TMpro web server and web service: transmembrane helix prediction through amino acid property analysis

Madhavi Ganapathiraju1, Christopher Jon Jursa2, Hassan A. Karimi2, and Judith Klein-Seetharaman1,3
1Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
2School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15213
3Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA

Summary

TMpro is a transmembrane (TM) helix prediction algorithm that uses language processing methodology for TM segment identification. It is primarily based on the analysis of statistical distributions of properties of amino acids in transmembrane segments. This article describes the availability of TMpro on the internet via a web interface. The key features of the interface are: (i) output is generated in multiple formats including a user-interactive graphical chart which allows comparison of TMpro predicted segment locations with other labeled segments input by the user, such as predictions from other methods. (ii) Up to 5000 sequences can be submitted at a time for prediction. (iii) TMpro is available as a web server and is published as a web service so that the method can be accessed by users as well as other services depending on the need for data integration.

1 INTRODUCTION

Membrane proteins are encoded by ~30% of the genes in typical genomes and play key functional roles especially as ion channels and as receptors in cell signal pathways. In eukaryotes, all known membrane proteins in the plasma membrane consist of alpha helical transmembrane (TM) bundles connected by loops. Accurate computational prediction of transmembrane helical segments is important for modeling membrane protein 3D structure and function. Early methods for TM helix prediction were based on the hydrophobicity of TM segments but suffered low accuracy (Kyte and Doolittle, 1982). Previous best models used hundreds or thousands of free parameters which were tuned to fit the relatively small number of proteins with known TM segments. Further, they were often restricted to the topology ‘cytoplasmic-transmembrane-extracellular’, and cannot be expected to adequately predict membrane proteins that do not conform to this topology. Recent crystal structures reveal an increasing number of novel architectures. Thus, new high-accuracy methods using less restrictive models are needed.
2 TMpro

TMpro is a computational method for TM helix prediction from the primary sequence of a protein through the analysis of the properties of the amino acids in the sequence (Ganapathiraju et al., 2007). The algorithm is analogous to latent semantic analysis used in text processing, and gains significant improvements over comparable methods. TMpro uses only 25 free parameters and achieves a 95% segment F-score corresponding to a 50% reduction in error rate in benchmark analysis as compared to TMHMM. Tested on more recent and larger data sets, the F-score is 93% corresponding to 30–45% reduction in error rate compared to recent methods (TMHMM, SOSUI and DAS). In only 1% of cases, a membrane protein is erroneously classified as a soluble protein, while achieving 90% accuracy in distinguishing membrane and soluble proteins. The Q^2 of TMpro is 80%, while other best methods achieve ~84% Q^2. Here, we describe the availability of TMpro as a web resource.

3 TMpro WEB SERVER

TMpro can be accessed through a traditional web interface where a user can submit up to 5000 protein sequences at a time and obtain the prediction of TM segments. The basic TM prediction requires only a web browser. However, to view the results with a user-interactive chart, a Java Run Time Environment (JRE) is required, which may be downloaded from http://www.java.com/en/download/.

3.1 Input

A single sequence can be submitted as raw or FASTA sequence or by its Swissprot or PDB id, or multiple sequences can be submitted as a FASTA file.

3.2 Plain text output

Textual output gives residue ranges of the predicted TM segments. Results are also sent as an email to the user if an address is provided.

3.3 User interactive chart output

In addition to a text output of predicted segments, a user-interactive chart is generated as a Java Applet® so that the user can visualize the positions of the predicted TM segments along the primary sequence. The chart shows the analog output of the TMpro neural network and the predicted helix segments as a plot. To allow the user to compare TMpro prediction with additional sequence annotations, we created the ability to enter such information, upon which the chart is updated and displays this user input data. Any number of sequence annotations can be added by the user to the plot (Fig. 1). These can be either analog value for the residue positions in the sequence (e.g. predicted disorder in the protein), or start and end positions of segment information (e.g. predicted TM helices). Information about the protein from numerous sources are presented visually for comprehensive analysis.

3.4 Standardized TMpro format

A text output file is created for each submitted protein. It contains the primary sequence information and TMpro prediction information. One additional line is added for every manual input provided by the user. The standardized TMpro format file created is useful for client-side computerized processing of the TMpro predictions and additional information input by the user, if any. This is a unique feature of our interface and is particularly important in TM structure prediction, where little data is available and user input is valuable in integrating different information sources for a specific membrane protein of interest.
4 OPEN AND INTEROPERABLE WEB SERVICE

Interoperability of TMpro with other applications is enabled through a web service, which is made adherent to current W3C standards including Extensible Markup Language (XML) based languages such as Simple Object Access Protocol (SOAP) and Web Services Description Language (WSDL). For instance, users can write client-side applications to send protein sequences and receive predictions through SOAP. The service can be invoked by calling operation “fetchResponse” with four arguments in the following order: Swissprot or PDB ID, sequence, your email address and personal notes. Please note either an ID or sequence is required but not both. The web service’s operations are also described in the online WSDL document, which are accessible by adding “?wsdl” to the web service URL.

Acknowledgments

This work was supported by the National Science Foundation grants 0225656, 0225636 and CAREER CC044917. M.G. and J.K.S thank Raj Reddy and N. Balakrishnan for the discussions during the development of TMpro. M.G. and J.K.S have developed the TMpro algorithm and designed the web interface. C.J.J. and H.A.K created the web interface and WSDL/SOAP extensions.

References


Interactive chart. TMpro generates a graphical chart as a Java Applet showing the analog output of TMpro neural network and its predicted TM segments. In the figure, predictions by TMpro of the K+ channel protein (Swissprot ID KCSA_STRCO) are shown in red. The remaining lines in dark blue, light blue, gray, green and yellow are experimentally known TM segments, experimentally known pore-forming helix, selectivity filter and predicted TM segments by other TM prediction algorithms, TMHMM (Krogh et al., 2001) and PRODIV-TMHMM (Viklund and Elofsson, 2004) as examples. Visualization of this information shows that TMHMM and PRODIV-TMHMM confuse the pore-forming helix and selectivity filter together to be an additional TM segment. This visualization can aid researchers of specific proteins in drawing conclusions by integrating information from multiple sources. Check boxes are provided to selectively view specific sources of information from among those entered.