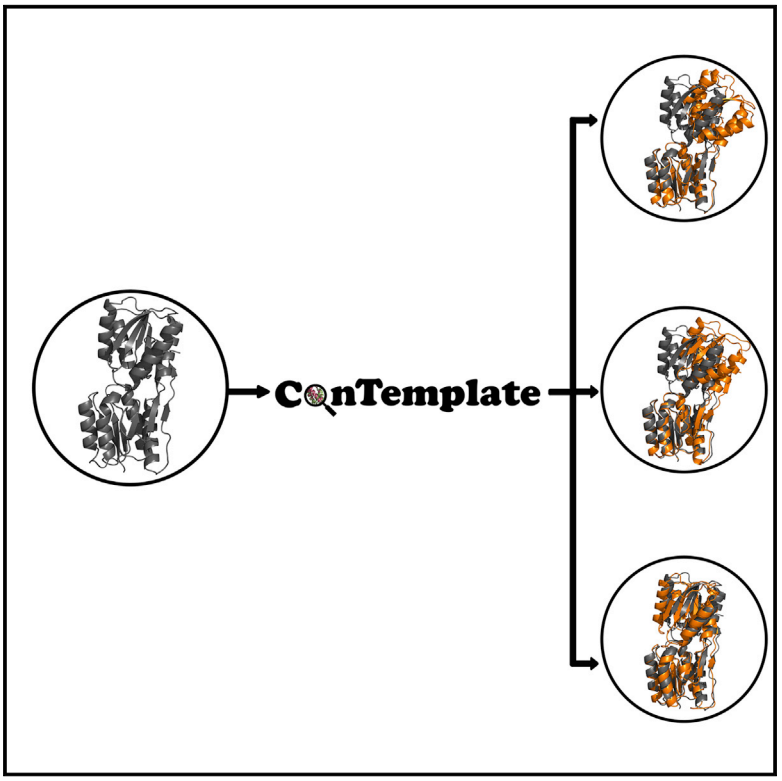
Resource

Structure

ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure

Graphical Abstract



Authors

Aya Narunsky, Sergey Nepomnyachiy, Haim Ashkenazy, Rachel Kolodny, Nir Ben-Tal

Correspondence

bental@ashtoret.tau.ac.il (N.B.-T.), trachel@cs.haifa.ac.il (R.K.)

In Brief

To conduct their function, proteins typically alternate between various conformations, but often only some of these important conformations are known. Narunsky et al. introduce the ConTemplate methodology and web server for inferring missing conformations of a query protein based on the structural repertoire in the PDB.

Highlights

- Most PDB proteins have multiple structures, often in various conformations
- Two proteins that share one conformation often share additional conformations
- ConTemplate suggests conformations for a query protein of known structure

Narunsky et al., 2015, Structure *23*, 2162–2170
November 3, 2015
http://dx.doi.org/10.1016/j.str.2015.08.018

CellPress

Structure
Resource

ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure

Aya Narunsky,[1] Sergey Nepomnyachiy,[2] Haim Ashkenazy,[3] Rachel Kolodny,[4,*] and Nir Ben-Tal[1,*]
[1]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[2]Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201, USA
[3]The Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
[4]Department of Computer Science, University of Haifa, Mount Carmel, Haifa 31905, Israel
*Correspondence: bental@ashtoret.tau.ac.il (N.B.-T.), trachel@cs.haifa.ac.il (R.K.)
http://dx.doi.org/10.1016/j.str.2015.08.018

SUMMARY

Protein function involves conformational changes, but often, for a given protein, only some of these conformations are known. The missing conformations could be predicted using the wealth of data in the PDB. Most PDB proteins have multiple structures, and proteins sharing one similar conformation often share others as well. The ConTemplate web server (http://bental.tau.ac.il/contemplate) exploits these observations to suggest conformations for a query protein with at least one known conformation (or model thereof). We demonstrate ConTemplate on a ribose-binding protein that undergoes significant conformational changes upon substrate binding. Querying ConTemplate with the ligand-free (or bound) structure of the protein produces the ligand-bound (or free) conformation with a root-mean-square deviation of 1.7 Å (or 2.2 Å); the models are derived from conformations of other sugar-binding proteins, sharing approximately 30% sequence identity with the query. The calculation also suggests intermediate conformations and a pathway between the bound and free conformations.

INTRODUCTION

Many proteins function by changing their conformations in response to ligand binding, pH alterations, or other changes in the environment (Kosloff and Kolodny, 2008). Thus, studying the conformational space of a given protein may further our understanding of its mechanism of action (Kessel and Ben-Tal, 2010). Normal mode analyses and methods based on physicochemical force fields, such as molecular dynamics simulations, have been used to create ensembles of conformations for proteins of interest (Adcock and McCammon, 2006; Eyal et al., 2015; Flores and Gerstein, 2007, 2011; Grant et al., 2010; Korkut and Hendrickson, 2009; Laughton et al., 2009). However, because it is difficult to accurately simulate protein dynamics, these approaches may result in unrealistic conformations.

Alternatively, one can take a knowledge-based approach and mine the wealth of data in the PDB to identify additional conformations. Previous efforts to this end include compilation of databases of conformational changes (Gerstein and Krebs, 1998; Juritz et al., 2011; Li et al., 2015; Monzon et al., 2013) and their analysis (Kosloff and Kolodny, 2008). In addition, homology-modeling tools, such as Swiss-Model (Biasini et al., 2014) and ModBase (Pieper et al., 2014), use various templates and may capture a given query at different conformations. Also relevant are the domain classifications of all PDB proteins (Berman et al., 2000). These classifications include SCOP (Fox et al., 2014), CATH (Knudsen and Wiuf, 2010), and ECOD (Cheng et al., 2014), which cluster proteins on the basis of their sequence and structure, as well as the Pfam domain sequence classification (Finn et al., 2014). Diverse structures in a given cluster of domains (e.g., in a SCOP fold) might represent alternative conformations for each of the domains included in the cluster. Thus, theoretically it should be possible to use a classification to build models of alternative conformations of a query domain. Yet there are several problems with doing so. (1) There are no automatic tools that implement this idea. (2) Using a domain-level classification obscures domain-domain motions, and these may be significant. For example, binding and catalysis may take place at the interface between domains. A more specific example is the ribose-binding protein, which includes two domains that change their relative positioning with respect to each other upon ribose binding (Groarke et al., 1983), whereas the CATH domains do not change throughout this conformational change. (3) The decomposition into domains also makes it difficult to detect remote allosteric effects in multi-domain proteins. For example, the elongation factor Ef-Tu from *Thermus aquaticus* comprises a nucleotide-binding domain, an elongation factor domain, and a C-terminal domain; guanosine triphosphate hydrolysis in the nucleotide-binding domain induces a change in the orientation between the other two domains (Kjeldgaard et al., 1993; Polekhina et al., 1996).

Here, we propose an approach to mining PDB information that bypasses the limitations associated with domain-based classifications. First, we show that most PDB proteins have more than one available structure. Furthermore, among proteins that undergo major conformational changes, many protein pairs that share one conformation share additional conformations as well. On the basis of these encouraging observations we have

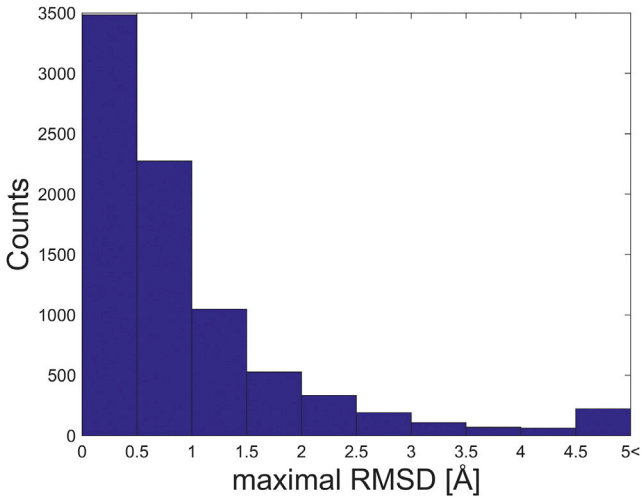


Figure 1. The Maximal RMSD between Two PDB Structures of the Same Protein

The size of the maximal RMSD between two structures of the same protein, in a set of 8,322 protein chains; less than 1 Å for the vast majority of proteins.

developed ConTemplate, a knowledge-based computational tool and web server for suggesting possible alternative conformations of a query protein deduced from conformations of available PDB structures. ConTemplate searches for proteins that share similar structures with the query, and uses different conformations of those proteins as templates to model the query protein in alternative (suggested) conformations.

RESULTS

Protein Pairs that Share One Conformation Often Share Additional Conformations

We counted the occurrences of protein chains in multiple conformations in the PDB by conducting a BLAST search at various sequence-identity thresholds (90%, 95%, 99%, and 100%). The vast majority of chains appear more than once, often in the same entry, and 66%–83% of chains appear in multiple entries (Figure S1).

We then carried out a search for conformational differences across identical chains. As minor conformational changes can have profound effects on function, it is not trivial to distinguish structural variability resulting from conformational changes from variability that is merely due to thermal fluctuations (or different experimental processes). For example, the distinct (and well-characterized) oxy and deoxy conformations of hemoglobin superimpose with a root-mean-square deviation (RMSD) of less than 1 Å (Perutz, 1970). Of all 77,663 high-quality chains in the PDB (SPACI score higher than 0.4 [Brenner et al., 2000]), we compiled sets of proteins, sharing at most 80% sequence identity with one another, that are each characterized by multiple conformations. Most of the proteins in each set featured only minimal conformational changes (41.8% <0.5 Å and 69.2% <1 Å; see Figure 1). The largest conformational change we observed was of c-Src, with 23-Å RMSD between the active (Cowan-Jacob et al., 2005) and inactive (Xu et al., 1999) conformations.

We derived datasets in which the conformations of the same protein differ from one another by RMSD of over 2, 3, 4, 5, and 6 Å (Table 1). Similar conclusions emerge from analyses of all of the datasets, and we describe here the analysis performed with the 4-Å set, which includes 246 proteins in 516 alternative conformations. Less than 1% of the protein pairs in this dataset are structurally similar to each other (Figure 2A), but 57.0% of the proteins that share one similar conformation with each other have additional conformations in common (Figure 2B).

We analyzed the proteins in the set that had at least one similar conformation, and compared their GO and ECOD annotations. The vast majority of proteins that share a conformation with each other also have the same ECOD X-group classification (marking structural similarity without convincing evidence for homology), as expected. In the few cases where two proteins did not have the same X-group classification, they had only one conformation in common. In a similar manner, structurally similar proteins often have similar function annotations (Table S1).

The main observation that emerges from analyzing all of the datasets is that 50% or more of the protein pairs that share one similar conformation with each other have other conformation(s) in common (Table 1). ConTemplate is based on this observation.

The ConTemplate Method

ConTemplate builds an ensemble of conformations for a query protein that has at least one known structure, using a three-step process (Narunsky and Ben-Tal, 2014): First, it searches for a set of proteins that are structural equivalents of the query, using the structure-aligner GESAMT (Krissinel, 2012) based on preset similarity thresholds (Figures 3A, 4A, and 4D). For efficiency, the search is only among the structural neighbors of the query, rather than the entire PDB. We define structural neighbors as proteins that have similar FragBag profiles; FragBag is a succinct representation of protein structure that can be used to carry out efficient structural comparison (Budowski-Tal et al., 2010). At the end of the first step, ConTemplate has a list of the known conformations of the query, together with alignments of the query and its structural equivalents (it first obtains structure alignments and subsequently applies those alignments to the sequences as well). In the second step, ConTemplate runs BLAST to identify additional PDB conformations for all structural equivalents (Altschul et al., 1990) (Figures 3B, 4B, and 4E). The conformations that are identified are subsequently clustered (Figure 3C). In the third and last step, the server calculates models of the query in various conformations. To this end, it uses the structures closest to the centers of the clusters found in the second step as templates, and relies on the structure-based sequence alignments found in the first step (Figures 4C and 4F).

The Web Server

Input

ConTemplate has a simple and intuitive user interface. It is fully automated, and the user only needs to upload the coordinates of the query protein or specify its PDB ID and chain. An advanced user can set the structural similarity thresholds between the query and its structural neighbors (used in the first step): similarity is quantified by threshold values for RMSD, alignment coverage, and the alignment quality measure Q-score (Krissinel,

Table 1. Proteins that Share One Similar Conformation Often Have Additional Conformations in Common

RMSD Threshold between Conformations (Å)	Structural-Similarity Thresholds			No. of Proteins	No. of Protein-Pairs Sharing One Conformation	No. of Protein-Pairs Sharing Additional Conformations	Percent of Protein-Pairs that Share One Conformation and Have Another Conformation in Common (%)
	Maximal RMSD (Å)	MINIMAL Q-Score	Minimal Coverage (%)				
2	1.5	0.5	80	1024	364	183	50.3
3	1.75	0.45	75	425	138	67	48.6
4	2.0	0.4	70	246	79	45	57.0
5	2.25	0.35	65	146	20	15	75.0
6	2.5	0.3	60	102	14	10	71.4

The sequence identity between each pair of proteins in the set is no more than 80%.

2012). The user can also specify thresholds for the second step: the minimal sequence identity to be used in the BLAST search, and the number of clusters (which determines the maximal number of proposed conformations; see Figures 3A and 4A). The ConTemplate interface provides a detailed example, explaining the parameters.

Output

ConTemplate's output includes known and suggested conformations of the query. The server builds structural models for the query protein in the suggested conformations, based on the templates that are closest to the centers of the clusters obtained in the second step. The number of models is equivalent to the number of clusters; when a cluster has more than one structure, we use the one closest to the cluster's center as a template. When the number of structures identified in the second step is smaller than, or equal to, the number of clusters designated by the user, all identified structures are used as templates in modeling. The output lists the models, including the PDB ID of each template and the structurally equivalent protein that was the origin of that template. For each model, ConTemplate also indicates the RMSD from the query and the size of the cluster it represents; the latter may indicate whether the conformation is biologically relevant. Using JSmol (an open-source HTML5 viewer for chemical structures in 3D, http://wiki.jmol.org/index.php/JSmol#JSmol), the user can inspect the structural alignments of the models and the original conformation of the query. The user can also download the models.

In addition, ConTemplate provides the user with the output of the second step, i.e., the list of all proteins that are structurally equivalent to the query, in their various known conformations, segregated into clusters. The user may then seek out more information on these proteins, including their crystallization conditions, conformational variability, function annotation, and so forth.

In addition to suggesting alternative conformations based on other proteins, ConTemplate looks for multiple occurrences of the query in the PDB. Each occurrence is listed in the output, including its PDB ID, sequence identity with the query, and RMSD from the query. ConTemplate also offers an overall view of all the available and suggested conformations of the query as a similarity network using Cytoscape (Saito et al., 2012) and CyToStruct (Nepomnyachiy et al., 2015). The conformations are represented as nodes, connected to each other by edges, the lengths of which correspond to the RMSDs between the conformations.

Case Study

We use ConTemplate to study conformational changes in the D-ribose-binding protein, a member of the periplasmic binding protein superfamily (Quiocho, 1991; Shilton et al., 1996). Members of this family participate in the membrane transport process, and some, including the ribose-binding protein, also serve as chemotaxis receptors. They are located in the periplasmic space of the bacteria cell, between the outer wall and the cytoplasmic membrane. The periplasmic binding proteins differ in sequence but have similar structures: two domains connected by a hinge. Ligand binding induces a rotation around this hinge, bringing the two domains closer together to a closed conformation. The protein in the closed conformation interacts with a transport complex in the inner membrane, and facilitates ligand entrance into the cell.

The D-ribose-binding protein was chosen as an example because of the high sequence diversity within the periplasmic binding protein superfamily; this enabled us to use an artificial cutoff of maximum 50% sequence identity between the query and structurally equivalent proteins (and thus to avoid the trivial outcome whereby ConTemplate selects the other conformation of the query protein itself). This cutoff significantly limits ConTemplate's ability to reproduce conformational changes, and is not the default setting of the web server.

Given a query structure in the open, ligand-free conformation (Bjorkman and Mowbray, 1998), and setting the number of clusters to two, ConTemplate reproduces the closed, ligand-bound conformation (Bjorkman et al., 1994), with RMSD of 1.7 Å from the actual closed conformation. Taking into account that the closed and open conformations superimpose structurally on each other with RMSD of 4.1 Å, this is a good model structure. Modeling is based on the query's structural similarity to the open conformation of a xylose-binding protein, and the known closed conformation of that protein, used as a template (Sooriyaarachchi et al., 2010) (Figure 4). Note that the query and the template proteins share only 27% sequence identity. When the number of clusters is set to more than two, ConTemplate produces a larger number of models, yet each model resembles either the ligand-bound or the ligand-free conformation, meaning that none of the models appear to be an intermediate on the pathway between the two conformations.

In the reverse scenario of querying ConTemplate with the ligand-bound conformation, reproducing the ligand-free conformation is more challenging. The closed (ligand-bound) conformation is far more abundant in the PDB than is the open

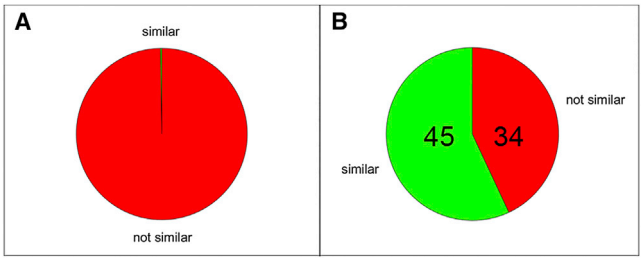


Figure 2. Proteins that Have One Conformation in Common Often Share Additional Conformations

(A) All protein pairs with at least one similar conformation found in a collection of 246 proteins. Out of 30,135 protein pairs, 79 (0.3%) were found to have a similar conformation.

(B) Of the 79 protein pairs that share a similar conformation, 45 (57%) have additional conformations in common.

conformation, and ConTemplate finds a large number of similar proteins and templates. The uneven distribution of the PDB conformations, as well as the presence of irrelevant, albeit similar, conformations, requires increasing the number of clusters to detect the rare open conformation. Only when the number of clusters is set to nine or more does one of the models capture the open conformation. Overall, as the number of clusters is increased, ConTemplate produces models that better describe the open conformation. For example, with nine clusters one model represents an open conformation, but the RMSD between this model and the actual open conformation is 4.2 Å. The template used to produce this model is a glucose/galactose-binding protein bound to an antagonist, which prevents the protein from adopting the closed conformation (Borrok et al., 2009). With 20 clusters, ConTemplate reproduces the open conformation with RMSD of 2.2 Å (Figure 3D). The proposed model is based on the query's structural similarity to the bound conformation of the D-allose binding protein, and the known open conformation of this protein (Chaudhuri et al., 1999). In this case, the sequence identity between the query and the template is 34%. Interestingly, when increasing the number of clusters, we obtain models that simultaneously resemble both the open and closed conformations. These novel structures may represent intermediate conformations on the pathway between the open and closed states, a pathway that can be visualized using the Cytoscape network analysis tool (Saito et al., 2012) (Figure 5 and Movie S1).

Reassuringly, the pathway between the open and closed conformations of the ribose-binding protein appears also as a dominant mode of motion in analysis using the anisotropic network model (Atilgan et al., 2001; Eyal et al., 2015). In particular, the first (internal) mode of motion of the closed conformation corresponds to a squeezing motion whereby the domains rotate in opposite directions; many of the suggested conformations in the vicinity of the closed conformation (the large cluster in Figure 5) can be aligned to represent this motion. The second and third modes of motion are, in essence, degenerate, and represent the pathway between the open and closed conformations. Anisotropic network model analysis using the open conformation shows very similar results, and the models proposed by ConTemplate are again compatible with the first modes of motion obtained in this analysis.

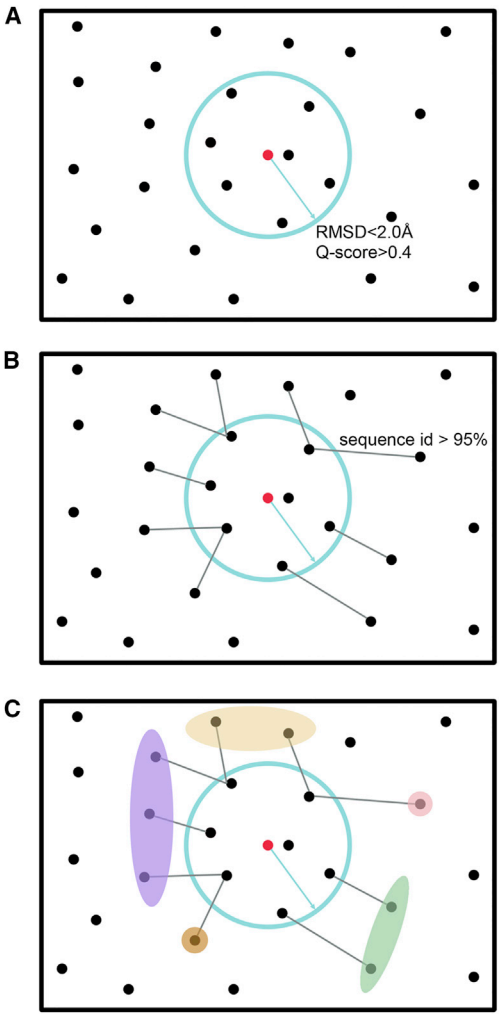


Figure 3. ConTemplate Methodology

The query protein is represented by a red point, and other PDB structures are represented by black points.

(A) In the first step, ConTemplate uses GESAMT (Krissinel, 2012) to identify proteins with structural similarity to the query. The blue circle represents the structural vicinity of the query; it encircles the collection of proteins sharing the query's structure. The number of proteins in this vicinity depends on the RMSD and Q score thresholds (the default values are listed); the user can configure these.

(B) Using BLAST (Altschul et al., 1990), ConTemplate suggests alternative conformations of the proteins detected in (A), including the query; the gray edges connect proteins that share, in essence, the same sequence (but have different structures). Note that some proteins may have several suggested conformations, while others have none; the number of these conformations depends on the sequence-identity threshold (the default value is listed). The user can configure this threshold.

(C) Clustering the proteins found in (B). Five clusters are shown. Their centers are used to model alternative (suggested) conformations of the query. The user can configure the number of clusters.

For comparison, the Swiss-Model (Biasini et al., 2014) and ModBase (Pieper et al., 2014) homology-modeling tools suggest only structures with the abundant, ligand-bound conformation of the ribose-binding protein as templates; templates that resemble the rare open conformation are not offered.

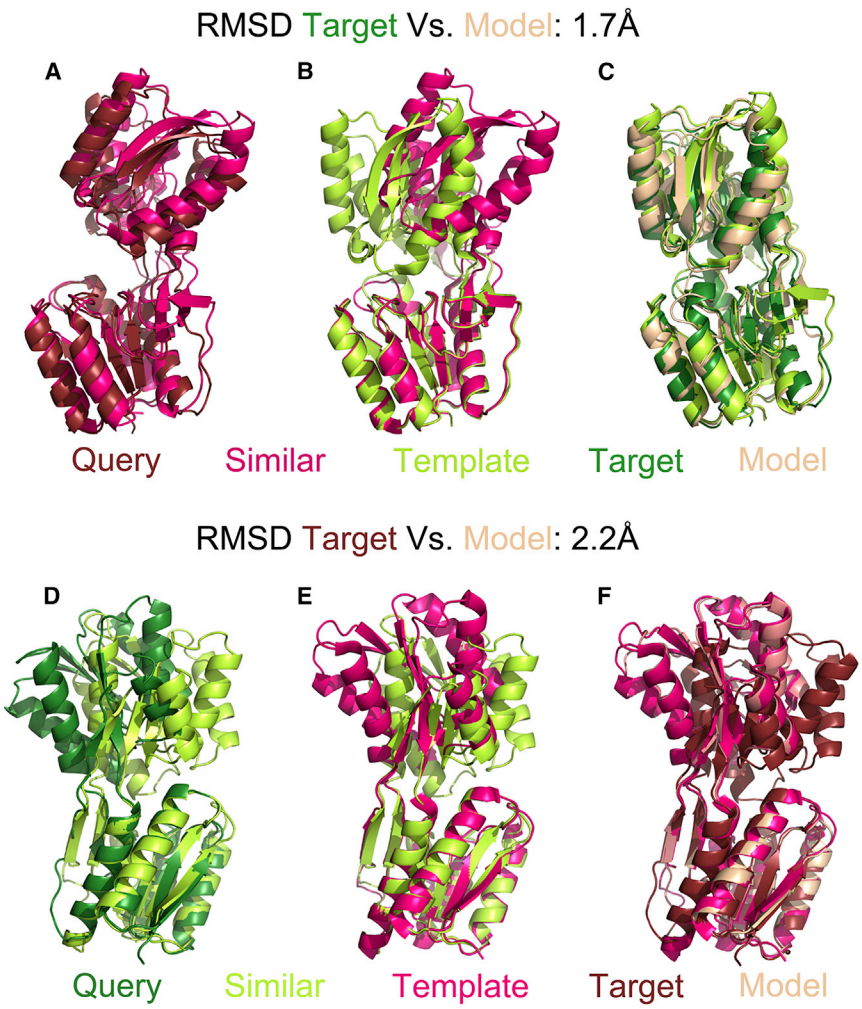


Figure 4. Modeling Conformational Changes in the Ribose-Binding Protein

Upper: Using the known structure of the open (ligand-free) protein as a query (Query, PDB: 1URP) and reproducing its closed, ligand-bound conformation (Target, PDB: 2DRI). Lower: Using the known structure of the closed conformation as a query and reproducing the open conformation. The RMSD between the two conformations is 4.1 Å. The maximal RMSD between the query and similar proteins is set to 2.0 Å, and the minimal Q score is set to 0.4. In the upper panel the number of clusters is set to 2. In the lower panel we consider 20 clusters because the target is a rare conformation.

(A and D) Selecting proteins with structural similarity to the query; only one is shown here (Similar, PDB: 3M9X, 1RPJ).

(B and E) Suggesting alternative conformations of the proteins detected in step 1; only one is shown here (Template, PDB: 3MA0, 1GUB).

(C and F) Modeling suggested conformations of the query using the conformations detected in step 2 as templates; only one is shown here (Model).

DISCUSSION

ConTemplate suggests conformations for a query protein that has (at least) one known conformation (or a model structure). The tool is the first automatic implementation of the commonly used concept of suggesting additional conformations for a protein on the basis of its similarity to other proteins. We hope that the web server, which has a simple interface and allows users to adjust the relevant parameters, will enable researchers to easily identify biologically relevant conformations.

The conformations proposed by ConTemplate are either known structures of the query (identified using sequence search), or conformations derived from proteins that share a conformation with the query. Users can configure the thresholds for both tracks: the percent sequence identity determines the level of sequence similarity used to identify known conformations, and the RMSD and Q-score thresholds determine which additional proteins will be considered structural equivalents. In particular, the user can choose to carry out a homology search only, by configuring very strict structural vicinity thresholds.

Although it is computationally demanding to search for proteins that share a conformation with a given query, this approach may yield a set of proteins whose other conformations are relevant to the query as well. In addition, in ConTemplate, pairwise sequence alignment between the query and each of the templates is derived from structural alignment (of their shared conformation). Thus, the accuracy of this alignment should be superior to that of alignments based on sequence similarity alone, especially for remote homologs (Yang and Honig, 2000). To ensure feasible run-time, we have limited the user's ability to change ConTemplate's settings, confining the parameter values within an acceptable range. For determining structural similarity, we have enabled the user to set ConTemplate's maximal RMSD threshold at 2.5 Å and its minimum Q score at 0.3. These cutoffs were chosen on the basis of our preliminary analysis of the PDB, in which we observed that for similar thresholds for structural vicinity, half or more of the proteins that share a similar conformation with the query share additional conformations with that query (Figure 2 and Table 1).

To demonstrate an application of ConTemplate, and to analyze its capabilities and limitations, we queried ConTemplate with a protein that has more than one known conformation. To avoid the trivial outcome whereby ConTemplate identifies only the actual well-known additional conformation, we limited the sequence identity between the query and structurally equivalent proteins to 50% or less. This limitation effectively removed the "obvious answer" from the database, and enabled us to show that ConTemplate can deduce a protein's conformation from other structures alone; clearly, most practical applications of the web server would not incorporate such a restriction. In addition to the ribose-binding protein test case described here in detail, we queried ConTemplate with additional proteins that undergo conformational changes of various magnitudes. Analysis of selected cases, namely a G protein-coupled receptor, an

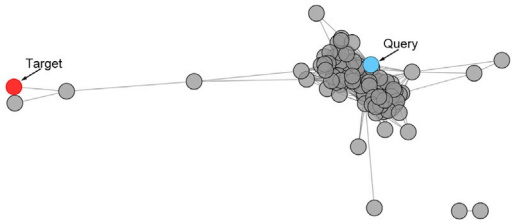


Figure 5. A Network Representation of the ConTemplate Results Could Suggest Pathways between Conformations and Help Eliminate Irrelevant Conformations

The network is presented using Cytoscape (Saito et al., 2012). The ribose-bound conformation (blue) was used as a query. The target (red) is the known ribose-free conformation. The gray nodes correspond to models obtained using ConTemplate. Two nodes are connected by an edge if the RMSD between the two corresponding structures is under 2.5 Å. The length of the connecting edge is proportional to the RMSD between the connected structures. The outlier nodes (bottom right) correspond to models obtained using templates that are unrelated to the ribose-binding protein; these are presumably irrelevant conformations.

ATPase transporter, and a kinase, can be found under the "Gallery" section in the ConTemplate web server. Furthermore, we have tested ConTemplate with a dozen or so additional proteins (without restricting the maximal sequence-identity level), and it successfully identified the known conformations of these proteins. We emphasize, however, that there are no guarantees: ConTemplate may miss true additional conformations of a query protein.

Furthermore, even though ConTemplate's suggested conformations are based on available structures, there is no guarantee that they are indeed relevant for the query protein. Rather, the predicted conformations should be considered as suggestions that can be used to raise testable hypotheses for further research. In this respect it is noteworthy that the anticipation, based on evolutionary design, would be for proteins of similar function to share, in essence, the same active form. However, there is much less restriction on the design of inactive forms, which could therefore differ from one another. Thus, suggestions for inactive conformations should be taken with a grain of salt. A literature review regarding the identified templates is one approach that may help to eliminate irrelevant suggestions. In our case study, for example, ConTemplate modeled one of the conformations for the ribose-binding protein on the basis of a DNA-bound conformation of a transcription factor. Being periplasmatic, the query is unlikely to interact with DNA, thus rendering this conformation irrelevant. Connecting a series of conformations to each other to suggest a trajectory between the initial conformation and a (remote) suggested conformation can also provide support for the relevance of the latter conformation. Disjoint conformations are less likely to be relevant. Indeed, some of the suggested conformations of the ribose-binding protein can be connected to form a pathway between the ligand-free and bound conformations, while the presumably irrelevant one, derived from the transcription factor template, appears as an outlier (Figure 5). Cytoscape (Saito et al., 2012) is a useful tool for visualizing the various conformations as a network, enabling the user to search for pathways among them. The CyToStruct applet (Nepomnyachiy et al., 2015) can be used to link the network to molecular viewers, such as VMD (Humphrey et al., 1996), PyMOL (Schrodinger LLC, 2010) and UCSF Chimera (Pettersen et al., 2004).

The proteins in the PDB are not sampled uniformly; some conformations are more common than others. For example, ligand-bound conformations of the ribose-binding protein are more abundant than ligand-free conformations. ConTemplate groups the suggested templates into clusters of uniform conformations, and represents each cluster by a single conformation. However, some regions in conformational space are represented by large clusters, which may mask less common conformations. Modifying the thresholds can help ConTemplate identify a greater number of proteins with additional known conformations, and these may ultimately be represented in the final set of models. Another important parameter is the number of clusters: increasing this number produces finer clustering, and facilitates the detection of less abundant conformations, which may be neglected otherwise. It can be informative to iteratively search while altering the thresholds for the level of similarity between the query and its structurally equivalent proteins, the sequence similarity thresholds, and the number of clusters. The default parameter values were determined on the basis of our experience thus far: balancing between strict thresholds that would yield a set of templates that are too similar to the query conformation, and lax thresholds that might bring irrelevant conformations. The default parameters are: RMSD cutoff of 2.0 Å, Q-score threshold of 0.4, at least 95% sequence identity between the structural equivalents and their additional conformations, and five clusters. The user can change these parameters, keeping the maximum RMSD under 2.5 Å, the Q-score above 0.3, the sequence identity above 70%, and the number of clusters up to 99.

The run-time for the full process varies, and depends on the parameters used and the number of collected templates. The run-time for a typical protein of 300 residues, using default parameter values, is about 2 hr. Once the process concludes, one can rerun it with different parameter values. When possible, to reduce the run-time of a resubmitted run, the server uses the structural alignments created in the first step of the original run. For example, if the only change is the number of clusters, the server will only recalculate the clusters and build their respective models. The process can be completed in several minutes, depending on the number of models produced.

Future Directions

The current ConTemplate approach entails searching for structural templates for a given query throughout the entire PDB. In many cases, experimental structure determination of a protein involves introducing modifications and mutations (e.g., His tags, deletion of unstructured segments, fusion of other proteins). In proteins that are known to alternate between conformations, it is common to induce mutations that stabilize certain conformations. For example, Kobilka and colleagues used several mutations to stabilize the protease-activated receptor 1 (PAR1) and enable it to be crystallized (Zhang et al., 2012). The common interpretation of such results is based on the population-shift model, according to which a given protein alternates between various states and the mutation stabilizes a specific conformation, which is less favorable otherwise. Nevertheless, these mutated structures are sometimes treated with caution, as the introduced mutations may theoretically yield irrelevant conformations. Had the

PDB been fully annotated, including data on mutations, ligand binding, post-translational modification, chimeras, etc., it would have been possible for ConTemplate to filter some of the structures on the basis of their level of "purity." It could also associate specific conformational changes with their corresponding alterations, e.g., acetylation. If PDB annotation eventually becomes sufficient for this purpose, we will incorporate these options into ConTemplate.

When two or more conformations of a protein are known, it is interesting to investigate the pathway(s) between them. Several works have studied the transition between two known conformations of a protein (Das et al., 2014; Enosh et al., 2008; Kim et al., 2002; Lei et al., 2004; Sfriso et al., 2013). The most popular tool for this purpose is the Morph server, which describes protein motion using molecular hinges (Flores et al., 2006). Combining ConTemplate with such methods is an obvious future research direction.

Conclusions

The PDB is highly redundant in that most of the protein chains appear in more than one entry. Two proteins sharing a conformation often have other conformations in common. Therefore, given a query protein with (at least) one known conformation, ConTemplate can suggest conformations for the query on the basis of its structural similarity to other proteins in the PDB, including proteins with low sequence similarity to the query. ConTemplate's output is based on previously observed structures, and may therefore be more biologically relevant compared with ensembles created solely on the basis of physicochemical considerations. The user can control all the parameters. ConTemplate can use either experimental structures or models as queries, and can be applied on a genome-wide scale. The server is simple to use even for non-experts. The output of this computational tool can provide insight into the function of the query protein, and thus help decipher its molecular mechanism.

EXPERIMENTAL PROCEDURES

Abundance of Conformations in the PDB: Assembling a Dataset of Proteins with Multiple Conformations

In Figure S1 we counted the number of conformations for each chain in the PDB. To do this, we ran a standard BLAST (Altschul et al., 1990) search against the PDB for each PDB chain and collected all the hits with 100%, 99%, 95%, and 90% sequence identity, and full coverage.

For Figures 1 and 2 we obtained from ASTRAL a dataset of 77,663 PDB chains, updated as of July 2014, with SPACI scores higher than 0.4 (Brenner et al., 2000). To search for alternative conformations of the proteins in the set, we ran BLAST to look for homologs in the PDB for which the product of the sequence identity and the mutual coverage was higher than 0.9. The search produced 56,255 chains that each appeared in more than one PDB entry. We structurally aligned the various structures of each protein chain using Kabsch's algorithm (Kabsch, 1978), on all the Cα atoms of the proteins.

For Figure 1, we ran PISCES (Wang and Dunbrack, 2005) to remove redundancy within the set of 56,255 protein chains that each have more than one available conformation. By default, when PISCES recognizes two proteins with sequence identity larger than the threshold, it removes the structure of lower quality. We used the PISCES logs to change that criterion, so that the protein removed from the set would be the protein that undergoes the smallest conformational change. We repeated this process until finally we were left with a set representing the proteins that undergo the largest conformational changes, sharing up to 80% sequence identity.

For Figure 2 and Table 1 we considered all the proteins that had two structures with RMSD above a certain threshold (2, 3, 4, 5, and 6 Å). We composed an all-against-all structure-alignment matrix including all the different structures of the protein, and clustered it to distinguish between the different conformations. Each cluster represented a single conformation. Using RMSD of 4 Å as the threshold for conformational change between two structures of the same protein, we were left with a set of 246 proteins, sharing up to 80% of their sequences. Each of these proteins had a collection of additional conformations. We clustered the various conformations of each protein, each cluster representing a single conformation. The centers of the clusters were used to form a set of 516 conformations (structures). We structurally aligned all-versus-all in this set using the GESAMT alignment tool. Two conformations were considered similar if they superimposed with RMSD less than 2 Å, Q score of more than 0.4, and coverage of more than 70%.

ConTemplate Methodology

Structural Neighbors

We have built a FragBag profile for each PDB chain longer than 40 residues, using a library of 400 fragments of 11 amino acids (Budowski-Tal et al., 2010). FragBag is a fast method for comparing protein structures. In this method, the protein backbone is structurally aligned to each of the library fragments and a profile is derived, which measures the number of appearances of each fragment in the protein (with overlaps). The profiles of proteins whose structures were determined using NMR methods (meaning that, for a given protein, there are several models of the structure in the PDB file) are built using the first model only. Structurally similar proteins share similar profiles. In ConTemplate, the profile of the query protein is compared with the profiles of all the proteins in the PDB to search for the nearest structural neighbors to be used in step 1 below, using cosine distance. The FragBag library was selected following the recommendation in Budowski-Tal et al. (2010).

Step 1: Collection of Proteins that Are Structurally Equivalent to the Query

ConTemplate searches for structural neighbors of a query protein, using the process described above. The query is then structurally aligned using GESAMT to each of the nearest neighbors (5,000 by default) (Krissinel, 2012) (Figures 3A and 4A). Through trial and error, we have found that an RMSD threshold of 2 Å and a Q-score threshold of 0.4 provide a large set of proteins that resemble the query conformation without including multiple conformations of the same structures. These thresholds are used by default; however, the user is free to modify them, within an acceptable range. Based on these structural superimpositions, ConTemplate derives a pairwise sequence alignment between the query and each of its structural equivalents to be used in the third step (see below).

Using a trial-and-error process, aimed at collecting all the relevant neighbors while excluding as many as possible proteins that are not sufficiently similar, we found that PDB chains with a cosine distance of 0.25 or less between their FragBag profiles can be considered *close* structural neighbors. Accordingly, when more than 5,000 close structural neighbors are found, and the user is not satisfied with the results of the run, he or she can repeat the process using all of the close structural neighbors. Note, however, that the run-time increases with the number of neighbors, as the server needs to carry out a larger number of structural alignments.

Step 2: Identifying and Clustering Additional Conformations

Searching for Additional Conformations. For each of the structurally equivalent proteins detected in step 1, ConTemplate identifies additional conformations of the protein by searching the PDB (Figures 3B and 4B) using BLAST (Altschul et al., 1990). A stringent similarity threshold of 95% sequence identity is recommended to make sure that the conformations are indeed related to the same protein, and yet tolerate minor differences, such as mutations or gaps. The user can change this threshold. The sequences of a given protein in its two corresponding conformations, i.e., the conformation originally identified as structurally equivalent to that of the query, and the newly identified conformation, referred to as a "template," may differ to some extent, and we use MUSCLE (Edgar, 2004) to align them. In addition, ConTemplate uses the same procedure (and sequence-identity threshold) to search for multiple occurrences of the query in the PDB.

Clustering the Suggested Conformations. ConTemplate clusters the templates and the available structures of the query (Figure 3C) using the *k*-means

clustering algorithm (Seber, 1984; Spath, 1985). The distance between templates is approximated using the distance between their Local Features Frequency profiles (LFF) (Choi et al., 2004); LFF is a fast method for comparing protein structures, which is sensitive to local changes and thus can be used to differentiate between conformations. The internal distance matrix of the protein is divided into overlapping submatrices. Each of the submatrices is compared to each of the features library elements, and a profile is derived. This profile measures the frequency of each feature in the protein. ConTemplate builds the profiles of the structural templates, using features library 100(10), i.e., 100 matrices of size 10 × 10, as recommended in Choi et al. (2004). As above, the profiles of templates whose structures were determined by NMR methods are built using the first model only. The profiling procedure is typically the most time-consuming phase of a ConTemplate run. The server identifies the representative template from each of the *k* clusters; the representatives are used in step 3 (model building). The value of *k* determines the maximal number of models in the resulting ensemble. A default value of 5 is suggested, which the user can change.

Step 3: Model Building

In the third ConTemplate step (Figures 4C and 4F), homology modeling using Modeller (Sali and Blundell, 1993) is used to build model structures of the query protein in various conformations, according to the representative templates selected in the second step. To obtain a model structure, ConTemplate provides Modeller with a representative template identified in step 2 and with its sequence alignment to the query, derived in step 1.

Detecting a Pathway between Conformations

In the usage example, the pathway between the model and the target was identified using the Cytoscape tool for network analysis (Saito et al., 2012). Two nodes are connected by an edge if the RMSD between them is below a preset threshold. The length of the edge is proportional to the RMSD (Figure 5). From our experience, the cutoff RMSD, under which two nodes will be connected, should be slightly smaller than the RMSD between the query and the model of interest. High cutoffs will show many distinct models and mask the pathway, whereas overly strict cutoffs may disrupt the connectivity of the network. A link to a Cytoscape view of the similarity network between the query and models is provided in ConTemplate's output. ConTemplate also generates a Cytoscape input file for users with local Cytoscape installation.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure, one table, and one movie and can be found with this article online at http://dx.doi.org/10.1016/j.str.2015.08.018.

ACKNOWLEDGMENTS

We thank Ron Diskin, Sarel Fleishman, Amit Kessel, and Meytal Landau for helpful discussions, and Vladimir Kadaner for technical support. This work was supported by grant No. 1775/12 of the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation. A.N. and H.A. were funded in part by the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

Received: March 31, 2015
Revised: July 31, 2015
Accepted: August 24, 2015
Published: October 8, 2015

REFERENCES

Adcock, S.A., and McCammon, J.A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. *106*, 1589–1615.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. *80*, 505–515.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. *42*, W252–W258.

Bjorkman, A.J., and Mowbray, S.L. (1998). Multiple open forms of ribose-binding protein trace the path of its conformational change. J. Mol. Biol. *279*, 651–664.

Bjorkman, A.J., Binnie, R.A., Zhang, H., Cole, L.B., Hermodson, M.A., and Mowbray, S.L. (1994). Probing protein-protein interactions. The ribose-binding protein in bacterial transport and chemotaxis. J. Biol. Chem. *269*, 30206–30211.

Borrok, M.J., Zhu, Y., Forest, K.T., and Kiessling, L.L. (2009). Structure-based design of a periplasmic binding protein antagonist that prevents domain closure. ACS Chem. Biol. *4*, 447–456.

Brenner, S.E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res. *28*, 254–256.

Budowski-Tal, I., Nov, Y., and Kolodny, R. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proc. Natl. Acad. Sci. USA *107*, 3481–3486.

Chaudhuri, B.N., Ko, J., Park, C., Jones, T.A., and Mowbray, S.L. (1999). Structure of D-allose binding protein from *Escherichia coli* bound to D-allose at 1.8 A resolution. J. Mol. Biol. *286*, 1519–1531.

Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., Kim, B.H., and Grishin, N.V. (2014). ECOD: an evolutionary classification of protein domains. PLoS Comput. Biol. *10*, e1003926.

Choi, I.G., Kwon, J., and Kim, S.H. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. Proc. Natl. Acad. Sci. USA *101*, 3797–3802.

Cowan-Jacob, S.W., Fendrich, G., Manley, P.W., Jahnke, W., Fabbro, D., Liebetanz, J., and Meyer, T. (2005). The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. Structure *13*, 861–871.

Das, A., Gur, M., Cheng, M.H., Jo, S., Bahar, I., and Roux, B. (2014). Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. PLoS Comput. Biol. *10*, e1003521.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Enosh, A., Raveh, B., Furman-Schueler, O., Halperin, D., and Ben-Tal, N. (2008). Generation, comparison, and merging of pathways between protein conformations: gating in K-channels. Biophys. J. *95*, 3850–3860.

Eyal, E., Lum, G., and Bahar, I. (2015). The anisotropic network model web server at 2015 (ANM 2.0). Bioinformatics *31*, 1487–1489.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Flores, S.C., and Gerstein, M.B. (2007). FlexOracle: predicting flexible hinges by identification of stable domains. BMC Bioinformatics *8*, 215.

Flores, S.C., and Gerstein, M.B. (2011). Predicting protein ligand binding motions with the conformation explorer. BMC Bioinformatics *12*, 417.

Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., and Gerstein, M. (2006). The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res. *34*, D296–D301.

Fox, N.K., Brenner, S.E., and Chandonia, J.M. (2014). SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. *42*, D304–D309.

Gerstein, M., and Krebs, W. (1998). A database of macromolecular motions. Nucleic Acids Res. *26*, 4280–4290.

Grant, B.J., Gorfe, A.A., and McCammon, J.A. (2010). Large conformational changes in proteins: signaling and other functions. Curr. Opin. Struct. Biol. *20*, 142–147.

Groarke, J.M., Mahoney, W.C., Hope, J.N., Furlong, C.E., Robb, F.T., Zalkin, H., and Hermodson, M.A. (1983). The amino acid sequence of D-ribose-binding protein from *Escherichia coli* K12. J. Biol. Chem. *258*, 12952–12956.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38, 27–38.

Juritz, E.I., Alberti, S.F., and Parisi, G.D. (2011). PCDB: a database of protein conformational diversity. Nucleic Acids Res. *39*, D475–D479.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. Acta. Crystallogr. *A34*, 2.

Kessel, A., and Ben-Tal, N. (2010). Introduction to Proteins: Structure, Function, and Motion (CRC Press).

Kim, M.K., Jernigan, R.L., and Chirikjian, G.S. (2002). Efficient generation of feasible pathways for protein conformational transitions. Biophys. J. *83*, 1620–1630.

Kjeldgaard, M., Nissen, P., Thirup, S., and Nyborg, J. (1993). The crystal structure of elongation factor EF-Tu from *Thermus aquaticus* in the GTP conformation. Structure *1*, 35–50.

Knudsen, M., and Wiuf, C. (2010). The CATH database. Hum. Genomics *4*, 207–212.

Korkut, A., and Hendrickson, W.A. (2009). A force field for virtual atom molecular mechanics of proteins. Proc. Natl. Acad. Sci. USA *106*, 15667–15672.

Kosloff, M., and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins *71*, 891–902.

Krissinel, E. (2012). Enhanced fold recognition using efficient short fragment clustering. J. Mol. Biochem. *1*, 76–85.

Laughton, C.A., Orozco, M., and Vranken, W. (2009). COCO: a simple tool to enrich the representation of conformational variability in NMR structures. Proteins *75*, 206–216.

Lei, M., Zavodszky, M.I., Kuhn, L.A., and Thorpe, M.F. (2004). Sampling protein conformations and pathways. J. Comput. Chem. *25*, 1133–1148.

Li, W., Kinch, L.N., Karplus, P.A., and Grishin, N.V. (2015). ChSeq: a database of chameleon sequences. Protein Sci. *24*, 1075–1086.

Monzon, A.M., Juritz, E., Fornasari, M.S., and Parisi, G. (2013). CoDNaS: a database of conformational diversity in the native state of proteins. Bioinformatics *29*, 2512–2514.

Narunsky, A., and Ben-Tal, N. (2014). ConTemplate: exploiting the protein databank to propose ensemble of conformations of a query protein of known structure. BMC Bioinformatics *15*, A5.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2015). CyToStruct: augmenting the network visualization of cytoscape with the power of molecular viewers. Structure *23*, 941–948.

Perutz, M.F. (1970). Stereochemistry of cooperative effects in haemoglobin. Nature *228*, 726–739.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pieper, U., Webb, B.M., Dong, G.Q., Schneidman-Duhovny, D., Fan, H., Kim, S.J., Khuri, N., Spill, Y.G., Weinkam, P., Hammel, M., et al. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. *42*, D336–D346.

Polekhina, G., Thirup, S., Kjeldgaard, M., Nissen, P., Lippmann, C., and Nyborg, J. (1996). Helix unwinding in the effector region of elongation factor EF-Tu-GDP. Structure *4*, 1141–1151.

Quiocho, F.A. (1991). Atomic structures and function of periplasmic receptors for active transport and chemotaxis. Curr. Opin. Struct. Biol. *1*, 922–933.

Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Lotia, S., Pico, A.R., Bader, G.D., and Ideker, T. (2012). A travel guide to cytoscape plugins. Nat. Methods *9*, 1069–1076.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Schrodinger LLC. (2010). The PyMOL Molecular Graphics System, Version 1.3r1 (Schrodinger LLC).

Seber, G.A.F. (1984). Dimension Reduction and Ordination in Multivariate Observations (John Wiley), pp. 175–278.

Sfriso, P., Hospital, A., Emperador, A., and Orozco, M. (2013). Exploration of conformational transition pathways from coarse-grained simulations. Bioinformatics *29*, 1980–1986.

Shilton, B.H., Flocco, M.M., Nilsson, M., and Mowbray, S.L. (1996). Conformational changes of three periplasmic receptors for bacterial chemotaxis and transport: the maltose-, glucose/galactose- and ribose-binding proteins. J. Mol. Biol. *264*, 350–363.

Sooriyaarachchi, S., Ubhayasekera, W., Park, C., and Mowbray, S.L. (2010). Conformational changes and ligand recognition of *Escherichia coli* D-xylose binding protein revealed. J. Mol. Biol. *402*, 657–668.

Spath, H. (1985). Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples (Prentice Hall).

Wang, G., and Dunbrack, R.L., Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. *33*, W94–W98.

Xu, W., Doshi, A., Lei, M., Eck, M.J., and Harrison, S.C. (1999). Crystal structures of c-Src reveal features of its autoinhibitory mechanism. Mol. Cell *3*, 629–638.

Yang, A.S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. *301*, 691–711.

Zhang, C., Srinivasan, Y., Arlow, D.H., Fung, J.J., Palmer, D., Zheng, Y., Green, H.F., Pandey, A., Dror, R.O., Shaw, D.E., et al. (2012). High-resolution crystal structure of human protease-activated receptor 1. Nature *492*, 387–392.

