BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 12 2008, pages 1469–1470
doi:10.1093/bioinformatics/btn202

Databases and ontologies

TOPDOM: database of domains and motifs with conservative location in transmembrane proteins

Gábor E. Tusnády*,†, Lajos Kalmár*,†, Hédi Hegyi, Péter Tompa and István Simon

Institue of Enzymology, BRC, Hungarian Academy of Sciences, H-1113 Karolina út 29, Budapest, Hungary

Received on March 21, 2008; revised on April 14, 2008; accepted on April 21, 2008

Advance Access publication April 23, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The TOPDOM database is a collection of domains and sequence motifs located consistently on the same side of the membrane in α-helical transmembrane proteins. The database was created by scanning well-annotated transmembrane protein sequences in the UniProt database by specific domain or motif detecting algorithms. The identified domains or motifs were added to the database if they were uniformly annotated on the same side of the membrane of the various proteins in the UniProt database. The information about the location of the collected domains and motifs can be incorporated into constrained topology prediction algorithms, like HMMTOP, increasing the prediction accuracy.

Availability: The TOPDOM database and the constrained HMMTOP prediction server are available on the page http://topdom.enzim.hu

Contact: tusi@enzim.hu; lkalmar@enzim.hu

1 INTRODUCTION

Transmembrane proteins play important roles in living cells by moving essential nutrients and metabolites across biological membranes, maintaining physiological concentration of ions, importing and exporting signaling molecules and function as receptors or toxin pumps. Their important role is reflected by their high numbers in various proteomes: sequence analysis of the known genomes showed that about 20–30% of the proteins coded by the genome are transmembrane proteins (Jones, 1998; Krogh *et al*., 2001). However, there are only a few hundred known structures due to the difficulties in the structure determination of these types of proteins both by X-ray crystallography and NMR techniques (Arora and Tamm, 2001). Therefore, theoretical investigation of the structure of these type of proteins by statistical and bioinformatic means has come to the forefront of the field.

Topology of transmembrane proteins can be regarded as a low-resolution structure for these proteins. The accuracy of the modern state-of-the-art topology prediction methods reaches 60–80%, depending on the number of transmembrane helices and the dataset used (Chen *et al*., 2002). It was first shown by Tusnady and Simon (2001) that prediction accuracy of transmembrane α-helices can be increased by incorporating prior topological information into the prediction method. Later this approach was used to determine the topology of 37 *Saccharomyces cerevisiae* membrane proteins (Kim *et al*., 2003), global topology analysis of *Escherichia coli* (Daley *et al*., 2005) and yeast (Kim *et al*., 2006) genomes and to improve the prediction accuracy by domain assignments (Bernsel and von Heijne, 2005). Recently, a new database, called TOPDB was established (Tusnady *et al*., 2008), containing experimentally derived topology information gathered from the literature and from public databases available on the internet for about 1500 transmembrane proteins. Using the collected topology information, constrained predictions were made for all proteins in the database by HMMTOP (Tusnady and Simon, 1998; Tusnady and Simon, 2001) as well.

Information collected in our novel database, called TOPDOM, can be used in constrained topology prediction methods. TOPDOM is based on a set of domains, motifs and other sequentially identifiable protein segments which can be found consistently on the same side of the membrane in α-helical transmembrane proteins. These data can be easily incorporated into Bayesian type topology prediction methods, determining the location of a certain part on the sequence. The database with the constrained HMMTOP prediction server is available on the page http://topdom.enzim.hu.

2 METHODS

A subset of α-helical transmembrane proteins was prepared from the SwissProt subset of the UniProt Knowledge Base sequence database (Release 55.0) (Bairoch *et al*., 2005) by choosing entries containing the word 'TRANSMEM' in the feature ('FT') line. Primary hits were further filtered by eliminating entries containing erroneous topological information, like two domains on the same side of the membrane, separated by an odd number of transmembrane segments (see e.g. MDR1_MOUSE). The filtering process resulted in 46 411 sequences with annotated topologies.

To establish a draft dataset, all 46 411 transmembrane proteins were searched locally for motifs and domains contained in Prosite 20.25 (by ps_scan, skipping frequently matching, unspecified patterns and profiles) (Sigrist *et al*., 2002); Prints 38.1 (by fingerPRINTScan v3.596, using *E*-value threshold) (Attwood *et al*., 2003); Smart 5.1 (by HMMER 2.3.2, using the model size and *E*-value threshold from the downloaded THRESHOLD file) (Letunic *et al*., 2004) and Pfam 22.0 (by HMMER 2.3.2, using gathered cutoff thresholds) (Finn *et al*., 2006). The hit results were projected back to the topology sequence

*To whom correspondence should be addressed.

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

of the appropriate entry, and the topology contents were collected to a temporary, draft XML database. This draft database contains all motifs and domains that appear in known transmembrane proteins.

To extract the conservatively located entries from this draft dataset, we used a filtering program, by straining the mixed (inside-transmembrane-outside) topologies. If the topology content did not contain any membrane region, and 90% of the members in the domain/motif set were located on the same side, the entry was deposited in the final database. The exact topology composition of the entry was indicated in the 'Support' attribute of the xml file <TOPDOM> tag.

The creation of the TOPDOM database is fully automatic, therefore the update of the database will be continuous, following the new SwissProt, and motif/domain database releases.

3 RESULTS AND DISCUSSIONS

We have identified 413, 162, 299 and 136 (altogether 1010) domains and/or motifs from Pfam, Prints, Prosite and Smart databases, respectively, which are part of transmembrane proteins and are located conservatively inside or outside of the membrane. We have launched a homepage of the TOPDOM database, located at http://topdom.enzim.hu. The collected data are primarily in XML format, and can be downloaded from our server either in a single file or in separate subsets. Browsing for individual entries is also allowed. A search engine was developed for the web page, which allows searching for entries by keywords or by identifiers of the various primary databases. The search engine can be fed by protein sequences as well. In this case the server performs domain and motif scan on the submitted sequences with the domain information of TOPDOM database. The color coded (red: inside, blue: outside) graphical output of the search result gives an overview of the topology. The squares representing domains and motifs are shaded according to the reliability of location (derived from the 'Support' attribute of <TOPDOM> tag). As a result, the domain/motif hits can be analyzed by making constrained topology prediction with HMMTOP.

We have compared the collected domain location information with two previously published datasets, one prepared from the Smart 4.0 database by Bernsel and von Heijne (2005) whereas the other, LocaloDom dataset (Lee *et al*., 2006) constructed from Pfam domains combined with SwissProt annotations and the results of the Phobius prediction server. Both datasets contain more one-side located domains than TOPDOM because these datasets were not filtered for transmembrane proteins and/or domains containing transmembrane regions were also included. There are several conflicts between the domain location in TOPDOM and the other two datasets. Upon careful investigation of the literature we found that those Smart and Pfam domains that were located on opposing sides of the membrane in these works and also in previous ones, were correctly localized by our method (see the TOPDOM website Documents/Comparisons menu for detailed information).

Thus, our novel database will promote studies of transmembrane proteins in two different, but interrelated directions. Reliable information on domain localization will enable to refine current views on the function of multidomain transmembrane proteins. Further, this information will also serve as an input to constrain topology prediction of unknown transmembrane proteins, which will significantly increase prediction accuracy and reduce the number of erroneous predictions. Because of the huge amount of available sequence data, automated annotation pipelines become ever more important for protein sequence analysis. The TOPDOM database represents an important improvement in this sense, as it is updated automatically, following the new releases of the protein sequence and topology database as well as the domain/motif databases. In all, these advantages will advance bioinformatics studies of transmembrane proteins, help correctly interpret the results of functional studies, and enable to formulate questions amenable for experimental verification.

ACKNOWLEDGEMENTS

Funding: This work was supported by grants from Hungarian research and development funds: OTKA K61684, K72569, NI68950; GVOP-3.1.1-2004-05-0195/3.0, and a Marie Curie reintegration grant (IRG-046572) from the European Commission to H.H. The Bolyai János Scholarship for G.E.T. and the Wellcome Trust ISRF GR067595 for P.T. are also gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Arora,A. and Tamm,L.K. (2001) Biophysical approaches to membrane protein structure determination. *Curr. Opin. Struct. Biol.*, **11**, 540–547.

Attwood,T.K. *et al*. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bernsel,A. and von Heijne,G. (2005) Improved membrane protein topology prediction by domain assignments. *Protein Sci.*, **14**, 1723–1728.

Chen,C.P. *et al*. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.

Daley,D.O. *et al*. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science*, **308**, 1321–1323.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Jones,D.T. (1998) Do transmembrane protein superfolds exist? *FEBS Lett.*, **423**, 281–285.

Kim,H. *et al*. (2003) Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and predictions. *J. Biol. Chem.*, **278**, 10208–10213.

Kim,H. *et al*. (2006) A global topology map of the Saccharomyces cerevisiae membrane proteome. *Proc. Natl Acad. Sci. USA*, **103**, 11142–11147.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lee,S. *et al*. (2006) Localizome: a server for identifying transmembrane topologies and TM helices of eukaryotic proteins utilizing domain information. *Nucleic Acids Res.*, **34**, W99–W103.

Letunic,I. *et al*. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.

Sigrist,C.J. *et al*. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.

Tusnady,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.

Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

Tusnady,G.E. *et al*. (2008) TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res.*, **36**, D234–D239.