Web-based toolkits for topology prediction of transmembrane helical proteins, fold recognition, structure and binding scoring, folding-kinetics analysis and comparative analysis of domain combinations

Hongyi Zhou[1], Chi Zhang[1], Song Liu[1] and Yaoqi Zhou[1,2,*]

[1]Department of Physiology & Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214, USA and [2]Department of Macromolecular Science, The Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China

ABSTRACT

We have developed the following web servers for protein structural modeling and analysis at http://theory.med.buffalo.edu: THUMBUP, UMDHMM[TMHP] and TUPS, predictors of transmembrane helical protein topology based on a mean-burial-propensity scale of amino acid residues (THUMBUP), hidden Markov model (UMDHMM[TMHP]) and their combinations (TUPS); SPARKS 2.0 and SP[3], two profile–profile alignment methods, that match input query sequence(s) to structural templates by integrating sequence profile with knowledge-based structural score (SPARKS 2.0) and structure-derived profile (SP[3]); DFIRE, a knowledge-based potential for scoring free energy of monomers (DMONOMER), loop conformations (DLOOP), mutant stability (DMUTANT) and binding affinity of protein–protein/peptide/DNA complexes (DCOMPLEX & DDNA); TCD, a program for protein-folding rate and transition-state analysis of small globular proteins; and DOGMA, a web-server that allows comparative analysis of domain combinations between plant and other 55 organisms. These servers provide tools for prediction and/or analysis of proteins on the secondary structure, tertiary structure and interaction levels, respectively.

BACKGROUND

In the post-genomics era, attention is now squarely focused on the interconnections between sequences, structures and function of proteins. As more sequences from genome-sequencing projects and more structures from structure-genomics projects become available, tools are urgently needed to extract the maximum amount of information from them in order to analyze and predict unknown structures and function. We present a number of web-based servers available at http://theory.med.buffalo.edu as shown in Table 1. They are THUMBUP, UMDHMM[TMHP] and TUPS for topology prediction of transmembrane helical proteins (1); SPARKS 2.0 (2) and SP[3] (3) for sequence-to-structure fold recognition and alignment; DFIRE energy function (4) for scoring structural monomer (DMONOMER) and loop conformations (DLOOP) (5), predicting mutant stability (DMUTANT) (4), binding affinity of protein–protein/peptide complexes (DCOMPLEX) (6) and protein–DNA complexes (DDNA) (7); TCD for analysis of folding kinetics (8,9) and DOGMA for comparative analysis of plant domain graph (10). These servers can be classified as the tools for prediction and analysis of the secondary

*To whom correspondence should be addressed at Department of Physiology & Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214, USA. Tel: +1 716 829 2985; Fax: +1 716 829 2344; Email: yqzhou@buffalo.edu

structures, tertiary structures and interactions of proteins as shown in Figure 1. Details are described below.

THUMBUP, UMDHMM^TMHP AND TUPS

Overview

Communications and regulation of the communications between the inside and the outside of cell membranes are controlled mostly by transmembrane (TM) proteins. Most TM proteins are helical (TMH) proteins. Many different methods have been developed to predict the topology of TMH proteins (11–13). The determination of the topology of a TMH protein is useful for the annotation of its function.

Description

THUMBUP uses a simple scale of burial propensity and a sliding window-based algorithm to predict TM helical segments, and a positive-inside rule (14) to predict N-terminal orientation. The use of burial propensity was based on the fact that helical membrane proteins are packed more tightly than helical soluble proteins (15). It was found that THUMBUP gives an excellent prediction for TM proteins with known structures (3D_helix database), but relatively poorer prediction for a 1D_helix database (topology information was obtained by gene fusion and other experimental techniques) (1). The latter was attributed in part to the high inaccuracy of 1D_helix database employed (16–18).

UMDHMM^TMHP uses a modified version of hidden Markov model software developed at University of Maryland (version 1.02, http://www.cfar.umd.edu/~kanungo/software/software.html) for transmembrane-helical-topology prediction. The program differs from typical HMM-based methods for TMH proteins in that the parameters in UMDHMM^TMHP were trained by the 3D_helix database only.

TUPS combines the prediction of THUMBUP and UMDHMM^TMHP for TM segments and PHOBIUS (19) for the identification of signal peptides. More specifically, TUPS first takes the results from UMDHMM^TMHP. Then, if a TM segment predicted by THUMBUP does not overlap with any TM segments predicted by UMDHMM^TMHP, the segment is included in the TUPS prediction. Finally, signal peptides identified by PHOBIUS are removed from the TUPS prediction. There is no additional parameter introduced in TUPS other than the parameters determined in THUMBUP and UMDHMM^TMHP.

Table 1. List of web-based toolkits on the services section of the website: http://theory.med.buffalo.edu

Name (reference)	Input[a]	Output
TM helical topology (secondary structure level)		
THUMBUP (1)	Sequence	TMH residue ranges
		N-terminal orientation (in or out)
UMDHMM^TMHP (1)	As above	As above
TUPS	As above	As above
Fold recognition, alignment and structure prediction (tertiary structure level)		
SPARKS 2.0 (2)	Sequence	Sequence-to-structure alignment
	No. of models to be built	Models built (in PDB format)
SP³(3)	As above	As above
Application of DFIRE energy function (interaction level)		
DMONOMER (4)	Structure file	Conformation energy score
DLOOP (5)	Structure file	Conformation energy score
	Loop location	
DMUTANT (4)	Structure file	Stability change
	Residue mutated	
DCOMPLEX (6)	Complex structure file	Binding affinity
	Two chain IDs	
DDNA (7)	Complex structure file	Binding affinity
	Two chain IDs	
Protein folding kinetics (interaction level)		
TCD (8,9)	Structure file	TCD, folding rate transition-state size
	Chain ID	
	Residue range	
Domain graph analysis (interaction level)		
DOGMA (10)	Organism name	Comparative domain graph
	List of domain names	Shortest path between domains
		Phylogenetic profiling of domain/combination
		Topology analysis of domain graph

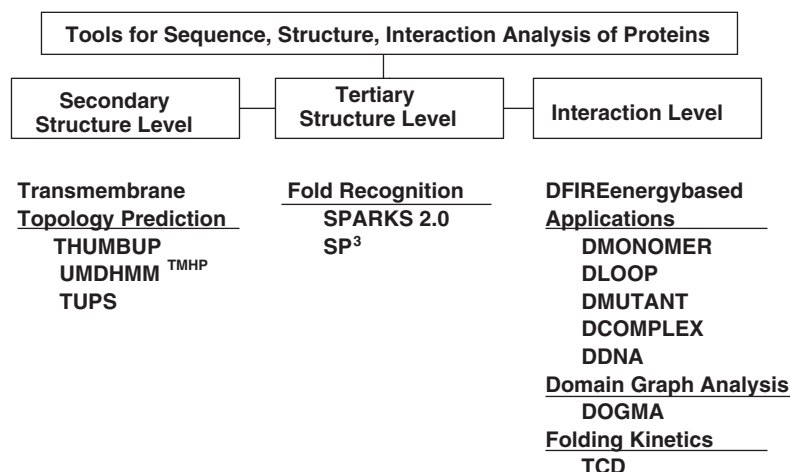[a]The formats for sequence and structural inputs are those of FASTA and PDB, respectively.

Figure 1. The classification of the web servers available on http://theory.med.buffalo.edu.

Performance

In addition to the 3D and 1D helix datasets tested in the original paper (1), we tested THUMBUP and UMDHMM[TMHP] in the static benchmark established by Kernytsky and Rost (20). UMDHMM[TMHP] and THUMBUP without any modification provides 86 and 80% per-segment accuracy for high-resolution dataset, respectively. The performances were ranked #1 and #3, respectively, among the methods compared in the static benchmark. Their performances on low-resolution dataset were only about average, as expected. The new TUPS server provides 88% per-segment accuracy for high-resolution dataset in this benchmark with significant lower rate for misidentifying signal peptides as TM helices (3 versus 70 in UMDHMM[TMHP] and 28 in THUMBUP). TUPS also provides a substantially better performance per topology accuracy on our 3D_helix test set (1) (86% versus 75% by THUMBUP and 78% by UMDHMM[TMHP]).

Input and output

The input is protein sequence in the FASTA format. Multiple sequences can also be submitted. The output provides information on the residue ranges of TM helices (if any) and the N-terminal orientation (Inside or Outside of membrane if the protein is a TMH protein) for every protein submitted. The output is now reported in a table format for easy understanding. A graphical interface will be built in near future for visualizing the TM region. Sample input and output with detailed line-to-line explanations are available online.

SPARKS 2.0 AND SP[3]

Overview

Fold recognition refers to recognition of structural similarity of two proteins with or without significant sequence identity. One way to detect structural similarity is to identify remote sequence homology via sequence comparison. Advances have been made from the pairwise to multiple sequence comparison, from sequence-to-sequence, sequence-to-profile to profile-to-profile comparison. Another way to detect structural similarity is via sequence-to-structure threading. More recent works attempt to optimally combine the sequence and structure information for a more accurate/sensitive fold recognition. For a recent review, see Ref. (21).

Description

Both fold recognition servers SPARKS 2.0 (2) and SP[3] (3) belong to the profile-based methods that provide sequence to structure alignment based on the sequence as well as the structure information of templates. SPARKS 2.0 and SP[3] differ in how structural information is integrated with the sequence profile of templates. The former uses a sophisticated knowledge-based, single-body score that includes torsion, contact energy and surface-accessible potentials. The structure score is calculated by threading the query sequence into template structure. The latter builds two separate sequence profiles from the sequence and structure of a template. The structure-derived sequence profile was derived from depth-dependent structural alignment of the fragments in the template structure with the fragments in a fragment library. SPARKS 2.0 an

upgraded version of SPARKS (2), takes the methods for parameter optimization, dynamic programming and template ranking from SP[3] (3). Both SPARKS 2.0 and SP[3] automatically make a weekly update for template and sequence libraries, i.e. based on new releases from the NCBI (sequences) and PDB (structures), respectively.

Performance

Testing on various benchmarks including LiveBench (22) indicates that SP[3] is slightly more accurate than SPARKS 2.0. SPARKS 2.0 and SP[3] are the two best servers for comparative modeling targets and are among the top single-method servers for all targets in the CASP 6 meeting that assessed 49 automatic webservers (http://predictioncenter.llnl.gov/casp6/meeting/presentations/talks.html).

Input and output

The input for both SPARKS 2.0 and SP[3] is the query sequence in the FASTA format and the number of structure models to be built is based on top ranked templates. The structure models are built by MODELLER (23). It usually takes 30 min to a few hours to complete the fold recognition of a sequence (depending on the size of the query protein and the load of the server computer). The output (in html format) contains the links to PSI-BLAST output for sequence profile, PSIPRED output for the secondary structure prediction, the top 10 sequence-to-structure alignments and the structure models (in PDB format) built based on the alignments. The significance of the sequence-to-structure alignment is indicated by the Z-score for each alignment. An alignment is significant if Z-score is >5.6 for SPARKS 2.0 and >6.3 for SP[3]. The thresholds were based on LiveBench 8 (22) for predicted models with MaxSub score (24) >0.01 when compared to their respective native structures. The output is now reported in a table format for easy understanding. Sample input and output with detailed line-to-line explanations are available online.

DFIRE ENERGY-BASED SERVERS

Overview

One bottleneck to the solution of the problems of how proteins fold, bind and function is the lack of an accurate energy function. The energy functions that are currently used by the computational biology community are obtained through either a physical-based (25) or a 'bioinformatics-based' statistical approach (26). Statistical energy functions are easy to produce and have been proven effective in many applications.

Description

Our group developed an all-atom statistical potential based on a new reference state named Distance-scaled, Finite, Ideal-gas REference (DFIRE). The DFIRE-based energy function has been successfully applied to structure (4) and docking selections (6), loop scoring (5), prediction of mutation-induced change in stability (4), and binding affinity of protein–protein (peptide) (6), protein–ligand (7) and protein–DNA complexes (7). These applications resulted in several servers: DMONOMER and DLOOP for scoring protein monomer and loop conformations, respectively; DMUTANT for predicting

mutant stability; DCOMPLEX and DDNA for predicting binding affinities of protein–protein/peptide complexes and those of protein–DNA complexes, respectively.

Performance

Comparisons between the DFIRE energy function and other knowledge-based or physical-based energy functions were made. For example, the DFIRE energy function was found to be comparable in accuracy to some physical-based energy functions equipped with various state-of-the-art solvation models [illustrated in loop selection (5)] or empirical energy functions with many adjustable terms [illustrated in docking (6) and prediction of protein–ligand binding affinities (7)]. The usefulness of the DFIRE energy-based servers was also independently verified in predicting protein stability of arc repressor mutants by using our webserver (27).

Input and output

The input for DMONOMER, DCOMPLEX and DDNA is the atomic coordinates file in PDB format and the chain ID, while DLOOP needs additional input for loop location. The outputs for these four servers are corresponding DFIRE energy scores and/or binding affinities. DCOMPLEX also gives an indication whether input complex is a genuine homodimer or crystal artifact. Inputs for DMUTANT is structure file, Chain ID and residue position. The output is the stability change due to the mutation of a specified residue into 19 other residues. Note that the binding affinities predicted by DCOMPLEX and DDNA were shifted and/or scaled based on test sets used in publication. Sample input and output with detailed explanations are available online for each server.

TCD

Our group developed a parameter called total contact distance (TCD) to predict folding rates of small two-state proteins (8). This parameter was built on the observation that either contact order (CO) or long-range order (LRO) parameter has a significant correlation with the logarithms of folding rates (28,29).

The TCD web-server takes the inputs of the structure file, chain ID and residue range of interest for a specific protein. Its output is the calculated value of TCD as well as the predicted folding rate. The auxiliary TCD transition-state server presents the predicted TCD, the approximate size of the folding transition state of a given protein (9).

DOGMA

Proteins are made of functional domains. One effective method to uncover the function of proteins on a genomic scale is by analyzing the network graph of domain–domain interactions (30). A domain graph consists of all domains found in a given proteome. Each vertex (node) represents a distinct domain and two vertices are linked by an edge if they occur together in at least one protein.

DOGMA is an online server implementing CADO (Comparative Analysis of Protein Domain Organization) algorithms (31) and applying it in the comparative analysis of domain graph between plant and other 55 organisms (9 eukaryote, 30 bacteria and 16 archae) (10). The input includes name(s) of Pfam domain(s) (32) and organism(s) to be compared with plant (taken *Arabidopsis* as representative). Depending on the option chosen, output can be domain graph, shortest path between two given domains, phylogentic profile, and others in both comparative and graphical format. Although the original paper is about comparison between plant and other proteomes, DOGMA could be used to analyze any one against other 55 proteomes.

REFERENCES

1. Zhou,H. and Zhou,Y. (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci.*, **12**, 1547–1555.
2. Zhou,H. and Zhou,Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.
3. Zhou,H. and Zhou,Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, **58**, 321–328.
4. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
5. Zhang,C., Liu,S. and Zhou,Y. (2004) Accurate and efficient loop selections using DFIRE-based all-atom statistical potential. *Protein Sci.*, **13**, 391–399.
6. Liu,S., Zhang,C., Zhou,H. and Zhou,Y. (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.
7. Zhang,C., Liu,S., Zhu,Q. and Zhou,Y. (2005) A knowledge-based energy function for protein–ligand, protein–protein and protein–DNA complexes. *J. Med. Chem.* in press.
8. Zhou,H. and Zhou,Y. (2002) Folding rate prediction using total contact distance. *Biophys. J.*, **82**, 458–463.
9. Bai,Y., Zhou,H. and Zhou,Y. (2004) Critical nucleation size in the folding of small apparently two-state proteins. *Protein Sci.*, **13**, 1173–1181.
10. Liu,S., Zhang,C. and Zhou,Y. (2005) Domains and domain combinations in *Arabidopsis thaliana* by comparative analysis. *J. Proteome Res.*, in press.
11. Eisenberg,D., Weiss,R.M., Terwilliger,T.C. and Wilcox,W. (1982) Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.*, **17**, 109–120.
12. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
13. Rost,B., Casadio,R. and Fariselli,P. (1996) Refining neural network predictions for helical transmembrane proteins by dynamic programming. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **4**, 192–200.
14. vonHeijne,G. (1994) Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167–192.
15. Eilers,M., Patel,A.B., Liu,W. and Smith,S.O. (2002) Comparison of helix interactions in membrane and soluble α-bundle proteins. *Biophys. J.*, **82**, 2720–2736.

16. Traxler,B., Boyd,D. and Beckwith,J. (1993) The topological analysis of integral cytoplasmic membrane proteins. *J. Membr. Biol.*, **132**, 1–11.
17. Jayasinghe,S., Hristova,K. and White,S.H. (2001) MPtopo: a database of membrane protein topology. *Protein Sci.*, **10**, 455–458.
18. Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci*, **11**, 2774–2791.
19. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
20. Kernytsky,A. and Rost,B. (2003) Static benchmarking of membrane helix predictions. *Nucleic Acids Res.*, **31**, 3642–3644.
21. Godzik,A. (2003) Fold recognition methods. *Methods Biochem. Anal.*, **44**, 525–546.
22. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Livebench-1: large-scale automated evaluation of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.
23. Marti-Renom,M., Stuart,A., Fiser,A., Sanchez,R., Melo,F. and Šali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
24. Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.
25. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
26. Tanaka,S. and Scheraga,H.A. (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.
27. deArmas,R.R., Diaz,H.G., Molina,R. and Uriarte,E. (2004) Markovian backbone negentropies: molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants. *Proteins*, **56**, 715–723.
28. Plaxco,K.W., Simons,K.T. and Baker,D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.
29. Grombiha,M.M. and Selvaraj,S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.*, **310**, 27–32.
30. Barabsi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Rev. Genet.*, **5**, 101–113.
31. Ye,Y.Z. and Godzik,A. (2004) Comparative analysis of portein domain organization. *Genome Res.*, **14**, 343–353.
32. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.