

SstmpDB: a database of single-spanning transmembrane proteins

Olga Bejleri, Zoi Litou, Stavros Hamodrakas✉

Department of Cell Biology and Biophysics, Faculty of Biology, Panepistimiopolis, University of Athens, Athens, Greece

Received 30 July 2013; Accepted 6 September 2013; Published 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Membrane proteins represent ca. 20–30% of both eukaryotic and prokaryotic proteomes. They play crucial roles in cell survival and cell communication, as they function as transporters, receptors, anchors and enzymes. More than 30% of all prescribed drugs are targeting membrane proteins. Transmembrane proteins are either single-spanning membrane proteins or multi-spanning proteins. Single-spanning proteins can be classified into four types I, II, III and IV, depending on their topology and membrane targeting. They are very important functionally, involved in the presentation of antigens to the immune system, they are calcium-dependent cell adhesion proteins, they play a role in septum formation and they have many more specific, crucial roles. The purpose of this work was the construction of a database containing all single-spanning membrane proteins and their functional classification. This database is available at <http://aias.biol.uoa.gr/sstmpdb>.

Motivation and Objectives

Membrane proteins represent ca. 20–30% of both eukaryotic and prokaryotic proteomes. They play crucial roles in cell survival and cell communication, as they function as transporters, receptors, anchors and enzymes. More than 30% of all prescribed drugs are targeting membrane proteins. Transmembrane proteins either span the membrane once (single-spanning membrane proteins) or several times (multi-spanning membrane proteins). Single-spanning proteins are classified into four types I, II, III and IV, depending on their topology and membrane targeting (Hedin *et al.*, 2011). They are very important functionally, involved in the presentation of antigens to the immune system, they are calcium-dependent cell adhesion proteins, they play a role in septum formation and they have many more specific, crucial roles.

The key objective of this project was the collection of all available to date single pass transmembrane proteins and the construction of a database and a web interface for storing and handling these proteins. Also, a functional clustering was performed, aiming at the creation/discovery of novel functional clusters/families, for all single-pass transmembrane protein types.

Methods

For data collection, the database used was UniProtKB/SwissProt, release 2012_11¹. From all initially collected data, fragments were removed and the remaining data set was further filtered by subcellular location, keeping only

single spanning proteins. Then all virus proteins were removed and the final data set contained only proteins with clear experimental evidence at protein and transcript level. Isoforms were not kept as separate entries in the database. Data was grouped by type, organism, and subcellular location. All data pre-processing has been done using Perl scripts. The main database was built using MySQL on a Apache server and the web interface for SSTMPdb, created with PHP and javascript, is located at <http://aias.biol.uoa.gr/sstmpdb/>.

For functional clustering, modern NLP algorithms (e.g., Latent Semantic Analysis, LSA) (Landauer *et al.*, 1998) and common techniques for statistical data analysis/clustering, such as k-means clustering using MATLAB (Zeimpekis *et al.*, 2006), were used. As input, pre-processed datasets of the field *Function* of the Uniprot/Swiss-Prot files, for all single-pass transmembrane proteins were utilised.

Results and Discussion

SstmpDB currently contains 10,250 proteins from 344 organisms and provides information such as their sequence, their type, the functional family they belong to, isoforms, etc. From the web interface of the database, the user has the ability to search entries by Uniprot AC, type and organism and a more advanced search is also available. All data are downloadable in FASTA, text and tab delimited format for each entry or several entries, at will. The web site also allows BLAST searches against the database and contains a detailed manual as supporting material. SstmpDB is the

¹ <http://www.uniprot.org/>

SSTMPDB

Home Blast Documentation Contact

Home Search Browse Download Manual Contact

Protein attributes [text](#) [fasta](#)

Entry ID 10429
 UniProt AC [P49830](#)
 Description Zona pellucida sperm-binding protein 3
 Gene Name ZP3
 Organism Bos taurus [Taxonomy](#)
 Type I
 Length 421AA
 Evidence at transcript level

Features

Type I

Sequence

Sequence
 MGPCSRFLVCFLLWGSTELCSQPFWDDETERFRPSKPPAVMVECCQAQLVVTYDKDLFGTGKLRPADLTGLPDCNCEPLASAD
 TDGVVRFVAVGLHEGNILOVTDNALVYSTFLHNPPAGNLSILRTNRAEVPTECHYPRQGNVSSWAIQPTWVPRTTTFSEKLA
 VFSRLMEENWSAEKMTPTFQLGDRHLQAOVHTGSHVPLRLFDHCVASLTPDWSTSPYHTIVDFHGLVDGLTDASSAFKA
 PRPRPEILQFTVDVFRFANDSRNMIYITCHLVTPVDRVPDQLNACSFSSNRWSPVEGPTDICRCCSKGRGCSGRSMRLSHR
 EGRPVPRSRRHVTEEDVTVGLIFLRKMNDRGVEGPTSSPPLV**MLGLGLATVMTLTLAATVLG**LTGLRAASHPVCPVSAQ

Transmembrane section 382-402
 SSTMPDB is located in Biophysics and Bioinformatics Laboratory

© 2013 - University of Athens

Figure 1. Display of data entry 10429. Protein attributes, features and sequence with transmembrane section (382-402) are shown.

first publicly available database that collects and provides information about single-spanning membrane proteins.

Acknowledgements

This work was funded by the SYNERGASIA 2009 PROGRAMME, co-funded by the European Regional Development Fund and National resources (Project Code 09SYN-13-999), General Secretariat for Research and Technology of the Greek Ministry of Education and Religious Affairs, Culture and Sports.

References

- Hedin L, Illergård K, *et al.* (2011) An introduction to membrane proteins. *J Proteome Res* **10**(8), 3324-3331. doi: [10.1021/pr200145g](#).
- Landauer T, Foltz P, *et al.* (1998) Introduction to Latent Semantic Analysis. *Discourse Processes* **25**, 259-284.
- The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). (2012) *Nucleic Acids Res.* **40**, D71-D75. doi: [10.1093/nar/gkr981](#).
- Zeimpekis D, Gallopoulos E (2006) TMG: A MATLAB Toolbox for Generating Term-Document Matrices from Text Collections. In: Kogan J, Nicholas C, Teboulle M, (Eds.), *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, Berlin, 187-210.