SOSUI: classification and secondary structure prediction system for membrane proteins

Takatsugu Hirokawa, Seah Boon-Chieng and Shigeki Mitaku[1]

Department of Biotechnology, Tokyo University of Agriculture and Technology, Nakacho, Koganei, Tokyo 184, Japan

Received on October 28, 1997; revised and accepted on January 23, 1998

Abstract

Summary: *The system SOSUI for the discrimination of membrane proteins and soluble ones together with the prediction of transmembrane helices was developed, in which the accuracy of the classification of proteins was 99% and the corresponding value for the transmembrane helix prediction was 97%.*
Availability: *The system SOSUI is available through internet access: http://www.tuat.ac.jp/~mitaku/sosui/.*
Contact: *sosui@biophys.bio.tuat.ac.jp.*

While genome sequencing has resulted in a huge increase in the number of sequences available for analysis (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995; Bult *et al.*, 1996; Mewes *et al.*, 1996; Tomb *et al.*, 1997), a significant percentage still have no homology with any other known protein. Therefore, new theoretical methods that do not depend on the sequence alignment, but on the physicochemical properties of amino acid sequences, still play an important role. For this purpose, we planned to construct a structure prediction system with three modules for: (i) the discrimination between soluble and membrane proteins; (ii) secondary structure prediction for soluble and membrane subsets; (iii) tertiary structure prediction. Here, we describe an application called SOSUI which distinguishes between membrane and soluble proteins from amino acid sequences, and predicts the transmembrane helices for the former. This application has very high accuracy for prediction and can be calculated very quickly.

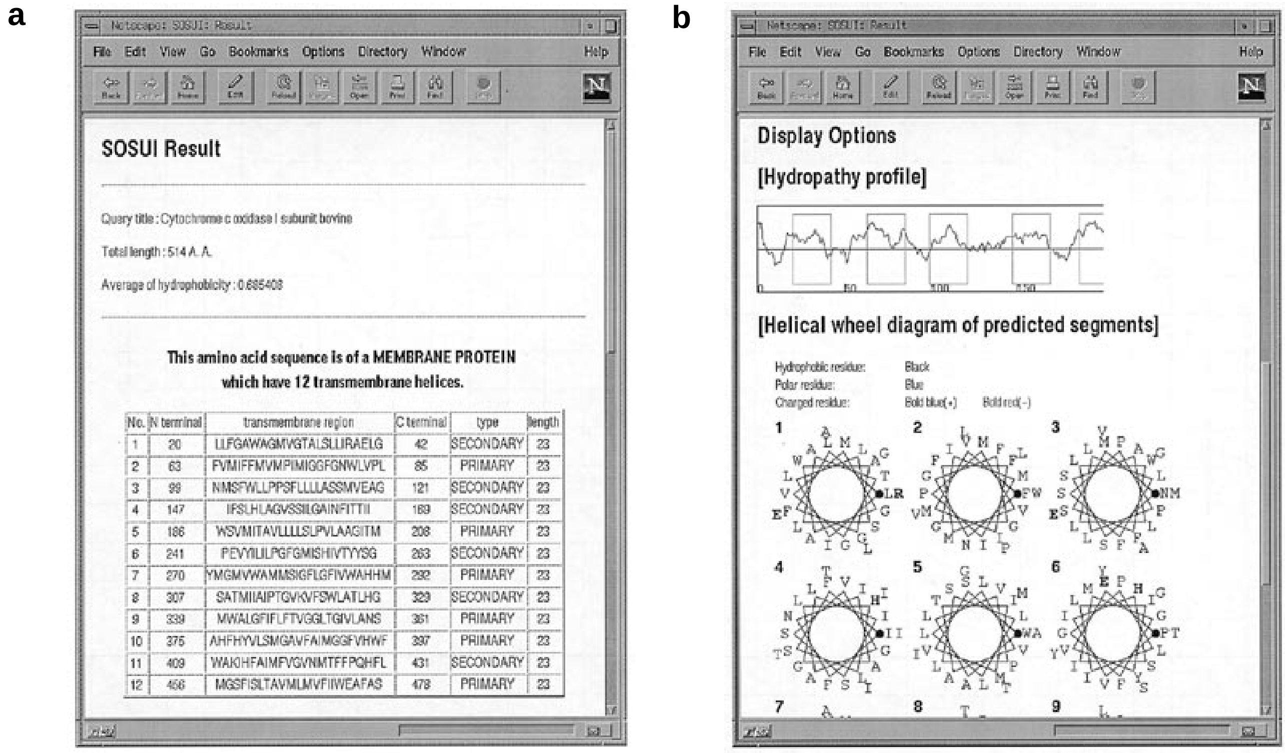
There are three basic assumptions in the SOSUI system. First, membrane proteins are characterized by at least one, particularly hydrophobic, primary transmembrane helix. Secondary hydrophilic transmembrane helices may also exist in multispanning membrane proteins even though their hydrophobicity is in fact similar to the hydrophobic segments of soluble proteins. The possible role of secondary transmembrane helices is the formation of active sites of proteins. Third, the primary transmembrane helices are stabilized by a combination of amphiphilic side chains at the helix ends as well as high hydrophobicity in the central region. When polar interactions are also found in the center of a primary helix, their existence is usually for the stabilization of transmembrane helices. For our software, we have used four physicochemical parameters: the hydropathy index of Kyte and Doolittle (Kyte and Doolittle, 1982), an amphiphilicity index, an index of amino acid charges, and the length of each sequence. The second parameter expresses the amphiphilicity of polar side chains and was devised by the calculation of the transfer energy of the hydrocarbon part of a polar side chain. The values of this parameter are finite for large polar residues: 3.67, Lys; 2.45, Arg; 1.45, His; 1.27, Glu; 1.25, Gln; 6.93, Trp; 5.06, Tyr.

A polypeptide can become a membrane protein if it contain at least one transmembrane helix. Membrane proteins are discriminated from soluble ones in the present method by predicting the existence of a primary transmembrane helix for each sequence. Analysis of the three-dimensional structure of membrane proteins indicated that about one-third of transmembrane helices are very hydrophobic helices. Those very hydrophobic helices could be discriminated from false candidates of transmembrane segments by using two parameters: the average hydrophobicity and the average amphiphilicity index in the end region of helices. The remaining transmembrane helices were predicted, adding the information on the distribution of polar residues without electric charges.

The SOSUI system is a WWW-based tool and users can input their query sequence on the submission. Results are typically returned in 1 min. Two predictions and two graphs are presented in the output page: (i) the type of protein; (ii) the region of transmembrane helices when the protein is a membrane type; (iii) a graph of the hydropathy plot; (iv) helical wheel diagrams of all transmembrane helices (Figure 1). Input sequence length is limited to the range of 20–5000 amino acids. Diagrams are displayed by a Java Applet program. SOSUI is available at http://www.tuat.ac.jp/~mitaku/sosui/.

To test SOSUI, we applied it to two kinds of data sets: 101 membrane proteins from Fariselli's data set (Fariselli and

[1] *To whom correspondence should be addressed*


a
b
Netscape: SOSUI: Result
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Edit Reload Images Open Print Find Stop

SOSUI Result

Query title : Cytochrome c oxidase I subunit bovine

Total length : 514 A. A.

Average of hydrophobicity : 0.685408

This amino acid sequence is of a MEMBRANE PROTEIN which have 12 transmembrane helices.

No.	N terminal	transmembrane region	C terminal	type	length
1	20	LLFGAWAGMVGTALSLLIRAELG	42	SECONDARY	23
2	63	FVMIFFMVMPIMIGGFGNWLVPL	85	PRIMARY	23
3	99	NMSFWLLPPSFLLLLASSMVEAG	121	SECONDARY	23
4	147	IFSLHLAGVSSILGAINFITTII	169	SECONDARY	23
5	186	WSVMITAVLLLLSLPVLAAGITM	208	PRIMARY	23
6	241	PEVYILILPGFGMISHIVTYYSG	263	SECONDARY	23
7	270	YMGMVWAMMSIGFLGFIVWAHHM	292	PRIMARY	23
8	307	SATMIIAIPTGVKVFSWLATLHG	329	SECONDARY	23
9	339	MWALGFIFLFTVGGLTGIVLANS	361	PRIMARY	23
10	375	AHFHYVLSMGAVFAIMGGFVHWF	397	PRIMARY	23
11	409	WAKIHFAIMFVGVNMTFFPQHFL	431	SECONDARY	23
12	456	MGSFISLTAVMLMVFIIWEAFAS	478	PRIMARY	23

Display Options

[Hydropathy profile]

[Helical wheel diagram of predicted segments]

Hydrophobic residue: Black
Polar residue: Blue
Charged residue: Bold blue(+) Bold red(-)

Fig. 1. (**a**) Output page of the SOSUI system on the WWW. A simple description of properties, the average hydrophobicity and length of the query sequence are first shown together with the result of the prediction for transmembrane helix regions. (**b**) A hydropathy plot of the amino acid sequence and helical wheel diagrams for predicted transmembrane helical segments.

Casadio, 1996) and 502 soluble proteins (Hobohm and Sander, 1994) for evaluation of system performance. Our system discriminated all proteins correctly, except for one in each set of data. In other words, one soluble protein was falsely predicted to be a membrane protein, and one membrane protein was assigned as a soluble protein. This results in an accuracy >99%. The data set of 101 membrane proteins contained 393 transmembrane helices, and the present system correctly predicted 373 of them, corresponding to an accuracy of ~97%.

Distribution of binary execute files of the SOSUI system for analysis of large data sets are now being prepared. The e-mail address for comments about SOSUI is sosui@biophys.bio.tuat.ac.jp.

Acknowledgements

The authors thank Dr M.Suwa and Dr M.Swindells for critical reading and revision of the manuscript. This work was supported by a grant-in-aid for scientific research on the priority area, genome science, from Monbusho of Japan.

References

Bult,C.J. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1072.

Fariselli,P. and Casadio,R. (1996) HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Comput. Applic. Biosci.*, **12**, 41–48.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fraser,C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Mewes,H.W. *et al.* (1997) Overview of the yeast genome. *Nature*, **387**, 7–65.

Tomb,J.-F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.