Bioinformatics, 32(17), 2016, 2725–2726
doi: 10.1093/bioinformatics/btw193
Advance Access Publication Date: 12 April 2016
Applications Note

OXFORD

Databases and Ontologies

TOPDOM: database of conservatively located domains and motifs in proteins

Julia Varga, László Dobson and Gábor E. Tusnády*

'Momentum' Membrane Protein Bioinformatics Research Group, Institute of Enzymology, RCNS, HAS, Budapest H-1518, Hungary

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 16, 2016; revised on March 22, 2016; accepted on April 4, 2016

Abstract

Summary: The TOPDOM database—originally created as a collection of domains and motifs located consistently on the same side of the membranes in α-helical transmembrane proteins—has been updated and extended by taking into consideration consistently localized domains and motifs in globular proteins, too. By taking advantage of the recently developed CCTOP algorithm to determine the type of a protein and predict topology in case of transmembrane proteins, and by applying a thorough search for domains and motifs as well as utilizing the most up-to-date version of all source databases, we managed to reach a 6-fold increase in the size of the whole database and a 2-fold increase in the number of transmembrane proteins.

Availability and implementation: TOPDOM database is available at http://topdom.enzim.hu. The webpage utilizes the common Apache, PHP5 and MySQL software to provide the user interface for accessing and searching the database. The database itself is generated on a high performance computer.

Contact: tusnady.gabor@ttk.mta.hu.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein domains can fold independently from other parts of proteins and have distinct functions. A function linked to a domain may be important in several processes, therefore one domain may appear in different proteins (Patthy, 1991). Since domains can only be functional in appropriate conditions, a particular domain in various proteins should be located in the same environment, i.e. in the same subcellular location or in organelles which interiors have similar physical–chemical properties. Identifying those protein domains that consistently appear in the same location may help the annotation of unknown proteins with the same or similar domains (Deng and Chen, 2015; Mitchell *et al.*, 2014; The UniProt Consortium, 2014) or the topology prediction of transmembrane proteins (TMPs; Dobson *et al.*, 2015a; Tusnády and Simon, 2001).

During process of cytosis, membranes keep their initial orientation; therefore, all cellular space can be defined as one of the followings: inside (cytoplasmic) or outside (extra-cytoplasmic) space. A detailed description of the various membranes and their sidedness can be found on the homepage of TOPDB database (http://topdb.enzim.hu; Dobson *et al.*, 2015b).

Here, we report the update of the TOPDOM database, originally developed for gathering domains consistently located on the same side of TMPs (Tusnády *et al.*, 2008). The update was performed by extending the examination with globular proteins and using exhaustive search for domains and motifs in all proteins either having a precise location relative to the membranes (e.g. the localization 'Mitochondrion' is not precise but 'Mitochondrion matrix' is) or being transmembrane. Moreover, we applied statistical significance test to select those domains and motifs that are

consistently on the same side of the membrane either in TMPs or in non-TMPs.

2 Methods and databases

Amino acid sequences were downloaded from UniProt/SwissProt (2016_01). CCTOP algorithm was used to discriminate between TMPs and non-TMPs and to predict topology for TMPs. For non-TMPs Swissknife (v1.72; Hermjakob *et al.*, 1999) was used to extract their subcellular localizations. Four domain and motif databases were utilized to identify domains and motifs in amino acid sequences with the appropriate search methods. They are Pfam 29.0 (hmmer 3.1b2; Finn *et al.*, 2015; Mistry *et al.*, 2013), Prints 42_0 (fingerPRINTscan v3.596; Attwood *et al.*, 2003), Prosite 20.122 (ps_scan rev. 1.79; Sigrist *et al.*, 2013) and Smart 06/08/2012 (hmmer 3.1b2; Letunic *et al.*, 2004). A bootstrap method was developed to filter significant hits by random shuffling the amino acid residues on the same side of the membrane in proteins and searching the domains in the randomized sequences as well. The randomization and searching was repeated ten times, and the average and standard deviation were calculated from the ten search results. A hit was reported significant if the real counts were greater than the average plus three times the standard deviation. The final database was transferred to our web server and imported into an SQL database. Along with the latter, an Apache server provides the user interface by using PHP and Bootstrap CSS to ensure functionality across different browsers, operating systems and platforms.

3 Results and discussions

The first version of TOPDOM (Tusnády *et al.*, 2008) was heavily based on the UniProt (The UniProt Consortium, 2014) annotation. The discrimination and topological information of TMPs as well as the identified domains and motifs were taken from the corresponding CC, FT and DR lines of the UniProt entry. Now, the TOPDOM database has been improved in three ways. First, instead of using the topology annotation of UniProt, we used CCTOP prediction method to differentiate between TMPs and non-TMPs and to predict the topology of TMPs. Running CCTOP resulted in 77.036 TMPs and among the rest of the proteins 194.875 non-TMPs with annotated and unambiguous location was identified using the localization information from UniProt. β–barrel proteins are not involved in the current release of the TOPDOM database, since the topology prediction of β-barrel proteins is far less accurate than that of α-helical TMPs. Second, an exhaustive search for domains and motifs were applied on all of these 271.911 proteins, bringing 1.538.118 hits. The locations of the hits could be determined using their sequence position and the topological and localization information provided by CCTOP and UniProt/SwissProt for TMPs and non-TMPs, respectively. The third enhancement concerns the generation of the final database. Formerly, those domains and motifs were selected that could be identified with more than 90% frequency on a particular side of the membrane. Now, we calculate the significance of the hits by a bootstrap method and select those domains and motifs whose count is greater than three and the significance level of the hits are over 99%.

Both the large-scale application of CCTOP (on the whole UniProt/SwissProt) and the scanning of motifs and domains requires massive computation, therefore these calculations are performed on our 256 CPU core HPC computer, and the final database is deposited to the publicly available web server. Moreover, since TOPDOM, CCTOP and the generation of the transmembrane subset of UniProt are circularly dependent on each other, the generation of the database was iterated twice. All of these steps can be done automatically; therefore, we plan to update the TOPDOM database following the UniProt updates. The logic and the physical places of data generation process are delineated in Supplementary Figure S1.

Altogether 6.374 consistently located domains and motifs were identified from the four source databases (Pfam: 4.134, Prints: 540, Prosite: 1.370, Smart: 330). 2.007 out of them can be found in TMPs. The new version of TOPDOM is more than six times larger than the previous one and more than two times larger regarding TMPs only.

Funding

This work was supported by grants from Hungarian Research and Developments Fund [OTKA K104586]. GET and the research group are supported by the Momentum Grant of Hungarian Academy of Science [LP2012-35].

Conflict of Interest: none declared.

References

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

Deng,L. and Chen,Z. (2015) An integrated framework for functional annotation of protein structural domains. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **12**, 902–913.

Dobson,L. *et al.* (2015a) CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.*, **43**, W408–W412.

Dobson,L. *et al.* (2015b) Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.*, **43**, D283–D289.

Finn,R.D. *et al.* (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Hermjakob,H. *et al.* (1999) Swissknife–'lazy parsing' of SWISS-PROT entries. *Bioinformatics*, **15**, 771–772.

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

Mistry,J. *et al.* (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**, e121.

Mitchell,A. *et al.* (2014) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.

Patthy,L. (1991) Exons–original building blocks of proteins? *Bioessays*, **13**, 187–192.

Sigrist,C.J.A. *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.

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

Tusnády,G.E. *et al.* (2008) TOPDOM: database of domains and motifs with conservative location in transmembrane proteins. *Bioinformatics*, **24**, 1469–1470.

UniProt Consortium. (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.