Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W99–W103***
doi:10.1093/nar/gkl351

Localizome: a server for identifying transmembrane topologies and TM helices of eukaryotic proteins utilizing domain information

Sunghoon Lee[1], Byungwook Lee[1], Insoo Jang[1], Sangsoo Kim[1,2,*] and Jong Bhak[1,*]

[1]National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, 52 Ueun-dong, Yuseong-gu, Daejeon 305-333, Republic of Korea and [2]Department of Bioinformatics, Soongsil University, 1-1 Sangdo-dong, Dongjak-gu, Seoul 156-743, Republic of Korea

Received February 14, 2006; Revised March 1, 2006; Accepted April 19, 2006

ABSTRACT

The Localizome server predicts the transmembrane (TM) helix number and TM topology of a user-supplied eukaryotic protein and presents the result as an intuitive graphic representation. It utilizes hmmpfam to detect the presence of Pfam domains and a prediction algorithm, Phobius, to predict the TM helices. The results are combined and checked against the TM topology rules stored in a protein domain database called LocaloDom. LocaloDom is a curated database that contains TM topologies and TM helix numbers of known protein domains. It was constructed from Pfam domains combined with Swiss-Prot annotations and Phobius predictions. The Localizome server corrects the combined results of the user sequence to conform to the rules stored in LocaloDom. Compared with other programs, this server showed the highest accuracy for TM topology prediction: for soluble proteins, the accuracy and coverage were 99 and 75%, respectively, while for TM protein domain regions, they were 96 and 68%, respectively. With a graphical representation of TM topology and TM helix positions with the domain units, the Localizome server is a highly accurate and comprehensive information source for subcellular localization for soluble proteins as well as membrane proteins. The Localizome server can be found at http://localizome.org/.

INTRODUCTION

Transmembrane (TM) proteins play important roles in biology. They are especially important in signal reception, molecular pumping and energy transduction. Their medical importance is also growing rapidly after the completion of many large genome projects. Genomics projects rendered a great need to analyze TM proteins bioinformatically. To analyze TM proteins *en masse* and automatically, it is required to predict the loop topologies between TM helices, as well as TM helix positions. Since the loops (or terminals) between TM helices play crucial roles in analyzing protein–protein interactions and finding drug targets, the accurate prediction of TM topology can significantly enhance the efficiency of TM protein analyses. It is also critical to accurately predict TM helix numbers and positions, because a mis-predicted TM helix can reverse the TM topology of a downstream region of proteins.

There are many TM helix prediction tools. Although most of them can predict topologies reasonably well (1–8), the prediction accuracy for the N-terminus region is only ~50–70% (9,10). Another key feature for TM topology is signal peptides. These are short N-terminal sequence stretches responsible for co-translational insertion into the lumen of endoplasmic reticulum. However, protein sequences predicted by gene predictors directly from genome sequences, such as Genscan (11), often contain incomplete N-termini (12,13). Truncated N-termini can lead to frequent mis-prediction by signal peptide predictors. These problems can be alleviated significantly by using the already established TM topologies of specific regions of query proteins as constraints. In previous reports (14,15), it was noticed that the subcellular compartment of a domain can be used as a good constraint to correct the mis-predicted TM topology.

Here, we introduce the Localizome server, a web-based system for more precise prediction of TM helix numbers and TM topologies of proteins. Its high prediction accuracy is based on a database called LocaloDom. LocaloDom is comprised of already established and relatively accurate

*To whom correspondence should be addressed. Tel: +82 42 879 8500; Fax: + 82 42 879 8519; Email: jong@kribb.re.kr
*Corresponcdence may also be addressed to Sangsoo Kim. Tel: +82 2 820 0457; Fax: +82 2 824 4383; Email: sskimb@ssu.ac.kr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

TM helix numbers and TM topologies of Pfam domains (16). It is used as a constraint source to correct theoretically predicted results from the Phobius algorithm (17). Phobius was the most accurate TM helix predictor in our in-house benchmarking. Another feature of Localizome server is that it can help the user deduce whether a query protein is a non-TM spanning peripheral membrane protein or not. Peripheral membrane proteins should have no TM helix or membrane localization, both of which can be predicted by the Localizome server. This feature is useful as these proteins often cause problems for TM helix predictors.

We describe the construction procedure of and the accuracy evaluation results of LocaloDom-based prediction.

CONSTRUCTION OF THE LocaloDom DATABASE

The prediction accuracy of the Localizome server depends on the reliability of the LocaloDom database. LocaloDom was constructed by combining experimentally determined subcellular localization information and the prediction results of TM helix numbers by the Phobius program. The LocaloDom construction procedure can be divided into two steps as described below. The first step is for determining the TM topology of soluble Pfam domains and the TM helix numbers of Pfam domains. The second step is for determining the TM topologies of Pfam domains with TM helix and remaining soluble domains not determined in the first step.

(1) Determining reliable TM helix numbers and subcellular compartments of eukaryotic domains

(1.1) Subcellular compartments of soluble protein domains using Swiss-Prot

The subcellular compartments of Pfam domains used in soluble proteins were determined manually for high accuracy using experimentally proven subcellular localization description fields in Swiss-Prot release 48 (18). To obtain experimentally verified information, we eliminated proteins belonging to non-experimental qualifiers, 'Probable', 'By similarity' and 'Potential', in the subcellular localization description. We divided 2956 subcellular localization descriptions of the soluble proteins into two categories: 1434 'IN' (cytoplasmic or nuclear) subcellular compartments including cytoplasm and nucleus, and 316 'OUT' (extracellular or lumenal) subcellular compartments. It includes extracellular space and the lumens of organelles. For the topology of Pfam domains, the decision was made according to the protein topology to which they belong. As a result, we determined the TM topology of ~35% (1781) of eukaryotic Pfam domains. Domains with conflicting subcellular compartment information found in proteins with both 'IN' and 'OUT' were discarded.

(1.2) Determining TM helix numbers for domains by prediction algorithms

Protein sequences in pfamseq were analyzed by applying the Phobius program to predict TM helices. To remove the bias caused by protein sequence redundancy, a CD-HIT algorithm (19) was used with an 80% identity threshold. The CD-HIT algorithm clustered protein sequences at a similarity threshold of 80% as sequence groups. It yielded the longest sequences as representative sequences. Next, the TM helix numbers of Pfam domains were determined by counting the number of TM helices in each Pfam domain. When inconsistency arose in deciding the TM helix number due to conflicting predictions, the dominant TM helix number (over a threshold of 85%, which was determined to include the seventh TM helix of the rhodopsin family, the most redundant and well defined 7-TM protein family, in LocaloDom) was selected. Through this procedure, we could determine the TM helix numbers of 4212 eukaryotic Pfam domains. The majority (93%, 3929) of these had no TM helices, and only 283 had one or more TM helices. In addition, ~300 domains were found to have more than one TM helix, but the numbers of TM helices could not be determined, because these domains did not have any dominant TM helix numbers. Considering these additional TM helix-containing domains, 15% of eukaryotic Pfam domains had at least one TM helix.

(2) TM topology determination of domains that were not processed in the first step: a salvage step

To increase the coverage of LocaloDom, a rule-based procedure was applied to annotate the TM topologies of Pfam domains not determined in the first step. The steps below were applied to the already extracted information containing the sites of the TM helix and the Pfam domain(s) of each protein from pfamseq sequences.

(2.1) The TM helix numbers in the domains were corrected according to the numbers determined in Step 1.2.

(2.2) Using the pre-determined domain information in Step 1.1 as a standard, the TM topologies of the remaining domains were assigned, considering the relationships of domains and TM helices. Soluble protein domains inherited the subcellular compartments of coexisting standard domains. In transmembrane proteins, the loop topologies between TM helices were assigned according to repeated occurrences of IN and OUT topologies. In any conflicting case, the dominant TM topology (over a threshold of 80%, determined empirically) was selected.

(2.3) The subcellular compartment information of domains in each protein sequence was corrected using the information acquired from Step 2.2. Then, Step 2.2 was iterated.

As a result, we could determine the TM topologies of a total of 2283 IN, 554 OUT and 183 TM domains. These Pfam domains were ported to LocaloDom. LocaloDom covers fairly high coverage of 60% (3020/5027) of eukaryotic Pfam domains, which were acquired from Pfam taxonomy search. These domains are used in 60% (20 519/33 869) of ENSEMBL human proteome (20) and 54% (4302/7916) of Phobius-predicted TM proteome.

(3) TM topology determination using LocaloDom information

Localizome identifies Pfam domains using the hmmpfam program and the Pfam database. If there is no Pfam domain in a user's query protein sequence, as a default, the server presents the protein TM topology predicted by Phobius. Otherwise, the server checks inconsistencies between the predicted TM topology and the pre-determined protein domain information (Figure 1). The inconsistencies are corrected by utilizing the information from their constituent domains. Resulting output is displayed as an intuitive graph showing

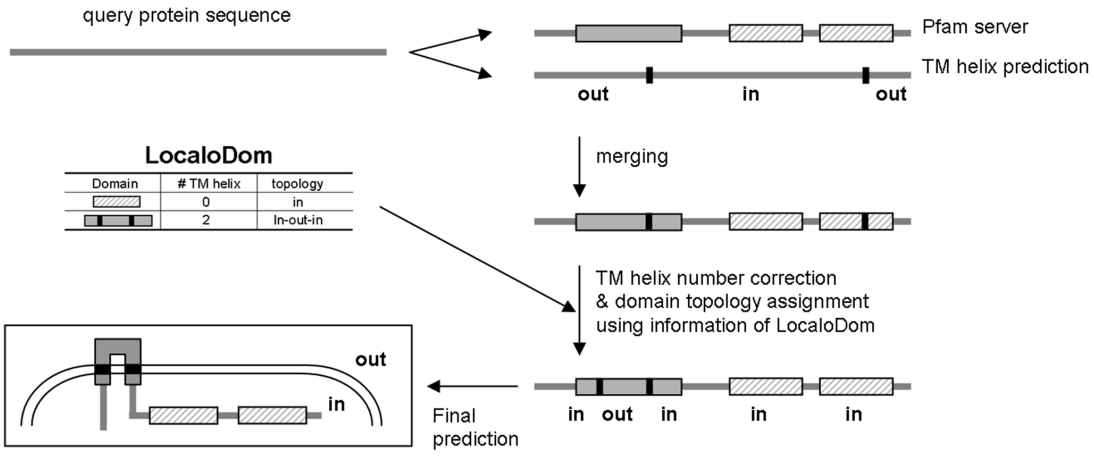

Figure 1. The procedure overview of determining TM topology and TM helix numbers of a query protein sequence. Long gray lines are proteins, vertical bold lines are TM helices, rectangles are various PfamA domains, and 'in' and 'out' represent the TM topology. The shaded types of rectangles represent the domain type. A curved line in the box represents the cell membrane.

the positions of domains and TM helices across biological membrane.

EVALUATION OF THE LocaloDom-BASED TM TOPOLOGY PREDICTION

We compared the results of the LocaloDom-based procedure with the experimentally verified TM topologies.

Evaluation for TM proteins

TM proteins with experimentally known TM topologies were compared with LocaloDom-based TM topology predictions. Eukaryotic sequences and topologies from TMPdb (21) and the Möller database (22) were combined, and the homology was reduced at a 70% threshold (23) using the CD-HIT algorithm. We excluded 12 mitochondria and chloroplast proteins because of the ambiguity of TM topology determination caused by the complex membrane systems of these organelles. Among the resulting 149 sequences, 101 (68%) had at least one LocaloDom domain. The LocaloDom-based procedure predicted the TM topologies of protein regions occupied by domains at the much higher accuracy of 96% (97/101), as compared with Phobius with 89% accuracy. In addition, the Localizome server detected 6 out of 18 TM helix prediction errors by Phobius (Supplementary Table 1).

Evaluation for soluble proteins

LocaloDom can predict soluble proteins' TM topologies. To evaluate soluble protein prediction capability, we used the following test sets. One was from http://yeastgfp.ucsf.edu (24) and the other was from the LOCATE database (25). There were 2628 yeast proteins that contain the IN topology (cytosolic and nuclear) from the yeastgfp site. There were 476 mouse proteins that contain OUT topology (secreted) information. Since all of these are soluble proteins, domains in these proteins must have no TM helix. Furthermore, all IN yeast proteins must consist of IN topology-domains and all secreted mouse proteins OUT topology-domains.

We combined these two sets into one. In the process, we removed redundant sequences using CD-HIT at the threshold of 70%. In the end, we acquired 2881 non-redundant soluble protein sequences. Out of the 2881, 720 (25%) proteins did not contain any domains that matched LocaloDom domains, and it was not possible to use LocaloDom for them. The remaining 75% (2161) used 1199 kinds of LocaloDom domains. Around 98% (1178) of these were proved to have correctly defined TM helix number (zero) and TM topologies according to the proteins to which they belong. Furthermore, 99% (2137/2161) of these proteins consisted exclusively of LocaloDom domains with the same TM topologies of proteins (Supplementary Table 2). The number of non-matching domains will be reduced as more protein domain and localization information becomes available.

IMPLEMENTATION

The Localizome server is composed of a web interface, a MySQL database management system (DBMS), and core computer programs. The web interface is implemented in static HTML pages and CGI scripts. MySQL DBMS is used to store the LocaloDom database. The core programs are written in Perl. They are divided into four main parts: (i) predicting TM helices of a query protein with the Phobius program and assigning Pfam domains with the hmmpfam program, (ii) building a TM topology model from the Phobius result, (iii) correcting the model according to domain TM topology information in the LocaloDom database and (iv) saving the final results as images and HTML. The GD library (http://www.boutell.com/gd/) is used to create a dynamic image of a TM topology. Its image is saved in a PNG format. The server is currently running on a machine with four AMD Opteron 875 CPUs and 16 GB of RAM, running Fedora core5 Linux version 2.6.15 and Apache web server.

INPUT AND OUTPUT

Input

The query interface allows the user to submit protein sequences in FASTA format. Sequences must be specified in the single-letter amino acid notation. The user has a choice

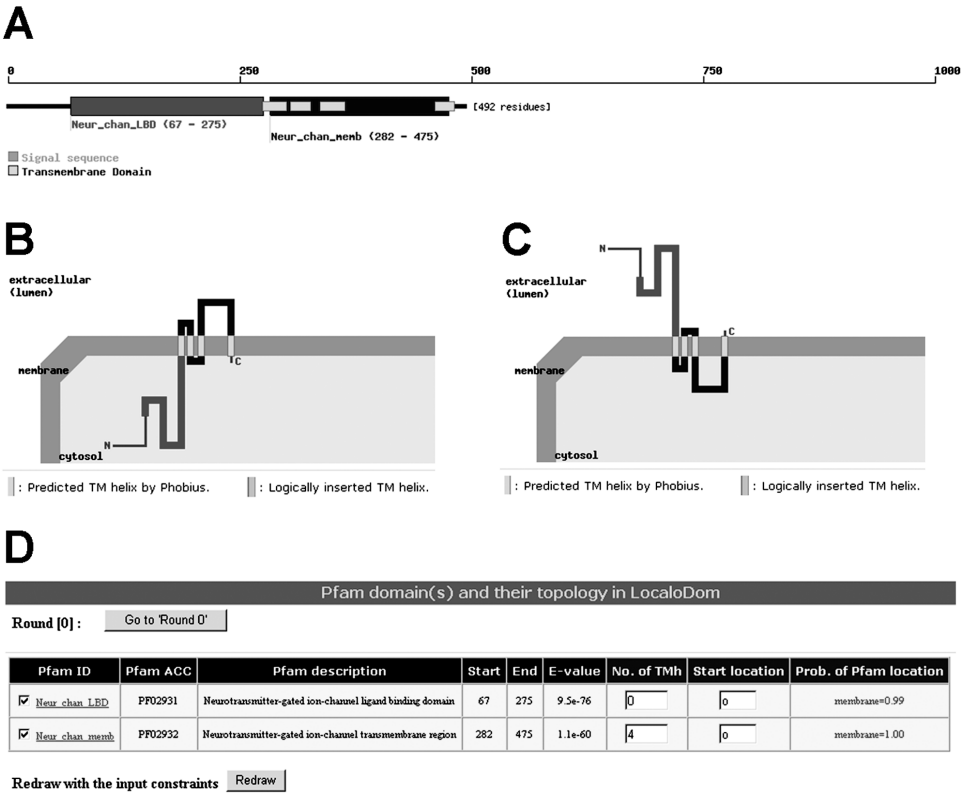

Figure 2. Screenshots of the output page of the Localizome server: **(A)** protein image, **(B)** falsely predicted TM topology by Phobius, **(C)** corrected TM topology by LocaloDom and **(D)** domain table containing domains and their TM topology information.

of inputting query sequences either by pasting the sequences directly into a sequence input form or by uploading a sequence file from the user's local disks. The maximum number of input protein sequences in a single submission is 500 proteins. The length of each sequence is limited up to 3000 residues. The user can select an *E*-value cutoff level for Pfam searching. In the case of submitting more than two proteins, the user has to input an Email address to receive the Localizome results.

Output

The output of the Localizome service is an HTML-formatted file, as shown in Figure 2. It consists of three parts: protein image, TM topology image and domain topology table. In the protein image, a protein is represented as a line, on which Pfam domains and Phobius-predicted TM helices are shown as rectangles according to their position. The TM topology image is created using the corrected TM topology information. In the TM topology image, the TM helices added after the correction are shown in a different color, distinguished from Phobius-predicted TM helices. The user can compare the corrected TM topology with the Phobius result by clicking on the 'Phobius output' button. The domain topology table displays the Pfam domain information of the query protein and its domain TM topology and localization information.

The results are sent to the user in different manners according to the number of input proteins. For single protein, the results are displayed on the same browser with an input web browser. For more than two proteins, the results are sent to the user via Email.

The single protein output page has two more functions, compared with the Emailed output, for modifying TM topology. First, the user can alter the TM topology image by changing the TM helix number or TM topology of a domain in the domain topology table. Second, the user can reset the whole process returning to the very first HTML page by clicking on the 'round 0' button. A detailed description of the output web-page is given in Supplementary Figure S1.

CONCLUSION

The Localizome server was able to predict the protein subcellular compartments with 96–99% accuracy, the highest accuracy among tested algorithms. The Localizome server can also detect and correct falsely predicted TM helices.

The Localizome server is not a pure prediction algorithm. However, using the LocaloDom database as a key information source and additional algorithm for predicting TM helices, it can provide comprehensive information about subcellular localization of query proteins. Since the Localizome server's logic utilizes useful domain information in a query protein, it will become a powerful tool with the increase of domain coverage and localization information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Maryana Bhak for editing the manuscript. This work was supported by the KRIBB Research Initiative Program. J.B. was supported by the Biogreen21 fund. J.B. thanks professional and honest scientists who do science for the sake and joy of science itself. Funding to pay the Open Access publication charges for this article was provided by the KRIBB Research Initiative Program.

Conflict of interest statement. None declared.

REFERENCES

1. Claros,M.G. and von Heijne,G. (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.*, **10**, 685–686.
2. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
3. Hofmann,K. and Stoffel,W. (1993) TMbase—A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, **374**, 166.
4. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.
5. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
6. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
7. Persson,B. and Argos,P. (1997) Prediction of membrane protein topology utilizing multiple sequence alignments. *J. Protein Chem.*, **16**, 453–457.
8. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
9. Ikeda,M., Arai,M., Lao,D.M. and Shimizu,T. (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.*, **2**, 19–33.
10. Melen,K., Krogh,A. and von Heijne,G. (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.*, **327**, 735–744.
11. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
12. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
13. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
14. Mott,R., Schultz,J., Bork,P. and Ponting,C.P. (2002) Predicting protein cellular localization using a domain projection method. *Genome Res.*, **12**, 1168–1174.
15. Bernsel,A. and Von Heijne,G. (2005) Improved membrane protein topology prediction by domain assignments. *Protein Sci.*, **14**, 1723–1728.
16. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–141.
17. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
18. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
19. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
20. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
21. Ikeda,M., Arai,M., Okuno,T. and Shimizu,T. (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucleic Acids Res.*, **31**, 406–409.
22. Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
23. Park,J., Holm,L., Heger,A. and Chothia,C. (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.
24. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
25. Fink,J.L., Aturaliya,R.N., Davis,M.J., Zhang,F., Hanson,K., Teasdale,M.S., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. *et al.* (2006) LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Res.*, **34**, D213–217.