Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3359–3363***
DOI: 10.1093/nar/gkg589

Servers for sequence–structure relationship analysis and prediction

Zsuzsanna Dosztányi, Csaba Magyar, Gábor E. Tusnády, Miklós Cserző[1], András Fiser[2] and István Simon*

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1518 Budapest, PO Box 7, Hungary, [1]Bioinformatics Group, Agricultural Biotechnology Center, Gödöllő, Szent-Györgyi A. út 4., H-2100 Hungary and [2]Department of Biochemistry and Seaver Foundation Center for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Received February 14, 2003; Revised and Accepted April 7, 2003

ABSTRACT

We describe several algorithms and public servers that were developed to analyze and predict various features of protein structures. These servers provide information about the covalent state of cysteine (CYSREDOX), as well as about residues involved in non-covalent cross links that play an important role in the structural stability of proteins (SCIDE and SCPRED). We also discuss methods and servers developed to identify helical transmembrane proteins from large databases and rough genomic data, including two of the most popular transmembrane prediction methods, DAS and HMMTOP. Several biologically interesting applications of these servers are also presented. The servers are available through http://www.enzim.hu/servers.html.

INTRODUCTION

Biochemical function of proteins is defined by their three-dimensional (3D) structures and under physiological conditions the 3D structures of proteins are defined by their primary sequences. Nowadays, amino acid sequences can be obtained in a fast and automated way, however solving the 3D structures of proteins is still a much more complicated and often unsuccessful endeavor. On average 20–30% of a typical genome is predicted to be partially or fully membrane immersed (1) and another 10–20% predicted to contain unstructured proteins (2,3). The experimental solution of structures, even for globular proteins, is far from trivial. Furthermore, the complex 3D structure of proteins makes it often complicated to extract the relevant biological information, despite knowing the atomic coordinates (4).

Therefore there is a strong demand for automatic servers which can predict structure related information from amino acid sequence alone or offer an analytical tool to understand the relationship between structural and functional properties of proteins. In this paper, we present a collection of algorithms and servers, providing information about the covalent state of Cys, identifying and predicting residues involved in non-covalent interactions and predicting putative transmembrane segments and their topology. All the discussed algorithms and servers were developed by current and former members of our group at the Institute of Enzymology, Hungarian Academy of Sciences with the exception of the DAS and DAS-TMfilter membrane protein prediction methods, which were developed as joint projects with laboratories in Austria, France and Sweden.

PREDICTING THE REDOX STATE OF Cys RESIDUES AND ELEMENTS OF COVALENT CROSS-LINKS IN PROTEINS

From the view point of covalent structure, the amino acid sequence determination from gene sequencing is not complete, because it does not distinguish cysteine and half cystine residues, the two different covalent structures coded by the same triplets in DNA, and it does not account for various other post translational covalent modifications. However, the covalent state of cysteines is often a major factor determining functional features of proteins: Cys is the second most frequent active site residue (5). In addition, the covalent state of Cys provides information about the possible cellular location of proteins and many small proteins simply cannot form stable structures without disulfide bonds. Information on Cys residues can also greatly enhance the performance of *ab initio* modeling studies (6).

Several methods have been developed to predict disulfide bond forming Cys residues from the sequence, by setting up disulfide bond forming statistical potentials or using neural network based approaches or a combination of neural network based approaches with hidden Markov models (7–9). Earlier methods could be significantly improved by incorporating evolutionary information of Cys residues from multiple sequence alignments (10,11). These approaches proved to be not only more effective but our analysis revealed that the biologically meaningful borderline lies rather in between

*To whom correspondence should be addressed. Tel: +36 14669276; Fax: +36 14665465; Email: simon@enzim.hu

the different oxidation states of Cys and not in between half cystines and cysteins (11,12). An automated server, CYSREDOX (http://alto.rockefeller.edu/cysredox/cysredox.html) (Table 1) was set up that implements this latter approach predicting the oxidation state of Cys residues with an accuracy of 82% (11).

The CYSREDOX server takes a single sequence as input, collects relevant homologous sequences from the NR database of NCBI using the Psi-Blast algorithm (13), and aligns them with the ClustalW program (14). Next, the conservation of residues in the resulting multiple sequence alignment is analyzed by calculating conservation scores according to the number of commonly shared or systematically neglected physicochemical features in each position. The scale of normalized conservation scores has been calibrated in our study and is used to estimate the probable redox form of Cys.

ANALYZING AND PREDICTING RESIDUES INVOLVED IN NON-COVALENT CROSS LINKS IN PROTEINS

Under physiological conditions, the native state of globular protein is dominated by a relatively well-defined conformation. The stability of this conformation is ensured by a large number of non-covalent interactions. Although the energy of individual interactions is small, comparable to the thermal energy of one degree of freedom, interactions can have a significant contribution to stability when they act cooperatively. Furthermore, the prevention of large-scale motions requires long-range interactions, which cross-link regions far apart in the sequence. These simple arguments highlight the importance of cooperative long-range interactions in ensuring the kinetic stability of proteins. Residues involved in these clusters have been shown to have a primary effect on the rate of spontaneous unfolding due to thermal fluctuation (15). It is clear that identifying these residue clusters would greatly enhance our understanding of stabilizing factors of proteins.

The concept of stabilization centers

We suggested an algorithm to select a subset of residues from known protein structures, which have the desired property of stabilizing interaction (i.e. they are involved in forming cooperative long-range interactions). These clusters, called stabilization centers, were defined in the following way. Two residues are part of a stabilization center if (i) they are involved in long range interactions (i.e. they are at least 10 residues apart in the primary structure and have at least one pair of non-hydrogen atoms which are closer than the sum of their van der Waals radii plus 1.0 Å) and (ii) two supporting residues can be selected from both of their flanking tetrapeptides, which together with the central residues form at least seven out of the possible nine contacts (16). The importance of residues involved in these contacts is reflected in their higher evolutionary conservation compared to the rest of the sequence. Some special roles played by stabilization centers in the structural stability of proteins has been reported for MHC proteins (17) and for the class of four helix bundle proteins (18).

Table 1. Name and internet address of servers

Server	Url
CYSREDOX	http://alto.rockefeller.edu/cysredox/cysredox.html
SCIDE	http://www.enzim.hu/scide
SCPRED	http://www.enzim.hu/scpred
DAS	http://www.sbc.su.se/~miklos/DAS
HMMTOP	http://www.enzim.hu/hmmtop

Locating stabilization center residues in proteins of known structure

The SCIDE server is devoted to the analysis of stabilization centers in known protein structures (19). The server takes a PDB coordinate file as an input—either specified by its PDB code or uploaded as a file—and locates the residues involved in stabilization centers. The analysis can be narrowed down to regions specified by the user but stabilization centers formed between different chains can be studied as well. This latter approach is especially important when a protein chain is not stable on its own but only by acquiring extra stability through interactions with other chains. The result is presented in text or graphical format. The text output returns the sequence of a protein indicating residues involved in stabilization centers. The graphical output shows the image of the contact map of the protein, highlighting the stabilization centers among other inter-residue interactions. Clicking to stabilization centers brings up a detail of the contact map, highlighting which contacts contribute to the formation of the given stabilization center. The graphical output is useful for investigating protein structures in detail, while the text output can be used for automated analysis of larger datasets. This program is available at http://www.enzim.hu/scide (Table 1).

Predicting stabilization center residues from amino acid sequences

The information about positions in sequence, which are likely to be involved in long-range contacts, can aid structure predictions. For example, low energy structures of small protein segments can easily be calculated by conformational energy minimization but putting the pieces of the puzzle together requires additional information on long range interactions (6,20,21). Another possible application involves protein engineering. It was suggested that amino acid replacements in these centers could significantly influence protein stability (22). For these and similar purposes a method was developed to predict residues of stabilization centers from amino acid sequences alone. The algorithm is based on our finding that the sequential neighborhood of residues involved in stabilization centers was different from residues not involved in stabilization centers. Although involvement in stabilization center refers to a global property of proteins, some information is also stored locally in agreement with the minimal frustration theory of proteins (23). Based on the difference in sequential neighborhood, an artificial neural network was trained to recognize the location of probable stabilization center residues from the sequence. It takes a 17 residue-long segment and predicts whether the

central residue is involved in stabilization center or not. Using a single sequence as an input, the accuracy of the prediction was 65%, which improved marginally using multiple alignments (68%) (16). The SCPRED server is available at http://www.enzim.hu/scpred (Table 1).

PREDICTION METHODS CONCERNING INTEGRAL MEMBRANE PROTEINS

Integral membrane proteins play a central role in material and information exchange between the cell and the outside world including the adjacent cells of the living organism. Therefore, these proteins are the primary targets of many basic research and pharmaceutical studies. Since it is extremely difficult to crystallize these proteins and they are much too big for NMR structure determination, theoretical approaches for obtaining structural information about these proteins are especially important. Several methods have been developed to predict the location of the transmembrane segments in the primary structure of proteins and the orientation of these segments, that is the topology from the amino acid sequences (reviewed in 24 and 25). Since the membrane interior is rather hydrophobic and contains no proton donor or acceptor, the transmembrane segments of the polypeptide chain should be composed mainly of hydrophobic residues. In order to minimize the free energy of the protein–membrane system, regions interacting with the membrane interior are dictated to form alpha helical or beta-barrel structures, which ensures that all proton donors and acceptors are fully satisfied in hydrogen bonds. So far, beta barrel structures have been found only in bacterial outer membrane (26), while the overwhelming majority of the transmembrane segments form alpha helices.

Transmembrane segments can be identified on the bases of the hydrophobic residue content above the average (27) but this simple method is rather inaccurate. A more accurate method can be obtained based on the observation that all transmembrane segments recognize each other as related ones in the course of sequence alignment (28).

The DAS and DAS-TMfilter algorithms of transmembrane segment predictions

The DAS ('Dense Alignment Surface') method is based on the very traditional dot-plot of a query sequence against a well-characterized reference. The algorithm differs from the original method in three important aspects. The various segments of the two sequences are compared and the stringency score is calculated in a sliding window. Scores exceeding a certain value are marked on the alignment surface as hits. In the DAS algorithm, this limit is set to very low level resulting in thousands of hits all over the surface. However, the distribution of the hits on the surface is not random but follows a characteristic 'chess-board' like pattern—due to our special substitution matrix (29). If the two sequences in question are transmembrane proteins, the hits are concentrated in the areas at the intersections of the transmembrane segment positions. The evaluation of the alignment surface is the third unique aspect of the DAS algorithm. The surface is scanned for hits at each position of the query along the reference sequence and the score is summed. This procedure converts the surface to a profile projecting it to the query axis. As the hits are concentrated around the transmembrane segments, these parts are represented by high values in the resulting profile. In the prediction procedure, the segments above an empirical limit are identified as helical transmembrane fragments. Please note: the endpoints of the segments do not correspond to the points where the protein enters into the membrane. The experimental database is not sufficiently accurate to verify this kind of prediction. The efficiency of the method in terms of number of correctly predicted transmembrane segments is >90% (25).

The public server implementing the DAS algorithm is available at http://www.sbc.su.se/~miklos/DAS/ (Table 1). The server takes an input sequence as the query and runs it against the library of 44 experimentally documented transmembrane sequences. It returns the average profile of the query profiles of the individual runs and lists the segments above the empirical limit as possible transmembrane locations (30).

Recently, it became obvious that while the most advanced transmembrane prediction methods perform with a success rate >90% for integral helical proteins, they tend to incorrectly identify other hydrophobic clusters in globular proteins as 'transmembrane segments'. It was shown that even the best transmembrane prediction methods falsely identify at least one transmembrane segment in >25% of the non-transmembrane protein sequences (31,32). Since servers often used to analyze rough genetic sequence data, we upgraded the DAS algorithm to be able to filter out false positive transmembrane segment results. This modified DAS-TMfilter method performs a DAS prediction first: runs the query against the updated TM-sequence library of well-documented TM-proteins. In the second step the query is used in a 'reverse prediction' cycle as 'reference' sequence and a set of known TM-proteins tested for TM-regions. True TM-proteins perform well in this test (i.e. most of the known TM-segments are detected) while these segments are missed when the set is tested with a globular sequence. This even happens if the globular sequence contains a false positive TM-segment prediction. In that case, the original DAS prediction is overruled and the sequence is identified as globular (33).

The HMMTOP server

Another transmembrane segment and topology prediction method developed in our laboratory is the HMMTOP method (34). This prediction method is based on the principle that the topology of the transmembrane proteins is determined by the maximum divergence of amino acid composition of sequence segments. These segments are located in different areas of the cell, for example in the membrane interior, in the membrane border areas, in the cytosol or in the extra-cytosolic space. The physicochemical properties of the cell within these compartments are different, therefore the amino acid composition of polypeptide segments passing through these areas has to be different. These amino acid compositions should differ the most—according to the maximum-likelihood law—considering all possible segmentations of a given sequence. This segmentation can be found by probabilistic methods, such as hidden Markov models (HMM). Because HMM is applied as an optimization method, there is no need to teach the model, and the parameters are independent of the available

databases. The optimization performed by the Baum–Welch algorithm provides the model parameters, which in turn can be used by the Viterbi algorithm to find the most probable segmentation of a sequence resulting in the topology prediction.

The improved version of HMMTOP program can take into account preliminary experimental information or any other knowledge about the topology as conditional probabilities (35). It is worth emphasizing that this method solves the segmentation problem (mentioned above) using conditional probability, to assess whether a certain sequence is located in a certain segment, which in turn can affect the predicted topology of other parts of the protein sequence. This means that the number, position and orientation of the transmembrane helices can change according to the given condition. According to our knowledge, this new version of HMMTOP is the first method that can incorporate this type of conditions in a prediction.

The segment accuracy of HMMTOP 2.0 was >97% on a dataset (36) containing 148 transmembrane proteins, while the accuracy of transmembrane helix and topology prediction at the protein level was 90% and 73%, respectively (35). HMMTOP is freely available to non-commercial users at http://www.enzim.hu/hmmtop (Table 1). Source code is available upon request to academic users. The web server can handle both 'post' and 'get' http requests, allowing external linking to prediction results. The server accepts sequence as plain text or in fasta and NBRF/PIR format. There are two submission forms, one for doing simple prediction on one sequence, and one for advanced users, where various options can be given to the server. The output format can be varied as well, from a very simple one line description of topology and transmembrane helices to a detailed output containing the full sequence(s), as well as the parameters of the HMM used for the prediction.

ALGORITHMS—USE NOT ONLY AS DIRECTED

The obvious field of application of the algorithms discussed in this paper is analyzing and predicting redox states of Cys residues, as well as residues composing stabilization centers and transmembrane helices. However, these algorithms can be applied in non-conventional ways, too. The different results of cysteine/half-cysteine and of redox state prediction methods triggered the proposed mechanism for prion protein polymerization (37). The analysis of stabilization center residues led to the proposed model of divergent evolution of certain restriction endonucleases (38). The application of transmembrane prediction methods in the structural study of a non-transmembrane protein, prion, led to our proposal that the very special structure feature of this protein is the consequence of its putative transmembrane origin (32). Therefore, we would like to encourage the potential users to apply these algorithms and servers in a wide area of research.

ACKNOWLEDGEMENTS

We would like to thank J.M. Bernassau and B. Maigret of University of Nancy; E. Wallin, G. von Heijne and A. Elofsson of Stockholm University; and B. Eisenhaber and F. Eisenhaber of IMP Bioinformatics, Vienna, for their contributions in the development of DAS and DAS-TMfilter algorithms and servers. Financial support from grants BIO-0005/2001, OTKA T34131 and F043609 are acknowledged. G.E.T. was supported by the Bolyai János Scholarship and OTKA (D42207) Fellowship, and C.M. by the OTKA Fellowship (D38487).

REFERENCES

1. Krogh,A., Larsson,B., von Heijne,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.
2. Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.
3. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
4. Chance,M.R., Bresnick,A.R., Burley,S.K., Jiang,J.S., Lima,C.D., Sali,A., Almo,S.C., Bonanno,J.B., Buglino,J.A., Boulton,S. *et al.* (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.*, **11**, 723–738.
5. Fiser,A., Simon,I. and Barton,G.J. (1996) Conservation of amino acids in multiple alignments: aspartic acid has unexpected conservation. *FEBS Lett.*, **397**, 225–229.
6. Simon,I., Glasser,L. and Scheraga,H.A. (1991) Calculation of protein conformation as an assembly of stable overlapping segments: application to bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA*, **88**, 3661–3665.
7. Fiser,A., Cserző,M., Tüdős,É. and Simon,I. (1992) Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. *FEBS Lett.*, **302**, 117–120.
8. Muskal,S.M., Holbrook,S.R. and Kim,S.H. (1990) Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng.*, **3**, 667–672.
9. Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.*, **11**, 2735–2739.
10. Fariselli,P., Riccobelli,P. and Casadio,R.(1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins*, **36**, 340–346.
11. Fiser,A. and Simon,I. (2000) Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, **16**, 251–256.
12. Fiser,A. and Simon,I. (2002) Predicting redox state of cysteines in proteins. *Methods Enzymol.*, **353**, 10–21.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
15. Abkevich,V.I., Gutin,A.M. and Shakhnovich,E.I. (1995) Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.*, **252**, 460–471.
16. Dosztányi,Zs., Fiser,A. and Simon,I. (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.*, **272**, 597–612.
17. Simon,Á., Dosztányi,Zs., Rajnavölgyi,É. and Simon,I. (2000) Function-related regulation of the stability of MHC proteins. *Biophys. J.*, **79**, 2305–2313.
18. Fuxreiter,M. and Simon,I. (2002) Role of stabilization centers in 4 helix bundle proteins. *Proteins*, **48**, 320–326.
19. Dosztányi,Zs., Magyar,Cs., Tusnády,G.E. and Simon,I. (2003) SCide: identification of stabilization centers in proteins. *Bioinformatics*, **19**, 899–900.
20. Panchenko,A.R., Luthey-Schulten,Z., Cole,R. and Wolynes,P.G. (1997) The foldon universe: a survey of structural similarity and self-recognition of independently folding units. *J. Mol. Biol.*, **272**, 95–105.

21. Kolodny,R., Koehl,P., Guibas,L. and Levitt,M. (2002) Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.*, **323**, 297–307.
22. Simon,Á., Dosztányi,Zs., Magyar,Cs., Szirtes,G., Rajnavölgyi,É. and Simon,I. (2001) Stabilization centers and protein stability. *Theor. Chem. Acc.*, **106**, 121–127.
23. Miyazawa,S. and Jernigan,R.L. (2003) Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space. *Proteins*, **50**, 35–43.
24. Simon,I., Fiser,A. and Tusnády,G.E. (2001) Predicting protein conformation by statistical methods. *Biochim. Biophys. Acta*, **1549**, 123–136.
25. Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.
26. Buchanan,S.K. (1999) Beta-barrel proteins from bacterial outer membranes: structure, function and refolding. *Curr. Opin. Struct. Biol.*, **9**, 455–461.
27. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
28. Cserző,M., Bernassau,J.M., Simon,I. and Maigret,B. (1994) New alignment strategy for transmembrane proteins. *J. Mol. Biol.*, **243**, 388–396.
29. Tüdős,É., Cserző,M. and Simon,I. (1990) Predicting isomorphic residue replacements for protein design. *Int. J. Pept. Protein Res.*, **36**, 236–239.
30. Cserző,M., Wallin,E., Simon,I., von Heijne,G. and Elofsson A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.*, **10**, 673–676.
31. Jayasinghe,S., Hristova,K. and White,S.H. (2001) Energetics, stability, and prediction of transmembrane helices. *J. Mol. Biol.*, **312**, 927–934.
32. Tompa,P., Tusnády,G.E., Cserző,M. and Simon,I. (2001) Prion protein: evolution caught en route. *Proc. Natl Acad. Sci. USA*, **98**, 4431–4436.
33. Cserző,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2002) On filtering false positive transmembrane protein predictions. *Protein Eng.*, **15**, 745–752.
34. Tusnády,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.
35. Tusnády,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
36. Möller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterized integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
37. Tompa,P., Tusnády,G.E., Friedrich,P. and Simon,I. (2002) The role of dimerization in prion replication. *Biophys. J.*, **82**, 1711–1718.
38. Fuxreiter,M. and Simon,I. (2002) Protein stability indicates divergent evolution of PD-(D/E)XK type II restriction endonucleases. *Protein Sci.*, **11**, 1978–1983.