49. Rath, A., Melnyk, R. A. & Deber, C. M. (2006). Evidence for assembly of small multidrug resistance proteins by a "two-faced" transmembrane helix. *J. Biol. Chem*, **281**, 15546–15553.
50. Curran, A. R. & Engelman, D. M. (2003). Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.* **13**, 412–417.
51. Kaback, H. R., Sahin-Toth, M. & Weinglass, A. B. (2001). The kamikaze approach to membrane transport. *Nature Rev. Mol. Cell. Biol.* **2**, 610–620.
52. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
53. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
54. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
55. Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, R. *et al.* (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, **20**, 1322–1324.

56. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
57. Fleishman, S. J., Schlessinger, J. & Ben-Tal, N. (2002). A putative activation switch in the transmembrane domain of erbB2. *Proc. Natl Acad. Sci. USA*, **99**, 15937–15940.
58. Kessel, A. & Ben-Tal, N. (2002). Free energy determinants of peptide association with lipid bilayers. In *Current Topics in Membranes* (Simon, S. & McIntosh, T., eds), vol. 52, pp. 205–253, Academic Press, San Diego.
59. Elbaz, Y., Tayer, N., Steinfels, E., Steiner-Mordoch, S. & Schuldiner, S. (2005). Substrate-induced tryptophan fluorescence changes in EmrE, the smallest ion-coupled multidrug transporter. *Biochemistry*, **44**, 7369–7377.
60. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
61. Merritt, E. A. & Bacon, D. J. (1997). Raster3D photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524.

Edited by J. Bowie

(Received 19 July 2006; received in revised form 25 August 2006; accepted 25 August 2006)
Available online 30 August 2006

Opinion TRENDS in Biochemical Sciences Vol.xx No.xx Monthxxxx

Has the code for protein translocation been broken?

Dalit Shental-Bechor*, Sarel J. Fleishman* and Nir Ben-Tal

Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University, 69978 Ramat Aviv, Israel

Polypeptides chains are segregated by the translocon channel into secreted or membrane-inserted proteins. Recent reports claim that an *in vivo* system has been used to break the 'amino acid code' used by translocons to make the determination of protein type (i.e. secreted or membrane-inserted). However, the experimental setup used in these studies could have confused the derivation of this code, in particular for polar amino acids. These residues are likely to undergo stabilizing interactions with other protein components in the experiment, shielding them from direct contact with the inhospitable membrane. Hence, it is our view that the 'code' for protein translocation has not yet been deciphered and that further experiments are required for teasing apart the various energetic factors contributing to protein translocation.

Introduction

Co-translational translocation is the process by which ribosomes that are attached to the endoplasmic reticulum (ER) extrude proteins through the translocon channel, giving rise to two different classes of proteins: those that are secreted or inserted into cellular membranes [1]. This crucial classification process is conducted on the basis of the sequence of the translated protein, which has led to the expectation that a 'sequence code' exists. If identified, this 'code' could be used to explain and predict which proteins would eventually reside within the membrane and which would be secreted into the ER lumen (and subsequently transported to various cellular compartments or expelled from the cell) [2]. This assumption was the foundation for several hydrophobicity scales (a ranking of the 20 amino acids according to their polarity), which were computed either from physical principles or from experiments that quantitatively compare the equilibrium distribution of amino acid residues in hydrophobic and hydrophilic media [3–8]. Such scales have been extremely useful, and have remained the principal means for identifying transmembrane (TM) segments in protein sequences for more than two decades [9].

An *in vivo* system for probing the energetics of translocation

Recently, Hessa *et al.* [10] carried out a series of experiments designed to decipher, for the first time, the translocation sequence code using an *in vivo* system containing ER membranes, ribosomes and the translocon channel – an approach that is far more realistic than the simple model systems previously employed. The experimental procedure is based on the use of an artificially designed variant of the leader peptidase (Lep) protein from *Escherichia coli*. This protein includes two endogenous TM helices (TM1 and TM2) and a soluble domain (P2) (Figure 1a). Using Lep as a host, an additional sequence segment (termed H) was engineered as a probe downstream of TM2 so that the equilibrium concentrations of the inserted versus the translocated H could be measured *in vitro* [10]. The procedure is attractive because of the clarity of the experimental readout, in addition to its simplicity, even though it addresses a highly complicated physiological system.

Hessa *et al.* [10] proceeded to read the 'sequence code' by translating the equilibrium concentrations of secreted and inserted Lep that contained various H probe segments into free-energy differences between the two states, inserted and translocated (Figure 1a). However, this treatment implicitly postulates two crucial, albeit unproven, thermodynamic assumptions: (i) that the H segment forms the same secondary structure, presumably an α-helix, in both the translocated and inserted states; and (ii) that the H segment is isolated from other protein components and contacts only the lipid molecules. Deviations from helicity or association with other protein components would mean that more thermodynamic states would need to be considered and that energetic contributions other than direct peptide–membrane interactions were involved, thus confounding the derivation of a hydrophobicity scale (Box 1).

In our opinion, the experimental setup used by Hessa *et al.* [10] cannot discriminate between the effects of the interactions of H with the membrane (hydrophobicity) and with TM1 and TM2, which are specific for the Lep host. Rather than the two thermodynamic states suggested by Hessa *et al.* [10] (Figure 1a), we believe that the H segment resides in at least five different states (Figure 1b), three corresponding to the classification of membrane-inserted [Figure 1b(i–iii)] and two to that of translocated [Figure 1b(iv–v)]. The inserted states differ from one another in the extent to which they expose the sidechains or backbone of their H segments to the lipid milieu. In the conformation in which the H segment is separate from the remainder of the TM domain [Figure 1b(i)], all of its sidechains are exposed to the lipid and its backbone is

Corresponding author: Ben-Tal, N. (nirb@tauex.tau.ac.il).
* Authors contributed equally.

 0968-0004/$ - see front matter doi:10.1016/j.tibs.2006.02.002

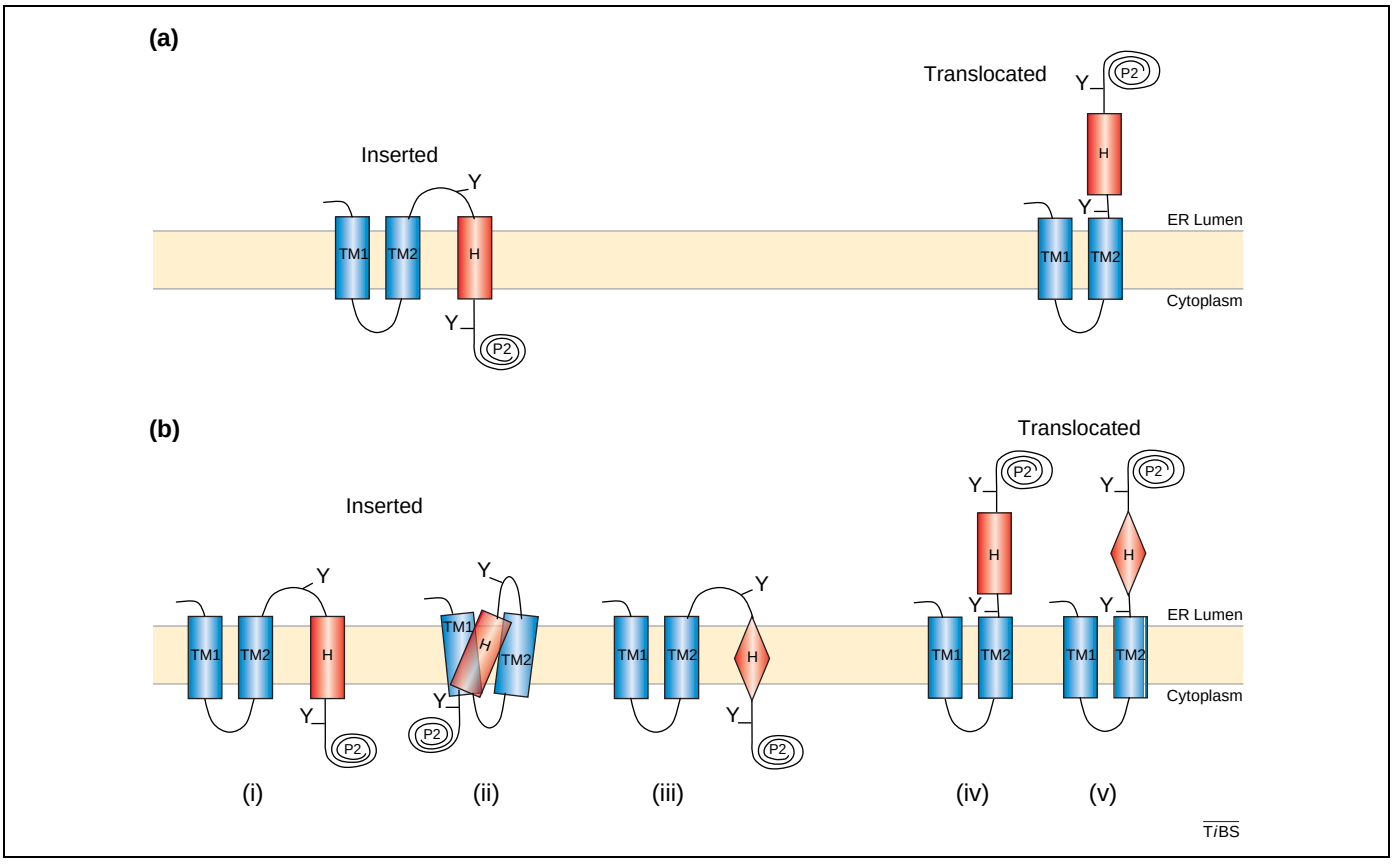


Figure 1. Schematic representation of the thermodynamic states associated with the *in vivo* system for probing the energetics of transloca tion. Hessa *et al.* [10] modified the leader peptidase (Lep) protein to include a probe H segment. Thus, the protein included two endogenous TM domains (TM1 and TM2), an extramembrane domain (P2) and the engineered H segment with two glycosylation sites on each end (Y). Glycosylation takes place only in the luminal side of the membrane, such that the inserted and translocated states can be differentiated from each other by the number of glycosylations that took place. **(a)** Following the suggestion of Hessa et al. [10], the translocation process can be described as a chemical equilibrium between two states – inserted and translocated. **(b)** Inserted Lep might assume a conformation in which the H segment is isolated from the remainder of the TM domain (i). However, other conformations, in which H is packed against TM1 and TM2 and interacts specifically with their sidechains and backbones (ii) or one in which H deviates from α-helicity [red rectangle in (i) versus diamond in (iii)] are also feasible. In the studies by Hessa *et al.* [10,16], it is impossible to distinguish between these three different thermodynamic states because all of them would be denoted as 'inserted Lep'. Similar confusion will arise between the two translocated states [(iv) and (v)], which differ from one another in the conformation of H.

maintained in an α-helical conformation. This is the only inserted state postulated by Hessa *et al.* [10]. But conformations in which H interacts with TM1 and TM2 shielding some of the sidechains of H from lipid [Figure 1b(ii)] or in which H deviates from α-helicity [Figure 1b(iii)], are also possible. Similarly, Hessa *et al.* [10] acknowledge the conformation in which an α-helical H segment is translocated [Figure 1b(iv)], but they ignore the fact that a non-helical conformation of the translocated H is also likely [Figure 1b(v)].

We perceive that the main flaw in the interpretation of experimental results by Hessa *et al.* [10] is that it is unlikely to apply to other TM proteins because the conformation in which H interacts directly with TM1 and TM2 [Figure 1b(ii)], which was not considered by Hessa *et al.*, is specific for Lep; therefore, the generality of the scale derived by Hessa *et al.* [10] requires substantiation. For example, among the inserted states, the conformation for membrane-inserted H considered by Hessa *et al.* [Figure 1b(i)] is likely to dominate in Lep variants with hydrophobic H segments because the interactions with the hydrophobic lipids are probably at least as favourable as the interactions with protein for such segments. However, as the polarity of H increases,

Box 1. The thermodynamic interpretation of the experimental results of Hessa *et al.*

Hessa *et al.* [10] engineered a probe H segment flanked by two *N*-linked glycosylation sites (see Figure 1 in the main text). Glycosylation took place only on the luminal side of the microsomes that were generated in the experiment, so the extent of Lep glycosylation is indicative of the state of the H segment. That is, inserted and translocated H are associated with singly and doubly glycosylated Lep, respectively. The proportion of singly (f_{1g}) and doubly (f_{2g}) glycosylated Lep were measured *in vitro* using SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) gels, and an 'apparent equilibrium constant' was assigned to their ratio according to equation 1.

$$K_{app} = \frac{f_{1g}}{f_{2g}} \qquad \text{(Eqn 1)}$$

The results were represented by converting K_{app} into apparent free energy according to the conventional thermodynamic definition: $\Delta G_{app} = -RT\ln K_{app}$, where R is the gas constant, T is the absolute temperature and ln is the natural logarithm. We contend that, in these experiments, H can be in an ensemble of at least five rather than two states (see Figure 1b in the main text) and that, therefore, it is overly simplistic to describe the equilibrium between inserted and translocated H using the ΔG_{app} formula used by Hessa *et al.* [10].

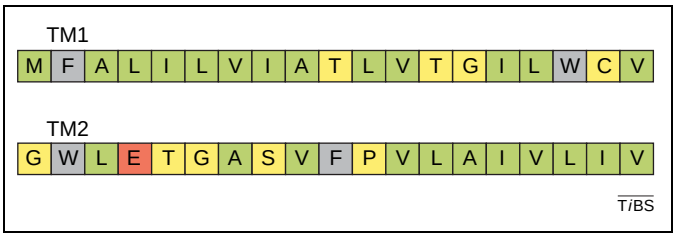


Figure 2. Amino acid sequences of the two transmembrane spans (TM1 and TM2) of Lep (SWISSPROT entry P00803). These segments contain several polar residues, which could form contacts with other polar residues on the probe H segment, shielding them from the hydrophobic membrane environment. Because the interactions of H with TM1 and TM2 are likely to be specific for Lep, they would alter its insertion propensity in a way that does not solely reveal the interactions of H with membrane and would, thus, limit the generality of the hydrophobicity scale derived by Hessa *et al.* [10]. Grey, aromatic residues; green, hydrophobic residues; yellow, small polar residues; red, negatively charged glutamic acid.

the conformation in which H interacts with TM1 and TM2 [Figure 1b(ii)] is likely to become the more populated state because of the known tendency of highly polar or charged residues (e.g. asparagine and glutamic acid) to drive the packing of their host helix against the polar backbone and sidechains of other TM helices [5,11–15]. The TM1 and TM2 segments of Lep contain one titratable (glutamic acid) and several residues that are small and polar (e.g. glycine and serine) (Figure 2), which would enable the other polar residues on the H segment to be shielded from the inhospitable membrane environment. If this conformation dominates [Figure 1b(ii)], then changing the position of the polar residues in the H segment would alter the stability of the protein in the membrane as a result of the interactions between H and the endogenous TM1 and TM2 of Lep. Indeed, the results reported by Hessa *et al.* [10,16] revealed that such positional dependence is observed for H segments that contain highly polar or charged residues. Moreover, the simple additivity of contributions to stability observed for apolar and mildly polar residues (e.g. leucine and serine, respectively) breaks down with the introduction of highly polar residues. One way to explain positional dependence and deviations from additivity, which is not refuted by Hessa *et al.*, is that the polar helices introduced into Lep as H segments formed stabilizing interactions with the endogenous TM helices of Lep. The expectation that the conformation in which H interacts with TM1 and TM2 [Figure 1b(ii)] dominates in H segments that contain charged residues might also explain the low values of ΔG_{app} penalties obtained by Hessa *et al.* [10] (Box 1) for the transfer of such residues from translocated to inserted states relative to other hydrophobicity scales [5,6,8]. Furthermore, although the original Lep protein is monomeric in membranes [17], it is not clear whether the Lep variants containing the more polar H segments oligomerize. If so, the equilibrium between inserted and secreted Lep would comprise many more states than the five suggested here (Figure 1b).

Membrane insertion of the S4 segment of voltage-gated K^+ channels

The limitations described here apply equally to the use of engineered Lep for studying the partitioning of natural TM segments. For instance, the same experimental

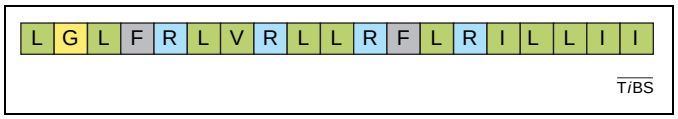


Figure 3. Amino acid sequence of the S4 segment of the voltage sensor of the KvAP channel. The sequence of S4 is mostly hydrophobic but four arginines (blue), at least three of which are charged, occupy conserved positions in the sequence. These charges could interact favourably with polar groups on TM1 and TM2 in a conformation schematically represented in Figure 1b(ii). Such interactions would stabilize the inserted conformation of S4, which would therefore not be exposed completely to membrane. Grey, aromatic residues; green, hydrophobic residues; yellow, small polar residues; blue, positively charged arginine residues.

framework has been used to study the free energy of transfer between translocated and inserted states of the S4 segment [16], which constitutes the core voltage-sensing element in the voltage-gated K^+ channel from *Aeropyrum pernix* (KvAP) [18]. This 19-residue segment is a hydrophobic cation consisting mostly of highly hydrophobic residues interspersed with four arginines at conserved positions (Figure 3), of which at least three are charged [19]. Until recently, it was anticipated – mainly on energetic grounds – that the S4 segment is packed against the other TM helices of the K^+ channel because of its high polarity [20]. By contrast, the first structure of the KvAP channel, which was crystallized in the absence of lipid, showed S4 to be exposed to the membrane [21]. This finding elicited considerable controversy. Hessa *et al.* [10] found that the net polarity of S4 was at the threshold that would enable its efficient insertion into the membrane, supporting the view that it is exposed to lipid.

To provide a thermodynamic explanation to the conclusion drawn by Hessa *et al.* [10] that isolated S4 could insert efficiently into membranes, Freites *et al.* [22] conducted molecular-dynamics simulations of an inserted α-helical segment bearing the S4 sequence, which was placed in a TM orientation and surrounded by lipid on all sides in isolation from other protein components. On the basis of these simulations, they suggested that the S4 segment is stabilized in a TM orientation despite its high polarity owing to contacts formed between the arginine sidechains on S4 and the phosphate headgroups and water molecules. These contacts can form according to the simulations because, in the immediate vicinity of S4, the thickness of the hydrocarbon core of the membrane shrinks from a steady-state width of ~30 Å [3] (Figure 4) to a mere 10 Å, which is considerably thinner than a lipid monolayer (Figure 4). In particular, Freites *et al.* [22] noted that one of the lipids assumes a conformation that spans the entire membrane in the region of S4, further demonstrating the enormous distortion of the membrane in these simulations. We note that a conservative estimate of the energetic penalty of such a large contraction of the membrane lipids that considers only the effects of dihedral-angle strains would be 12 kcal mol^{-1} [5,23]. Furthermore, a snapshot provided by Freites *et al.* [22] reveals that, for some of the lipids, the distortion is so large that their aliphatic chains form contacts with water molecules. Taken together, the strains to the aliphatic chains and the solvation penalty on direct contacts between polar and aliphatic groups that were observed

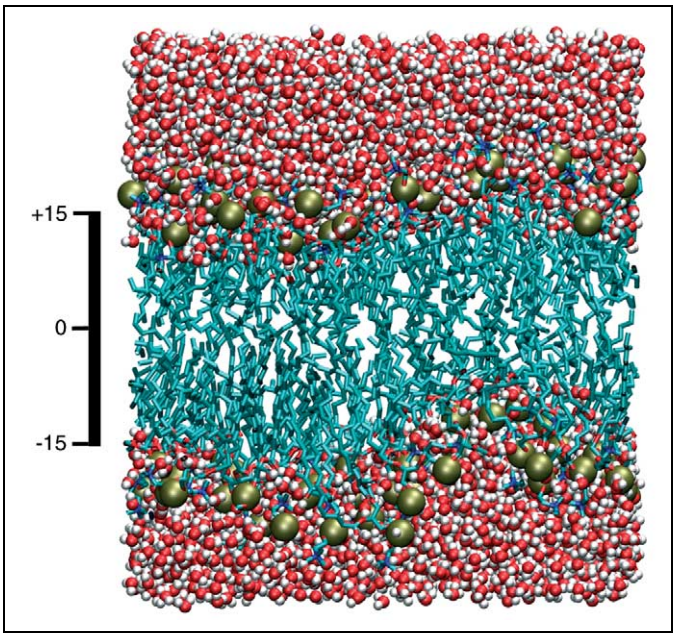


Figure 4. An equilibrated membrane bilayer composed of dimyristoylphosphocholine molecules embedded in water. Cyan, aliphatic chains; gold, phosphates in the headgroups; red, oxygen atoms in water. The bar shows the approximate span of the aliphatic chains in Ångstroms, with an approximation of the membrane midplane marked by 0. The hydrophobic core of each leaflet of the bilayer spans 15 Å. The membrane hydrophobic core fluctuates by a few Ångstroms around an equilibrium width of 30 Å.

in these simulations make membrane contraction an unlikely explanation for efficient insertion of S4.

The suggestion that a TM orientation for S4 is stabilized by such large-scale membrane distortion is made even more implausible in view of biophysical studies of model TM segments [5,24]. The experimental data collected on many different peptides show that the membrane width might decrease by several Ångstroms to better match the hydrophobic length of a peptide, but that peptides with hydrophobic lengths that are considerably shorter than the width of the hydrocarbon region of the bilayer (as in S4) do not partition into the membrane but, instead, reside on the membrane surface. Although S4 was not directly targeted by these experiments, the data provide an indication that the membrane would not undergo contraction of 20 Å (as suggested by Freites *et al.* [22]) to stabilize the inserted conformation of S4. Similar to our criticism of the results on the membrane insertion of polar residues [10], a more likely explanation for the observed tendency of S4 towards membrane insertion in the experimental setup of Hessa *et al.* is that S4 forms stabilizing contacts with the endogenous TM1 and TM2 of Lep, and that lipid contraction has a much smaller role than suggested by the simulations of Freites *et al.* [22]. In this respect, it is notable that a more recent structure of the voltage-gated K^+ channel, which was crystallized in the presence of lipids and is therefore considered a more faithful representation of the physiological structure than the previous crystal structure, showed that two of the four arginines in S4 are buried at a helix–helix interface where they are partially shielded from the inhospitable lipid environment [25].

Concluding remarks

We have raised several points that question the validity of deriving thermodynamic quantities for the interactions between amino acids and peptides with membranes using the *in vivo* system of Hessa *et al.* [10]. The key problems are the lack of experimental controls for the α-helicity of the H segment and, most importantly, whether H associates directly with membranes without forming stabilizing contacts with other protein components in the system. To address these issues, it still needs to be shown that: (i) physical contacts are not formed between H and the two endogenous TM segments of Lep and that engineered Lep does not oligomerize (this could be achieved using fluorescence-labelling techniques [26], for example); and (ii) that the H segment retains its secondary structure (presumably an α-helix) in both the inserted and the translocated states for substitutions of all 20 amino acids. The generality of the scale derived from the results of the study [10] could also be validated by using hosts other than the Lep protein to observe that the same thermodynamic quantities are obtained. The partitioning of isolated S4 segment from KvAP could also be monitored using solid-state NMR [24].

It should be appreciated that Hessa *et al.* [10] have focused on a highly complicated physiological system comprising a plethora of different proteins, which is subject to a complex environment that includes water, protein and membrane. The experimental setup devised by these authors certainly represents a major step towards probing the energetics of protein translocation within a physiologically relevant framework, and could prove useful in future studies of protein–protein interactions within the membrane. However, the readout from this system probably reflects a mixture of hydrophobicity and various contributions stemming from interactions with the host protein Lep, especially for the more polar segments tested. Building on this setup, more experiments will be needed before it can be safely concluded that 'the fundamental code used by the translocon to select polypeptide segments for insertion as TM helices has been broken' [27].

Acknowledgements

We thank I.T. Arkin for kindly providing Figure 4. We also thank I.T. Arkin, J.U. Bowie, J.A. Hirsch, Y. Ofran and O. Yifrach for critical reading of this article. This study was supported by grant 222/04 from the Israel Science Foundation. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation.

References

1 Osborne, A.R. *et al.* (2005) protein translocation by the Sec61/SecY channel. *Annu. Rev. Cell Dev. Biol.* 21, 529–550
2 Bowie, J.U. (2005) Cell biology: border crossing. *Nature* 433, 367–369
3 White, S.H. and Wimley, W.C. (1999) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* 28, 319–365
4 Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132
5 Kessel, A. and Ben-Tal, N. (2002) Free energy determinants of peptide association with lipid bilayers. In *Current topics in membranes* (Vol. 52) (Simon, S. and McIntosh, T., eds), pp. 205–253, Academic Press

6 Engelman, D.M. *et al*. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 15, 321–353

7 Eisenberg, D. *et al*. (1986) Hydrophobicity and amphiphilicity in protein structure. *J. Cell. Biochem.* 31, 11–17

8 von Heijne, G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 225, 487–494

9 Chen, C.P. *et al*. (2002) Transmembrane helix predictions revisited. *Protein Sci.* 11, 2774–2791

10 Hessa, T. *et al*. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433, 377–381

11 Stevens, T.J. and Arkin, I.T. (1999) Are membrane proteins 'inside-out' proteins? *Proteins* 36, 135–143

12 Fleishman, S.J. and Ben-Tal, N. (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices. *J. Mol. Biol.* 321, 363–378

13 Faham, S. *et al*. (2004) Side-chain contributions to membrane protein structure and stability. *J. Mol. Biol.* 335, 297–305

14 Curran, A.R. and Engelman, D.M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.* 13, 412–417

15 Sternberg, M.J. and Gullick, W.J. (1989) Neu receptor dimerization. *Nature* 339, 587

16 Hessa, T. *et al*. (2005) Membrane insertion of a potassium-channel voltage sensor. *Science* 307, 1427

17 Zwizinski, C. *et al*. (1981) Leader peptidase is found in both the inner and outer membranes of *Escherichia coli*. *J. Biol. Chem.* 256, 3593–3597

18 Ruta, V. *et al*. (2003) Functional analysis of an archaebacterial voltage-dependent K^+ channel. *Nature* 422, 180–185

19 Schoppa, N.E. *et al*. (1992) The size of gating charge in wild-type and mutant Shaker potassium channels. *Science* 255, 1712–1715

20 Miller, C. (2003) A charged view of voltage-gated ion channels. *Nat. Struct. Biol.* 10, 422–424

21 Jiang, Y. *et al*. (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature* 423, 33–41

22 Freites, J.A. *et al*. (2005) Interface connections of a transmembrane voltage sensor. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15059–15064

23 Fattal, D.R. and Ben-Shaul, A. (1993) A molecular model for lipid-protein interaction in membranes: the role of hydrophobic mismatch. *Biophys. J.* 65, 1795–1809

24 Bechinger, B. (2000) Understanding peptide interactions with the lipid bilayer: a guide to membrane protein engineering. *Curr. Opin. Chem. Biol.* 4, 639–644

25 Long, S.B. *et al*. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science* 309, 897–903

26 Wallrabe, H. and Periasamy, A. (2005) Imaging protein molecules using FRET and FLIM microscopy. *Curr. Opin. Biotechnol.* 16, 19–27

27 White, S.H. and von Heijne, G. (2005) Transmembrane helices before, during, and after insertion. *Curr. Opin. Struct. Biol.* 15, 378–386

Review TRENDS in Biochemical Sciences Vol.xx No.xx Monthxxxx

Transmembrane protein structures without X-rays

Sarel J. Fleishman[1], Vinzenz M. Unger[2] and Nir Ben-Tal[1]

[1]Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel
[2]Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208024, 333 Cedar Street, New Haven, CT 06520-8024, USA

Transmembrane (TM) proteins constitute 15–30% of the genome, but <1% of the structures in the Protein Data Bank. This discrepancy is disturbing, and emphasizes that structure determination of TM proteins remains challenging. The challenge is greatest for proteins from eukaryotes, the structures of which remain intractable despite tremendous advances that have been made towards structure determination of bacterial TM proteins. Notably, >50% of the membrane protein families in eukaryotes lack bacterial homologs. Therefore, it is conceivable that many more years will elapse before high-resolution structures of eukaryotic TM proteins emerge. Until then, integrated approaches that combine biochemical and computational analyses with low-resolution structures are likely to have increasingly important roles in providing frameworks for the mechanistic understanding of membrane-protein structure and function.

Introduction

It is estimated that transmembrane (TM) proteins constitute ~15–30% of eukaryotic genomes [1–4]. Owing to their strategic localization at the interfaces between the interior and exterior of the cell and between cellular compartments, membrane proteins have pivotal roles in many cellular processes, including cell-to-cell signaling events, solute transport and cellular organization. For this reason, membrane proteins are by far the most attractive targets for drug discovery. Despite their importance, however, only a few distinct folds of TM proteins have been solved to date by high-resolution methods such as X-ray crystallography [5] and nuclear magnetic resonance (NMR) [6]; therefore, TM protein structures constitute <1% of the entries in the Protein Data Bank (PDB). Disturbingly, only two of the current entries represent a membrane protein from human origin [7,8], whereas the majority of entries are of bacterial membrane proteins (Figure 1).

Part of the reason why progress has been faster for bacterial membrane proteins stems from the fact that they can more easily be expressed in large quantities in bacterial hosts, and that they lack many of the post-translational modifications that potentially complicate crystallization. Moreover, the fast pace at which bacterial genomes are sequenced provides an almost unlimited repertoire of target proteins including homologs from thermophilic bacteria that are often more stable during detergent solubilization, purification and crystallization. By contrast, eukaryotic membrane proteins are more difficult to express than their bacterial homologs, are subject to post-translational modifications and, often, only few candidate genes are available for screens to identify the ideal target protein. It thus comes as no surprise that, over the past few years, efforts have been focused on identifying bacterial homologs of eukaryotic membrane proteins, and pursuing their structure determination by 'brute-force' approaches, sometimes using thousands of combinations of homologs of the protein and different crystallization conditions [9]. This strategy has begun to bear fruit (Figure 1) and, indeed, the recent growth in novel TM-protein structures was estimated to be exponential, as it is for soluble proteins, suggesting that, over the next few years, many new structures will emerge [5]. However, this growth has not been steady over the years and, more importantly, has been restricted mostly to TM proteins from bacteria; the pace of discovery of novel TM proteins from eukaryotes, however, has remained low (Figure 1). Notably, the use of bacterial homologs for eukaryotic TM proteins does not represent the ultimate solution because many eukaryotic membrane proteins do not have bacterial homologs. In fact, a search in the Pfam-A database of protein families [10] shows that only 47% of the eukaryotic TM protein families have bacterial or archaeal homologs.

In an attempt to overcome the problem of there being such a large proportion of eukaryotic proteins for which direct structure determination is likely to have to wait many years, data-based modeling approaches were developed that rely on inferences derived from biochemical, computational, evolutionary and intermediate-resolution structural methods. Here, we focus on the methods that have been used to model helical membrane proteins before their experimental structure determination at high resolution. Notably, helical proteins are the dominant class of TM proteins in eukaryotes and in bacterial inner membranes. We also delineate potentially productive venues for future research. We will not deal with comparative or homology modeling applied to TM proteins (but see recent reviews [11,12]).

Corresponding author: Ben-Tal, N. (nirb@tauex.tau.ac.il).

 0968-0004/$ - see front matter doi:10.1016/j.tibs.2005.12.005

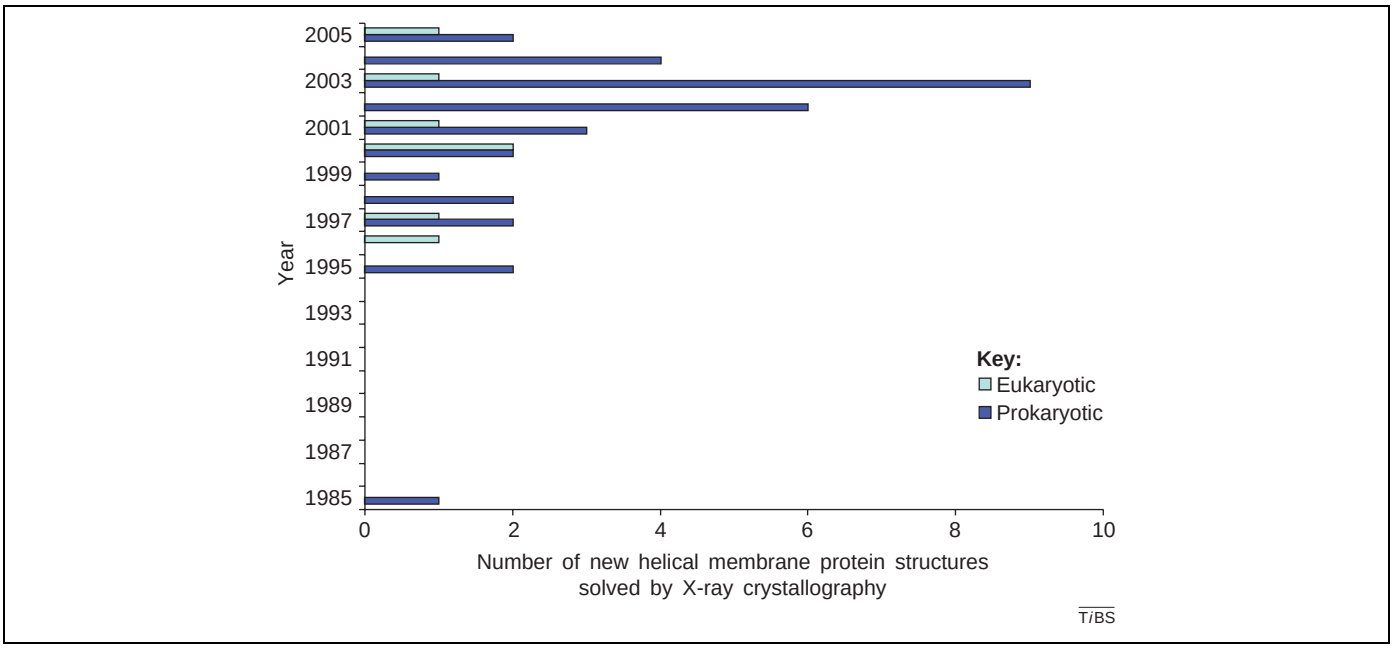


Figure 1. Number of new helical membrane-protein folds solved in recent years. Tremendous progress has been made over the past few years in crystallization of TM proteins from bacteria, although the growth in the discovery of novel structures has not been steady. Moreover, crystallization of eukaryotic TM proteins still lags far behind, and only a handful of structures have been obtained. The entry for 2005 includes structures up to and including November 2005.

Architecture of helical TM proteins

A simple rule that has guided many of the approaches to modeling helical TM proteins is the two-stage model of folding [13]. According to this model, hydrophobic segments are first inserted into the plasma membrane in the form of helices, which engage the polar carbonyl and amide groups on the backbone of the peptide chain through hydrogen bonds, and shield them from the hydrophobic lipid bilayer. Next, these helices associate with one another to shape the tertiary structure of the protein. One of the implications of the two-stage model for computational modeling is that each of the hydrophobic segments comprising the TM domain can be approximated as an energetically stable canonical α helix, the polar backbone and N and C termini of which are shielded from the membrane environment. Hence, TM-protein-structure prediction can concentrate on the relative configurations of preformed α helices. This constraint considerably reduces the number of degrees of freedom that must be explored computationally.

This quite simple picture of TM-protein architecture was supported by the first few membrane proteins to be solved [14–17] (e.g. that shown in Figure 2). Moreover, the extramembrane loops are short in these proteins, dictating that consecutive domains in the sequence are proximal in the 3D structure [18]. However, this simplistic picture collapsed when the first ion-channel structures revealed that helices need not span the entire width of the bilayer [19], and can be extremely long and highly tilted with respect to the membrane normal [20] (Figure 3a,b). Recent transporter structures have also shown marked deviations from α helicity; it has been suggested that these deviations have a role in the conformational changes underlying transporter functions by destabilizing the structures [21] (Figure 3c). All of these structural features are still beyond what can be reliably predicted by computational methods, raising the question of how many membrane domains might have gone unnoticed by contemporary methods for the detection of TM spans [22]. More importantly, however, the observation that not all consecutive hydrophobic domains form physical contacts [19,20] heralded the end of naïve modeling of TM proteins, and underscored the importance of a joint experimental–computational approach to structure prediction. Over the past several years, two sources of experimental data have proven valuable in aiding most modeling exercises of membrane proteins: low-resolution structures obtained by

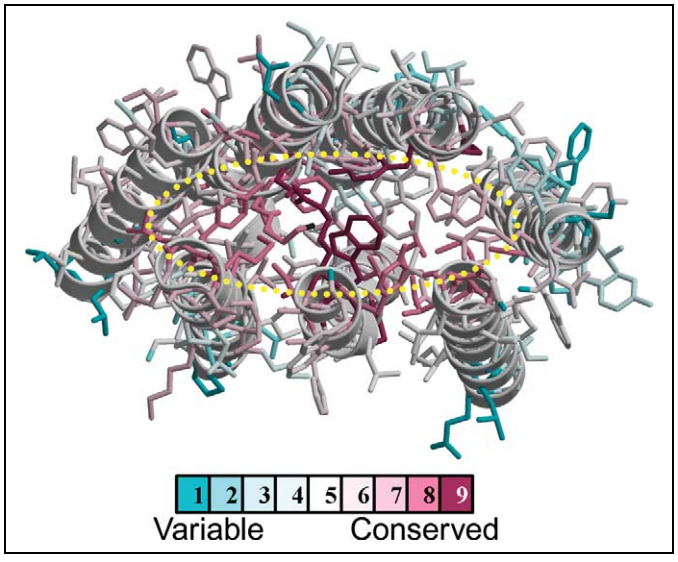


Figure 2. Evolutionary conservation can aid the orientation of TM helices. Evolutionary conservation is projected on the bacteriorhodopsin structure viewed from the direction vertical to the membrane plane, showing that the core of the protein (within the yellow ellipse) is more conserved than its periphery. Conservation was computed using the ConSurf webserver (http://consurf.tau.ac.il) [66].

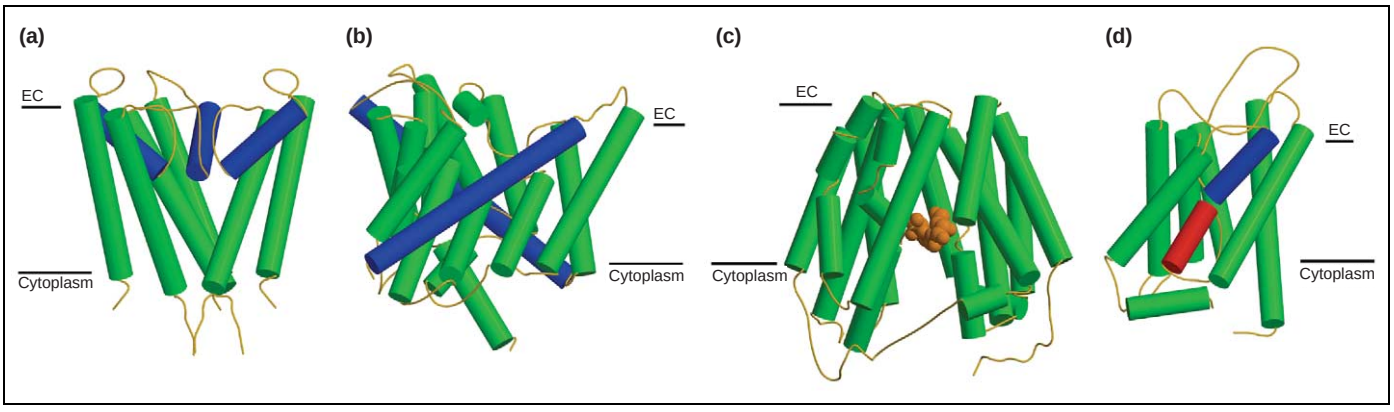


Figure 3. Recent structures reveal many discrepancies from the view that TM helices are canonical and span the entire lipid bilayer. **(a)** For clarity, only three of the four monomers comprising the K^+ ion channel are shown [19]. Blue cylinders represent the pore helix, which spans only half of the membrane width. **(b)** A monomer of the Cl^- channel [20]. The blue cylinders represent the locations of helices B and J, which are highly tilted with respect to the membrane normal and comprise ~35 amino acids each. **(c)** Structure of the transporter lac permease [21]. Some of the helices are kinked. Orange spheres represent a lactose analog. **(d)** Structure of the aquaporin 1 water channel [7]. Blue and red cylinders represent two half helices that meet at the mid-point of the membrane. (EC, extracellular.)

cryo-electron microscopy (cryo-EM) and mutational analyses of structure–function relationships.

Cryo-EM of 2D crystals of TM proteins

In contrast to the difficulties usually experienced in obtaining 3D crystals of TM proteins, in some cases, membrane proteins readily form 2D arrays in the membrane (e.g. bacteriorhodopsin [23], photosystem II [24], the gap junction [25], the bacterial translocon complex secYEG [26] and the bacterial multidrug-resistance transporter EmrE [27]). Added advantages of 2D crystals are that they mimic the native environment of the protein more closely than 3D crystals do, including interactions with the surrounding lipid molecules, which sometimes have important roles in determining the physiological structure [28]. For instance, substantial differences were observed between the cryo-EM map of EmrE [27] and a structure of the protein derived from X-ray analysis of 3D crystals [29]. Another demonstration of the importance of maintaining a membrane-like environment is provided by the differences between two recent X-ray structures of the voltage-gated K^+ channel [30,31], one of which was crystallized in the presence of lipids. In addition, it is sometimes possible to induce crystal formation in 2D, even when the proteins are dispersed in the membrane [32], and small and poorly ordered crystals can be used to derive data in the 5–10-Å resolution range thanks to digital-image-processing protocols that enable crystals to be corrected for translational disorder [14,33,34].

However, cryo-EM of 2D crystals usually produces structures at limited resolutions (typically, >4.5 Å in the plane of the membrane) so that individual amino-acid sidechains are not visible and, often, flexible loops and extramembranous domains are unresolved owing to lack of crystallographic order. Moreover, the resolution in the direction vertical to the lipid bilayer is worse than the in-plane resolution. This reduced resolution entails an uncertainty regarding the actual length of each helical segment, and might obscure the helical register. The lower vertical resolution might also limit the detection of helices that do not span the entire bilayer. In the case of the aquaporin-1 water channel for instance, an initial map at 6-Å in-plane resolution [35] did not reveal the surprising architecture of the channel, whereby two half-helices meet midway through the membrane (Figure 3d): misleadingly, these half-helices seemed to be one. A subsequent cryo-EM map at 4.5-Å resolution uncovered the two half-helices [36], and enabled a combination of sequence-based methods to be used to predict a model structure [37,38], which was found to be in agreement with the subsequently solved high-resolution structure [7]. The initially incorrect interpretation underscores the importance of improving resolution even marginally within the intermediate-resolution range to ascertain the general architecture of the protein.

Despite these shortcomings of intermediate-resolution maps, the fact that they provide an overall description of the protein architecture and the approximate packing of TM helices tremendously reduces the degrees of freedom for conformational search and the extent of uncertainty in constructing model structures. In fact, by assuming that ideal α helices occupy the locations observed in the map, the conformation search for the backbone positions can be limited to identifying the native-state orientation of each helix around its principal axis [39].

Building on this realization, and using further constraints obtained from multiple-sequence alignments and biochemical data (Box 1), Baldwin *et al.* [40] pioneered a structure-based modeling approach to derive the first model of the G-protein-coupled receptor (GPCR) rhodopsin based on a structure at 7-Å in-plane resolution [41]. Although rough, this model served as a template for modeling other GPCRs, which then provided a framework for interpreting the effects of mutations in the context of the receptor structure (see, for example, Refs [42,43]). Three years later, the first high-resolution structure of rhodopsin was solved by X-ray crystallography of 3D crystals [44], and showed that the previous model approximated the native-state structure to within 3.2-Å root-mean-square deviation. The orientations of all of the helices were predicted quite accurately by Baldwin *et al.* [40], and the main structural differences were due to

Box 1. Combinations of methods used in TM-protein-structure prediction

Many of the modeling applications for TM-protein structures have used at least some of the data sources and analyses shown in Figure I [37,40,47,67]. For many TM proteins, sufficient biochemical and biophysical data are available, specifying, for example, which sequence segments form helices [54] and make contact with other helices [48]. These data can be used to predict or verify the model. By contrast, cryo-EM maps at resolutions that enable the helix-packing arrangement to be discerned (typically better than 10 Å) have so far been obtained for only a few TM proteins, but more are expected to follow. The last two stages of modeling, in which modeling is refined by direct experimentation, have not yet been implemented in structure prediction of TM proteins. Generating atomic-resolution models (final step in Figure I) is complicated by the fact that even minor differences from the native-state structure often result in energetically unfavorable steric clashes and the abrogation of favorable polar bonds [68]. Because the positions of the backbone atoms inferred in the former steps of the flowchart are, at best, approximations of the native-state structure [39], conformational searching must also explore backbone degrees of freedom. Nevertheless, atomic resolution could considerably increase the quality and utility of TM-protein-structure prediction.

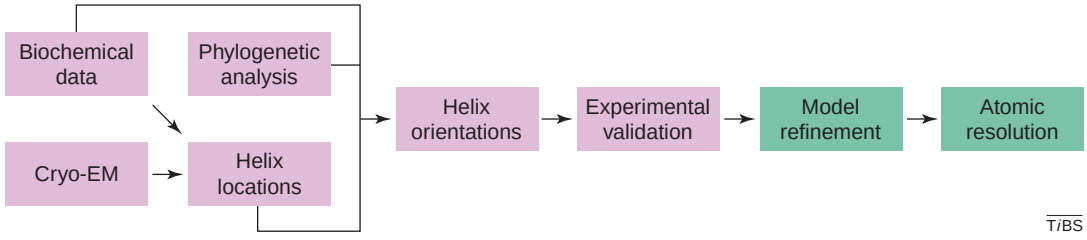


Figure I. A flowchart for the modeling of TM-protein structures. Various sources of experimental and computational information are often integrated to model TM proteins. The last two stages (green) have, so far, not been implemented in TM-protein structural modeling.

deviations in the positioning of the kinked helices (Figure 4).

The successful combination of cryo-EM and computational methods for the modeling of rhodopsin served as a basis for developing automatic tools for modeling based on phylogenetic analysis (Box 2) and intermediate-resolution structures [39,45]. These methods were then used to predict the structure of the TM domain of the gap junction, which is a channel that connects neighboring cells in a tissue, and lacks bacterial homologs [46]. A map of the gap junction was solved initially at a resolution of 7.5 Å in the membrane plane [25], and was subsequently improved to 5.7 Å [47]. The intermediate-resolution structure revealed a large pore (~15-Å diameter at the point of constriction), and clearly distinguished the four helices (M1–M4) that comprise each of the six gap-junction forming connexin monomers [25]. Because the intermediate-resolution map did not reveal the connectivities between the TM helices, the four hydrophobic segments in connexin sequences (M1–M4) were assigned to the four helices seen in the structure based on a combination of experimental and computational data. Subsequently, the four helices were oriented using evolutionary conservation and evolutionarily correlated mutations (Box 2; Figure 5).

Using this combination of approaches and data sources, a 5.7-Å resolution map (in-plane resolution), evolutionary conservation and correlated mutations, the final model structure predicted previously undetected interactions between pairs of polar residues in the structure. The model also suggested a molecular cause for almost 30 disease-related mutations. Although not taken into account during modeling, most of these mutations were revealed as mapping to structurally packed regions of the helix bundle, whereas two physico-chemically radical polymorphisms localized to the more spacious regions of the structure facing the lipid or the pore lumen [47] (Figure 5). Although it is clearly a model, it seems worthwhile to point out that the gap junction is an example of a eukaryotic membrane protein, the intermediate-resolution structure of which has not been superseded by a high-resolution crystal structure even six years after its original publication. Given the difficulties in obtaining well-ordered 3D crystals of eukaryotic membrane proteins, it seems likely that more cases are to follow, emphasizing why structure-based modeling is important and how it can help to generate a

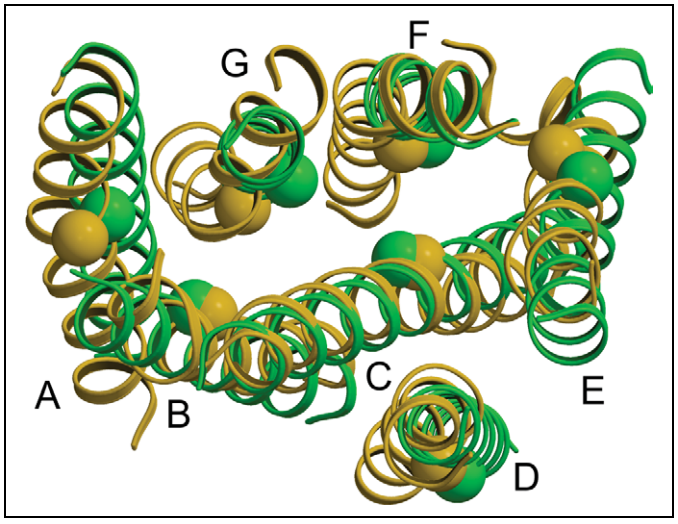


Figure 4. Comparison of the hypothetical and high-resolution structures of rhodopsin. The crystal and the hypothetical structures of rhodopsin are superimposed (yellow and green, respectively). The hypothetical structure was modeled on the basis on an electron-density map at 7-Å in-plane resolution [40]. The two structures deviate by 3.2-Å root-mean-square. Spheres are included to aid identification of identical positions in the hypothetical and crystal structures. The orientations of all of the helices are similar, and the main differences are in the locations of the helices within the plane of the membrane, particularly in the kinked helices F and G.

Box 2. Phylogenetic analysis used in TM-protein-structure prediction

Phylogenetic inference, and particularly conservation analysis, has found many applications in TM-protein-structure prediction [37,40,47]. Based on a multiple-sequence alignment of homologs of the target protein, individual amino-acid positions that show a low degree of sequence variation are considered important for protein structure or function [64], and are placed at strategic locations in the model structure, for example, at the interfaces between helices. Conversely, variable positions are considered to be unimportant, and are placed in lipid-facing positions. This type of sequence-based analysis is analogous to a large-scale mutagenesis scan conducted by evolution. To further refine the role of individual sidechain contributions, determination of evolutionarily co-varying sites can provide clues for contacts between positions. That is, if two positions form contact in 3D space, then a substitution in one site could be compensated by a substitution in the other. In this sense, determination of co-varying sites by any of several methods (see, for example, Refs [45,69]) can be regarded as an *in-silico* second-site suppression screen.

framework for planning and interpreting biochemical studies.

Biochemical and biophysical assays provide restraints for modeling

Mutagenesis and cross-linking assays have long been used to probe structure–function relationships in TM proteins, where high-resolution structures were not available (for reviews see Refs [46,48,49]) (Box 1). One aspect in which these techniques can aid modeling is validation because models make specific predictions regarding physical contacts between pairs of residues. Mutation analyses can also be used in the earlier stages of modeling. For instance, they have been used to identify the packing interfaces between helices [50] and the positions of pore-lining residues in channels [51], and cross-linking data have been used to constrain distances between pairs of positions [52]. Biochemical and biophysical analyses can also be used to assign the hydrophobic domains in the protein sequence to the helices seen in low-resolution structures [47,53], and to identify the secondary structure and tilt angles of the helices with respect to the membrane normal [54]. However, a major pitfall – which has obscured structural interpretation of some of these data – is the fact that mutagenesis assays cannot be used to discriminate between direct and indirect effects on helix association.

Glycophorin A (GpA), which is a small and extensively characterized bitopic protein that forms homodimers in the plasma membrane [55], is a good example to illustrate this problem. Much work has been conducted by Engelman and co-workers to explore the determinants of stability in GpA dimerization, and to gain insights into the process of helix association in the membrane.

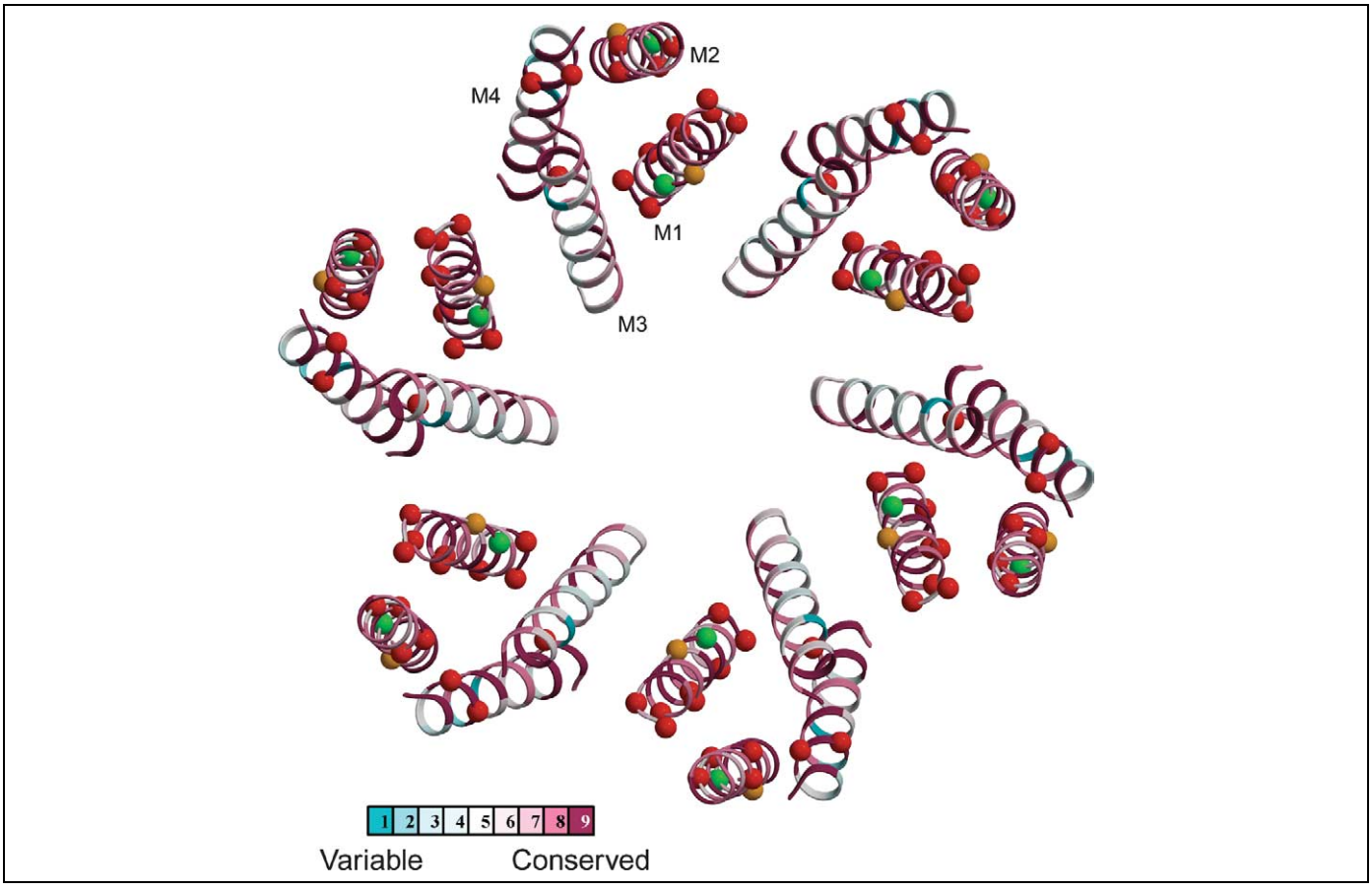


Figure 5. The model structure of the gap junction TM domain. The model structure is viewed from the cytoplasm of one cell looking in the direction vertical to the membrane [47]. Evolutionary conservation is color-coded on the structure according to the key. The positions of physico-chemically mild mutations that cause diseases and are packed within the bundle core are shown as red spheres; physico-chemically mild mutations that are not packed within the bundle core are shown as orange spheres; and physico-chemically radical polymorphisms are shown as green spheres. Almost all of the mild disease-causing mutations pack inside the bundle core, whereas the radical polymorphisms face the lipid or the pore lumen [47].

Systematic mutagenesis work by Lemmon *et al.* [50] identified a short sequence motif consisting of two glycine residues (Gly79 and Gly83), in which even physico-chemically mild substitutions abrogated dimerization. It was suggested that two glycine residues, which are small and polar, separated by three residues in the amino-acid sequence (Gly-Xaa-Xaa-Xaa-Gly) facilitate a closer approach of the two interacting α helices. It was later found that the Gly-Xaa-Xaa-Xaa-Gly motif can drive the dimerization of hydrophobic segments [56], and that it is statistically over-represented in TM sequences [57]. This and other sequence motifs were shown to have structural and functional roles in various TM proteins [58].

Although the role of the two glycine residues in the dimerization of GpA was deduced correctly from the mutagenesis assays, the same assays initially led to wrong conclusions with regard to Thr87 [50]. This position was also shown to be crucial for dimerization, and two different structural models based on molecular-dynamics-simulated annealing were suggested that supported the important roles of this triad of residues in dimerization [59,60]. One model consisted of an asymmetric right-handed supercoil [59]; the other model, suggested four years later, showed symmetric right-handed packing of the two helices [60]. The models agreed that the two glycine residues mediate much of the inter-helix contact. However, whereas Thr87 made a direct contribution to helix association in the earlier model by forming an inter-helical hydrogen bond, the residue stabilized the interface indirectly in the later model by forming an intra-helical hydrogen bond. The structure of GpA solved subsequently by NMR [8] supported the latter model (Figure 6). It is interesting to note that, because mutation analyses alone cannot discriminate between these two types of contributions to helix interaction, the interpretation of the experimental results led initially to the acceptance of an incorrect model [59].

A major problem with molecular dynamics (which was used in the prediction of GpA [59,60]) is that it is computationally demanding, essentially restricting its application to small homo-oligomers. A different approach for predicting the structures of pairs of TM helices was recently suggested that is based, in essence, on an integration of the experimental data on the stability of TM oligomers [61]. Studies of model TM proteins such as GpA highlighted the important role of small and polar sidechains in mediating inter-helix contacts [56–58]. Thus, a simple scoring function was suggested that favored contact formation between such residues, and penalized contacts mediated by large residues. The scoring function can discriminate between decoys and the conformations of several native-state pairs of tightly packed helices with known structures, including GpA. Using this function, it has been found that the TM domains of the receptor tyrosine kinase ErbB2 could exist in two stable alternative conformations [62], which is in agreement with *in vitro* studies [63]. These results were used to suggest a model of activation for this receptor that is coupled to a switch between the two conformations of the TM domain. However, a major drawback of this method [61] is that it assumes that the pairs of helices under study are closely packed (<9-Å separation between the principal axes of the helices), thus, in effect, precluding its applicability to most polytopic proteins [61].

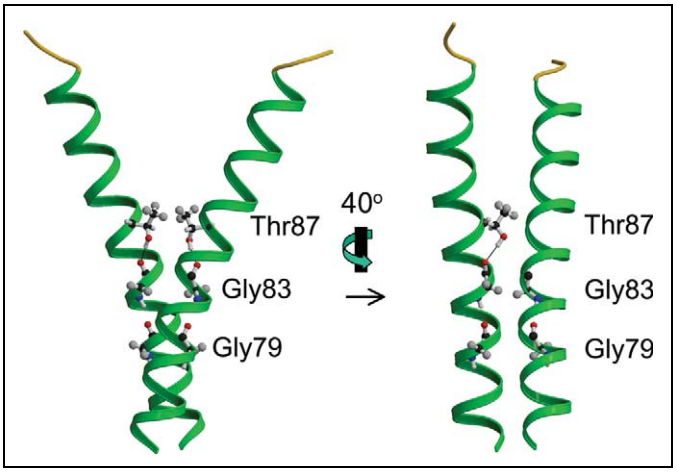


Figure 6. NMR Structure of GpA [8]. The two glycine amino acids enable the helices to pack tightly. In the view rotated by 40° (right), the intra-helical hydrogen bond between Thr87 and Gly83 is marked with a solid line.

Recently, a different modeling strategy, based on a combination of biochemical and biophysical data, was applied to the lac-permease [53]. This 12-membrane-spanning bacterial protein catalyzes the stoichiometric transport of galactosides with a proton across the membrane. The transporter was extensively investigated using a combination of single-site mutagenesis, double-cysteine mutants, second-site suppressors and biophysical methods [48]. In modeling, the data on the membrane-spanning segments were interpreted as constraints that approximate helical structures, and other experimental results were employed to provide 99 long-range constraints (between residues that are not sequence neighbors). Several other constraints were derived from data about the residues that participate in the binding of ligand. These distance constraints were then used in modeling the protein structure based on algorithms that are employed in NMR studies [53]. Once the structure was solved at atomic resolution by X-ray crystallography of 3D crystals [21] (Figure 3c), it was possible to compare it to the model. The comparison revealed many global discrepancies but also confirmed many local interactions (e.g. residues that interact directly with sugar and positions of residues involved in proton translocation). A closer look at the constraints derived from the cross-linking experiments [52] showed that the distances implied by these data agreed with the crystal structure on the compact side of the protein that faces the periplasm (Figure 3c). However, many of the constraints in the cytoplasmic-facing part of the protein consistently underestimated the distances seen in the crystal structure by ~10 Å [64].

Two main reasons were suggested for the discrepancies between the model and the crystal structure [21]: (i) the transporter is a dynamic structure with alternating cytoplasmic- and periplasmic-facing conformations, and, consequently, results from the mutation and cross-linking analyses reflect a superposition of several conformational sub-states; (ii) using disulfide-bond formation as an indication for inter-residue proximity tends to

underestimate the distances in the native state because the conformational changes of the protein bring into proximity residue pairs that might nevertheless be distal in the native structure. It is interesting to note that the cross-linking results indicated a model structure for the alternative periplasmic-facing conformation based on rotation of part of the structure with respect to the other [64]. In this alternative conformation, many of the experimental constraints in the periplasmic domain were consistent with the structure.

Concluding remarks

Structure determination of eukaryotic membrane proteins remains too slow to sustain hypothesis-driven experimentation aimed at understanding structure–function relationships in integral membrane proteins. Here, we have given examples for how mutational and computational techniques can be used to overcome this bottleneck by exploiting the information that is contained in intermediate-resolution structures obtained by cryo-EM. Although none of the techniques in isolation can provide anything more than clues, the sum of the different approaches yields insights into the structures of the targeted proteins at the level of individual amino-acid residues. Undoubtedly, as more intermediate-resolution structures emerge in the future, modeling techniques will be further refined and might be extended to include modeling of sidechains and non-canonical structures (Box 1) such as bulges and kinks [65]. Such refinements and extensions are likely to become crucial in the field of TM-protein structural studies, and will present researchers with a treasure trove of testable hypotheses to gain mechanistic insights into the function of integral membrane proteins.

Acknowledgements

We thank I.T. Arkin, D.M. Engelman, S. Harrington, H.R. Kaback and P.L. Sorgen for many useful comments. We regret that, owing to the focus on methods that have been used to predict novel structures, several new approaches are not reviewed. This study was supported by grant 222/04 from the Israel Science Foundation (ISF) to N.B-T., and NIH grants GM66145 and GM071590 to V.M.U. S.J.F. was supported by a doctoral fellowship from the Clore Israel Foundation.

References

1 Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.* 10, 1970–1979
2 Rost, B. *et al*. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* 5, 1704–1718
3 Mitaku, S. *et al*. (1999) Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the SOSUI prediction system. *Biophys. Chem.* 82, 165–171
4 Krogh, A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580
5 White, S.H. (2004) The progress of membrane protein structure determination. *Protein Sci.* 13, 1948–1949
6 Opella, S.J. and Marassi, F.M. (2004) Structure determination of membrane proteins by NMR spectroscopy. *Chem. Rev.* 104, 3587–3606
7 Murata, K. *et al*. (2000) Structural determinants of water permeation through aquaporin-1. *Nature* 407, 599–605
8 MacKenzie, K.R. *et al*. (1997) A transmembrane helix dimer: structure and implications. *Science* 276, 131–133
9 Chang, G. *et al*. (1998) Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* 282, 2220–2226
10 Bateman, A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.* 32 (Database issue), D138–D141
11 Fanelli, F. and De Benedetti, P.G. (2005) Computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev.* 105, 3297–3351
12 Oliveira, L. *et al*. (2004) Heavier-than-air flying machines are impossible. *FEBS Lett.* 564, 269–273
13 Engelman, D.M. *et al*. (2003) Membrane protein folding: beyond the two stage model. *FEBS Lett.* 555, 122–125
14 Henderson, R. *et al*. (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* 213, 899–929
15 Baldwin, J.M. (1993) The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* 12, 1693–1703
16 Kuhlbrandt, W. and Wang, D.N. (1991) Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. *Nature* 350, 130–134
17 Deisenhofer, J. *et al*. (1995) Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J. Mol. Biol.* 246, 429–457
18 Bowie, J.U. (1997) Helix packing in membrane proteins. *J. Mol. Biol.* 272, 780–789
19 Doyle, D.A. *et al*. (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* 280, 69–77
20 Dutzler, R. *et al*. (2002) X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* 415, 287–294
21 Abramson, J. *et al*. (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* 301, 610–615
22 Chen, C.P. *et al*. (2002) Transmembrane helix predictions revisited. *Protein Sci.* 11, 2774–2791
23 Unwin, P.N. and Henderson, R. (1975) Molecular structure determination by electron microscopy of unstained crystalline specimens. *J. Mol. Biol.* 94, 425–440
24 Rhee, K.H. *et al*. (1998) Three-dimensional structure of the plant photosystem II reaction centre at 8 Å resolution. *Nature* 396, 283–286
25 Unger, V.M. *et al*. (1999) Three-dimensional structure of a recombinant gap junction membrane channel. *Science* 283, 1176–1180
26 Breyton, C. *et al*. (2002) Three-dimensional structure of the bacterial protein-translocation complex SecYEG. *Nature* 418, 662–665
27 Ubarretxena-Belandia, I. *et al*. (2003) Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer. *EMBO J.* 22, 6175–6181
28 Fujiyoshi, Y. (1998) The structural study of membrane proteins by electron crystallography. *Adv. Biophys.* 35, 25–80
29 Ma, C. and Chang, G. (2004) Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 2852–2857
30 Jiang, Y. *et al*. (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature* 423, 33–41
31 Long, S.B. *et al*. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* 309, 897–903
32 Hasler, L. *et al*. (1998) 2D crystallization of membrane proteins: rationales and examples. *J. Struct. Biol.* 121, 162–171
33 Henderson, R. *et al*. (1986) Structure of purple membrane from *Halobacterium halobium*: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* 19, 147–178
34 Amos, L.A. *et al*. (1982) Three-dimensional structure determination by electron microscopy of two-dimensional crystals. *Prog. Biophys. Mol. Biol.* 39, 183–231
35 Walz, T. *et al*. (1997) The three-dimensional structure of aquaporin-1. *Nature* 387, 624–627
36 Mitsuoka, K. *et al*. (1999) The structure of aquaporin-1 at 4.5-Å resolution reveals short α-helices in the center of the monomer. *J. Struct. Biol.* 128, 34–43
37 Heymann, J.B. and Engel, A. (2000) Structural clues in the sequences of the aquaporins. *J. Mol. Biol.* 295, 1039–1053
38 de Groot, B.L. *et al*. (2000) The fold of human aquaporin 1. *J. Mol. Biol.* 300, 987–994

39 Fleishman, S.J. *et al*. (2004) An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data. *Biophys. J.* 87, 3448–3459
40 Baldwin, J.M. *et al*. (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* 272, 144–164
41 Unger, V.M. *et al*. (1997) Arrangement of rhodopsin transmembrane α-helices. *Nature* 389, 203–206
42 Latronico, A.C. *et al*. (1998) A unique constitutively activating mutation in third transmembrane helix of luteinizing hormone receptor causes sporadic male gonadotropin-independent precocious puberty. *J. Clin. Endocrinol. Metab.* 83, 2435–2440
43 Scheer, A. *et al*. (2000) Mutational analysis of the highly conserved arginine within the Glu/Asp-Arg-Tyr motif of the α_{1b}-adrenergic receptor: effects on receptor isomerization and activation. *Mol. Pharmacol.* 57, 219–231
44 Palczewski, K. *et al*. (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* 289, 739–745
45 Fleishman, S.J. *et al*. (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J. Mol. Biol.* 340, 307–318
46 Harris, A.L. (2001) Emerging issues of connexin channels: biophysics fills the gap. *Q. Rev. Biophys.* 34, 325–472
47 Fleishman, S.J. *et al*. (2004) A C-α model for the transmembrane α-helices of gap-junction intercellular channels. *Mol. Cell* 15, 879–888
48 Kaback, H.R. *et al*. (2001) The kamikaze approach to membrane transport. *Nat. Rev. Mol. Cell Biol.* 2, 610–620
49 Karlin, A. (1993) Structure of nicotinic acetylcholine receptors. *Curr. Opin. Neurobiol.* 3, 299–309
50 Lemmon, M.A. *et al*. (1992) Sequence specificity in the dimerization of transmembrane α-helices. *Biochemistry* 31, 12719–12725
51 Karlin, A. and Akabas, M.H. (1998) Substituted-cysteine accessibility method. *Methods Enzymol.* 293, 123–145
52 Kwaw, I. *et al*. (2000) Thiol cross-linking of cytoplasmic loops in the lactose permease of *Escherichia coli*. *Biochemistry* 39, 3134–3140
53 Sorgen, P.L. *et al*. (2002) An approach to membrane protein structure without crystals. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14037–14040
54 Torres, J. *et al*. (2001) Site-specific examination of secondary structure and orientation determination in membrane proteins: the peptidic $^{13}C{=}^{18}O$ group as a novel infrared probe. *Biopolymers* 59, 396–401
55 Furthmayr, H. and Marchesi, V.T. (1976) Subunit structure of human erythrocyte glycophorin A. *Biochemistry* 15, 1137–1144
56 Russ, W.P. and Engelman, D.M. (2000) The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.* 296, 911–919
57 Senes, A. *et al*. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *J. Mol. Biol.* 296, 921–936
58 Sternberg, M.J. and Gullick, W.J. (1990) A sequence motif in the transmembrane region of growth factor receptors with tyrosine kinase activity mediates dimerization. *Protein Eng.* 3, 245–248
59 Treutlein, H.R. *et al*. (1992) The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry* 31, 12726–12732
60 Adams, P.D. *et al*. (1996) Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins* 26, 257–261
61 Fleishman, S.J. and Ben-Tal, N. (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices. *J. Mol. Biol.* 321, 363–378
62 Fleishman, S.J. *et al*. (2002) A putative activation switch in the transmembrane domain of erbB2. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15937–15940
63 Mendrola, J.M. *et al*. (2002) The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J. Biol. Chem.* 277, 4704–4712
64 Abramson, J. *et al*. (2003) The lactose permease of *Escherichia coli*: overall structure, the sugar-binding site and the alternating access model for transport. *FEBS Lett.* 555, 96–101
65 Yohannan, S. *et al*. (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.* 101, 959–963
66 Glaser, F. *et al*. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19, 163–164
67 Beuming, T. and Weinstein, H. (2005) Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT. *Protein Eng. Des. Sel.* 18, 119–125
68 Schueler-Furman, O. *et al*. (2005) Progress in modeling of protein structures and interactions. *Science* 310, 638–642
69 Gobel, U. *et al*. (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18, 309–317

ELSEVIER

Progress in structure prediction of α-helical membrane proteins

Sarel J Fleishman and Nir Ben-Tal

Transmembrane (TM) proteins comprise 20–30% of the genome but, because of experimental difficulties, they represent less than 1% of the Protein Data Bank. The dearth of membrane protein structures makes computational prediction a potentially important means of obtaining novel structures. Recent advances in computational methods have been combined with experimental data to constrain the modeling of three-dimensional structures. Furthermore, threading and *ab initio* modeling approaches that were effective for soluble proteins have been applied to TM domains. Surprisingly, experimental structures, proteomic analyses and bioinformatics have revealed unexpected architectures that counter long-held views on TM protein structure and stability. Future computational and experimental studies aimed at understanding the thermodynamic and evolutionary bases of these architectural details will greatly enhance predictive capabilities.

Addresses
Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel-Aviv University Ramat Aviv 69978, Israel

Corresponding author: Ben-Tal, Nir (nirb@tauex.tau.ac.il)

Current Opinion in Structural Biology 2006, **16**:496–504

This review comes from a themed issue on
Membranes
Edited by Roderick MacKinnon and Gunnar von Heijne

Available online 5th July 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.06.003

Introduction

Transmembrane (TM) proteins comprise ~20-30% of the genome [1,2] and are involved in many crucial cellular processes, such as cell-to-cell signaling, metabolite transport and energy production. Solving the structures of these proteins is therefore imperative for clear mechanistic understanding of central processes in physiology. However, despite recent advances in production of TM protein crystals, membrane protein structures are difficult to obtain and comprise less than 1% of the entries in the Protein Data Bank (PDB) [3].

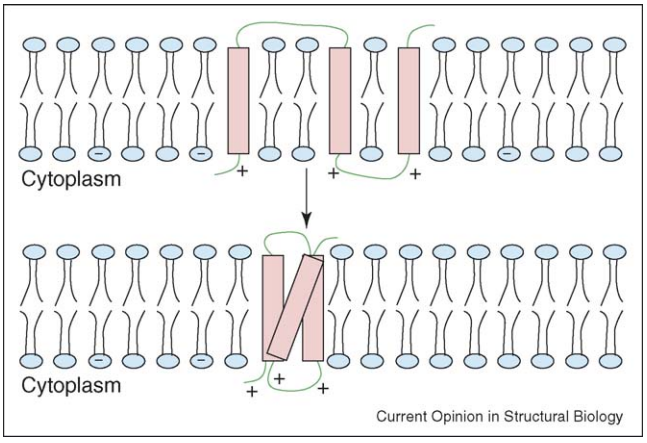
Comparative- or homology-based approaches to structure prediction have been immensely successful with soluble proteins [4]. These methods require a homologous protein, for which a structure has been solved. Because of this requirement, homology modeling has been most useful for the few TM protein families, for which at least one member has been crystallized. A recent analysis of homology-modeling accuracy for membrane proteins has shown that the protocols that are successful in comparative modeling of soluble proteins reach similar achievements for membrane proteins [5•]. However, because at present only few representative atomic-resolution structures of TM protein families are available, homology modeling cannot serve as a general purpose approach for structural modeling. In this review, we will therefore focus on recent advances in structure prediction that do not rely on homology to solve structures (subject covered in [6,7•]).

Membrane protein folding can be conceptually decomposed into two consecutive steps: folding of the individual hydrophobic segments into helices followed by helix association (Figure 1) [8]. Accordingly, the problem of predicting the structure of α-helical TM proteins has been approached by breaking it down into the following steps: (i) delineating the boundaries of the TM segments, each of which will assume a helical conformation; (ii) determining the topology of the protein (i.e. which extramembrane segments reside inside the cytoplasm and, conversely, which segments reside outside the cell); and (iii) predicting the tertiary conformation of the protein (i.e. the way in which the helices are packed with respect to one another). The past few years have seen considerable advances in all of these steps. In this review, we will describe some of these advances and emphasize the discovery of novel features of TM protein folds that bear on the goal of structure prediction.

Identification of TM α-helices in the protein sequence

Early attempts for predicting the locations in the sequence of membrane-integral segments were based on the notion that a sequence segment would partition into the membrane if it were sufficiently long and hydrophobic. Starting with the method of Kyte and Doolittle [9], various algorithms for detecting membrane-embedded sequence segments were proposed on the basis of experimental and computational data. At the core of these methods lies a hydrophobicity scale that assigns to each amino acid residue a score that can be roughly interpreted as the free energy of its transfer from hydrophilic to hydrophobic media, corresponding to its insertion probability into the membrane. The typical approach would then be to search the sequence for a sufficiently hydrophobic stretch of residues comprising

Figure 1



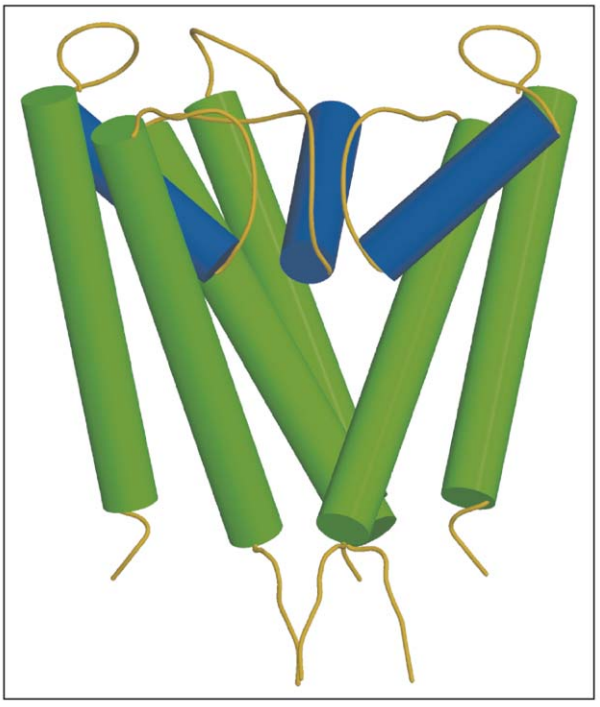
TM protein folding can be thought to proceed in two stages [8]: the folding of individual TM segments into helices (top) followed by helix packing (bottom). The topology of the protein is often determined by the positive-inside rule [17], with the cytoplasmic loops tending to be enriched by positively charged residues in comparison with the extracellular loops.

approximately 20 amino acids, which is the minimal length necessary for an α-helix to traverse the 30 Å hydrophobic core of the membrane [10].

During the 90s, there was a departure from physicochemically based approaches to methods that rely on statistical inference, such as hidden Markov models, support vector machines and neural nets, all of which make use of the existing knowledge on the partitioning of particular sequence segments to the membrane. These methods appeared at first to be superior to the simple hydrophobicity-based methods, with success rates of 90% and above [1]. However, a fundamental difficulty in the validation of statistical methods is to obtain sufficiently disparate datasets for training and validation. Indeed, when Rost and co-workers recently revisited the problem of TM sequence prediction [11] using datasets that were carefully constructed with the aim of decreasing redundancy, they found that the success of the statistical approaches was overrated, and they in fact achieved results that were not much better than those that were obtained by some of the hydrophobicity-based methods. In this respect it is important to emphasize that an overlap of only three amino acids between the predicted and observed helices is considered sufficient for being an accurate prediction [11]. Thus, in a recent survey it was demonstrated that, on average, the best-performing prediction methods were in error by a little more than two turns at the helix termini [12]. Because most structural modeling approaches rely on the correct identification of the helical segments in the sequence (see below), these large errors are likely to propagate in subsequent modeling stages, requiring manual intervention. A more alarming conclusion made in this survey concerned the inability of current prediction methods to identify 'irregular' structures, such as half helices and re-entrant loops, as those seen in the structure of the potassium channel (Figure 2) [13] and the aquaporin family [14]. Hopefully, with the likely increase in the number of proteins exhibiting such irregularities over the next few years, some unifying principles will emerge from their sequences, enabling prediction of these features.

Recently, the hydrophobicity-based approach to detecting membrane-embedded segments was given another boost from the experimental studies by von Heijne and co-workers [15••]. The authors reported a series of experiments that attempted to obtain a hydrophobicity scale using an experimental setup that is far closer to the physiological system than previous experimental reports, including the translocon protein-conducting channel and membranes from the endoplasmic reticulum (ER). Concerns were raised regarding the possibility that some of the measured partitioning energies encompass contributions from interactions between the probe sequence segments and other protein components in the system, thus limiting the generality of the scale produced by these measurements [16•]. Nevertheless, this experimental

Figure 2



The potassium channel [13] is one of the several structures of membrane proteins that show structural 'irregularities', such as half helices (blue) and re-entrant loops. These irregularities cannot be identified from the sequence by current methods [12]. For clarity, only three out of four of the subunits comprising the potassium channel are shown. Figure generated with MolScript [70] and rendered with Raster3d [71]. Figure reproduced with permission from [37].

approach is promising, raising hope that the prediction of the location of TM helices in the sequence of membrane proteins will eventually be based on algorithms that account for the various factors that affect protein translocation in biological systems.

Topology

Determining the topology of a membrane protein is a crucial preliminary step to modeling its structure as it constrains the way individual TM segments could associate within the membrane, as well as subunits within complexes. The positive-inside rule (i.e. the observation that the segments in the cytoplasmic loops and the TM segments that are adjacent to the cytoplasm are often enriched in the positively charged lysine (K) and arginine (R) residues when compared with the extracellular loops (Figure 1) [17]) has remained the most powerful tool for predicting the topology of a protein from its sequence for almost two decades. The factors contributing to the (K + R) bias are under intense study, and it is still unclear whether the bias originates from properties of the translocon [18] or the cytoplasmic membrane [19], but a recent statistical survey of 107 genomes reconfirmed the validity of this empirical rule [20]. The (K + R) bias can serve as a rule for predicting topology, by requiring that more positively charged residues face the cytoplasm [1].

Recently, von Heijne and co-workers have conducted a whole-proteome experimental analysis of the topology of TM proteins in the *Escherichia coli* inner membrane [21••]. They used two reporter proteins that were linked to the C-terminus of each putative membrane-integral protein in *E. coli*. One of these reporters is only active in the cytoplasm, whereas the other is exclusively activated in the periplasm. By measuring the activities of the reporters, the authors assigned the topology of 601 out of 700 predicted TM proteins in the *E. coli* genome. Comparing these data to the predictions of a widely used algorithm that is based on a hidden Markov model called TMHMM [2], the authors found that roughly 80% of the predictions were in accord with the experimentally determined topologies. This correlation shows that the major aspects affecting protein topology are captured by contemporary computational methods, but that these still have significant room for improvement. These experimental results can serve as a much-needed large-scale benchmark for validation and comparison of future topology prediction algorithms.

The vast majority of proteins in von Heijne and co-workers' analysis exhibited unique topology [21••], whereby their C-terminus was found to be either cytoplasmic or periplasmic. However, for five out of 601 proteins both reporters were activated, implying that for each of these five proteins, some of the protein copies inserted with one topology, and the others with the reverse topology [21••,22]. The five proteins with dual topology are relatively small in size, comprising ~100 amino acid residues and are predicted to contain four TM domains. Furthermore, as expected, all five exhibit very small (K + R) biases. For at least one of these proteins, the prototypical small multidrug resistance antiporter EmrE, the suggestion of dual topology was already made in the past on the basis of structural data and the lack of clear (K + R) bias [23]. Nevertheless, it is important to note that a previous study based on a different biochemical assay reported a unique topology for this protein [24]. This conflict between two lines of experimental evidence still needs to be resolved, but the suggestion that some TM proteins insert with opposite topology has significant implications for understanding structures and functions of these proteins.

Threading and *ab initio* structure prediction

On the one hand, integral membrane proteins exhibit much higher uniformity of secondary structure (mostly α-helical bundles) than soluble proteins, and are highly constrained in their conformations because of the presence of the membrane [25]. It could therefore be expected that *ab initio* structure prediction, whereby the protein structure is predicted without resorting to homology with other proteins or to experimental data, should be a more feasible goal for TM than for soluble proteins. On the other hand, as sampling significant portions of conformation space remains a very challenging aspect of *ab initio* structure prediction [26], success in soluble protein structure prediction has been restricted to small proteins, consisting of approximately 80 amino acid residues [27]. Membrane proteins are usually much larger; for instance, visual rhodopsin, which serves as a prototype for the large family of 7-TM GPCRs, consists of more than 300 amino acid residues.

Two similar methods, MembStruk [28–31] and PREDICT [32,33], were specifically tailored to predict the structures of GPCRs on the basis of physicochemical principles. For both methods, a full-atom model of the GPCR is automatically obtained, based on the amino acid sequence of the protein alone. In the first step, the boundaries of the seven TM helices are predicted by means of hydrophobicity scales. A preliminary (tentative) coarse-grained model of the packing of these helices into a compact and closed structure is constructed, and various conformations in the vicinity of this state are sampled at random, favoring conformations in which hydrophobic residues face the lipid. Full-atom models of the TM domains of these structures are built and subjected to several cycles of optimization using molecular dynamics (MD) simulations. The outcome is a full-atom model of the entire protein, including the extra-membrane loops. The methods produced 3D models of bovine rhodopsin, the only GPCR structure available in the PDB, with ~3 Å root-mean-square deviation (RMSD) from the native structure in the TM region. Further validation of this

approach includes *in silico* docking of known drug-like compounds to the receptors. Model structures of several GPCRs, including the β2 adrenergic [30] and D2 dopamine [28] receptors, were built this way and used successfully for drug design [32]. This suggests that important structural aspects of the ligand-binding site were accurately captured by these methods. However, it was not shown unambiguously that the remainder of the structure is correct too.

Another potentially promising approach utilizes the two-step TASSER method that threads the sequence on parts of solved protein structures, and then refines the resulting template [34•]. Validation on a set of 38 nonhomologous TM protein structures yielded 17 structures for which the RMSD to native was less than 6.5 Å, but many others with RMSD to native greater than 10 Å. When applied to predicting the structure of bovine rhodopsin, TASSER produced a model with a low 2.1 Å RMSD from native on the C^{α} coordinates of the TM domain. Subsequently, the method was applied to model the structures of most of the ~900 human GPCRs, and a few of these models were examined and appeared to be consistent with the available experimental data. It is important to note that although the method's success in modeling rhodopsin is promising, only a few other GPCRs showed substantial similarity (>30% sequence identity) to bovine rhodopsin [7,34•], and it is therefore uncertain that the other models are as faithful to the native state as the model of rhodopsin. Also, it is not known yet whether TASSER's GPCR models are likely to be closer to the receptors' inactive or active form, the latter of which is pharmaceutically more interesting [7]. Nevertheless, the models generated by TASSER might provide an important resource for probing structure–function relationships in this important class of receptors, as many of the current approaches to modeling GPCR structures rely on homology to bovine rhodopsin [6], despite the low sequence identity.

Recently, the Rosetta algorithm for structure prediction, which has been successful in the free-modeling category of the community-wide experiment on critical assessment of structure prediction (CASP) [35], was adopted and implemented for TM protein structures [36•]. Inter-residue contact potentials were derived from a set of solved protein structures, and enriched with their sequence homologues. Validation on a set of solved TM protein structures showed that the performance of this implementation of Rosetta (below 4 Å for 51–145 of the superimposed residues) is comparable to that of Rosetta for soluble proteins in the same size range. Although full-atom prediction was shown to produce significant improvements in prediction accuracy of soluble proteins [27], it was not tested in this implementation of Rosetta, partly because of the prohibitive computational load associated with full-atom prediction for large proteins.

Structure prediction based on experimental constraints

One potential venue for obtaining novel structures, which has been explored by several groups in recent years, is the exploitation of functional and low-resolution structural data on TM proteins to constrain models [37•]. Such data could involve site-specific mutagenesis, chemical crosslinking, intermediate-resolution structures and biophysical data, such as NMR, EPR and FTIR. These heterogeneous data are interpreted as constraints on the positions of individual amino acid residues or on the structural relationships among them. For instance, positions that are intolerant to substitution are likely to be packed inside the protein core, and positions that crosslink are likely to be vicinal. In addition to these experimental data, the modeling methods assume that the hydrophobic sequence segments form α-helices that traverse the membrane.

The pioneering work of Herzyk and Hubbard [38] employing such disparate data sources produced very promising results, with a model of bacteriorhodopsin matching the native-state structure by a low 1.87 Å RMSD. However, further modeling attempts that relied primarily on mutation and crosslinking data demonstrated that it is difficult to interpret many of these data in a structurally unequivocal way [37•]. Recent implementations of this approach have therefore relied on more limited data sources. For instance, a method was suggested recently that employs data that can be interpreted as distance constraints between amino acid residues from EPR, FTIR and chemical crosslinking [39]. Models consisting of α-helices were sampled using a Monte Carlo strategy. The conformations were scored according to the extent to which they satisfied the experimental distance constraints and structural parameters derived from a set of solved TM proteins, including preferred helix-packing angles and distances, pairwise amino acid contact preferences and overall structural compactness. Encouragingly, this method was shown to produce a model of rhodopsin, which was 3.2 Å RMSD from the native-state structure, based on only 27 experimentally derived distance constraints (taken from published studies), demonstrating that it might be possible to obtain close-to-native models of large membrane proteins on the basis of a limited set of experimental constraints.

Several groups have recently suggested methods that employ data from cryo-electron microscopy (cryo-EM) intermediate-resolution structures, together with data on hydrophobicity, evolutionary patterns and the lengths of the loops that connect neighboring TM segments [37•]. For several proteins, cryo-EM structures are available at in-plane resolutions of 5–10 Å (e.g. the gap junction [40] and EmrE [23]). At this resolution, it is impossible to either position individual amino acid residues, or even unambiguously identify the assignment of TM segments

to the helices observed in the cryo-EM structure. Hence, structure prediction based on cryo-EM is typically comprised of helix assignment, followed by orientation of the helices around their principal axes.

To solve the helix assignment problem, various studies used biochemical data on the functional roles of individual TM segments [41,42]. A complementary approach relies on the fact that some of the loops that connect TM helices are quite short (less than eight amino acid residues). Such short loops constrain the distance between the helix termini that they connect. Based on this constraint, an algorithm was recently suggested, which, for a given cryo-EM structure and the lengths of each of the interconnecting loops, scans all possible assignments (potentially $n!$ permutations, where n is the number of helices in the map), and ranks them by their compatibility with the cryo-EM structure [43]. The performance of the algorithm was found to be sensitive to the exact delineation of the helix start and end points, which are difficult to predict with accuracy. Another proposed method that suffers less from such sensitivity ranks each TM sequence segment according to its overall hydrophobicity and evolutionary conservation [44]. Highly conserved and hydrophilic segments were ranked as helices that are likely to be buried within the protein core, and more variable and hydrophilic segments were assigned to lipid-exposed positions.

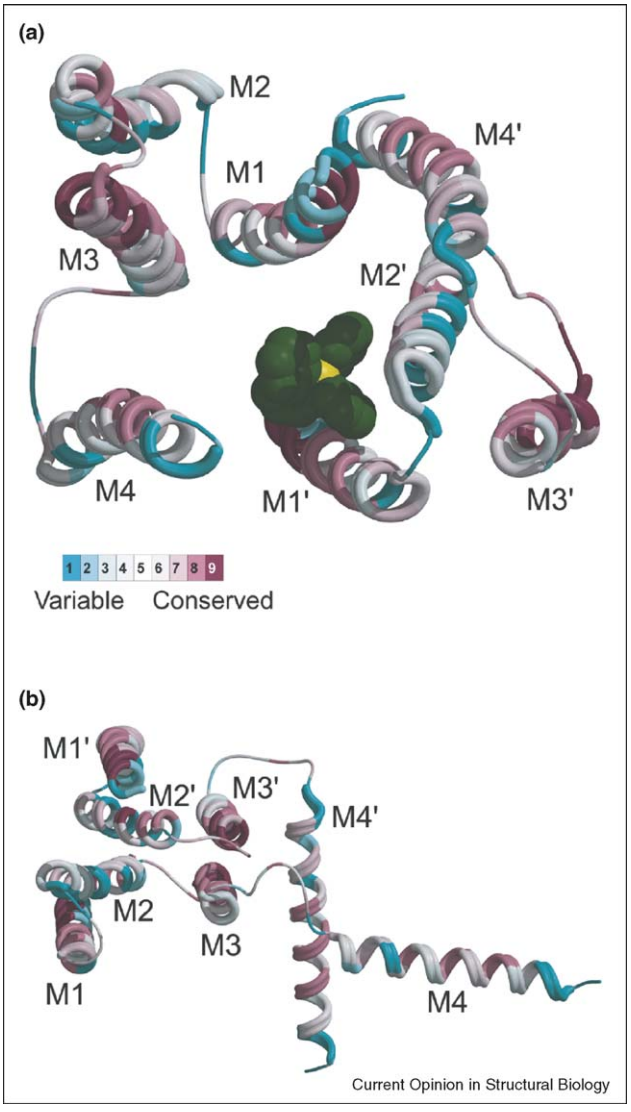
Once the helix assignment problem is solved for a given protein, canonical α-helices are constructed to fit the data in the cryo-EM map, and are rotated around their principal axes to identify the native state conformation. Following the work of Baldwin *et al.* [45] on the prediction of the structure of the TM domain of rhodopsin based on its cryo-EM structure and sequence analysis, recently two similar methods [46,47] were independently suggested. It was shown that the cores of many TM protein structures are much more evolutionarily conserved than their peripheries, and tend to pack the most polar residues [48]. These observations can be framed as predictive rules, according to which orientations that pack conserved and hydrophilic positions in the helix bundle are more favored than others. One of the methods generates only C^{α} models [47], whereas the other adds sidechains and uses manual refinements and minimization to generate full-atom models [46]. It should be noted, however, that often the energy landscape for full-atom models is extremely rugged and even 1 Å differences in the atom positions from the native-state structure can result in large energy penalties [26]; thus, it still remains to be seen whether the addition of sidechains improves the resulting models. The two methods were applied to intermediate-resolution structures of TM proteins, for which atomic-resolution data were not available: the oxalate transporter OxlT [46] and the gap junction [49••]. Because the evolutionary-conservation pattern on two of the helices of the gap-junction forming protein, connexin, was not informative enough to constrain their orientations, another sequence analysis method [50] was employed that identified correlated amino acid positions, thus predicting which pairs of amino acid residues could interact. Part of the attractiveness of an approach to structure prediction, which uses information from sequences and cryo-EM structures, lies in the fact that it does not necessarily rely on large amounts of previously published functional data. Hence, it is possible to subsequently use these data for validation. In the modeling of the gap junction TM domain, for instance, it was shown that, although the model was not constrained by clinical data, it placed almost 30 disease-causing but physicochemically mild mutations in the core of the helix bundle, where they would disrupt folding, whereas two physicochemically radical polymorphisms were placed in more spacious regions of the protein structure [49••]. Similarly, the model structure of OxlT placed residues that were found to crosslink in experimental assays in proximal positions [46].

Kinks in TM proteins are known to have important functional roles [51,52] but, until recently, could not be predicted from sequence information. Recently, it was shown that, in many cases where a kink is present in a TM protein structure, prolines are observed in the multiple-sequence alignment, even if the solved protein structure does not contain a proline at that position [53•]. The direction and magnitude of the kink might also be predicted from local sequence features [54]. Accordingly, it might be possible to model kinks where these have been observed in low-resolution structures, as in EmrE [23], or to bias the *ab initio* predictions to produce kinks and, thus, generate more native-like models.

Computational validation of structures

Recently, a small number of atomic resolution structures of membrane-integral proteins were suggested to represent conformations that are distorted with respect to the native-state structure [55,56]. Atomic resolution structures inspire a large amount of (usually very productive) work aimed at understanding structure–function relationships. Conversely, physiologically irrelevant structures might cause much work to be done in vain, on top of supplying a wrong view of the protein. Usually, the ultimate test for the physiological relevance of a structure is its compatibility with carefully crafted biochemical and biophysical analysis. However, such analyses are often difficult to conduct. Because some of the computational analyses described above can be used to predict the structures of membrane-integral proteins, it is reasonable to expect that they might provide grounds for doubting structures that have not been sufficiently supported by biochemical data. As an example of this approach, Figure 3 shows two structures of the bacterial multidrug resistance protein EmrE obtained by X-ray crystallography at 3.8 Å and 3.7 Å resolution [57,58]. Both structures

Figure 3



Two recently solved structures of homodimers of the multidrug resistance protein EmrE from *E. coli* are shown, which are incompatible with the observation that amino acid residues at the core of many membrane-integral proteins tend to be evolutionarily conserved, whereas those on the periphery are variable. **(a)** The structure of substrate-bound EmrE [58] exhibits highly variable residues on helix M2 forming tight contacts with M3, whereas highly conserved positions on M1, M2, M3, M3' and M4' are exposed to lipid. The substrate tetraphenylphosphonium molecule is shown in space-fill mode, with the phosphate colored in yellow, and carbon atoms in green. The structure is viewed perpendicular to the proposed membrane plane. **(b)** Similarly, the structure of EmrE without bound substrate [57] locates highly variable residues in the tight interface formed between M2 and M2', and highly conserved residues on M1, M4, M1', M3', and M4' in lipid exposed positions. The incompatibility between the conservation pattern and the burial of amino acid residues parallels the observation that both structures have many features that are in contradiction with biochemical data on EmrE [61]. Evolutionary conservation was computed using a multiple-sequence alignment of 99 small multidrug resistance proteins with the ConSurf webserver [72]. Figure generated with MolScript [70] and rendered with Raster3d [71].

are clearly at odds with the observation made on many TM protein structures that evolutionarily conserved positions tend to be packed in the core of the α-helix bundle, whereas the variable residues face the lipid environment [46,47,59,60]. The discrepancy between the conservation pattern and the packing of residues parallels an analysis, reported in this issue of *Current Opinion in Structural Biology* [61], that compares these structures with the known biochemical and biophysical data on EmrE, concluding that they most likely do not represent the physiological native state of the protein.

Future directions

In recent years, computational methods have been implemented for the prediction of TM protein structures. However, the roles of different energetic factors in contributing to TM protein folding are still poorly understood [25,62] and therefore difficult to predict. For instance, it was proposed that in low-dielectric environments polar bonds would make a large contribution to protein stability [10]. Indeed, in engineered systems, hydrogen bonds were shown to drive the interaction between TM helices [63,64], but recent measurements of the strengths of polar interactions in membrane proteins have yielded smaller magnitudes [65,66] than anticipated by computations on ideal hydrogen bonds [67,68]. Based on these and other measurements of the energetics of helix association in the membrane, it has been suggested that the primary contribution to helix interactions in the membrane comes from van der Waals packing and originates from buried surface area as in soluble proteins [69••]. This suggestion, which requires additional experimental support, is crucial because it implies that the major factors that are currently embodied in *ab initio* methods for structure prediction in soluble proteins, such as steric packing [27], might be equally useful in membrane-integral proteins. It is likely that the relative contributions of polar and van der Waals interactions to membrane protein stability will continue to be a matter of intense experimental investigation over the next few years, and that the lessons learned from these studies will be incorporated into the force fields of *ab initio* and threading algorithms for membrane proteins [34•,36•]. The use of these lessons could reduce, in part, the need for deriving pairwise contact potentials from the small number of solved TM protein structures.

One impediment on the way to the application of *ab initio* techniques to membrane proteins is the fact that these proteins are very large in comparison with soluble proteins, to which these methods were successfully applied, thus making full-atom prediction impractical [36•]. However, as modeling approaches that make use of experimental information, such as cryo-EM low-resolution structures and distance constraints, have been clearly successful in identifying near-native although coarse-grained conformations of TM proteins [38,39,45–47], a

synergy might be attainable from combining these methods with full-atom predictions. This would result in reliable atomic models at a computationally feasible cost.

With the advent of new structures and the application of novel biochemical assays to membrane-integral proteins, the last few years have seen a large increase in the qualitative understanding of TM protein folds. This improved understanding has gone hand-in-hand with more sophisticated prediction and modeling attempts. Undoubtedly, the new structures and structure–function analyses that will be conducted over the next few years will teach us many lessons on the possible architectures of TM proteins and their governing thermodynamic principles, further increasing our predictive capabilities.

Update

Recently, the Rosetta membrane methodology [36•] was adapted and applied to study the voltage-induced conformational changes in the voltage-dependent potassium (Kv) channels [73]. Open and closed conformations were computed for the eukaryotic Kv1.2 channel and for the bacterial KvAP on the basis of the published methodology, the homology to X-ray structures of these channels and several experimental constraints. The computed open conformation of Kv1.2 was close to its crystal structure, thus serving as partial validation for the approach. Interestingly, the results suggest that the conformational changes in the voltage-sensor domain of the bacterial protein are larger than the changes in Kv1.2, which could explain the large inconsistencies between functional studies of the bacterial and eukaryotic channels.

Acknowledgements

The authors thank SE Harrington, JU Bowie, J Skolnick, O Kalid and CG Tate for critical reading, and B Honig, LR Forrest, L Adamian, J Liang, CG Tate and DT Jones for providing manuscripts before publication. This study was supported by a grant 222/04 from the Israel Science Foundation to N B-T. SJF was supported by a doctoral fellowship from the Clore Israel Foundation.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rost B, Fariselli P, Casadio R: **Topology prediction for helical transmembrane proteins at 86% accuracy**. *Protein Sci* 1996, **5**:1704-1718.

2. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**:567-580.

3. White SH: **The progress of membrane protein structure determination**. *Protein Sci* 2004, **13**:1948-1949.

4. Petrey D, Honig B: **Protein structure prediction: inroads to biology**. *Mol Cell* 2005, **20**:811-819.

5. • Forrest LR, Tang CL, Honig B: **On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins**. *Biophys J* 2006, in press.

An evaluation of homology modeling applied to TM proteins that segregated the available TM protein structures into families according to sequence homology, and attempted to predict the structures of proteins using their homologues as templates. It was concluded that those methods that were shown to work well for soluble proteins work equally well for TM proteins.

6. Fanelli F, De Benedetti PG: **Computational modeling approaches to structure–function analysis of G protein-coupled receptors**. *Chem Rev* 2005, **105**:3297-3351.

7. • Oliveira L, Hulsen T, Lutje Hulsik D, Paiva AC, Vriend G: **Heavier-than-air flying machines are impossible**. *FEBS Lett* 2004, **564**:269-273.

An extensive evaluation of modeling approaches applied to GPCRs, particularly to the use of rhodopsin's structure as a template.

8. Popot JL, Engelman DM: **Membrane protein folding and oligomerization: the two-stage model**. *Biochemistry* 1990, **29**:4031-4037.

9. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein**. *J Mol Biol* 1982, **157**:105-132.

10. White SH, Wimley WC: **Membrane protein folding and stability: physical principles**. *Annu Rev Biophys Biomol Struct* 1999, **28**:319-365.

11. Chen CP, Kernytsky A, Rost B: **Transmembrane helix predictions revisited**. *Protein Sci* 2002, **11**:2774-2791.

12. Cuthbertson JM, Doyle DA, Sansom MS: **Transmembrane helix prediction: a comparative evaluation and analysis**. *Protein Eng Des Sel* 2005, **18**:295-308.

13. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R: **The structure of the potassium channel: molecular basis of K^+ conduction and selectivity**. *Science* 1998, **280**:69-77.

14. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, Stroud RM: **Structure of a glycerol-conducting channel and the basis for its selectivity**. *Science* 2000, **290**:481-486.

15. •• Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G: **Recognition of transmembrane helices by the endoplasmic reticulum translocon**. *Nature* 2005, **433**:377-381.

This article reports the use of an experimental system to probe the energetics of the transfer of peptides between translocated and membrane-inserted forms, using an experimental setup very close to physiological conditions. Thus, the authors derive a hydrophobicity scale.

16. • Shental-Bechor D, Fleishman SJ, Ben-Tal N: **Has the code of protein translocation been broken?** *Trends Biochem Sci* 2006, **31**:192-196.

A critique of the thermodynamic quantities obtained by Hessa *et al.* [15••] in their analysis of peptide insertion into the membrane. It is argued that the more polar peptides might be stabilized by other protein components in the experiment, causing the energetic penalty on the transfer for polar amino acid residues to appear lower than it actually is.

17. von Heijne G, Gavel Y: **Topogenic signals in integral membrane proteins**. *Eur J Biochem* 1988, **174**:671-678.

18. Goder V, Junne T, Spiess M: **Sec61p contributes to signal sequence orientation according to the positive-inside rule**. *Mol Biol Cell* 2004, **15**:1470-1478.

19. van Klompenburg W, Nilsson I, von Heijne G, de Kruijff B: **Anionic phospholipids are determinants of membrane protein topology**. *EMBO J* 1997, **16**:4261-4266.

20. Nilsson J, Persson B, von Heijne G: **Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes**. *Proteins* 2005, **60**:606-616.

21. •• Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G: **Global topology analysis of the *Escherichia coli* inner membrane proteome**. *Science* 2005, **308**:1321-1323.

A whole-proteome analysis of the topology of proteins in *E. coli* that are predicted to be transmembrane. The data could serve as a benchmark for future studies and evaluations of topology prediction algorithms. Five out of 601 proteins were identified as having putative dual topology, with

some of the protein copies inserting into the membrane with one topology and others with the reverse topology.

22. Rapp M, Granseth E, Seppala S, von Heijne G, Daley DO, Melen K, Drew D: **Identification and evolution of dual-topology membrane proteins**. *Nat Struct Mol Biol* 2006, **13**:112-116.

23. Ubarretxena-Belandia I, Baldwin JM, Schuldiner S, Tate CG: **Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer**. *EMBO J* 2003, **22**:6175-6181.

24. Ninio S, Elbaz Y, Schuldiner S: **The membrane topology of EmrE — a small multidrug transporter from *Escherichia coli***. *FEBS Lett* 2004, **562**:193-196.

25. Bowie JU: **Solving the membrane protein folding problem**. *Nature* 2005, **438**:581-589.

26. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D: **Progress in modeling of protein structures and interactions**. *Science* 2005, **310**:638-642.

27. Bradley P, Misura KM, Baker D: **Toward high-resolution de novo structure prediction for small proteins**. *Science* 2005, **309**:1868-1871.

28. Kalani MY, Vaidehi N, Hall SE, Trabanino RJ, Freddolino PL, Kalani MA, Floriano WB, Kam VW, Goddard WA III: **The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists**. *Proc Natl Acad Sci USA* 2004, **101**:3815-3820.

29. Trabanino RJ, Hall SE, Vaidehi N, Floriano WB, Kam VW, Goddard WA III: **First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin**. *Biophys J* 2004, **86**:1904-1921.

30. Freddolino PL, Kalani MY, Vaidehi N, Floriano WB, Hall SE, Trabanino RJ, Kam VW, Goddard WA III: **Predicted 3D structure for the human β2 adrenergic receptor and its binding site for agonists and antagonists**. *Proc Natl Acad Sci USA* 2004, **101**:2736-2741.

31. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA III: **Prediction of structure and function of G protein-coupled receptors**. *Proc Natl Acad Sci USA* 2002, **99**:12622-12627.

32. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S: **G protein-coupled receptors: *in silico* drug discovery in 3D**. *Proc Natl Acad Sci USA* 2004, **101**:11304-11309.

33. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M *et al.*: **Predict modeling and *in-silico* screening for G-protein coupled receptors**. *Proteins* 2004, **57**:51-86.

34. • Zhang Y, Devries ME, Skolnick J: **Structure modeling of all identified G protein-coupled receptors in the human genome**. *PLoS Comput Biol* 2006, **2**:e13.
An adaptation of the TASSER algorithm for threading and refinement of protein structures to membrane proteins. The algorithm was validated on several proteins of solved structure, and then applied to predicting the structure of most human GPCRs. The resource of predicted structures is available at http://cssb.biology.gatech.edu/skolnick/files/gpcr/gpcr.html.

35. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D: **Free modeling with Rosetta in CASP6**. *Proteins* 2005, **61**:128-134.

36. • Yarov-Yarovoy V, Schonbrun J, Baker D: **Multipass membrane protein structure prediction using Rosetta**. *Proteins* 2006, **62**:1010-1025.
An adaptation of the Rosetta algorithm for *ab initio* protein structure prediction to membrane proteins. The quality of the predicted models was similar to that obtained for soluble proteins. Full-atom prediction was not attempted because of the computational cost of such implementations in large proteins.

37. • Fleishman SJ, Unger VM, Ben-Tal N: **Transmembrane protein structures without X-rays**. *Trends Biochem Sci* 2006, **31**:106-113.
A review of approaches for modeling TM protein structures based on intermediate resolution data. Some experimental data, particularly from crosslinking, are sometimes found to bias models away from the native state structures.

38. Herzyk P, Hubbard RE: **Automated method for modeling seven-helix transmembrane receptors from experimental data**. *Biophys J* 1995, **69**:2419-2442.

39. Sale K, Faulon JL, Gray GA, Schoeniger JS, Young MM: **Optimal bundling of transmembrane helices using sparse distance constraints**. *Protein Sci* 2004, **13**:2613-2627.

40. Unger VM, Kumar NM, Gilula NB, Yeager M: **Three-dimensional structure of a recombinant gap junction membrane channel**. *Science* 1999, **283**:1176-1180.

41. Hirai T, Heymann JA, Maloney PC, Subramaniam S: **Structural model for 12-helix transporters belonging to the major facilitator superfamily**. *J Bacteriol* 2003, **185**:1712-1718.

42. Baldwin JM: **The probable arrangement of the helices in G protein-coupled receptors**. *EMBO J* 1993, **12**:1693-1703.

43. Enosh A, Fleishman SJ, Ben-Tal N, Halperin D: **Assigning transmembrane segments to helices in intermediate-resolution structures**. *Bioinformatics* 2004, **20**:I122-I129.

44. Adamian L, Liang J: **Prediction of buried helices in multispan α helical membrane proteins**. *Proteins* 2006, **63**:1-5.

45. Baldwin JM, Schertler GF, Unger VM: **An α-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors**. *J Mol Biol* 1997, **272**:144-164.

46. Beuming T, Weinstein H: **Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the TM domains of the oxalate transporter OxlT**. *Protein Eng Des Sel* 2005, **18**:119-125.

47. Fleishman SJ, Harrington S, Friesner RA, Honig B, Ben-Tal N: **An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data**. *Biophys J* 2004, **87**:3448-3459.

48. Hurwitz N, Pellegrini-Calace M, Jones DT: **Towards genome-scale structure prediction for transmembrane proteins**. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**:465-475.

49. •• Fleishman SJ, Unger VM, Yeager M, Ben-Tal N: **A C-α model for the transmembrane α-helices of gap-junction intercellular channels**. *Mol Cell* 2004, **15**:879-888.
A cryo-EM map of the gap junction was used together with evolutionary-conservation and correlated-mutations analyses to predict a model structure of the TM domain. The model puts disease-causing point mutations in structurally packed regions of the model.

50. Fleishman SJ, Yifrach O, Ben-Tal N: **An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels**. *J Mol Biol* 2004, **340**:307-318.

51. Ubarretxena-Belandia I, Engelman DM: **Helical membrane proteins: diversity of functions in the context of simple architecture**. *Curr Opin Struct Biol* 2001, **11**:370-376.

52. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S: **Structure and mechanism of the lactose permease of *Escherichia coli***. *Science* 2003, **301**:610-615.

53. • Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU: **The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors**. *Proc Natl Acad Sci USA* 2004, **101**:959-963.
This analysis finds that in most cases where a proline is not observed in a kinked region of a TM protein structure, the multiple-sequence alignment exhibits a proline in several sequence homologues. This observation provides an approach for predicting the locations of kinks in protein structures.

54. Deupi X, Olivella M, Govaerts C, Ballesteros JA, Campillo M, Pardo L: **Ser and Thr residues modulate the conformation of pro-kinked transmembrane α-helices**. *Biophys J* 2004, **86**:105-115.

55. Lee SY, Lee A, Chen J, MacKinnon R: **Structure of the KvAP voltage-dependent K^+ channel and its dependence on the lipid membrane**. *Proc Natl Acad Sci USA* 2005, **102**:15441-15446.

56. Davidson AL, Chen J: **Structural biology. Flipping lipids: is the third time the charm?** *Science* 2005, **308**:963-965.

57. Ma C, Chang G: **Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli***. *Proc Natl Acad Sci USA* 2004, **101**:2852-2857.

58. Pornillos O, Chen YJ, Chen AP, Chang G: **X-ray structure of the EmrE multidrug transporter in complex with a substrate**. *Science* 2005, **310**:1950-1953.

59. Donnelly D, Overington JP, Ruffle SV, Nugent JH, Blundell TL: **Modeling α-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues**. *Protein Sci* 1993, **2**:55-70.

60. Briggs JA, Torres J, Arkin IT: **A new method to model membrane protein structure based on silent amino acid substitutions**. *Proteins* 2001, **44**:370-375.

61. Tate CG: **Comparison of three structures of the multidrug transporter EmrE**. *Curr Opin Struct Biol* 2006, **16**: this issue.

62. Mottamal M, Zhang J, Lazaridis T: **Energetics of the native and non-native states of the glycophorin transmembrane helix dimer**. *Proteins* 2006, **62**:996-1009.

63. Zhou FX, Cocco MJ, Russ WP, Brunger AT, Engelman DM: **Interhelical hydrogen bonding drives strong interactions in membrane proteins**. *Nat Struct Biol* 2000, **7**:154-160.

64. Choma C, Gratkowski H, Lear JD, DeGrado WF: **Asparagine-mediated self-association of a model transmembrane helix**. *Nat Struct Biol* 2000, **7**:161-166.

65. Arbely E, Arkin IT: **Experimental measurement of the strength of a C^{α}–H. . .O bond in a lipid bilayer**. *J Am Chem Soc* 2004, **126**:5362-5363.

66. Yohannan S, Faham S, Yang D, Grosfeld D, Chamberlain AK, Bowie JU: **A C^{α}–H. . .O hydrogen bond in a membrane protein is not stabilizing**. *J Am Chem Soc* 2004, **126**:2284-2285.

67. Vargas R, Garza J, Dixon D, Hay B: **How strong is the C^{α}–H. . .O=C hydrogen bond?** *J Am Chem Soc* 2000, **122**:4750-4755.

68. Scheiner S, Kar T, Gu Y: **Strength of the C^{α}–H. . .O hydrogen bond of amino acid residues**. *J Biol Chem* 2001, **276**:9832-9837.

69. •• Faham S, Yang D, Bare E, Yohannan S, Whitelegge JP, Bowie JU: **Side-chain contributions to membrane protein structure and stability**. *J Mol Biol* 2004, **335**:297-305.

An analysis of the contributions to stability of individual amino acid residues on helix B from bacteriorhodopsin. It is found that the contribution correlates with the amount of buried surface area rather than the ability to provide hydrogen-bonding interactions, roughly as seen for soluble proteins. Surprisingly, a mutation of a kink-inducing proline to alanine did not decrease stability significantly, and only elicited minor changes in secondary structure.

70. Kraulis PJ: **MolScript: a program to produce both detailed and schematic plots of protein structures**. *J Appl Cryst* 1991, **24**:946-950.

71. Merritt EA, Bacon DJ: **Raster3D: photorealistic molecular graphics**. *Methods Enzymol* 1997, **277**:505-524.

72. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N: **ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information**. *Bioinformatics* 2003, **19**:163-164.

73. Yarov-Yarovoy V, Baker D, Caterall WA: **Voltage sensor conformations in the open and closed states in structural models of K^{+} channels**. *Proc Natl Acad Sci USA* 2006, **103**:7292-7297.

Discussion

The last five years have seen a tremendous increase in the pace of structure determination of TM proteins (M11:Fleishman et al., 2006). However, this increase has not been consistent (see Figure 1 of Introduction), and the rate of structure determination of membrane proteins lags far behind that of soluble proteins. Most importantly, only a handful of TM-protein structures from eukaryotes have been determined so far. These persisting challenges provide the motivation for the methodological developments in TM-protein structure prediction over the last few years (reviewed in M12:Fleishman and Ben-Tal, 2006). Many insights into the stability and the folding process of TM proteins have been obtained over the course of the past few years (reviewed in M12:Fleishman and Ben-Tal, 2006). However, these insights have not yet been translated to significant improvements in *ab-initio* predictions, where the structure of the protein is predicted solely on the basis of its amino-acid sequence. This lag between qualitative understanding and prediction capabilities might stem from the difficulties in inferring general principles from the few available TM-protein structures, and might be alleviated as structures accumulate. Although *ab-initio* structure prediction remains an important long-term goal, the fact that this class of methods have still not been shown to yield close-to-native models (Pellegrini-Calace et al., 2003; Yarov-Yarovoy et al., 2006) strongly implies that in order to generate reliable models one must still make extensive use of experimental data (M11:Fleishman et al., 2006).

The main objective of the work that is the subject of this dissertation has been the development and implementation of tools that would be capable of producing reliable model structures that can be used to plan and interpret experiments on

structure-function relationships in TM proteins. Hence, much of the work has focused on developing methodologies that have proven useful and accurate in the past, such as the use of evolutionary conservation to guide the orientation of membrane-spanning α-helices (M4:Fleishman et al., 2004) and the detection of correlated substitutions in the evolutionary history of protein families to constrain distances between amino-acid positions (M5:Fleishman et al., 2004). I have studied three specific cases that represent three broad functional classes in TM proteins: the receptor tyrosine kinase (RTK) ErbB2 (M1:Fleishman et al., 2002), the gap-junction intercellular channel (M7:Fleishman et al., 2004), and the small multidrug resistance transporter EmrE (M9:Fleishman et al., 2006). Hence, taken as a whole, this dissertation demonstrates the general applicability of the bioinformatics-based structure-prediction tools developed here to diverse classes of TM proteins. This approach does not provide new insights on the protein-folding process, and cannot be used in large-scale structure prediction. However, in contrast to most *ab-initio* methods, which produce models of TM proteins that are still too remote from the native state to be useful for motivating experiments (Pellegrini-Calace et al., 2003; Yarov-Yarovoy et al., 2006), the bioinformatics-based approach has generated models that have a good correspondence with biochemical data as well as with high-resolution structures (Baldwin et al., 1997; M4:Fleishman et al., 2004). In all of our reports on structural models (M9:Fleishman et al., 2006; M1:Fleishman et al., 2002; M7:Fleishman et al., 2004), we have formulated specific hypotheses that can be tested experimentally in order to validate the models and to gain additional mechanistic understanding. To demonstrate how these models can be used to inspire structure-function studies, we have collaborated with experimentalists to produce the first data on interactions that stabilize the TM domain of the gap junction intercellular channel (M8:Fleishman et al., 2006).

The modeling approach reported here treats the membrane-spanning segments as canonical α-helices, in spite of the fact that significant deviations from α-helicity have been observed in membrane-protein structures (see Figure 4 of Introduction) and linked to important mechanistic features (Abramson et al., 2003; Ubarretxena-Belandia and Engelman, 2001). Although there are no general methods that can predict these deviations (M11:Fleishman et al., 2006), studies of TM-protein folds and the factors that produce deviations from helicity can engender some improvements in the quality of the resultant models. In this connection, a recent bioinformatics analysis has shown that the locations of kinks within TM α-helices, might be predictable from sequence features (Yohannan et al., 2004). This demonstration has helped us to position a kink in the model structure of EmrE, and thus to produce a structurally more realistic model (M9:Fleishman et al., 2006).

An important aspect that is lacking in the methods reported in this dissertation is the prediction of locations of sidechain atoms. Reliable modeling of sidechain atoms will be a significant advance because it will allow careful inspection of the physicochemical soundness of the models and provide a framework for direct analysis of the factors that stabilize the protein structure, such as packing and polar interactions. Since bioinformatics-based methods produce near-native models (Baldwin et al., 1997; M4:Fleishman et al., 2004), it might be expected that they would serve as a convenient platform for full-atom modeling using energy minimization. Although this notion has been explored recently, it has not been shown whether the resultant full-atom models are more accurate than the preliminary Cα-trace models, on which they were based (Beuming and Weinstein, 2004). It should be borne in mind that due to the roughness of the energy landscape of protein conformations even a 1Å RMS difference in the atom positions from the energetically

optimal conformation might engender very high energies (Schueler-Furman et al., 2005); since the bioinformatics-based models vary by 1-3Å from the native state (M4:Fleishman et al., 2004), adding full atoms to the Cα-trace models is not a trivial undertaking. Considering all of the above, one reasonable route in order to improve the quality of the structural models, which has not been tested so far, is to employ the *ab-initio* approach (Yarov-Yarovoy et al., 2006) to search for full-atom models around the conformations predicted by the bioinformatics-based method (M12:Fleishman and Ben-Tal, 2006). This proposition makes use of the strengths of both approaches, and is likely to result in reliable full-atom models at computationally reasonable costs, which is one of the most important goals of the structural biology of membrane proteins.

Bibliography

Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., and Iwata, S. (2003). Structure and mechanism of the lactose permease of Escherichia coli. Science *301*, 610-615.

Baldwin, J. M., Schertler, G. F., and Unger, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J. Mol. Biol. *272*, 144-164.

Beuming, T., and Weinstein, H. (2004). A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics *20*, 1822-1835.

Fleishman, S. J., and Ben-Tal, N. (2006). Progress in structure prediction of alpha-helical membrane proteins. Curr. Opin. Struc. Biol. *in press*.

Fleishman, S. J., Harrington, S., Friesner, R. A., Honig, B., and Ben-Tal, N. (2004). An automatic method for predicting the structures of transmembrane proteins using cryo-EM and evolutionary data. Biophys. J. *87*, 3448-3459.

Fleishman, S. J., Harrington, S. E., Enosh, A., Halperin, D., Tate, C. G., and Ben-Tal, N. (2006). Cryo-EM-based model structure of the bacterial multidrug transporter EmrE. in preparation.

Fleishman, S. J., Sabag, A. D., Ophir, E., Avraham, K. A., and Ben-Tal, N. (2006). The Structural Context of Disease-causing Mutations in Gap Junctions. in preparation.

Fleishman, S. J., Schlessinger, J., and Ben-Tal, N. (2002). A putative activation switch in the transmembrane domain of erbB2. Proc. Natl. Acad. Sci. USA *99*, 15937-15940.

Fleishman, S. J., Unger, V. M., and Ben-Tal, N. (2006). Transmembrane protein structures without X-rays. Trends Biochem. Sci. *31*, 106-113.

Fleishman, S. J., Unger, V. M., Yeager, M., and Ben-Tal, N. (2004). A C-alpha model for the transmembrane alpha-helices of gap-junction intercellular channels. Mol. Cell *15*, 879-888.

Fleishman, S. J., Yifrach, O., and Ben-Tal, N. (2004). An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J. Mol. Biol. *340*, 307-318.

Pellegrini-Calace, M., Carotti, A., and Jones, D. T. (2003). Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. Proteins *50*, 537-545.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. (2005). Progress in modeling of protein structures and interactions. Science *310*, 638-642.

Ubarretxena-Belandia, I., and Engelman, D. M. (2001). Helical membrane proteins: diversity of functions in the context of simple architecture. Curr. Opin. Struc. Biol. *11*, 370-376.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. Proteins *62*, 1010-1025.

Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P., and Bowie, J. U. (2004). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc. Natl. Acad. Sci. USA *101*, 959-963.

Appendices

doi:10.1016/S0022-2836(02)00590-9 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 321, 363–378

JMB

A Novel Scoring Function for Predicting the Conformations of Tightly Packed Pairs of Transmembrane α-Helices

Sarel J. Fleishman and Nir Ben-Tal*

Department of Biochemistry George S. Wise Faculty of Life Sciences, Tel-Aviv University 69978 Ramat-Aviv, Israel

Pairs of helices in transmembrane (TM) proteins are often tightly packed. We present a scoring function and a computational methodology for predicting the tertiary fold of a pair of α-helices such that its chances of being tightly packed are maximized. Since the number of TM protein structures solved to date is small, it seems unlikely that a reliable scoring function derived statistically from the known set of TM protein structures will be available in the near future. We therefore constructed a scoring function based on the qualitative insights gained in the past two decades from the solved structures of TM and soluble proteins. In brief, we reward the formation of contacts between small amino acid residues such as Gly, Cys, and Ser, that are known to promote dimerization of helices, and penalize the burial of large amino acid residues such as Arg and Trp. As a case study, we show that our method predicts the native structure of the TM homodimer glycophorin A (GpA) to be, in essence, at the global score optimum. In addition, by correlating our results with empirical point mutations on this homodimer, we demonstrate that our method can be a helpful adjunct to mutation analysis. We present a data set of canonical α-helices from the solved structures of TM proteins and provide a set of programs for analyzing it (http://ashtoret.tau.ac.il/~sarel). From this data set we derived 11 helix pairs, and conducted searches around their native states as a further test of our method. Approximately 73% of our predictions showed a reasonable fit (RMS deviation <2 Å) with the native structures compared to the success rate of 8% expected by chance. The search method we employ is less effective for helix pairs that are connected *via* short loops (<20 amino acid residues), indicating that short loops may play an important role in determining the conformation of α-helices in TM proteins.

Keywords: empirical energy function; ridges into grooves; transmembrane helices database; tight packing; structure prediction

*Corresponding author

Introduction

Transmembrane (TM) proteins are crucial mediators of cell-to-cell signaling and of transport processes. This makes them attractive targets for drug discovery as well as for improving our understanding of cellular processes. Despite their importance, however, only about a dozen distinct folds of TM proteins have been solved to date by such high-resolution methods as crystallography and NMR. Attempts to determine the structure of this class of proteins by these methods are hampered seriously by technical problems related to their purification and crystallization. It would therefore be advantageous if these technical difficulties could be bypassed, and the structure of these proteins inferred by computational means.

Because the number of TM proteins whose structures have been solved at high resolution is small, an energy-like contact potential cannot be constructed by straightforward statistical means. Our approach has been to construct a quasi-energy scoring function based on qualitative analyses of TM protein structures carried out over the past decade. We hope that this work may be used also

Abbreviations used: GpA, glycophorin A; RTK, receptor tyrosine kinase; TM, transmembrane.

E-mail address of the corresponding author: bental@ashtoret.tau.ac.il; http://ashtoret.tau.ac.il

as an evaluation of the current level of understanding of the factors driving helix association in TM proteins.[1]

Structure prediction in soluble proteins by computational methods is considered extremely difficult, largely because of the variety of possible folds, which implies a vast number of degrees of freedom. In contrast, TM proteins may be grouped into two classes, the α-helix bundle and the β-barrel. This considerably reduces the number of degrees of freedom that determine the structures of these proteins. Here, we concern ourselves only with the α-helix bundle class, which is the only one known to inhabit the plasma membrane.

According to the widely accepted two-stage model,[2] the first step in TM protein folding is the insertion into the membrane of the TM domains as α-helices. Only in the second stage do these helices associate to form helix bundles. (For recent reviews of this and other thermodynamic models of membrane protein folding, see Popot & Engelman[3] and White & Wimley.[4])

One of the implications of the two-stage model is that, overall, the stability of individual TM domains is independent of that of other domains. Hence, prediction of TM protein structure may begin with prediction of TM helix locations on amino acid sequences. The past few years have seen much progress in computational methods devised for this purpose.[5] Algorithms for determining the topology of these segments in the membrane, i.e. for establishing whether the N terminus is inside or outside the cell, have been successful.[6] There is room for improvement in the understanding of this stage of protein folding, but essentially it has been well explored. Here, we reduce the problem of TM protein structure prediction to the problem of predicting the correct packing of rigid α-helices. Deviations from ideal α-helicity, such as kinking and uncoiling, are indeed encountered in TM proteins, and are known to have functional importance.[7] However, since there are no known methods for predicting these phenomena from sequences, we do not address them here.

Some early attempts were made to predict helix orientations in relation to each other by using the hydrophobic moment concept.[8,9] However, in view of the hydrophobic nature of the membrane, the hydrophobic driving force is probably less important in this medium than in soluble proteins, and the hydrophobic moment has proved to be of limited use in TM structure prediction.[10,11] The main driving force for the folding process is thus considered to be the efficient packing of helices.[12]

Attempts have been made to predict the structure of specific TM proteins.[13–20] For high-resolution structure prediction of pairs of TM α-helices Adams *et al*.[18] developed a method based on molecular dynamics, utilizing data derived from mutational analyses. Briggs *et al*.[19] extended this method by using phylogenetic data instead of mutational analyses. Pappu *et al*.[20] showed that the computational load associated with searches in conformational space, using models in atomic detail, may be reduced considerably by the use of a potential-smoothing technique. They demonstrated the competence of the approach by successfully retrieving the structure of glycophorin A (GpA). Based on their experience, we explore the possibility of reducing the computational burden further by using low resolution from the outset. This allows us to carry out an exhaustive search of conformational space, and it enables us to systematically test the method on many examples.

The number of solved TM protein structures is relatively small, and the factors driving helix association in the membrane are still poorly understood.[1] Nevertheless, several studies have offered substantial qualitative insight into TM helix–helix dimerization. It was shown that TM helices are at least as tightly packed as helices of soluble proteins, and that small residues (Ala and Gly) and small hydroxyl-containing residues (Ser and Thr) are often buried deeply in TM proteins.[12,21] An important role was ascribed to Gly in mediating helix–helix contacts in TM proteins.[22,23] In an attempt to overcome some of the limitations that are inherent in the analysis of residue propensities in TM proteins because of the small number of solved TM protein structures, Senes *et al*.[24] carried out a statistically more extensive study on the sequences of TM proteins. Their results reinforce the conclusions of the qualitative studies, and suggest that TM helix interactions are often mediated by β-branched amino acid residues (Ile and Val) and, to a lesser extent, by the γ-branched amino acid residue Leu.

Lemmon & Engelman[25] offer an explanation for these dimerization-related phenomena in terms of the so-called lipophobic effect. They argue that the presence of small residues on the face of the helix leads to the formation of cavities, should the helix interact with the "cylindrical" lipid chains. Cavity formation is considered costly in terms of energy. On the other hand, these cavities may be eliminated by another helix with an accommodating pattern of large and small residues. The β-branched amino acids are thought to be preferable for dimerization because their rotamers are constrained within the context of an α-helix.[26] This reduces the entropy loss that usually accompanies the association of protein parts.

Recently, Senes *et al*.[27] showed that hydrogen bonding, with C^{α} acting as hydrogen donor, may be an important factor in driving helix–helix association in TM proteins. Their analysis provides a thermodynamic justification for the important role of amino acid residues with small side-chains (Ala, Gly, Ser and Thr) in mediating helix–helix dimerization; their small volume makes the backbone atoms more accessible. Recent work has shown that hydrogen bonds between polar side-chains, e.g. Asn–Asn, play a significant role in stabilizing helix association in model TM

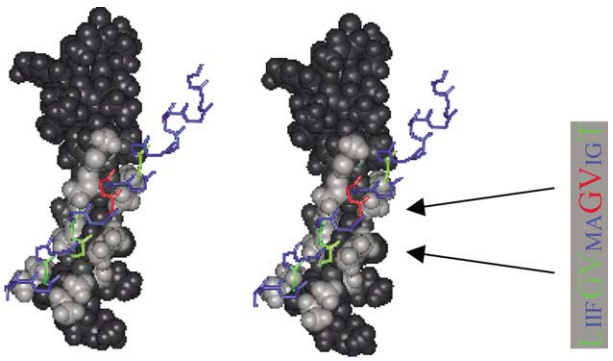


Figure 1. Stereo view of the TM segment of the human homodimer GpA in its native state (PDB code: 1afo). Only the first model of the collection of NMR structures is displayed. One monomer is presented in all-atom CPK rendering and the other shows only the backbone atoms. The light-colored residues on the CPK rendering show the pattern of two ridges on the face of a monomer. The ridge on the right-hand side is formed by the side-chains of Ile76, Val80 and Val84. Gly79 and Gly83 on the other monomer pack against this ridge, allowing for close interactions among backbone atoms. The ridge on the left-hand side is not continuous due to the presence of Gly79 and Gly83. The monomer represented by backbone atoms is colored differentially according to the burial scores of the residues; blue signifies residues that are not buried in the other monomer, green indicates intermediately buried residues, and red indicates residues that are well buried. The burial score of an amino acid residue is computed according to the distance and the angular orientation of the C^{α} and the axis of the other helix (see Methods). The two helices are packed against each other at a crossing angle of about $-40°$. On the right is a portion of the amino acid sequence of GpA's TM domain in one-letter code, colored according to the burial scores *B*. Large letters signify residues that mediate contact. Note that the middle part of the sequence contains the GxxxG motif.[24]

proteins.[28–30] Hydrogen bonds have also been found among side-chains in solved TM protein structures.[31]

In a preliminary study, we examined whether the burial of amino acid residues that are frequently observed at helix–helix interfaces may by itself provide a criterion for determining the native-state packing of two helices. We found that this criterion leads, in almost all cases, to the formation of helix dimers with their principal axes essentially parallel with each other, so that the crossing angle between the two helices is negligible. This contrasts with the findings in many solved TM structures, which show a preference for different crossing angles,[32] and predominantly for a crossing angle of approximately $+20°$.

We therefore employed the "ridges-into-grooves" structural motif described by Chothia *et al.*,[33] which is typical for tightly packed helix pairs, e.g. GpA (Figure 1). Chothia *et al.* argue that α-helices are not smooth cylinders, as their side-chains form protrusions on the face of the helix. Residues that are separated by one, three or four residues on the sequence may form continuous ridges on the face of the helix. These ridges are separated by grooves. In order to maximize the hydrophobic surface area making contact between these side-chains, and to minimize cavities, a ridge on one helix may be inserted into a groove on the face of another. For example, a conformation in which a ridge formed by residues separated by four amino acid residues in the sequence of one helix is associated with a ridge separated by three amino acid residues in the sequence of another helix is called 4–3 class packing. Analysis of this model had helped explain why certain crossing angles predominate in the packing of helices.[33]

Recently, Bowie[34] and Walther *et al.*[35] showed that in the case of globular proteins, much of the preference for the packing angles predicted by the ridges-into-grooves model is actually a result of statistical bias. Their results demonstrate that when the packing-angle propensities are normalized, the preference for these packing angles is not as pronounced as expected from the ridges-into-grooves model alone.[33] Nevertheless, the case of TM helices is different, since the area mediating contact between the helices in TM proteins is usually larger than that in globular proteins.[34] Therefore, steric packing is likely to play a more important role in determining the structure of these proteins.[34] We stress that in tightly packed helices (defined here as helices in which the distance between the principal axes is less than 9 Å), steric packing is likely to play an important role. This is because the very short distance between the principal axes of the helices essentially compels the side-chains in the contact region on one helix to be accommodated by the grooves of the other helix.

Our method in constructing the scoring function for discriminating conformations that would allow tight packing from those that would not was to formulate the qualitative insights pertaining to solved TM protein structures, as presented above, in a quantitative manner. We then tested this formulation against a selection of helix pairs from the solved structures of membrane proteins.

Here, we present our scoring function for contact between TM helices. As a case study, we examine the TM homodimer GpA, and discuss at length our computational results on this protein in the light of empirical mutation analyses[36] and its structure determination.[26,37] We present our results of searches for optimal structures of 11 TM helix pairs derived from TM proteins of known structure.

The Proposed Model

Our aim in this work was to construct a scoring function to distinguish conformations that allow tight packing of a pair of helices from conformations that impede such packing. The helices are reduced to their C^{α} trace. Our function attaches a score based on the amino acid composition of the

Table 1. The maximum score that can be contributed to the total conformational score by a pair of contacting residues

	Gly	β-Branched (I,T,V)	Small (A,C,S)	Constrained (L,N,P)	Others
Gly	−1	−1	−1/2	−1/4	0
β-Branched (I,T,V)	−1	0	−1/2	0	0
Small (A,C,S)	−1/2	−1/2	−1/2	0	0
Constrained (L,N,P)	−1/4	0	0	0	0
Others	0	0	0	0	0

Residues are grouped according to their steric characteristics, and reported in the one-letter code. I, Ile; T, Thr; V, Val; A, Ala; C, Cys; S, Ser; N, Asn; L, Leu; P, Pro. The category Others includes all other amino acids, which contribute zero to the total score. The values reflect the structural analyses described in Introduction.

helices and the space coordinates of their C^{α} atoms to each conformation of two helices. The function is defined such that its minima are associated with tightly packed conformations. For tightly packed pairs of helices, one of these minima should be the native state.

Our approach in computing the score of a given conformation of helices is to maximize the number of contacts between residues that promote helix interactions and to penalize the burial of large amino acid residues. We score any conformation of a given pair of helices as the sum of two terms: a negative term contributing to the score for contact between pairs of residues known to promote close-packing among helices, and a positive term penalizing the burial of large residues in the interface. The optimal score is thus expected to be a global minimum:

$$\text{Score} = \sum (B^i + B^j) M^{(i,j)} + 10 \sum B^l \ : \ (i,j) \in P, l \in L \tag{1}$$

where P is the set of all pairs of residues forming contact in a given conformation, and B^i and B^j are approximations of the burial of the two residues i and j forming that contact between two different helices, as described in Methods. Values for B range from 0, signifying no burial, to 1, signifying complete burial (Figure 1). L is the set of all amino acid residues l with large side-chains (Arg, His, Lys, Met, Phe, Trp and Tyr) that appear on either helix, and are well buried in the interface between the helices ($B^l \geq 0.9$). We penalize the burial of large residues only if they are buried to a large extent in the interface; in other cases, large residues may often assume accommodating conformers, and not form steric hindrances.

$M^{(i,j)}$ is the maximal score contributed by each pair of amino acid residues when mediating contact between a pair of helices (Table 1). To determine these contributions we considered four classes of amino acid residues: Gly; the small residues (Ala, Cys and Ser); the β-branched residues (Ile, Thr and Val); and residues with constrained side-chains (the γ-branched residues Asn and Leu, and Pro). In the absence of a direct statistical method to calculate the relative contribution of each pair of residues to the formation of contacts among helices, we used only four values (0, $-\frac{1}{4}$, $-\frac{1}{2}$ and -1) to reflect the relative contribution of each pair to dimer formation. These values are a crude approximation of the qualitative data available in the literature and presented in Introduction. Thus, since Gly–Gly and Gly–Val contacts have been shown to be favorable for promoting helix contact formation,[23,24,26,38,39] their respective classes contribute substantially to the overall score.

It was recently suggested that the C^{α}–H···O hydrogen bond[27] is a driving force for TM helix contact formation. By promoting contacts between Gly and small residues such as Ala and Cys, as well as contacts between Gly and the hydroxyl-containing amino acid residues Ser and Thr, the scoring function favors contacts between the C^{α} atom of Gly to either the backbone or side-chain hydrogen-bond acceptors. Interhelical hydrogen bonds among polar side-chains were recently shown to strongly promote association of model helices in the membrane.[1,28–30] Though our scoring function is concerned mainly with tight packing of helix pairs, some of the reported hydrogen-bond interactions are included implicitly, e.g. Ser–Ser, and Ser–Thr. Contacts among residues that do not belong to any of the above mentioned classes make no contribution to the overall score.

Results

Glycophorin A

The TM protein on which we have focused most attention as a representative of tightly packed TM proteins is the human GpA.[40] GpA is a monotopic sialoglycoprotein, which is abundant as a homodimer in erythrocyte membranes (Figure 1). In the past decade, the relationship of its amino acid composition to its dimerization characteristics has been scrutinized by a combination of mutational[36,38,41] and computational analyses.[42,43] Recently, its structure was solved both in micelles[26] and in membrane bilayers.[37] The structure conforms to many of the conclusions derived by the mutational analyses. The essential elements of the dimerization of GpA are interactions between Gly and Val residues. Apart from that, the two helices form the ridges-into-grooves class 4–4 packing motif.[33] Because the movement of the two helices comprising the homodimer is not constrained by

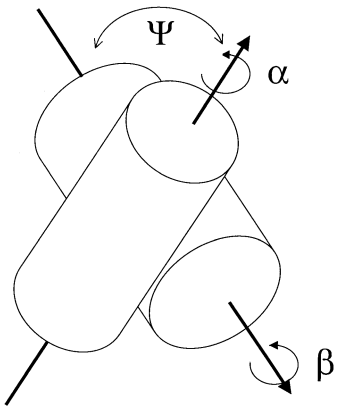


Figure 2. Six degrees of freedom are associated with each helix pair: Three rotational degrees of freedom are indicated in the Figure (α and β represent rotations of the helices around their principal axes; Ψ represents the pair's crossing angle). α and β were set arbitrarily to zero in the native-state conformation. Three translational degrees of freedom set the geometric center of one helix with respect to the other, corresponding to the interhelical distance (y), the height of one helix with respect to the other (z), and a sliding movement of one helix across the face of the other (x). In all our analyses, y was restricted to the value in the native-state conformation of the helix pair. The large arrows mark the principal axes of the helices. For homodimers such as GpA (Figure 1), we may assume that the structure is symmetrical and therefore force $\alpha = \beta$, and $z = 0$.

an interconnecting loop, the helices may be considered free to sample any configuration. This protein is therefore a particularly suitable example on which to test our scoring function.

The native-state conformation of GpA is situated at a global score optimum

The fact that GpA is a homodimer guarantees that its helices will form a nearly symmetrical tertiary structure.[44] By enforcing symmetry, we substantially diminish the number of degrees of freedom examined in our search method to three: the crossing angle (Ψ); one rotational degree of freedom around the axes of the helices (α); and one translation (x) (Figure 2). This allows us to use an extensive search range and obtain a fine resolution.

Using InsightII/Biopolymer (Accelrys, San Diego), we constructed an approximation of the C^{α} trace of GpA as a homodimer composed of two ideal α-helices. We explored the scoring function for this structure using a high-resolution "score surface" (in analogy with the commonly used term potential surface) in a cross-section, such that x is set at its value in the native state (Figure 3). The local minimum shown in Figure 3 at $\alpha = 0°$,

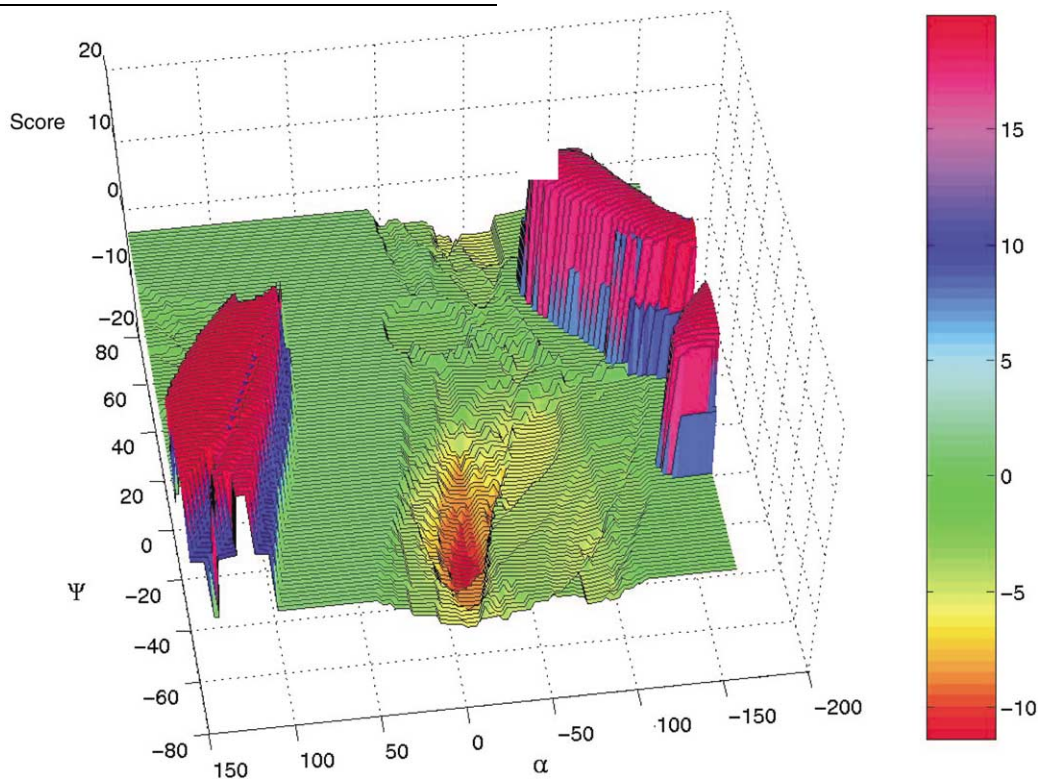


Figure 3. "Score surface" for the homodimeric TM protein GpA around its native state. The structure used for generating this surface is based on two ideal α-helices. The surface was generated by fixing x at its value in the native-state conformation (3.88 Å). While enforcing symmetry ($\alpha = \beta$; see Figure 2), the crossing angle Ψ and the rotation around each monomer's principal axes were modulated. The ranges and step sizes used are: α in the range of $-180°$ to 150° with 2.5° step size; and Ψ in the range of $-60°$ to 80° with 1.5° step size. Note that the native state[26] ($\alpha = 0°$, $\Psi = 40°$) is situated in a score well. The score peaks are associated with the burial of large amino acid residues in the interface. The landscape, as expected, is discontinuous. It is noteworthy, however, that the discontinuity is not very great, and that using a rather coarse step size of 10° for α and 5° for Ψ should suffice to capture its major features.

$\Psi = 40°$ is also the global minimum. Its RMS deviation from the native structure of GpA is 1.41 Å, indicating that ideal α-helices may be used in case the secondary structure is not known with certainty.

It is notable that the region around the minimum in Figure 3 seems to be distinct and large. The scoring function is expected to be discontinuous, but the score surface demonstrates that it is not extremely so, and that searching with a rather coarse resolution (step sizes of 10° for α and 5° for Ψ) would probably not miss the major features of the score surface. We also conducted a search based on the GpA protein database (PDB) structure (1afo)[26] without enforcing symmetry. This resulted in a near-symmetrical structure with an RMS deviation of less than 0.9 Å from the native-state structure (Table 2).

We examined the effect of modulating *y*, representing the distance between the helices' axes of symmetry, on the optimal structure and its score (Table 3). In general, increasing *y* results in a less favorable score. These results can be grouped into three categories on the basis of similarity between the optimal conformations: those obtained for interhelical distances between 6 and 7.5 Å; those between 7.5 and 8 Å; and those between 8 and 9 Å. This indicates that in the cases where the interhelical distance is not known with certainty, configuration space can be searched at two or three distinct interhelical distances, e.g. below 7.5 Å and above 8 Å.

Computational results correlate with empirical mutation analysis

We proceeded to determine whether our method could distinguish mutations that hinder dimer formation from those that do not. For this purpose, we analyzed all 106 non-redundant non-polar point mutations carried out by Lemmon *et al*.[36] On the assumption that the mutant GpA monomers form an ideal α-helical secondary structure, we built a C^{α} trace model for these structures using InsightII/Biopolymer (Accelrys, San Diego). By treating only symmetric conformations, we reduced the number of degrees of freedom to three, as described above. This decrease in the number of degrees of freedom allowed us to carry out a fine-grained search of the structures across much of the conformation space. The search ranges and step sizes are specified in Figure 4.

Figure 4 shows a comparison of our results with the mutation analysis conducted by Lemmon *et al*.[36] We define as a disruption to dimer formation any change in the score of the optimal structure or any deviation of its configuration relative to the optimal structure obtained for the wild-type sequence. Lemmon *et al*.[36] classify their empirical results according to four categories based on the ability of the point mutants to form dimers as well as the wild-type, in significant quantity, in detectable quantity, and no dimer formation. For the purposes of this comparison, we group the classes defined by Lemmon *et al*.[36] as same as wild type and in significant quantity (categories 1 and 2, respectively, in Figure 4), and compare them to our dimer formation class. The other two classes defined by Lemmon *et al*.[36] are compared to our dimer disruption category. It should be noted that our treatment does not allow us to make the distinction made by empirical mutation analyses with regard to the extent of dimer formation.

Our results show a positive correlation with those of Lemmon *et al*.[36] ($r^2 = 0.201$). Significantly, the characteristic mutation Gly83Ala, which abolishes dimer formation *in vitro*,[36] is also disruptive according to our analysis†.

A database of helix pairs

To examine whether our method could distinguish the native-state conformation of pairs of helices from near-native conformations, we analyzed 11 helix pairs chosen from various TM proteins according to automatic procedures as elaborated in Methods (Table 2). The helix pairs were used as they appear in the PDB, i.e. with their deviations from α-helix ideality maintained. We used a five-dimensional lattice to map the conformation space around the native state of each helix pair chosen (Figure 2). The search ranges and step sizes are specified in Table 2.

Our use of a lattice places considerable limitations on the conformation space examined and hence on the range of expected RMS deviation values. We therefore compared the RMS deviation values we obtained for the set of 11 helix pairs to a set of randomly generated structures (Figure 5). We constructed the random set of structures by generating 2000 conformations of the helix pair 1,7 of bacteriorhodopsin (PDB code: 1c3w) throughout the range defined in Table 2 of the five-dimensional lattice with uniform probability. We then calculated the RMS deviation of each of these structures from the helix pair's native-state conformation.

The results presented in Figure 5 indicate that our method yields optimal structures that are close to the native state (<2 Å RMS deviation) in 73% of the cases, as opposed to 8% expected by chance. In some cases (marked with an asterisk) the optimal results are at the end of the search range, and may therefore be underestimates of the real RMS deviation. We did not conduct searches across a larger part of the conformation space, because we treat a helix pair independently of the contacts it forms with other helices. In reality, TM helices in polytopic proteins often form contacts with more than one helix.[21] Such contacts constrain the helix pair from exploring conformations that are far from its native state.

† For supplementary material, see: http://ashtoret.tau.ac.il/~sarel

Table 2. Results of a search around the native state conformation of 11 helix pairs sorted according to the RMS deviation of the optimal structure from the native-state structure

PDB code	Helices	RMS deviation (Å)	Native state crossing angle (deg.)	Interhelical distance (Å)	Δx (Å)	Δz (Å)	α (deg.)	β (deg.)	Ψ (deg.)	Optimal score	Number of interhelical contacts		
											True positives	False negatives	False positives
1afo	1, 2	0.89	−40	7.4	−0.50	0.00	−20	−20	−45.5	−12.63	8	0	0
1fx8	2, 11	0.90	−26	7.2	0.00	−1.50	20	0	−31.5	−8.66	8	0	0
1eul	31, 36	1.65	−46	6.4	−1.00	−1.00	30	−40	−38.5	−10.06	6	3	0
1eul	5, 12	1.65	24	8.2	−1.50	−2.00	20	−30	14	−3.62	4	0	0
1occ[a]	108, 131	1.70	24	7.1	−3.00	0.00	40	−20	31.5	−6.12	6	0	3
1occ[a]	32, 54	1.80	14	8.5	1.00	−1.50	20	−60	10.5	−3.97	4	3	1
1c3w[a]	1, 7	1.88	−6	8.9	3.00	−1.50	30	60	21	−4.12	4	2	3
1bl8[a]	10, 12	1.93	13	8.6	−3.00	2.50	−10	60	10.5	−4.46	4	2	5
1qla[a]	1, 8	2.27	23	8.7	−1.50	2.00	0	−60	0	−2.56	2	2	2
1occ[a]	45, 47	2.88	7	7.6	2.50	−2.50	60	60	17.5	−3.89	2	4	2
1fx8[a]	9, 15	5.22	−41	6.5	0.50	−1.00	60	−20	56	−9.51	2	2	5

For each pair of helices, the interhelical distance was maintained at the value in the native-state conformation and the five other degrees of freedom were modulated. Δx and Δz describe the change in the x and z values in the optimal conformation relative to the native-state conformation. x, z, α, β, and Ψ are the degrees of freedom defined for the search in Figure 2. x and z were modulated between −3 and 3 Å with a step size of 0.5 Å; α, β were modulated between −60 and 60° with a step size of 10°; and Ψ was modulated between −77 and 77° with a step size of 3.5°. It should be noted that the crossing angles (Ψ) of the optimal structures are near their native-state values almost throughout the dataset. The native and predicted structures were also compared visually to determine the number of predicted true positive, false positive and false negative residue contacts. Figure 5 shows a distribution of the RMS deviation values reported here.

[a] Structures whose optimal results are at the ends of the search range. The RMS deviations reported for these structures should be regarded as underestimates.

Table 3. Search results for the structure of glycophorin A (GpA) at different interhelical distances (y)

Interhelical distance(Å)	α(°)	β(°)	Ψ(°)	Δx(Å)	Δz(Å)	Score
6.44	-30	-30	49	1.5	0.5	-13.32
6.94	-20	-20	49	2	-0.5	-13.14
7.44	-20	-20	45.5	1	0	-12.63
7.94	-60	10	28	-2	1	-11.87
8.44	-50	-10	21	-2	1.5	-9.66
8.94	-60	20	-7	-2	-1	-8.59

The parameters, their search ranges, and step sizes are defined as in Table 1. The scoring function simply increases with the interhelical distance. Note that the structures obtained throughout the range of 6.44–7.44 Å are essentially similar; as are the structures between 7.94 Å and 8.44 Å.

It is noteworthy that the success of our method is not restricted to a particular packing class. Most helix pairs examined in Table 2 are packed according to the 4–3 class packing,[33] which is more prevalent in TM proteins.[32] Nevertheless, our method shows considerable success with these as well as with pairs packed in the 4–4 packing class.[33]

We also tested a subset of helix pairs whose interhelical distance is beyond the 9 Å range,

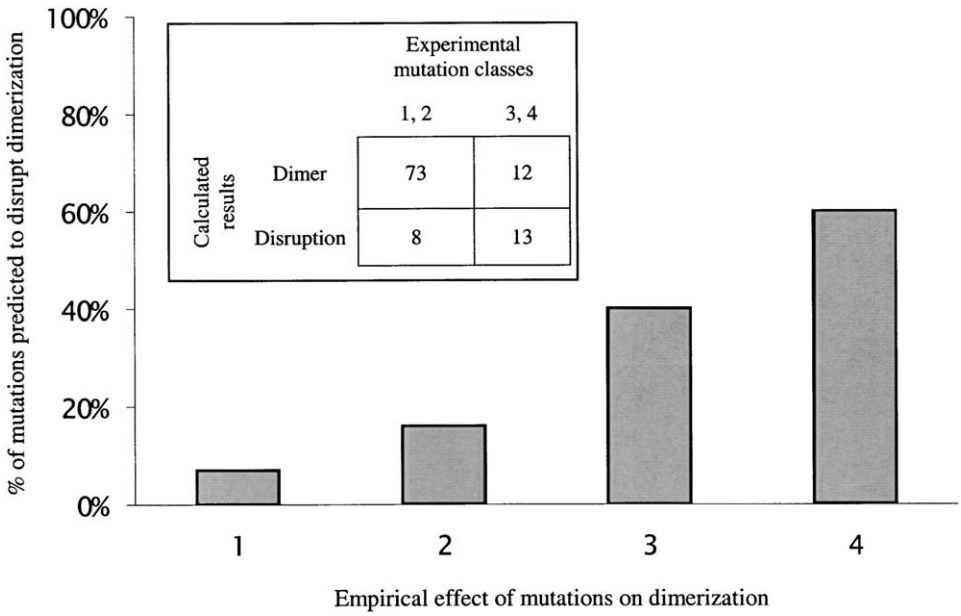


Figure 4. Comparison of computational and empirical results for 106 non-polar point mutations of GpA. The mutations are classified according to four groups, in keeping with the results reported by Lemmon *et al.*[36] Mutations in group 1 dimerized as well as the wild-type; those in group 2 dimerized significantly; group 3 mutants dimerized in detectable amounts; and group 4 mutations showed no detectable dimerization. Inset: a Table showing the correlation between the experimental results reported by Lemmon *et al.*[36] (horizontal) and our computational results (vertical). The correlation coefficient r^2 for these data is 0.201. Significantly, our computational results found only a small percentage (<10%) of mutations in classes 1 and 2 to be disruptive, whereas a large percentage (>50%) of the mutations in classes 3 and 4 were predicted to be disruptive. The search ranges and step sizes used with each mutant were Ψ in the range of −75 to 75° with a step size of 3°. The rotation around the principal axes (α = β) was carried out throughout the range 0 to 360° with a step size of 5°. x was searched in the range of −15 to 15 Å with a step size of 0.5 Å.

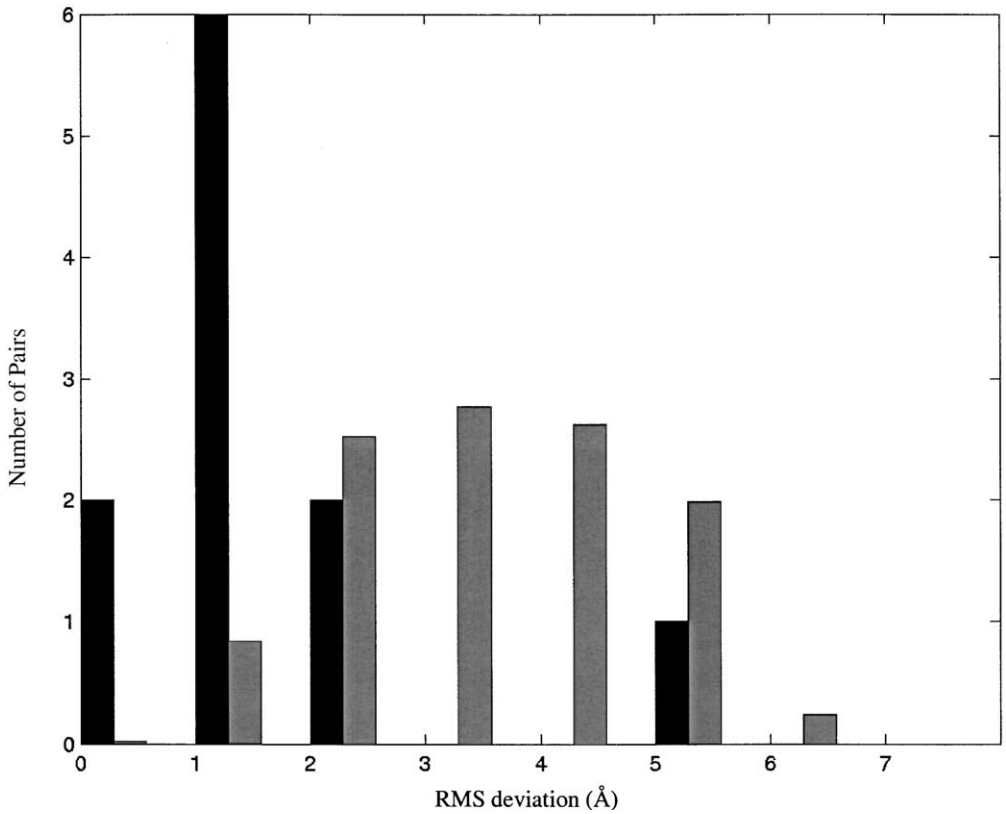


Figure 5. Distribution of RMS deviation values for the search results in Table 4 compared to a random set of structures. Dark bars indicate the distribution of RMS deviation values of a selection of 11 helix pairs to their native-state conformations. Light bars indicate the distribution of RMS deviation values expected by chance (see the text). The expected values were normalized according to the number of helix pairs. Note that 73% of the optimal structures obtained are within 2 Å of the native-state conformation in contrast to the expected value of approximately 8%.

but our results in those cases were much poorer (data not shown). We conclude that the interactions captured by this method are related more closely to those of tightly packed helix pairs.

Discussion

This work had two related goals. The first was to demonstrate the value of a simple rule; small residues go inside, for structure prediction in membrane proteins. To this effect, we used the rule in a simple though exhaustive search method and tested it in 11 carefully chosen TM helix pairs found in the dataset of 11 membrane proteins of known structure (see Methods). As discussed below, the results demonstrated the predictive power of this simple rule. However, they also showed that certain problems in the current implementation of the methodology need to be resolved for it to be potent in structure prediction in polytopic TM proteins. The second goal was to demonstrate the predictive power of the current methodology for tightly packed TM proteins such as GpA.

Ideally, a method for predicting the packing of TM helices would be based on calculating the free energy change occurring upon helix association (ΔG_{ass}). A step in this direction was recently taken by MacKenzie *et al.*[45] who used the data of Lemmon *et al.*[36] concerning the effect of point mutations on the dimerization of GpA to construct an energy-like function for predicting the effects of mutations on dimerization of GpA.

The approach we use here to derive the score function differs from that used by MacKenzie *et al.*[45] in some important respects. It is much humbler, in that it is aimed at discriminating only those conformations of helix pairs in TM proteins that are tightly packed from those that are not. At the same time, it is more ambitious, in that it is derived from general structural considerations, and should therefore be applicable, in principle, to all tightly packed TM helix pairs. The scoring function of Table 1, which constitutes the basis of our method, is a rudimentary construction reflecting contemporary knowledge of tightly packed TM helices; essentially no attempt was made to fit the values to improve the predictions. As anticipated, our results (Figure 4) do not correlate with the data of Lemmon *et al.*[36] as well as do the results of MacKenzie *et al.*[45] ($r^2 = 0.201$ and $r^2 = 0.760$, respectively). We were encouraged to note, however, that the optimal structure of GpA obtained in our calculations was within less than 1 Å RMS deviation from the native structure (Table 2).

To demonstrate the general applicability of our approach, i.e. the predictive power of the simple rule small residues go inside, we compiled a data-set of 11 helix pairs from TM proteins whose 3D structure is known (Table 2). Since we do not consider the current methodology suitable for structure prediction in polytopic TM proteins, we used the known 3D structure of each helix rather than using the corresponding canonical α-helix. The RMS deviations between the optimal and native structures (Table 2) thus reflect purely the quality of the score function, rather than a mixture of the score function with deviations from ideal α-helicity. Our calculations produced conformations that agreed with the known structures significantly better than expected by chance alone (Figure 5). However, a detailed analysis of the success rate is inherently complicated. It is possible that interactions that are not taken into account in our method, such as interactions with other helices, determine the stability of a particular helix pair within the context of the intact protein. In any event, these carefully chosen helix pairs represent only 28% of tightly-packed (distance of 9 Å or less between the principal axes of the helices) helix pairs in the TM proteins in our database, and approximately 10% of all helix pairs forming significant contact in the membrane (see Methods). This illustrates the restricted range of helix pairs our method can currently tackle.

Further complications in analyzing the success rate in Figure 5 arise from the search method we used. Our goal here was to demonstrate the potential use of the scoring function in TM protein structure predictions. Therefore, we did not treat α-helices that deviate markedly from ideality, e.g. pronounced kinks. These deviations may eliminate certain conformations and allow others. Our method is further limited by the fact that we do not model the interconnecting loops. Short loops place a considerable constraint on the conformation space that a pair of helices may explore. This restriction is therefore an important consideration in structure prediction. In the data set presented in Table 2 we included only helix pairs that are at least 20 residues apart from each other (thus excluding approximately 40% of relevant helix pairs from our analysis). Because of the constraints imposed by short loops on the conformations available for a helix pair to explore, the excluded pairs are capable, at least in principle, of exploring a rather restricted range in conformation space. The success rate indeed dropped significantly when helix pairs connected by shorter loops were analyzed also (data not shown). Relaxing these two limitations by using a more sophisticated search methodology may make it possible to study many more helix pairs.

To avoid introducing steric constraints into our calculations, we maintained the distance (y) between the principal axes of the helices at its value in the native state. Obviously, when *de novo* prediction of tertiary structure is attempted, y would also need to be modulated. Our results for the modulation of y (Table 3) indicate, however, that when *de novo* prediction is attempted, it may be possible to set y at two or three different values and obtain different optimal structures. In any case, by using a more detailed model, in which each residue is represented as two or three interaction sites, these limitations may be eliminated altogether.

The overall picture emerging from studies of TM helix dimers is that the specific factors driving contact formation among helices are qualitatively different in various ranges of interhelical distance. Eilers *et al*.[46] recently showed that the distribution of pairwise contacts between helices separated by large distances is different from that of tightly packed pairs of helices. As an example, interactions among aromatic residues, which are known to stabilize contact between some helix pairs,[47] are unlikely to occur in the dimerization region of helix pairs whose axes are separated by short distances, whereas backbone–backbone interactions are not possible when the interhelical distance is large. Other types of contacts that stabilize and specify TM protein structure may emerge in the future.[1] Thus, the scoring function presented in Table 1 may be considered as a basis for improvement as more knowledge about these factors accumulates.

Overall, despite the simplicity of our approach, the results demonstrate that it captures the salient features driving association between tightly packed TM helices. Significantly, the approach we employed uses a lower resolution than that of other methods for TM structure prediction.[18–20] However, in GpA, which is so far the only case in which direct comparison between the different approaches is possible, our results are comparable with those obtained by other methods. In our opinion this shows, above all, that the efforts to develop predictive tools for the tertiary structure of proteins can harness the knowledge derived from structural analyses of proteins in a straightforward manner. Furthermore, the lower computational burden associated with such low-resolution computations allows us to treat asymmetric helix pairs.

In conclusion, while our approach appears to capture at least certain aspects of tight packing between TM helices, many changes need to be introduced for our method to be robust in TM protein structure prediction. Nevertheless, the results indicate that the method can be used for structure predictions of TM dimers that resemble GpA, where the limitations described above are of secondary importance. One important class of proteins for which this is the case is the receptor tyrosine kinases (RTK).[48] It is well known that a critical step in the activation of these receptors is dimerization, and recent evidence has indicated that at least in some cases, e.g. ErbB2 (HER2), this dimerization is mediated by a specific interface on its TM domain.[49] Significantly, the

Sternberg–Gullick motif,[50] which is believed to promote dimerization of TM domains in RTKs, is similar in its general features to the GxxxG motif driving the dimerization of GpA.[24] Recently, Mendrola *et al*.[51] showed *in vivo* that the TM domains of ErbB receptors dimerize in cell membranes, and that the Sternberg–Gullick motifs are mediators of this dimerization. Our initial calculations on the TM domains of ErbB homo- and heterodimers match these observations (unpublished results).

Methods

The search method

Conformation space

We used the coordinates of the C^{α} traces of the individual helix pairs that were selected on the basis of the criteria specified below. Different conformations were examined by the scoring function defined in Equation (1). As with any two-rigid-body system, any configuration of two helices is defined completely in terms of three translational and three rotational degrees of freedom (Figure 2).

The computational load of the score calculations is relatively low. We therefore used an exhaustive search method rather than other search heuristics. The scoring function is not suitable for modulation of the interhelical distance (y); if the helices were brought closer together it would simply increase the score of a favorable conformation, regardless of steric clashes that would probably form in reality. We therefore searched conformation space, while maintaining the interhelical distance at the value given by the native-state packing of the helices. This reduces the number of degrees of freedom from six to five. We therefore mapped configuration space onto a five-dimensional lattice, such that each coordinate defines a unique conformation of a helix pair.

Comparison of minima with the native state

We estimated the dissimilarity between the score minima obtained and the conformation of the native state by calculating the RMS deviation between the C^{α} trace of the predicted conformation and that of the native-state conformation using InsightII (Accerlrys, San Diego). We compared by visual inspection the interhelical contacts formed in the native state conformation with those formed in our predicted structure, and classified the latter as true positive, false negative, or false positive contacts (Table 2). Comparison with the RMS deviation measured between the native state and the predicted structures reveals a good correlation between the two criteria. It also shows that a cut-off RMS deviation value of 2 Å constitutes a reasonable threshold, below which the predicted structures fit well with those of the native state.

Implementation

The search for the lowest score was implemented in MatLab (MathWorks, Natick MA), using completely vectorized code to improve performance, and run on parallel Origin 2000 *SGI* processors. The main computational load is the determination of the score of a conformation. This averaged approximately 14 ms per conformation on each of the Origin 2000 processors.

Quantifying the burial of each amino acid residue

The score function defined in equation (1) is based on quantification of the burial of amino acid residues that mediate contact between the helices. In measuring the extent of burial B^i of amino acid residue i we consider two criteria. The first is the distance between the residue and the principal axis of the other helix; the smaller the distance, the more deeply buried the residue. The second criterion is the orientation of the amino acid with respect to the principal axis of the other helix. The more the amino acid residue is directed towards the other helix, the better its burial (Figure 1).

Formally, we consider two parameters: the distance D^i between amino acid residue i and the axis of the other helix, and the angular orientation A^i of amino acid residue i with respect to the axis of the other helix. We define the burial of an amino acid residue as the intersection of these two criteria:

$$B^i = S(D^i)S(A^i) \tag{2}$$

where $S(D^i)$ and $S(A^i)$ are transformations of the distance and angular criteria as defined below.

The effect of increasing the distance or the angular orientation of an amino acid residue on its burial is quantified as a sigmoidal transformation. Clearly, the burial of a residue at close contact and the correct orientation are not much altered by small changes, as is the burial of an amino acid residue that is poorly buried. However, at a certain cut-off distance and orientation, the extent of a residue's burial changes rapidly.

We therefore use a sigmoidal relation of the form:

$$S(D^i) = \frac{1}{\left(\frac{D^i}{t}\right)^p + 1} \tag{3}$$

for the distance, and a similar expression for $S(A^i)$.

The above sigmoidal function produces values ranging from 0, signifying no burial, to 1, signifying complete burial. It approaches unity for small values of D^i and zero for large D^i. Note that $S(D^i = t)$ is 0.5, and that p controls the smoothness of the sigmoid, i.e. for large p the function approaches the form of a step function, in which the step occurs at $D^i = t$. Thus, t and p control the position of the threshold, where the function assumes half-value, and the contour of the function, respectively.

We used an ideal model of a helix pair, whose axes are separated by approximately 7.5 Å, to examine different parameter combinations. Thus, we found the parameter values $t = 60°$ and $p = 4$ to be suitable for transformation of the angle A^i. For transformation of the distance, we first subtract 4.3 Å from the value of D^i calculated for the distance between the amino acid residue and the axis of the other helix. This value approximates the smallest distance possible between an amino acid residue and another helix (the radius of an α-helix to its C^{α} atoms is 2.3 Å plus 2 Å for two exclusion radii), and produces a value of 1 for $S(D^i)$ if the amino acid residue is as close as possible to the axis of the other helix. The parameter values chosen for the transformation of the distance are $t = 2.5$ Å and $p = 6$. Thus the two

transformations for amino acid i are:

$$S(D^i) = \frac{1}{\left(\frac{D^i - 4.3}{2.5}\right)^6 + 1} \quad (4)$$

$$S(A^i) = \frac{1}{\left(\frac{A^i}{60}\right)^4 + 1} \quad (5)$$

where A^i and D^i are given in units of degrees and Å, respectively.

It should be noted that B^i (equation (2)) is sensitive to the choice of t and p values in these S relations. Thus, changes in these parameters may lead to substantial differences in the burial function, and hence in the scoring function defined in equation (1).

Measuring the distance and angular orientation of each residue with respect to the helix opposing it

To allow for some deviations from α-helical ideality, we employed a method presented by Chothia *et al.*[33] for defining a local helical axis rather than the global one. Local axes coincide with the actual curvature of the helical axis. Due to local deformations, the curvature may differ in places from that of the helix's principal axis. The local helical axis v^i of residue i is defined as the cross-product of the vectors Q^i and Q^{i+1}:

$$v^i = Q^i \times Q^{i+1} \quad (6)$$

where:

$$Q^i = C^i + C^{i+2} - 2C^{i+1} \quad (7)$$

and C^i is the position vector of the C^α of residue i. At the helix terminus, the local axes are defined as extensions of the last local axis calculated according to equation (6).

For each residue i we determine the space coordinates of a point p^i nearest to it on the helical axis, according to the method of Walther *et al.*[52] by calculating the geometric center of four consecutive C^α coordinates around i (C^{i-1} to C^{i+3}). Points on the local axis in both termini of the helix are calculated by extending a vector of length 1.5 Å, corresponding to the average helical rise, in the direction of the local axis calculated for those termini, v, defined in equation (6).

The distance D^i between residue i and the axis of the opposing helix is defined in our method as the distance between i and the nearest point p^j on the opposing helix's axis. The angular orientation of i (A^i) with respect to the other helix is then measured as the residue's orientation with respect to p^j. For a residue i, let us formally define a set P of all the points on the axis of the helix opposing residue i as defined above. The distance D^i between residue i and the axis of the opposing helix is defined as the distance between this residue and a point $p^j \in P$:

$$D^i = |C^i - p^j| \quad (8)$$

where p^j, a point on the helix axis, is defined as:

$$p^j = \min_{p \in P}(|C^i - p^j|) \quad (9)$$

The angular orientation of residue i with respect to the axis of the other helix is then defined as the angle formed between two vectors: $p^i - C^i$, a vector in the direction assumed in space by residue i, and $p^j - p^i$, the vector connecting residue i to the axis of the other helix.

Finding the dimerizing residues

An important implication of the ridges-into-grooves structural motif is that contact between tightly packed helices is mediated by amino acid residues that are relatively close to each other on the sequence,[33] i.e. the residues forming contact are all contained in a stretch of not longer than ten residues. This is corroborated, at least partially, by the results of the experiments reported by Mingarro *et al.*[41] which showed that an insertion mutation incorporating four Ala residues in the middle of the dimerization motif of the human GpA does not extend the length of its dimerization motif. We used this implied condition as a criterion for deciding which residues actually form close contact. It is interesting to note that without this criterion, the optimal structures obtained by the method consist of helix pairs with their principal axes parallel with each other (results not shown). This is in accordance with the argument made by Chothia *et al.*[33] that the assumption that helices are smooth cylinders leads to parallel orientations of helix pairs as the most favorable conformation.

To find such a stretch of buried amino acid residues, we examine windows of ten residues on the sequence of each helix for all relatively buried amino acid residues ($B^i \geq 0.2$). Of these windows, we pick the one in which the total burial of its residues is maximal. Formally, let us define W as the set of all contiguous ten residue stretches on each of the helices. We first look for $w' \in W$ such that:

$$w' = \max_{w \in W}\left(\sum_{i \in w\ :\ B^i \geq 0.2} B^i\right) \quad (10)$$

Then, the residues that form the contact between the helices w_{con} are the residues within w' whose burial score B^i indicates that they are well buried ($B^i \geq 0.2$):

$$w_{con} = \{i \in w'\ :\ B^i \geq 0.2\} \quad (11)$$

We thus obtain two such sets of buried residues, one for each helix in the pair.

Finding the pairs of residues that mediate contact

We were interested in finding a set of pairs P of residues, one from each of the w_{con} terms defined above, that form contact between the two helices. Two residues (i, j), such that i is located on one helix and j on the other, are said to form contact if both are buried (i.e. they are both members of the sets w_{con} defined above), and the distance between i and j is not greater than 5.5 Å. This cut-off should be regarded as rather low, since a choice of larger values, e.g. 6 Å, leads at times to structures conforming to class 3–3 helix packing, which were not identified using the assay described by Chothia *et al.*[33] Almost all contacts between residues in this 3–3 class conformation were relatively long-range (over 5.5 Å).

Construction of a database of helix pairs from solved TM protein structures

To test the validity and the applicability of our scoring function, we set up a dataset of helix pairs from the solved structures of TM proteins. The dataset was constructed using tailor-made programs written in MatLab (MathWorks, Natick MA) and Perl, which are available from our website†, and can easily be modified to suit

† http://www.ashtoret.tau.ac.il/usarel

Table 4. Proteins used in this work and their PDB identifiers

Protein name	PDB identifier
Bacteriorhodopsin	1c3w
Calcium ATPase	1eul
Cytochrome *c* oxidase	1occ
Fumarate reductase	1qla
Glycerol facilitator	1fx8
Glycophorin A	1afo
Light-harvesting complex II	1lgh
Mechanosensitive channel	1msl
Photosynthetic reaction center	1prc
Potassium channel	1bl8
Rhodopsin	1f88

other analytical needs besides those described here. We applied strict criteria for the inclusion of pairs of helices in our dataset. Briefly, we constructed an initial data set of 39 non-redundant helix pairs, whose interhelical distance is within the range of 6–9 Å. A further restriction on this data set is that pairs of helices are excluded if they are tilted against each other. This restriction is imposed because the effective contact area made by tilted helix pairs is usually rather small.[33] Of these 39 pairs, ten were eliminated because one or both of the helices did not conform to strict α-helicity; a further 14 pairs were eliminated because the pair constituted sequence neighbors separated by fewer than 20 amino acid residues; and four more pairs were eliminated after visual inspection because the interface actually formed by the helices was judged to be small, or produced by kinking and coiling of the helices. We obtained a total of 11 helix pairs, which are presented in Table 2.

Data

We obtained 11 structures of TM proteins from the Protein Data Bank (PDB‡) (Table 4). The helical parts in each protein were determined automatically according to the data supplied in the PDB, where available, and taken from the literature in the case of 1afo,[26] for which the data were not included in the PDB. All other parts of the proteins were ignored.

Elimination of α-helices that are far from canonical

With the object of excluding from our analysis any helices that deviate significantly from ideal α-helicity, we determined the characteristic helical rise and radius according to the structure of bovine cytochrome *c* oxidase (PDB code 1occ). The average helix radius is 2.52 Å ($\sigma = 0.14$ Å) and the average helical rise is 1.56 Å ($\sigma = 0.09$ Å). These values are comparable to those obtained by Walther *et al.*[52] For each helix, we also calculated the global geometric center G and a vector in the direction of its principal axis $\mathbf{u}$.

The search method for optimal conformations is sensitive to deviations from ideal α-helicity. We therefore eliminated substantial deviations from α-helicity by using a 99% confidence limit around the helical rise and radius. Only helices whose rise and radius fell within both limits were allowed into the subsequent analysis.

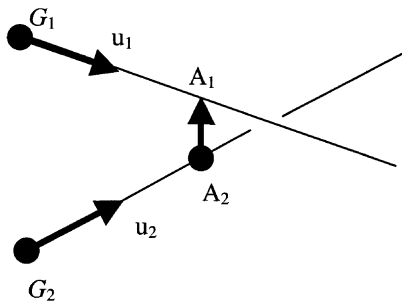


Figure 6. A representation of the method used to find the points of closest approach on two helical axes. G_1 and G_2 are the two helical geometric centers; $\mathbf{u}_1$ and $\mathbf{u}_2$ are two unit vectors pointing in the direction of the helical axes; A_1 and A_2 are the two points of closest approach that we seek. The scheme is adapted from Sunday†.

In addition, we wanted to guarantee that the helical rise and radius were maintained throughout each helix. We therefore selected only those helices in which the standard deviations of the rise and radius did not exceed twice the value of the standard deviation derived above for that parameter.

Selection of pairs of helices making contact

Many of the PDB structures we analyzed contained oligomers of the same subunit. To avoid redundancy, we identified all duplicate helix pairs according to their sequences, and eliminated them. In this way, we obtained, for each protein structure, a non-redundant set of helix pairs that are close to ideal α-helicity. For each possible pair of helices in these sets, we calculated the points of closest approach and the distance between their axes according to the method of Sunday† (Figure 6).

We treat the axes of the helices as infinite lines in 3D space. We know the coordinates of a point on each of the axes (the helix's geometric center G) and the direction of the helix axis ($\mathbf{u}$). Let us mark the two geometric centers as G_1 and G_2, and unit vectors in the direction of their respective helical axes as $\mathbf{u}_1$ and $\mathbf{u}_2$. Let us also mark A_1 and A_2, the respective points of closest approach, which we seek. Then:

$$A_1 = G_1 + s\mathbf{u_1} \qquad \text{and} \qquad A_2 = G_2 + t\mathbf{u_2} \tag{12}$$

where s and t are scalars. By definition, the line connecting A_1 and A_2 is uniquely orthogonal to the two axes, i.e.:

$$\mathbf{u_1}\boldsymbol{w_c} = 0 \qquad \text{and} \qquad \mathbf{u_2}\boldsymbol{w_c} = 0 \tag{13}$$

where $\boldsymbol{w_c} = A_1 - A_2$. Another way of formulating $\boldsymbol{w_c}$ is:

$$\boldsymbol{w_c} = \boldsymbol{w_0} + s\mathbf{u_1} - t\mathbf{u_2} \tag{14}$$

where $\boldsymbol{w_0} = G_1 - G_2$. Substituting equation (14) into the two simultaneous equations defined in equation (13) we obtain:

$$\begin{aligned} (\mathbf{u_1}\mathbf{u_1})s - (\mathbf{u_1}\mathbf{u_2})t &= -\mathbf{u_1}\boldsymbol{w_0}, \qquad \text{and} \\ (\mathbf{u_2}\mathbf{u_1})s - (\mathbf{u_2}\mathbf{u_2})t &= -\mathbf{u_2}\boldsymbol{w_0} \end{aligned} \tag{15}$$

‡ http://www.rcsb.org

† http://geometryalgorithms.com

For compactness, let us mark $a = \mathbf{u_1}\mathbf{u_1}$, $b = \mathbf{u_1}\mathbf{u_2}$, $c = \mathbf{u_2}\mathbf{u_2}$, $d = \mathbf{u_1}w_0$ and $e = \mathbf{u_2}w_0$. We can solve equation (15) for s and t:

$$s = \frac{be - cd}{ac - b^2} \text{ and } t = \frac{ae - bd}{ac - b^2} \quad (16)$$

By substituting s and t obtained from equation (16) into equation (12), we finally arrive at the points of closest approach on both helices. In cases where the denominator $ac - b^2$ is zero, the two axes are parallel and the distance between them is simply the distance between a point on one axis and the other axis.

This method thus allows us to limit our dataset to those helices whose distance does not exceed 9 Å. Apart from divulging the distance between the principal axes of the helices, this method allows us to determine whether the points of closest approach (A_1 and A_2) fall inside or outside the span of the helices. We regard pairs whose points of closest approach fall outside the helix span as tilted against each other, forming little if any contact. These pairs are therefore automatically eliminated from the list.

Contact-forming helices of TM proteins are often sequence neighbors.[32] In our initial set of 39 non-redundant helix pairs with interhelical distance in the range of 6–9 Å, we found 15 (38%) that are separated by fewer than 20 amino acid residues. In a preliminary study, we found that our method works considerably better for pairs of helices that are separated by 20 or more residues on the sequence. This is because short loops do not allow the helix pair to explore conformation space freely.[52] We therefore removed from the subsequent analysis all pairs of helices that are connected *via* such short loops.

The structures of all helix pairs were then inspected visually to eliminate helices that were kinked, coiled, or tilted against each other. Helices that exhibited considerable deviations from ideal α-helicity at their ends were split manually or shrunk to produce α-helices closer to the ideal.

To recapitulate: we automatically compiled a non-redundant data set comprised of pairs of helices forming close contact (6–9 Å) that do not deviate considerably from α-helicity, are not tilted against each other, and are separated by loops of 20 or more amino acid residues. We then manually pruned those pairs whose helices were tilted against each other. We also eliminated the ends of helices that deviated from α-helicity.

Calculation of average helix parameters

We computed the average helix rise and radius for each helix by an extension of the method described above for determining the space coordinates of points on the helical axis. For each helix, the average helical rise was computed by taking the average of the distances between subsequent points on the helical axis. The average helical radius was computed by taking the mean of the distances between the space coordinates C^i of residue i and the point on the helical axis p^i associated with it. In helices that contained 20 or more amino acid residues, we disregarded the three terminal residues at both ends, where deviations from ideal α-helicity often occur.

Acknowledgments

We thank Mark A. Lemmon for providing us with a complete list of the dimerization results of GpA point mutants. We acknowledge helpful discussions with Isaiah T. Arkin, Dalit Bechor-Shental and Doron Chema. We thank Shirley Smith for editing the manuscript. This work was supported by a Research Career Development Award from the Israel Cancer Research Fund. We acknowledge the Bioinformatics Unit and the Computation Center at Tel Aviv University for providing us with infrastructure.

References

1. Bowie, J. U. (2000). Understanding membrane protein structure by design. *Nature Struct. Biol.* **7**, 91–94.
2. Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, **29**, 4031–4037.
3. Popot, J. L. & Engelman, D. M. (2000). Helical membrane protein folding, stability, and evolution. *Annu. Rev. Biochem.* **69**, 881–922.
4. White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365.
5. von Heijne, G. (1996). Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* **66**, 113–139.
6. Tusnady, G. E. & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489–506.
7. Ubarretxena-Belandia, I. & Engelman, D. M. (2001). Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* **11**, 370–376.
8. Rees, D. C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science*, **245**, 510–513.
9. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142.
10. Pilpel, Y., Ben-Tal, N. & Lancet, D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **294**, 921–935.
11. Stevens, T. J. & Arkin, I. T. (1999). Are membrane proteins inside-out proteins? *Proteins: Struct. Funct. Genet.* **36**, 135–143.
12. Eilers, M., Shekar, S. C., Shieh, T., Smith, S. O. & Fleming, P. J. (2000). Internal packing of helical membrane proteins. *Proc. Natl Acad. Sci. USA*, **97**, 5796–5801.
13. Hirokawa, T., Uechi, J., Sasamoto, H., Suwa, M. & Mitaku, S. (2000). A triangle lattice model that predicts transmembrane helix configuration using a polar jigsaw puzzle. *Protein Eng.* **13**, 771–778.
14. Zhdanov, V. P. & Kasemo, B. (2001). Folding of bundles of alpha-helices in solution, membranes, and adsorbed overlayers. *Proteins: Struct. Funct. Genet.* **42**, 481–494.
15. Taylor, W. R., Jones, D. T. & Green, N. M. (1994). A method for alpha-helical integral membrane protein fold prediction. *Proteins: Struct. Funct. Genet.* **18**, 281–294.

16. Tuffery, P. & Lavery, R. (1993). Packing and recognition of protein structural elements: a new approach applied to the 4-helix bundle of myohemerythrin. *Proteins: Struct. Funct. Genet.* **15**, 413–425.
17. Baldwin, J. M., Schertler, G. F. & Unger, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* **272**, 144–164.
18. Adams, P. D., Arkin, I. T., Engelman, D. M. & Brunger, A. T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nature Struct. Biol.* **2**, 154–162.
19. Briggs, J. A., Torres, J. & Arkin, I. T. (2001). A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins: Struct. Funct. Genet.* **44**, 370–375.
20. Pappu, R. V., Marshall, G. R. & Ponder, J. W. (1999). A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nature Struct. Biol.* **6**, 50–55.
21. Adamian, L. & Liang, J. (2001). Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311**, 891–907.
22. Arkin, I. T. & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta*, **1429**, 113–128.
23. Javadpour, M. M., Eilers, M., Groesbeek, M. & Smith, S. O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys. J.* **77**, 1609–1618.
24. Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921–936.
25. Lemmon, M. A. & Engelman, D. M. (1994). Specificity and promiscuity in membrane helix interactions. *Quart. Rev. Biophys.* **27**, 157–218.
26. MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, **276**, 131–133.
27. Senes, A., Ubarretxena-Belandia, I. & Engelman, D. M. (2001). The Calpha–H···O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc. Natl Acad. Sci. USA*, **98**, 9056–9061.
28. Choma, C., Gratkowski, H., Lear, J. D. & DeGrado, W. F. (2000). Asparagine-mediated self-association of a model transmembrane helix. *Nature Struct. Biol.* **7**, 161–166.
29. Zhou, F. X., Cocco, M. J., Russ, W. P., Brunger, A. T. & Engelman, D. M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nature Struct. Biol.* **7**, 154–160.
30. Dawson, J. P., Weinger, J. S. & Engelman, D. M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.* **316**, 799–805.
31. Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins: Struct. Funct. Genet.* **47**, 209–218.
32. Bowie, J. U. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780–789.
33. Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215–250.
34. Bowie, J. U. (1997). Helix packing angle preferences. *Nature Struct. Biol.* **4**, 915–917.
35. Walther, D., Springer, C. & Cohen, F. E. (1998). Helix–helix packing angle preferences for finite helix axes. *Proteins: Struct. Funct. Genet.* **33**, 457–459.
36. Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**, 12719–12725.
37. Smith, S. O., Song, D., Shekar, S., Groesbeek, M., Ziliox, M. & Aimoto, S. (2001). Structure of the transmembrane dimer interface of glycophorin A in membrane bilayers. *Biochemistry*, **40**, 6553–6558.
38. Lemmon, M. A., Flanagan, J. M., Hunt, J. F., Adair, B. D., Bormann, B. J., Dempsey, C. E. & Engelman, D. M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J. Biol. Chem.* **267**, 7683–7689.
39. Russ, W. P. & Engelman, D. M. (2000). The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.* **296**, 911–919.
40. Furthmayr, H., Galardy, R. E., Tomita, M. & Marchesi, V. T. (1978). The intramembranous segment of human erythrocyte glycophorin A. *Arch. Biochem. Biophys.* **185**, 21–29.
41. Mingarro, I., Elofsson, A. & von Heijne, G. (1997). Helix–helix packing in a membrane-like environment. *J. Mol. Biol.* **272**, 633–641.
42. Treutlein, H. R., Lemmon, M. A., Engelman, D. M. & Brunger, A. T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry*, **31**, 12726–12732.
43. Adams, P. D., Engelman, D. M. & Brunger, A. T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins: Struct. Funct. Genet.* **26**, 257–261.
44. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*, 2nd edit., Garland Publishing Inc, New York.
45. MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix–helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA*, **95**, 3583–3590.
46. Eilers, M., Patel, A. B., Liu, W. & Smith, S. O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.* **82**, 2720–2736.
47. Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L. *et al.* (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69–77.
48. Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell*, **103**, 211–225.
49. Burke, C. L. & Stern, D. F. (1998). Activation of Neu (ErbB-2) mediated by disulfide bond-induced dimerization reveals a receptor tyrosine kinase dimer interface. *Mol. Cell. Biol.* **18**, 5371–5379.
50. Sternberg, M. J. & Gullick, W. J. (1990). A sequence motif in the transmembrane region of growth factor receptors with tyrosine kinase activity mediates dimerization. *Protein Eng.* **3**, 245–248.

51. Mendrola, J. M., Berger, M. B., King, M. C. & Lemmon, M. A. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J. Biol. Chem.* **277**, 4704–4712.

52. Walther, D., Eisenhaber, F. & Argos, P. (1996). Principles of helix–helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* **255**, 536–553.

Edited by G. von Heijne

(Received 19 February 2002; received in revised form 8 June 2002; accepted 11 June 2002)

Comment on "Network Motifs: Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks"

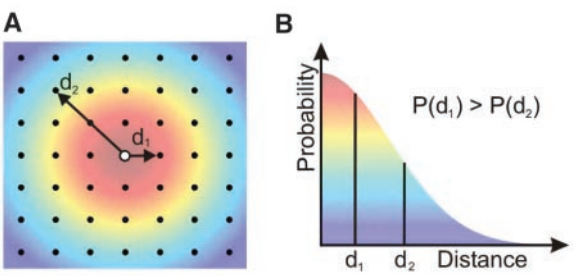
Recently, excitement has surrounded the application of null-hypothesis approaches for identifying evolutionary design principles in biological, technological, and social networks (*1*–*13*) and for classifying diverse networks into distinctive superfamilies (*2*). Here, we argue that the basic method suggested by Milo *et al*. (*1*, *2*) often has limitations in identifying evolutionary design principles.

The technique is relevant for any network that can be notated schematically as a directed graph of N nodes (for example, representing neurons) and a set of edges or links between pairs of nodes (for example, synaptic connections). In particular, the approach is able to identify unusually recurring "network motifs"—patterns of interconnections among a small number of nodes (typically three to five) that are significantly more common in real networks than expected by chance (*1*–*13*). Overabundance is taken to mean that the motifs are the manifestation of evolutionary design principles favored by selection in biological or synthetic systems (*1*–*8*).

In statistical parlance, the basic method [which has a long history in theoretical biology (*10*–*13*)] tests a "random null hypothesis" by statistically comparing the distribution of motifs in an observed network with that found in a computer-generated ensemble of appropriately randomized networks. Over and above the realistic constraint that the degree distribution of incoming and outgoing links to every node must be maintained (*14*), the edges in the randomized network are connected between nodes completely at random and without preference. Such randomized networks are considered null in that their structure is generated by a process free of any type of evolutionary selection acting on the network's constituent motifs. Rejection of the null hypothesis has thus, in many studies, been taken to represent evidence of functional constraints and design principles that have shaped network architecture at the level of the motifs through selection (*1*–*13*).

However, the method outlined above can lead to the wrong interpretations if the underlying null hypothesis is not posed carefully. For example, using this approach, Milo *et al*. (*1*) identified several significant network motifs in the neural-connectivity map of the nematode *Caenorhabditis elegans*. However, in the case of *C. elegans*, neurons are spatially aggregated and connections among neurons have a tendency to form in local clusters (*15*). Two neighboring neurons have a greater chance of forming a connection than two distant neurons at opposite ends of the network. This feature of local clustering, though, is not reflected in the baseline randomized networks used by Milo *et al*. (*1*, *2*), in which the probability of two neurons connecting is completely independent of their relative positions in the network (Fig. 1). The test is not null to this form of localized aggregation and will thus misclassify a completely random but spatially clustered network as one that is nonrandom and that has significant network motifs.

Analysis of a "toy network" (Fig. 1) illustrates what can go wrong. In this network, the nodes are randomly connected preferentially to nearby neighbors, but with a probability that falls off for more distant neighbors (a Gaussian distribution is used). Although the toy network is built devoid of any rule selecting particular motifs for their functions, we find that the same network motifs identified by Milo *et al*. (*1*) for *C. elegans* are present, and the random null hypothesis must be rejected (Fig. 1). Thus, the statistically significant motifs found in *C. elegans* (*1*) are more likely to be the result of the inherently localized partitioning of the nematode's connectivity network than a property that emerges from the action of evolutionary forces selecting particular motifs for their specific functions. It is not our goal in this case to construct a model that realistically captures the distribution of motifs as found in *C. elegans*, but merely to explore the implications of choosing an incomplete null model. Having said that, it is still somewhat surprising that the simple "toy model" reproduces the distribution (significance profile) of all three-node motifs with reasonable realism.



C

Motif	Feed-forward	Bi-fan	Bi-parallel
Observed	42	347	65
Mean	27.1	18.83	43.32
Std.	5.33	8.63	6.4
Z score	2.79	38.02	3.38

Fig. 1. (**A**) Construction of Gaussian "toy network." We used a 30 by 30 grid of 900 nodes. Edges were added on the basis that the probability P of two nodes being connected reduces with the distance d between them. Thus, $P(d_1) > P(d_2)$ when $d_1 < d_2$. This feature will be present to some degree in neural networks such as that of *C. elegans* (*14*). (**B**) Color-coded probability $P(d)$ of connecting to a node as a function of distance for the Gaussian toy network. (**C**) Overrepresentation of motif patterns in the Gaussian toy network. We focused on three motif patterns (feedforward, bi-fan, and bi-parallel) found in (*1*) to be significantly overrepresented in the *C. elegans* neural map. The observed number of each motif, as counted in the Gaussian toy network of (A), was compared with the mean number of motifs counted in 2000 randomized networks (*14*). For all three cases, the Z scores $\left(\frac{\text{Observed - Mean}}{\text{Std.}}\right)$ were larger than 2, signifying that the null hypothesis can be rejected and all motifs are significantly overrepresented.

Many biological and synthetic networks, such as the metabolic and transcription networks (*9*) and the World Wide Web (*16*), are characterized by a scale-free distribution of links to every node. In scale-free networks, the probability of a node having k connections obeys the power law $p(k) \sim k^{-\gamma}$ (with $\gamma > 2$)—that is, most nodes have few connections and a few nodes have many connections. It has been argued (*16*) that some biological scale-free networks are generated by the rule of preferential attachment, a

rule that in itself does not include any type of selection for or against particular motifs. We have used two variants of the preferential-attachment rule (*17*) to generate toy networks, and have then analyzed their motif structure. Using the first variant, we find that the feedforward loop (FFL, shown schematically in Fig. 1C) is always significantly overrepresented ($>2\sigma$ from the mean) compared with the randomized null networks, which implies that the motif has been favored by evolution. In contrast, for the second variant, the FFL is significantly underrepresented, which indicates that the motif has been disfavored. As such, the actual process by which a network is generated, even if it is free of selection for or against particular motif functions, can strongly bias an analysis that seeks to determine the quantitative significance of motifs.

Similar problems arise when applying the approach to studying complex ecological food webs (*10*–*13*). In these systems, each node represents an organism, and an edge between two organisms indicates that one feeds on the other. Food webs are nonrandom structures largely governed by trophic relationships; randomizing feeding links in a food-web network and testing the random null hypothesis serves at best only to trivially prove this point. Unsurprisingly, Milo *et al*. (*1*) find nonrandom overrepresented network motifs that are consistent with simple trophic relationships such as predator–prey–resource interactions. From an ecological perspective, little can be learned from rejecting the possibility that the food web is random. It may be worthwhile in the future to seek ways of posing the null hypothesis in a more sophisticated ecological framework (*10*–*13*).

In summary, for all of these examples, the null hypothesis test suggested the involvement of evolutionary design principles in random toy networks that were generated without the involvement of any fitness-based selection process. The only possible resolution to this problem is to reformulate the test in a manner that is able to identify functional constraints and design principles in networks and to discriminate them clearly from other likely origins, such as spatial clustering.

There is no denying that the network randomization approach has a certain charm in facilitating diverse and multidisciplinary cross-system comparisons in the search for common universal network motifs, design principles, and characteristics defining distinctive network superfamilies (*1*, *2*). Indeed, this approach has stimulated theoretical and experimental work that has demonstrated the utility of certain motifs in tasks such as information processing (*18*, *19*). However, given the dangers sketched above, any cross-system analysis may be very fragile and will be prone to comparing network motifs that are found to be statistically significant because of an ill-posed null hypothesis. Moreover, the method described in (*2*) forces a common reference frame for comparing motif significance profiles (distribution and significance of all possible motifs) of networks, even if they are of different origins—for example, neural networks, for which a null model based on spatial clustering may be justified, versus transcription networks, for which such a null model would be unsuitable. Thus, comparisons mediated through a common but inappropriate reference frame may give the wrong impression that different networks are in fact similar with respect to their motif significance profile. Clearly, these techniques need to be developed further before design principles can be deduced with confidence (*20*).

Yael Artzy-Randrup*
Biomathematics Unit
Department of Zoology
Tel Aviv University
Ramat Aviv, Tel Aviv, 69978, Israel

Sarel J. Fleishman*
Department of Biochemistry
Tel Aviv University

Nir Ben-Tal
Department of Biochemistry
Tel Aviv University

Lewi Stone†
Biomathematics Unit
Department of Zoology
Tel Aviv University
**These authors contributed equally to this work.*

†*To whom correspondence should be addressed. E-mail: lew521@yahoo.com*

References and Notes

1. R. Milo *et al*., *Science* **298**, 824 (2002).
2. R. Milo *et al*., *Science* **303**, 1538 (2004).
3. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
4. T. I. Lee *et al*., *Science* **298**, 799 (2002).
5. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nature Genet*. **31**, 64 (2002).
6. G. C. Conant, A. Wagner, *Nature Genet*. **34**, 264 (2003).
7. N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, *Bioinformatics*, in press.
8. S. Wuchty, Z. N. Oltvai, A. L. Barabasi, *Nature Genet*. **35**, 176 (2003).
9. N. Guelzim, S. Bottani, P. Bourgine, F. Kepes, *Nature Genet*. **31**, 60 (2002).
10. E. F. Connor, D. Simberloff, *Ecology* **60**, 1132 (1979).
11. A. Roberts, L. Stone, *Oecologia* **83**, 560 (1990).
12. L. Stone, A. Roberts, *Oecologia* **85**, 74 (1990).
13. N. J. Gotelli, G. R. Graves, *Null Models in Ecology* (Smithsonian Institution Press, Washington, DC, 1996).
14. Randomized networks were generated by randomly shuffling edges in the graph while leaving the number of ingoing and outgoing edges of every node unchanged. This was achieved (*1*–*13*) by randomly selecting a pair of edges, U→V and X→Y, and switching them to U→Y and X→V if these edges did not already exist. The switching procedure was implemented typically thousands of times to create a randomized matrix. The random switching ensures that the probability of two nodes being connected is effectively independent of the distance between them.
15. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, *Philos. Trans. R. Soc. London Ser. B* **314**, 1 (1986).
16. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
17. The preferential-attachment rule builds up networks so that each new node added to the system connects preferentially to well-connected nodes (hubs). In the first variant of the rule we used, toy networks were built up with older nodes directed to newer ones; in the second variant, edges were directed randomly.
18. S. Mangan, A. Zaslaver, U. Alon, *J. Mol. Biol*. **334**, 197 (2003).
19. S. Mangan, U. Alon, *Proc. Natl. Acad. Sci. U.S.A*. **100**, 11980 (2003).
20. A possible resolution to the problem in the context of molecular networks would require waiting for the availability of sufficient data on networks from several organisms. It might then be possible to test (*8*) whether some functions, such as the transcriptional control of a particular protein, in diverse organisms are preferentially governed by a certain motif, which in turn would strengthen the case for the role of selection.
21. We thank A. Ayali for his very helpful advice and suggestions. We are grateful for the support of the James S. McDonnell Foundation and the Internal Tel Aviv University Research Fund.

19 April 2004; accepted 21 July 2004

תקציר

חלבונים חוצי-ממבראנה מהווים 15-30% מהגנום, אך פחות מ-1% מהמבנים במאגר הנתונים על חלבונים (Protein Data Bank – PDB). הפער הזה מדגיש עד כמה מהווה פיתרון המבנה של חלבונים ממבראנליים בעיה קשה. הקושי חמור ביותר בהקשר של חלבונים אאוקריוטים, שעבורם קביעת המבנה נותרה בעיה כמעט בלתי-פתירה, על-אף ההתקדמות הניכרת שנעשתה בקביעת המבנה של חלבונים בקטריאליים. יש לציין, שלמעלה ממחצית מהחלבונים הטרנס-ממבראנליים מאאוקריוטים הם ללא הומולוגים בקטריליים, כך שסביר, ששנים רבות יחלפו עד אשר מספר ניכר של חלבונים טרנס-ממבראנליים מאאוקריוטים יפורסם. עד אז, גישות אינטגרטיביות, המשלבות אנליזות ביוכימיות וחישוביות עם מידע מבני ברזולוציה בינונית יספקו מסגרות להבנה מכניסטית של מבנה ותפקוד של חלבונים חוצי-ממבראנה. במסגרת זו, פיתחתי מספר שיטות לחיזוי מבנה בחלבונים חוצי-ממבראנה, העושות שימוש באנליזה פילוגנטית ובמידע ביוכימי ומבני ברזולוציה בינונית. השיטות החישוביות החדשות שפיתחתי במסגרת הדוקטורט ואשר שימשו לחיזוי מבנה כללו אנליזת שימור אבולוציוני לזיהוי של חומצות אמינו קבורות בליבת החלבון, וכן זיהוי של קורלציות באבולוציה של חומצות אמינו כדי לאתר אינטראקציות בין עמדות. בעזרת שיטות אלה ואחרות חקרתי את המבנה ומנגנון הפעולה של הקולטן ErbB2, התעלה הבין-תאית, gap junction, והאנטיפורטר החיידקי EmrE. בכל המקרים הללו, נמצאה התאמה משמעותית בין מידע ניסיוני לבין המבנים, וכן הגענו לתובנות חדשות לגבי תפקוד החלבונים, לרבות אודות תפקודם במחלות של שני החלבונים הראשונים. בהתבסס על המודל של ה-gap junction, הצענו מספר היפותזות על אינטראקציות בין חומצות אמינו, אשר אומתו ניסיונית.

מבנה, תפקוד ותנועה בחלבונים חוצי-ממבראנה: מחקר חישובי של קולטנים, משאבות ותעלות.

תיזה לקראת תואר
"דוקטור בפילוסופיה"

מאת שראל י. פליישמן

הוגש לסנאט של אוניברסיטת תל-אביב
יולי, 2006

העבודה בוצעה בהנחייתו של
פרופ' ניר בן-טל