Review

Israel Journal of Chemistry

DOI: 10.1002/ijch.201200096

ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function

Gershon Celniker,[a, c, f] Guy Nimrod,[a, b] Haim Ashkenazy,[c] Fabian Glaser,[d] Eric Martz,[e] Itay Mayrose,[f] Tal Pupko,[c] and Nir Ben-Tal*[a]

Abstract: Many mutations disappear from the population because they impair protein function and/or stability. Thus, amino acid positions that are essential for proper function evolve more slowly than others, or in other words, the slow evolutionary rate of a position reflects its importance. ConSurf (http://consurf.tau.ac.il), reviewed in this manuscript, exploits this to reveal key amino acid positions that are important for maintaining the native conformation(s) of the protein and its function, be it binding, catalysis, transport, etc. Given the sequence or 3D structure of the query protein as input, a search for similar sequences is conducted and the sequences are aligned. The multiple sequence alignment is subsequently used to calculate the evolutionary rates of each amino acid site, using Bayesian or maximum-likelihood algorithms. Both algorithms take into account the evolutionary relationships between the sequences, reflected in phylogenetic trees, to alleviate problems due to uneven (biased) sampling in sequence space. This is particularly important when the number of sequences is low. The ConSurf-DB, a new release of which is presented here, provides precalculated ConSurf conservation analysis of nearly all available structures in the Protein DataBank (PDB). The usefulness of ConSurf for the study of individual proteins and mutations, as well as a range of large-scale, genome-wide applications, is reviewed.

Keywords: bioinformatics · databases · genomics · protein models · protein structures

1 Introduction

Evolutionary data can often supplement our incomplete understanding of structure-function relationships in proteins, DNA and RNA. For example, methods of protein structure prediction that make use of existing structures, either in the form of full protein templates or short fragments, are generally more accurate than current molecular dynamics simulations and other methods that are based on first principles (perhaps due to limited computer capacity). Evolutionary data are also useful for highlighting important positions in the protein (or nucleic acid): slowly evolving, (i.e., evolutionarily conserved) amino acids in proteins are often important (e.g., reference [1]). Why else would they be conserved if not for maintaining the structure and function, be it catalysis or interaction with ligands, cofactors, DNA/RNA or other proteins? Rapidly evolving amino acids may also be crucial to function. For example, microbial surfaces evolve rapidly to evade host immune defenses, while host defense molecules such as antibody recognition sequences change rapidly in order to keep pace with microbial evasion. Thus, accurate estimates of the evolutionary rate can be very informative to the biologist. The ConSurf methodology and web server provide just that.[2] In the following sections we survey the methodology and review applications within the context of the prediction of protein structure and function and the effect of mutations, as well as systems biology and structural genomics. The term "protein" is used for convenience but the methodology is also applicable to nucleic acids.

[a] G. Celniker, G. Nimrod, N. Ben-Tal
The Department of Biochemistry and Molecular Biology
Tel Aviv University, 69978 Tel Aviv (Israel)
e-mail: NirB@tauex.tau.ac.il

[b] G. Nimrod
Present address: Biolojic Design
Akiva Aria 25, Tel Aviv 6215425 (Israel)

[c] G. Celniker, H. Ashkenazy, T. Pupko
The Department of Cell Research and Immunology
Tel Aviv University, 69978 Tel Aviv (Israel)

[d] F. Glaser
Bioinformatics Knowledge Unit
The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering
Technion – Israel Institute of Technology, Haifa 32000 (Israel)

[e] E. Martz
Department of Microbiology
University of Massachusetts, Amherst, MA (USA)

[f] G. Celniker, I. Mayrose
The Department of Molecular Biology and Ecology of Plants
Tel Aviv University, 69978 Tel Aviv (Israel)

2 Methodology

ConSurf's estimate of evolutionary conservation is based on sequence alone, but the input can be the sequence or 3D structure of the query protein. When the latter is used as input, ConSurf parses the PDB entry to extract the sequence that corresponds to the structure. A detailed description of the methodology is provided in the "Overview" section of the ConSurf web server (http://consurf.tau.ac.il). Given a sequence, ConSurf searches for closely related sequences. Detection of homologous sequences and their alignment is key for accurate estimation of evolutionary conservation. Because no single procedure guarantees success in all cases, ConSurf offers the user a selection of search methods and databases, and the ability to specify criteria for defining similarity. The defaults reflect our experience. Thus, the user may search for related proteins using CSI-BLAST[3] (default) or PSI-BLAST.[4] The user may also manually select the desired sequences from the collected hits, for example to limit the sequences to proteins whose function is identical to the query protein. The hits are clustered, and highly similar sequences are removed using CD-HIT.[5] A multiple sequence alignment (MSA) of the related sequences is constructed using MAFFT (default),[6] PRANK,[7] T-COFFEE,[8] MUSCLE,[9] or CLUSTALW.[10] A phylogenetic tree, reflecting the inferred evolutionary history, is built using the MSA and the neighbor-joining algorithm[11] as implemented in the Rate4Site program.[2a] Position-specific conservation scores are computed using the empirical Bayesian[12] or maximum-likelihood (ML)[13] paradigms. The inference of evolutionary conservation relies on a specified probabilistic model, either for amino acid replacements or nucleic acid substitutions. The server offers a selection of several such models, thus allowing the accurate description of the evolutionary dynamics of both coding and non-coding sequences. The continuous conservation scores are divided into a discrete scale of nine grades, each mapped to a color for visualization. The colors are projected on the MSA and the query sequence/structure. When the query is provided as a sequence, ConSurf outlines a list of related proteins of known structure (sharing at least 35% sequence identity with the query and 50% coverage), if available in the PDB. The user may select one of these, and the conservation grades are presented on the structure. Coloring a 3D structure by conservation is particularly powerful, because it enables identification of clusters of highly conserved (or highly variable) residues in the natively folded protein. Such clusters are more significant than are isolated residues.

Gershon Celniker was born in Russia and raised in Israel, where he completed his undergraduate studies in molecular biology and genetic engineering at the Technion Institute and then a master's degree in bioinformatics and genomics at the Hebrew University of Jerusalem (supervisor: Prof. Amiram Goldblum). His MSc. thesis focused on the study of flexible protein-DNA interactions, using computational methods to investigate DNA recognition processes, docking, and protein flexibility in its encounter with DNA. Gershon participated in the Oscar Getz research program at the Weizmann institute under the supervision of Prof. Doron Lancet and was part of the GeneCards (an integrated database of human genes) research team, focusing on genomics and the alternative splicing process. Currently, he is a bioinformatics developer in the Ben-Tal, Mayrose and Pupko groups at Tel Aviv University, where he is working on the development of bioinformatics tools for various research areas such as molecular evolution and structural biology.

Nir Ben-Tal completed his bachelor's degree in Biology, Chemistry and Physics at the Hebrew University of Jerusalem in 1988, and his DSc. in Chemistry at the Technion in 1993 (advisor: Prof. Nimrod Moiseyev). He later did his postdoctoral training in biophysical chemistry at Columbia University, New York (Prof. Barry Honig's lab). In 1997 he accepted a faculty position at the Department of Biochemistry and Molecular Biology, Tel Aviv University, and became full professor in 2008. Within the area of computational structural biology he specializes in protein structure, function and dynamics, with special interest in membrane proteins. He has coauthored over 110 peer-reviewed publications. In 2010 he jointly authored the textbook *Introduction to Proteins: Structure, Function, and Motion* with Dr. Amit Kessel.

3 Conservation and Importance

Originally, ConSurf was developed as a quick means to highlight functionally important regions on protein surfaces, using the 3D structure of the protein and MSA of homologues.[2b] The correlation between evolutionary conservation and biological importance in biopolymers has been well established in structural biology (e.g., reference [1]). In particular, functional regions (i.e., clusters of amino acids in spatial proximity to each other on the protein surface, which are involved in catalysis and interactions) are often evolutionarily conserved.[15] The most common approach for the detection of functional importance has been based on invariance, i.e., positions that feature the exact same amino (or nucleic) acid throughout the MSA. While invariance is certainly indicative of importance (provided that the list of similar sequences is sufficiently diverse), this strict definition overlooks many important positions. Another popular approach for the

Israel Journal of Chemistry

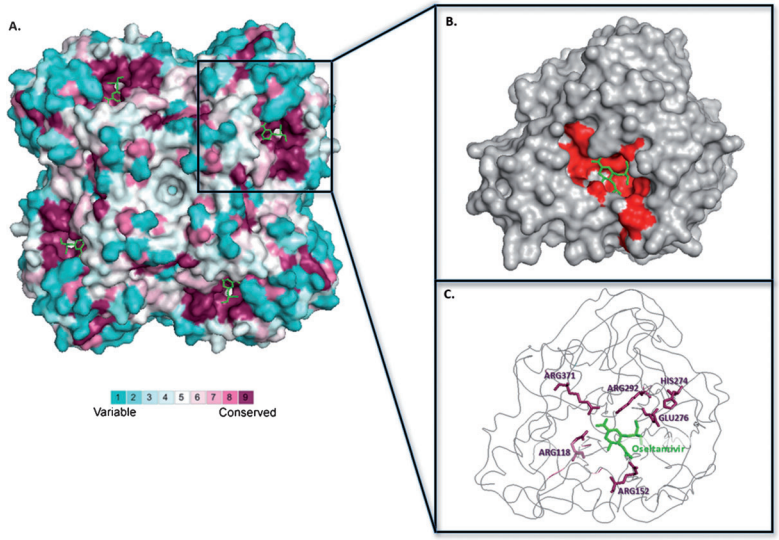

Figure 1. ConSurf analysis of the influenza neuraminidase protein. A) The 3D structure of the tetramer presented using a surface model. Amino acids are colored by their conservation grades using the color-coding bar, with turquoise-through-maroon indicating variable-through-conserved. The figure reveals the functionally important regions. B) A zoomed-in view of one of the monomers bound to the anti-flu drug oseltamivir (trade name: Tamiflu), presenting (in red) the conserved patch of highest likelihood found by PatchFinder.[21] C) A close-up view of the binding site. Highly conserved residues (ConSurf scores of 8 or 9) known to be crucial for binding the sialic acid substrate: Arg118, Arg292 and Arg371 bind the carboxylate, Arg152 interacts with the acetamido substituent, and Glu276 forms hydrogen bonds with the 8- and 9-hydroxyl groups of the substrate. PDB ID 2HU4 was used and the figure was generated using the PyMol[41] script output generated by ConSurf.

detection of functional importance has been based on the degree of consensus amongst the similar sequences concerning the identity of the dominant amino acid, e.g., 80% consensus (invariance corresponds to a consensus of 100%). A more sophisticated approach to estimate conservation is based on the Shannon entropy as a measure of the information content at each position in the homologues.[16] Both approaches can be informative when a large number of similar sequences are available, if they sample sequence space evenly. But what if they do not? For example, what if ninety of a hundred similar sequences are highly similar to each other and only the other ten diverge? The common solution has been to delete all but one of the close sequences and estimate consensus (or entropy). However, this solution discards information, and the result depends critically on the criteria used to define closeness (e.g., sequence identity). In addition, such an approach tends to ignore the physicochemical similarity between different amino acids, as it focuses mainly on searching for identity rather than similarity.

ConSurf alleviates the problem of uneven sampling in sequence space by making explicit use of the evolutionary relationships between the similar sequences, as reflected in the phylogenetic tree, an approach that was first introduced with the development of the evolutionary trace method.[15c] The first ConSurf method, presented a decade ago,[2b] was based on a single phylogenetic tree, obtained by the parsimony principle;[17] the tree topology and the ancestral sequences were reconstructed so as to minimize the number of changes during evolution. The more advanced ConSurf releases that followed have been based on more accurate inference of phylogenies and on explicit continuous-time Markov processes to model sequence evolution. In addition, it was realized that conservation can be estimated by explicit modeling of site-specific evolutionary rate, and that the latter can be reliably estimated using the maximum-likelihood[2a] and Bayesian[2d] paradigms. The statistical robustness of the ConSurf methodology provides not only accurate estimates of the evolutionary rate at each position, but also confidence intervals around these estimates. The server makes it easy to disregard evolutionary rate results that are unreliable (those with excessively large confidence intervals) by coloring them yellow ("caution").

An example of the usefulness of ConSurf for highlighting function is provided in Figure 1. The influenza virus membrane includes two glycoproteins: hemagglutinin and neuraminidase. The former mediates viral entry into the host cell by binding sialic acid receptors, and the latter is responsible for removing the sialic acid to facilitate virus release.[18] The ConSurf calculations on neuraminidase (PDB ID: 2HU4)[19] demonstrate that the functional regions of this enzyme are indeed highly conserved. This is particularly the case for residues that bind the sialic acid substrate (ConSurf scores of 8 or 9): Arg118, Arg292, Arg371, Arg152, His274 and Glu276. His274 is of particular interest because of the emergence of an oseltamivir-resistant mutant, His274Tyr. The larger tyrosine causes

a displacement of Glu276, occluding the drug-binding pocket and reducing oseltamivir's binding affinity, rendering it ineffective against the mutant virus. Unfortunately, the mutant enzyme is still functional because there is still enough room for sialic acid binding.[20]

ConSurf analysis of a protein 3D structure makes it easy to see clusters of highly conserved amino acids in close proximity to each other on the surface. Such clusters are likely to be functional (e.g., the substrate-binding pocket in Figure 1B). But what about more quantitative predictions? For example, how likely is a cluster of, say, five positions with a certain average conservation score, to actually be a functional region? And what is the exact boundary of the region, i.e., which positions are included in the cluster? The PatchFinder methodology and web server were designed in order to answer these questions statistically.[21] Given the protein 3D structure and ConSurf scores, the conservation scores are reshuffled many times and assigned to each position randomly. The results are used to formulate a null hypothesis concerning the likelihood of obtaining a cluster (patch) of a given average conservation. Following the maximum-likelihood approach, up to three clusters of spatially close and highly conserved residues, which were assigned the highest likelihood, are reported. For example, Figure 1B shows the patch of highest likelihood obtained for neuraminidase.

ConSurf is commonly used by the structural community to detect and present functional regions. A recent example is a report of a new structure of the cytoplasmic membrane protein TatC.[22] TatC is the central component of the twin-arginine translocation (Tat) pathway, a prokaryotic protein-transport system. Two evolutionarily conserved regions were revealed at opposite ends of the membrane, and it has been suggested that they mediate TatC interaction with other proteins in the pathway.[22] This is a good example of how ConSurf analysis often succeeds in raising testable hypotheses about protein function.

4 Mutation Design

ConSurf calculations may be used to design mutations in biopolymers with several goals in mind. At a very basic level, mutations in a highly conserved cluster could be designed where the goal is to impede a particular function, thereby attributing it to the cluster. In this context, the amino acid frequencies amongst the related proteins may be useful in designing mutations, as they represent the amino acid repertoire that a certain position can tolerate.

ConSurf analysis may also be used to study specificity within a family of homologous proteins. For example, a subfamily may share a unique binding specificity in a certain region. ConSurf analysis of the sequences in the subfamily may reveal a highly conserved region that is not shared by the rest of the family, suggesting that it is associated with the unique binding specificity of the subfamily. Mutations in members of the subfamily may be used to examine this hypothesis. Again, the amino acid profile in the family may guide the mutagenesis study. In this case it is advisable to choose the type of mutation based on the proteins that share the function versus those that do not.

ConSurf could also be used in rational protein design, for example by suggesting amino acid positions for mutations that could alter an existing function, which is far from trivial.[23] An easier task would be to point out highly variable regions, which often are not important for existing functionalities. These could be used for adding new functions without interfering with existing ones, or for the attachment of labels.[1]

5 Genetic Mutations and Single-Nucleotide Polymorphism

Geneticists often encounter the need to assess the likelihood of a mutation to be associated with a disease. This is particularly difficult with missense mutations, which change the nature of the amino acid leaving the rest of the protein sequence unaltered. ConSurf may aid in the discrimination between neutral and deleterious mutations. The former are more common in variable positions, while the latter are more frequent in conserved positions.[24] However, many exceptions are known, which is why many other qualities of the amino acid position and its vicinity are often required in order to improve the prediction accuracy.[25] Overall, it is very challenging to discriminate between deleterious mutations and harmless single-nucleotide polymorphisms, even when the 3D structure of the protein is known.[24]

6 ConSurf-DB and Systems Biology

ConSurf-DB is a database of precalculated ConSurf conservation profiles covering nearly all protein structures in the PDB.[26] We present here a new release of the database, which now covers 73,278 protein structures. Table 1 provides ConSurf-DB statistics. A detailed description of the updated ConSurf-DB methodology is provided in the "Overview" section in the web server (http://consurfdb.tau.ac.il) and a flowchart is shown in Figure 2. A four-step procedure was used to construct ConSurf-DB:

(i) Generating a non-redundant list of sequences in the PDB: the first step involved scanning the PDB repository to generate a protein sequence list according to the PDB entry and chain ID. Non-redundant structures were extracted from the list using the PISCES web server.[27]

(ii) Finding related sequences and constructing the MSA: a unique procedure was used for building an MSA for each protein, which balanced the need for sequence

Israel Journal of Chemistry

diversity while avoiding the inclusion of non-related proteins as much as possible. For that we relied as much as possible on the SWISSPROT database,[28] a small curated database of annotated proteins, and referred to the larger and noisier Uniref90 database[29] only when necessary. Initially, a CS-BLAST[3] search against the SWISSPROT database was conducted with the goal of detecting at least 50 unique hits. In cases of failure to meet the threshold, we searched the Uniref90 database using CS-BLAST, and CSI-BLAST with three iterations. The list of collected sequences was subsequently filtered by coverage (minimum 80 %) and sequence identity (between 30–95 %). The remaining sequences were filtered again using CD-HIT with a 95 % sequence identity clustering threshold.[5] The decision of whether to proceed with the search for related sequences, or abort and move to the next step, was based on the number of sequences after filtration. An MSA of the sequences was constructed using MAFFT.[6]

(iii) Conservation calculation: the MSA was used to build a phylogenetic tree using the neighbor-joining algorithm[11] as implemented in the Rate4Site[2a] program. Position-specific conservation scores were computed using the Bayesian paradigm[12] and JTT replacement model.[30]

(iv) Results formatting: continuous conservation scores were divided into a discrete scale of nine grades for visualization, from the most variable positions (grade 1, turquoise), through intermediately conserved positions (grade 5, white), to the most conserved positions (grade 9, maroon). Finally, the conservation colors were mapped onto the protein 3D structure and the MSA for visualization.

ConSurf-DB provides the biologist with precalculated conservation profiles of proteins of interest, allowing instantaneous initial evaluation of the results. ConSurf-DB is linked to other databases and interactive tools. One example is Proteopedia,[31] where the ConSurf-DB colored structure can be visualized interactively in Jmol on the

Table 1. Build statistics for the updated version of ConSurf-DB (August 2012).

Total number of chains located within 73,278 protein structures	192,647 chains
Total number of non-redundant chains processed	54,509 chains
First step: CS-BLAST on the SWISSPROT database generated	19,834 MSAs
Second step: CS-BLAST on the UniRef90 database generated	28,536 additional MSAs
Third step: CSI-BLAST (three iterations) on the UniRef90 database generated	2,418 additional MSAs
Number of chains left with less than 50 unique sequences (no calculations)	3,721 chains
Median number of unique sequences collected	142
Minimum and maximum number of unique homologues were set to	50 and 300

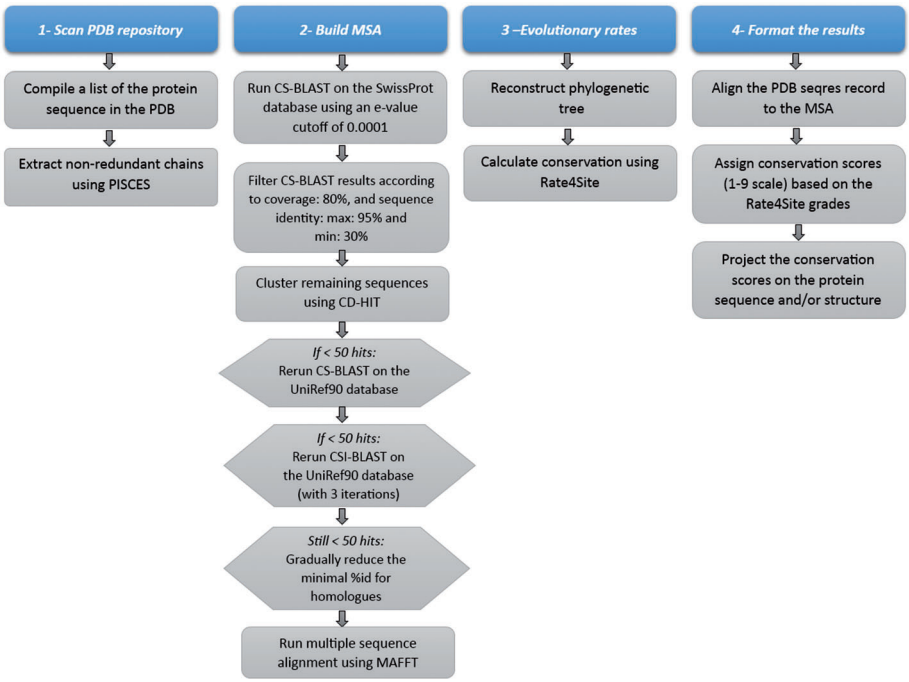

Figure 2. A flowchart of the process used to construct ConSurf-DB. A four-step procedure was used: scanning the PDB, building an MSA, calculating the conservation scores, and formatting the results.

same page as the structure publication title and abstract, identification of ligands and non-standard residues, and other information. Other examples are the PDBsum[32] and MarkUs,[33] a server to navigate sequence-structure-function space.

Convenient as ConSurf-DB is, it is important to remember that it is usually possible to further improve the results for a particular protein of interest by the use of tailor-made procedures for similarity detection, manual selection of related sequences (made easy in the ConSurf web server), as well as other means to reconstruct the alignment or phylogeny.

ConSurf-DB may also be useful for genome-wide and other large-scale studies of proteins where the goal is to deduce general characteristics. In this respect, ConSurf-DB could also be useful within the context of systems biology. It is noteworthy, though, that the ConSurf calculation provides only the relative evolutionary rate at each position with respect to the average within the examined family, as represented in the particular sequence collection used. Because ConSurf gives relative, rather than absolute, conservation scores, it can be used to compare two different protein families only if there is reason to believe that the families evolve at similar rates; one cannot group together all the highly conserved positions in two different protein families. For such purposes, and also to differentiate between purifying, neutral, and adaptive selective patterns, codon-based analyses should be used such as those implemented in the PAML package[34] or the SELECTON web server.[35] The additional information obtained through codon-based methods comes at the expense of a more tedious sequence similarity search and longer running times. Additionally, such analyses are possible only under the assumption that the rate of silent substitutions reflects the rate of neutral evolution and is similar across the studied protein-coding genes.[14]

7 Evaluation of the Quality of 3D Models

The correlation of conservation with solvent-accessibility profiles can be used to evaluate theoretical models of protein 3D structure (protein-folding predictions).[36] The expectation is that the protein core (i.e., buried amino acid positions) would be highly conserved, while the periphery (i.e., exposed positions) would be variable. This idea has been implemented in the ConQuass methodology, which is readily useful for examination of model structures.[37] Systematic examination using various datasets showed that the ConQuass score correlates with the quality of the model structure. In particular, results with a set of 11,686 models of 75 targets from CASP8 (http://predictioncenter.org) showed that when the conservation information is reliable, the method's performance is comparable and complementary to that of the other single-structure quality assessment methods.[37] The same conclusion emerged from the subsequent double-blind examination of ConQuass within the CASP9 competition (http://predictioncenter.org).

8 Structural Genomics: Infer Function from Structure

Due mostly to the worldwide structural genomics effort, we see the emergence of 3D structures of proteins with unknown function or incomplete function annotation.[38] ConSurf analysis may reveal highly conserved surface clusters of amino acids, which are presumably functionally important. In this respect, ConSurf may provide a first step towards function annotation. To this end, we have established the N-Func database of 757 structures of proteins of unknown function and their predicted functional regions.[21b]

9 Challenges

The exponential growth of sequence databases is a blessing but also raises theoretical and practical challenges. Let us start with the theoretical. A large database is advantageous in that it is likely to include many sequences that are truly related to our query protein. The problem is that it is more difficult to find these because of the high false discovery rate. Advanced statistical methods and/or systematic reorganization of sequence databases are required in order to resolve this issue. On the practical level, we will need to develop tools and strategies for dealing with hundreds and thousands of truly related sequences. A key question would be: when the number of truly related sequences exceeds a certain threshold, is it sufficient to estimate conservation based on more simple methods, such as consensus or the Shannon entropy? If not, methods that are based on phylogeny, like ConSurf, will have to be improved both in speed and memory usage.

10 Summary and Outlook

Evolutionary analysis is useful for many purposes. The ConSurf web server has been designed to make it readily accessible to the community. It is particularly useful for proteins, starting from sequence or structure queries, but it is also applicable to the analysis of DNA and RNA.

ConSurf can easily reveal highly conserved, presumably functional, surface regions in proteins. However, additional tools are needed in order to suggest what the function is. Analysis of the physicochemical nature of the conserved regions could provide hints (e.g., protein-protein or nucleotide binding regions). It can also be possible to

infer function by matching of the conserved surface regions with databases of annotated functional regions.[33,39]

It is noteworthy that while evolutionary conservation is indicative of importance, not all functionally important surface regions are highly conserved. As mentioned above, antigen-binding sites of antibodies and major histocompatibility complex (MHC) molecules of the immune system, as well as the surfaces of the proteins of many infectious agents (e.g., influenza hemagglutinin) provide good counterexamples. The hypervariability of these regions is crucial to support their functions.[40]

Acknowledgements

We are thankful to Yana Gofman for many helpful comments and discussions, and for critical comments of many ConSurf users. This work was supported by grant number 3-7935 from the Ministry of Science and Technology. G. C. and H. A. were funded in part by the Edmond J. Safra Center for Bioinformatics at Tel Aviv university. I. M. is supported by the Marie Curie Career Integration grant (FP7-PEOPLE-2011-CIG-293878) and the Israeli Science Foundation grant 1265/12.

References

[1] A. Kessel, N. Ben-Tal, *Introduction to Proteins: Structure, Function, and Motion*, CRC Press, Boca Raton, **2010**.
[2] a) T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, *Bioinformatics* **2002**, *18 (suppl 1)*, S71–S77; b) A. Armon, D. Graur, N. Ben-Tal, *J. Mol. Biol.* **2001**, *307*, 447–463; c) C. Berezin, F. Glaser, J. Rosenberg, I. Paz, T. Pupko, P. Fariselli, R. Casadio, N. Ben-Tal, *Bioinformatics* **2004**, *20*, 1322–1324; d) I. Mayrose, D. Graur, N. Ben-Tal, T. Pupko, *Mol. Biol. Evol.* **2004**, *21*, 1781–1791; e) H. Ashkenazy, E. Erez, E. Martz, T. Pupko, N. Ben-Tal, *Nucleic Acids Res.* **2010**, *38*, W529–W533.
[3] C. Angermüller, A. Biegert, J. Söding, *Bioinformatics* **2012**, *28*, 3240–3247.
[4] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
[5] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, *Bioinformatics* **2010**, *26*, 680–682.
[6] K. Katoh, H. Toh, *Bioinformatics* **2010**, *26*, 1899–1900.
[7] A. Loytynoja, N. Goldman, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10557–10562.
[8] C. Notredame, D. G. Higgins, J. Heringa, *J. Mol. Biol.* **2000**, *302*, 205–217.
[9] R. C. Edgar, *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
[10] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, *Bioinformatics* **2007**, *23*, 2947–2948.
[11] N. Saitou, M. Nei, *Mol. Biol. Evol.* **1987**, *4*, 406–425.
[12] See Ref. [2d].
[13] See Ref. [2a].
[14] N. D. Rubinstein, I. Mayrose, A. Doron-Faigenboim, T. Pupko, *Mol. Biol. Evol.* **2011**, *28*, 3297-3308.
[15] a) X. Gallet, B. Charloteaux, A. Thomas, R. Brasseur, *J. Mol. Biol.* **2000**, *302*, 917–926; b) O. Lichtarge, H. R. Bourne, F. E. Cohen, *J. Mol. Biol.* **1996**, *257*, 342–358; c) O. Lichtarge, H. R. Bourne, F. E. Cohen, *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 7507–7511; d) O. Lichtarge, K. R. Yamamoto, F. E. Cohen, *J. Mol. Biol.* **1997**, *274*, 325–337; e) R. Landgraf, I. Xenarios, D. Eisenberg, *J. Mol. Biol.* **2001**, *307*, 1487–1502; f) A. del Sol, F. Pazos, A. Valencia, *J. Mol. Biol.* **2003**, *326*, 1289–1302; g) W. S. Valdar, *Proteins* **2002**, *48*, 227–241.
[16] C. Sander, R. Schneider, *Proteins* **1991**, *9*, 56–68.
[17] J. Felsenstein, *Methods Enzymol.* **1996**, *266*, 418–427.
[18] R. J. Russell, P. S. Kerry, D. J. Stevens, D. A. Steinhauer, S. R. Martin, S. J. Gamblin, J. J. Skehel, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17736–17741.
[19] R. J. Russell, L. F. Haire, D. J. Stevens, P. J. Collins, Y. P. Lin, G. M. Blackburn, A. J. Hay, S. J. Gamblin, J. J. Skehel, *Nature* **2006**, *443*, 45–49.
[20] P. J. Collins, L. F. Haire, Y. P. Lin, J. Liu, R. J. Russell, P. A. Walker, J. J. Skehel, S. R. Martin, A. J. Hay, S. J. Gamblin, *Nature* **2008**, *453*, 1258–1261.
[21] a) G. Nimrod, F. Glaser, D. Steinberg, N. Ben-Tal, T. Pupko, *Bioinformatics* **2005**, *21 (suppl 1)*, i328–i337; b) G. Nimrod, M. Schushan, D. M. Steinberg, N. Ben-Tal, *Structure* **2008**, *16*, 1755–1763.
[22] S. E. Rollauer, M. J. Tarry, J. E. Graham, M. Jääskeläinen, F. Jäger, S. Johnson, M. Krehenbrink, S.-M. Liu, M. J. Lukey, J. Marcoux, M. A. McDowell, F. Rodriguez, P. Roversi, P. J. Stansfeld, C. V. Robinson, M. S. Sansom, T. Palmer, M. Högbom, B. C. Berks, S. M. Lea, *Nature* **2012**, *492*, 210–214.
[23] M. Goldsmith, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **2012**, *22*, 406–412.
[24] G. Wainreb, H. Ashkenazy, Y. Bromberg, A. Starovolsky-Shitrit, T. Haliloglu, E. Ruppin, K. B. Avraham, B. Rost, N. Ben-Tal, *Nucleic Acids Res.* **2010**, *38*, W523–W528.
[25] Y. Bromberg, B. Rost, *Nucleic Acids Res.* **2007**, *35*, 3823–3835.
[26] O. Goldenberg, E. Erez, G. Nimrod, N. Ben-Tal, *Nucleic Acids Res.* **2009**, *37*, D323–D327.
[27] G. Wang, R. L. Dunbrack Jr. , *Bioinformatics* **2003**, *19*, 1589–1591.
[28] The UniProt Consortium, *Nucleic Acids Res.* **2012**, *40*, D71–D75.
[29] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, *Bioinformatics* **2007**, *23*, 1282–1288.
[30] D. T. Jones, W. R. Taylor, J. M. Thornton, *CABIOS, Comput. Appl. Biosci.* **1992**, *8*, 275–282.
[31] E. Hodis, J. Prilusky, E. Martz, I. Silman, J. Moult, J. L. Sussman, *Genome Biol.* **2008**, *9*, R121.
[32] R. A. Laskowski, *Nucleic Acids Res.* **2009**, *37*, D355–D359.
[33] M. Fischer, Q. C. Zhang, F. Dey, B. Y. Chen, B. Honig, D. Petrey, *Nucleic Acids Res.* **2011**, *39*, W357–W361.
[34] Z. H. Yang, *Mol. Biol. Evol.* **2007**, *24*, 1586–1591.
[35] A. Stern, A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach, T. Pupko, *Nucleic Acids Res.* **2007**, *35*, W506–W511.
[36] a) S. J. Fleishman, N. Ben-Tal, *Curr. Opin. Struct. Biol.* **2006**, *16*, 496–504; b) M. Schushan, N. Ben-Tal, in *Introduction to Protein Structure Prediction: Methods and Algorithms* (Eds.: H. Rangwala, G. Karypis), John Wiley and Sons, New Jersey, **2010**, pp. 369–401.
[37] M. Kalman, N. Ben-Tal, *Bioinformatics* **2010**, *26*, 1299–1307.

[38] L. Jaroszewski, Z. W. Li, S. S. Krishna, C. Bakolitsa, J. Wooley, A. M. Deacon, I. A. Wilson, A. Godzik, *PloS Biol.* **2009**, 7, e1000205.

[39] a) A. Shulman-Peleg, R. Nussinov, H. J. Wolfson, *Nucleic Acids Res.* **2005**, *33*, W337–W341; b) Y. Y. Tseng, J. Dundas, J. Liang, *J. Mol. Biol.* **2009**, *387*, 451–464.

[40] P. A. Reche, E. L. Reinherz, *J. Mol. Biol.* **2003**, *331*, 623–641.

[41] The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC. Taken from http://www.pymol.org/citing.

Received: December 14, 2012
Accepted: March 10, 2013