*doi: 10.1093/nar/gkw939*

# TSTMP: target selection for structural genomics of human transmembrane proteins

**Julia Varga[1], László Dobson[2], István Reményi[2] and Gábor E. Tusnády[2,*]**

[1]Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rakpart 3, H1111 Hungary and [2]'Momentum' Membrane Protein Bioinformatics Research Group, Institute of Enzymology, RCNS, HAS, Budapest PO Box 7, H-1518 Hungary

Received August 10, 2016; Revised October 04, 2016; Editorial Decision October 06, 2016; Accepted October 07, 2016

## ABSTRACT

**The TSTMP database is designed to help the target selection of human transmembrane proteins for structural genomics projects and structure modeling studies. Currently, there are only 60 known 3D structures among the polytopic human transmembrane proteins and about a further 600 could be modeled using existing structures. Although there are a great number of human transmembrane protein structures left to be determined, surprisingly only a small fraction of these proteins have 'selected' (or above) status according to the current version the TargetDB/TargetTrack database. This figure is even worse regarding those transmembrane proteins that would contribute the most to the structural coverage of the human transmembrane proteome. The database was built by sorting out proteins from the human transmembrane proteome with known structure and searching for suitable model structures for the remaining proteins by combining the results of a state-of-the-art transmembrane specific fold recognition algorithm and a sequence similarity search algorithm. Proteins were searched for homologues among the human transmembrane proteins in order to select targets whose successful structure determination would lead to the best structural coverage of the human transmembrane proteome. The pipeline constructed for creating the TSTMP database guarantees to keep the database up-to-date. The database is available at http://tstmp.enzim.ttk.mta.hu.**

## INTRODUCTION

Transmembrane proteins (TMPs) act as the gatekeepers of cells as they control the transport of nutrients and drugs into and out of cells, filter or amplify signals in neurotransmission and perception. About 50% of all marketed pharmaceutical drugs (1,2) target transmembrane proteins. Although 25–30% of the proteomes are TMPs (3,4), the ratio of TMPs in the PDB (5) database is still under 2% (6) despite of the efforts of the various structural genomics projects. Determining the 3D structures of membrane proteins is essential to understand their functions and to develop more effective drugs.

In the pioneer international structural genomics initiatives, membrane proteins were simply excluded from targets, since they were known to be complicated targets for structure determination (7). Later, in the various TMP specific structural genomics projects, TMPs were identified as proteins with two or more transmembrane segments (TMS) predicted by the TMHMM method (8) in the selected genome or in Pfam families, while proteins that had homologues in the PDB database were excluded (9–11). Since expression, purification and crystallization are regarded to be easier for prokaryote membrane proteins, the various membrane specific structural genomics consortia focused on developing experimental pipelines to determine prokaryote TMPs rather than eukaryote TMPs (11).

The prediction of protein crystallizability is an additional step for target selection pipelines. XtalPred (12) uses the logarithmic opinion pool method to combine nine biochemical or biophysical features extracted from TargetDB (13) into a score. Prediction accuracy of crystallizability prediction methods were increased by machine learning approaches such as PPCPred (14) that is based on TargetDB as well, however this application incorporated PepcDB (15) and the source data were filtered more rigorously. PredPPCrys (16) and Crysalis (17) were developed to overcome the problem of the overfitting of supervised machine learning techniques by feature selection and they were shown to be the most accurate among other methods. G-protein-coupled receptors (GPCR), for which the used structures resulted from mainly engineered proteins and short fragments that have limited usefulness for further modeling, were ordered by utilizing special propensity score (18) but it was emphasized that GPCRs are highly challenging to crystallize. A common strategy in these methods is to incorporate the results of disordered prediction methods, such as IUPred (19,20), to exclude proteins with longer disordered region(s) from the

[*]To whom correspondence should be addressed. Tel: +36 1 382 6709; Fax: +36 1 382 6295; Email: tusnady.gabor@ttk.mta.hu

potential targets. In our recent paper (21) we investigated the usability of several disorder prediction algorithms on TMPs and found that *in silico* methods overpredict disordered regions in TMPs, especially in the N- and C-terminal regions, therefore their usage in target selection protocols may eliminate potential good targets. Another bottleneck could be the influence of the low number of solved TMP structures on the training sets. Moreover, all of the methods listed above were developed to predict the crystallizability of water soluble, globular proteins, therefore utilizing them on TMPs might result in limited prediction accuracy.

While homology modeling and *ab initio* structure prediction of globular proteins achieved high levels of structural genomics tasks, such as design of enzymatically active proteins (22,23), in case of TMPs they lag behind in this trends. This can be reasoned with the small amount of structural information and inappropriate usage of bioinformatical methods needed for *ab initio* structure prediction algorithms, such as prediction of TMSs or generating alignments, where TMSs are in record. Nevertheless, *ab initio* structure prediction methods using information of correlated mutations in protein families, such the Evfold method (24), can give accurately modeled structures of the TMPs, but for good accuracy enormous amount of homologues is needed.

Here, we report a new target selection database for human TMPs based on the combination of a transmembrane specific fold recognition algorithm and a highly accurate profile–profile sequence comparison. We compared our data with previously published results of various target selection protocols. The strategy applied in the pipeline for the development of TSTMP database is 'transmembrane protein specific', and we believe data in TSTMP database can be a useful source of structural genomics projects or for any laboratories who want to solve the structure of human TMPs that have a great impact on the structural coverage. Moreover, using the data collected in the TSTMP database for structure modeling of TMPs can increase the accuracy of these algorithms. The database is available at http://tstmp.enzim.ttk.mta.hu.

## MATERIALS AND METHODS

### Data resources

The HTP (v1.4) database (4) was used as source of human TMPs and their topology definition. This version of the database was created by running the CCTOP algorithm (25) on the UniProt human reference proteome (2016_06). CCTOP was shown to be the most accurate among the state-of-the-art methods for discriminating between TMPs and globular proteins and for topology prediction. The PDTBM (26) database was utilized to search for TMPs with solved 3D structures and for structures that could be used as templates for homology modeling for other TMPs. TargetTrack (27) database was incorporated into the pipeline in order to facilitate monitoring of the crystallization progress of each protein. For more details, see the Supplementary Material.

## RESULTS AND DISCUSSION

### Data processing

*The human transmembrane proteome.* The HTP database is a collection of human TMPs with predicted topologies. To achieve the most accurate topology prediction, all predictions were enhanced with available structural and experimental information from TOPDB (28). To ensure all information is up-to-date, we updated all source databases before the creation of TSTMP or used the most up-to-date version of them. HTP consists of 5609 TMPs whence 3102 TMPs were predicted to have two or more TMSs. These proteins were examined to see whether they have a solved structure or could be modeled, and were used to build the TSTMP database. Bitopic TMPs were excluded from this database to remain in the safe site of defining collection of TMPs. A recently launched database, called Membranome (http://membranome.org) of single TMS proteins provides models for these TMPs (29).

*Generating clusters for HTP.* Cluster generation resulted in 346 clusters with two or more members, containing altogether 2117 proteins. In TSTMP, we define clusters as groups of proteins that are connected to each other if sequence identity is supported by HHBlits, a profile based sequence searching and alignment method. In a cluster, protein structures may help to model each other in case any of them has solved structure and there is a direct connection between them. This definition is different from the more widely used protein family interpretation. The largest cluster consisted of 685 GPCR proteins mostly from Olfactory protein family, with 24, 239 and 422 proteins with evidence level of '*3D*', '*modelable*' and '*target*', respectively. Olfactory receptor 8S1 from this cluster is on the top of 'The Most Wanted' list (see below), suggesting that the solved structure of this protein would help to model almost 400 proteins of the cluster.

*The '3D' set.* Although the first structure of a TMP was reported more than thirty years ago (30), structure of TMPs are still underrepresented in PDB. Regarding the fold space of TMPs only one fifth of all TMP's fold are revealed so far (31). This figure is even worse taking into consideration human TMPs. Altogether, 60 TMPs can be found in the '*3D*' set, 24 from the same GPCR cluster and two proteins with no homologues from HTP database. Thirty two 3D structures have seven TMSs – still unable to cover the structures of all seven TMS proteins by reliable modeling—while there are 14 structures with four TMSs and seven with six TMSs. The length distribution of the solved TMP structures shows a biased picture: almost two thirds of the structures are between 300 and 500 amino acids in length.

*The 'modelable' set.* The '*modelable*' set contains 606 proteins with at least one template structure that could be potentially used to create a model of the protein. To determine whether a protein is suitable for modeling we used both sequential and structural information, and only accepted hits where the query is likely to be a relative of a structurally solved protein based on sequence identity and a low energy structure could be built using a relative's structure ac-

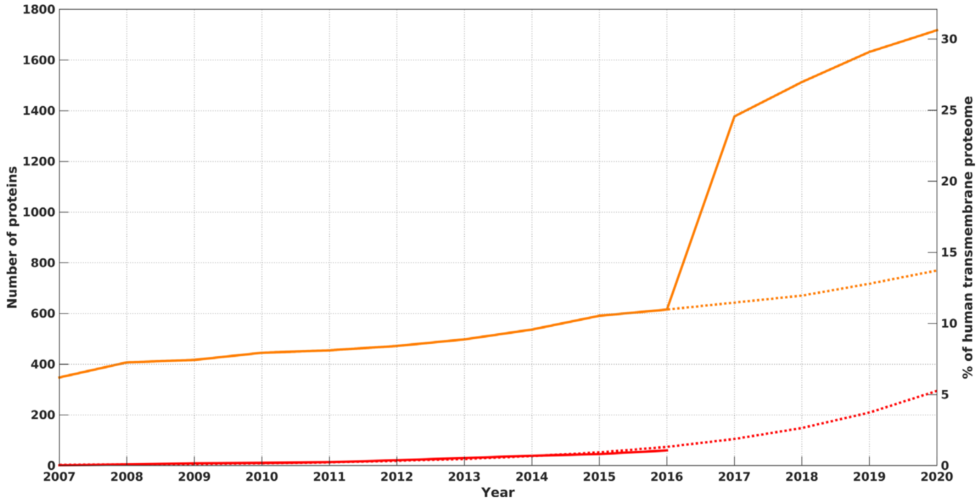

**Figure 1.** The growth of polytopic proteins having solved structure (red) or those that could be modeled using existing structures (orange) until 2016. The former curve was extrapolated until 2020 (red, dotted). The number of modelable proteins was estimated for cases when the target structures were chosen randomly (orange dotted) or according to the order of 'The Most Wanted' set of TSTMP database (orange continuous).

cording to the TMFoldRec algorithm (32,33). According to the data in the TSTMP database, the size of the '*modelable*' set would be increased by 797 (from 606 to 1403) if the first 50 structures from 'The Most Wanted' set would have been solved, and an additional 134 structures would become modelable by solving the next 50 structures from this set (see 'Statistics/by evidence' menu on the web page of TSTMP). The modelable set can be downloaded as a whole, or separately according to the level of how easy or hard task are their modeling. We defined proteins as easily modelable if they had at least one PDBTM structure with an identity of 50% or above regarding only transmembrane segments. Hard targets are defined as those modelable proteins that only have PDBTM structures with a maximum of 25% identity in transmembrane segments. In the Supplementary Material, we have provided two case studies on how the information in the modelable set can help homology modeling and *ab initio* structure prediction.

*The 'target' set.* The '*target*' set may be a good starting point to choose proteins with high number of homologues that could be modeled by determining the structure of the selected protein. A more successful attempt might involve selecting close homologues with similar number of relatives as it might increase the chance of successful crystallization of the selected proteins (34). The '*target*' set contains 2436 proteins. Besides providing a list of targets, we created 'The Most Wanted' list that shows how the human transmembrane proteome could be modeled with the minimum number of steps (i.e. with the smallest number of crystallized proteins). We have sorted the proteins based on how many '*target*' homologues they have, and selected the one with the highest number of relative targets. Then, we removed all entries from the list that could be modeled based on the selected protein and iterated this process until every human transmembrane proteins were covered.

*Prediction of crystallizability.* We have tried several protein crystallization prediction methods listed in the Introduction to check their prediction accuracy on TMPs. Since these methods were trained on globular proteins, we did not expect that they could be used on TMPs. Indeed, none of them could properly distinguish the crystallization propensity of membrane proteins; regardless we have checked it on proteins from the '*3D*' or '*target*' sets. As a consequence of the poor prediction accuracies of these methods on TMPs, we did not include the results of crystallizability predictions into the TSTMP database.

## The home page of the TSTMP database

The home page of the TSTMP database was created by using the common Apache+PHP+MySQL triplets with the Bootstrap library. The database can be downloaded either by selecting the whole database or various subsets of the database in XML format. It can be searched by identifiers, by number of homologous entries either in the Human Transmembrane Proteome or in the TargetTrack database. Users can also search by the status of the corresponding TargetTrack entries. We provide several statistics about the rate of the growth in the number of solved TMP structures as well as the yearly distribution of the number of modelable proteins either by following our target selection strategy or by randomly choosing TMPs from target set (see 'Statistics/by number of homologues' menu). The page of 'Statistics by TargetTrack status' shows the current distribution of TargetTrack (selected, cloned, expressed, etc.) statuses in the three evidence levels defined by our database ('*3D*', '*modelable*', '*target*'). We also provide a map created from the distribution of the number of homologues over the various statuses defined in the TartgetTrack database in order to help structural genomics laboratories to simplify their target selection. For example, according to this table, there are 107 TMPs in 'selected' status with only one target homologue. Solving the structures of these TMPs would not increase the '*modelable*' protein set significantly. As an opposite example, there is one protein in 'purified' status that has 22 homologues, the Solute carrier family 22 member 1,

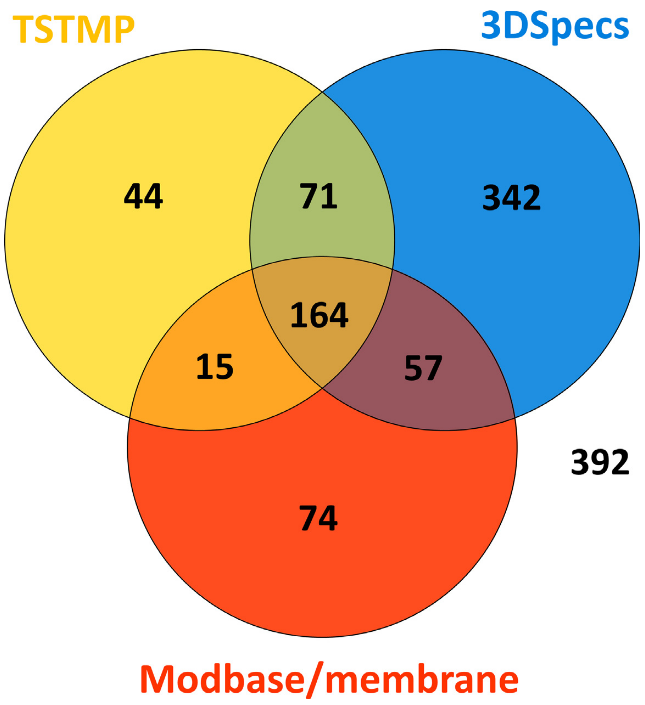

**Figure 2.** Comparison of the results of various target selection protocols. Proteins having '*3D*'/'*modelable*' evidence are inside the Venn diagram (38). Outside the circles, the number of proteins marked as targets by all three methods is shown.

solving the structure of this protein would have higher impact on the '*modelable*' set of human TMPs.

### Current and past statistics of the TSTMP database

We used the same pipeline with earlier structure database versions (for PDB and PDBTM, from 2007 to 2016) to investigate the changes of different statistics. Most interestingly, none of the structures of 'The Most Wanted' set from different years was determined in a later time, however they became '*modelable*' eventually. Laboratories seemed to prefer to solve structures that already have a relative in PDB: since 2007, only 26.7% of the proteins' status changed right from '*target*' to '*3D*', in the rest of the cases their structure were determined only after they could already be modeled, which suggests these laboratories prefer to solve proteins where there are preliminary knowledge about the structure. We examined clusters containing at least two members in which all proteins are above '*target*' status (i.e. the whole cluster could be modeled). Apparently, the number of the proteins in these clusters grows linearly over the years, but it is the result of determining new 3D structures in the clusters rather than growing the number of fully modelable clusters (Supplementary Figure S5).

The ratio of TMPs in PDB is around 2%, and this proportion has not changed in the past years. Since this data is highly redundant, (e.g. the same protein was often crystallized multiple times), the number of different proteins related to these structures were also investigated (Supplementary Figure S6A). Counting only the unique (all species or only human) proteins, these values are even worse. We fitted exponential curves to each plot as it was described by Dickerson in PDB Newsletter—1978. Even if the growth of all solved membrane protein structures is exponential as it was shown earlier by others (35), the rise of the curve is rather linear than exponential considering only unique structures. On Supplementary Figure S6B the same statistics are shown considering only polytopic TMPs.

We extrapolated the number of solved structures until 2020 and examined the ratio of '*modelable*' proteins in case the determined structures are chosen randomly or from 'The Most Wanted' set. Although crystallizability is affected by a huge amount of unforeseen or not investigated factors, in an optimistic case the number of '*modelable*' proteins could be doubled by using our ranking (Figure 1).

### Comparison to other target selection databases and methods

A comparison was made to similar databases/methods to reveal the extent of the overlap of proteins marked as '*target*' between different resources. For this purpose, we selected ModBase/membrane (36) and 3DSpecs (37). Both databases suggest targets for crystallizing and provide a list of proteins with corresponding PDB structure (if any). Since these databases were not updated since 2013 and 2011 (respectively), for a fair comparison we have used the hypothetical '2011' version of TSTMP (i.e. we used the same pipeline with the 2011 versions of the source databases). Only proteins listed in all three databases were taken into account.

3DSpecs categorizes proteins as '*Solved*', '*Not Solved and Template Found*' and '*Not Solved and No Template*' which corresponds to our '*3D*', '*modelable*' and '*target*' definition. They further categorize proteins by their type, for this comparison only integral membrane proteins were selected. ModBase/membrane does not have such evidence level, therefore we simply used a 25% sequence identity threshold (a transmembrane protein was assumed to be solved or modeled >25% identity with the assigned PDB entry). To overcome the differences caused by short crystallized fragments, only those proteins were considered as '*modelable*'/'*3D*', where all transmembrane helices were covered by the corresponding PDB entry.

Figure 2 shows how many of the proteins were listed as '*3D*'/'*modelable*' by these resources and the extent of the overlap between them. Outside the circles, the number of the proteins marked as '*target*' by all databases is shown. 3DSpecs highly overestimates '*3D*'/'*modelable*' proteins, and the agreement between the three methods is only 14%. Only 21% of the proteins could be assigned in all methods, which can be explained by the superficial TMP filtering. Both methods used TMHMM to predict the initial transmembrane proteome by defining a protein to be TMP if TMHMM predicted two or more and three or more TMS in ModBase/membrane and 3DSpecs, respectively. Although selecting polytopic membrane proteins from the human reference proteome is regarded as sufficiently addressed by existing methods (34), Supplementary Figure S7 clearly shows that using TMHMM for this task leads to different starting sets compared to the result of the CCTOP algorithm. Moreover, TSTMP exploits both structural and sequential information to define the evidence level of proteins, leading to a more accurate estimation of '*modelable*' proteins.

### Conclusions and future directions

We would like to help speeding up the process of structural genomics of the human TMPs, since despite the huge striving of the nine structural genomic centres of TMPs, there are still only sixty 3D structures of human TMPs known, and about 606 ones are modelable. In this database, we examined the possibility of modeling a protein structure assuming prior knowledge about its relatives and provided a list of target proteins from the human transmembrane proteome, as well as a strategy to achieve a higher structural coverage. Extending the number of modelable TMPs can help to understand their function, to develop drugs and to improve TM specific homology modeling, threading or *de novo* structure prediction algorithms. Although membrane protein crystallization is very challenging and crystallization propensity of different targets are obviously not equal, we believe our database could serve as a guide to select those proteins, whose structures would have a great impact on the structural coverage of human transmembrane proteome. Since preliminary knowledge about a protein helps the structure determination, we also provide list of modelable proteins, based on how easy to homology model or predict their structure de novo. We would like to update TSTMP database regularly, following the updates of the source databases (HTP, PDBTM, UniProt).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank András Horváth for the critical reading of the manuscript.

## FUNDING

Hungarian Scientific Research Fund [K104586, http://www.otka.hu]; New National Excellence Programme funded by the Ministry of Human Resources [ÚNKP-16-2_VBK-016]; 'Momentum' Program of the Hungarian Academy of Sciences [GET]. Funding for open access charge: Hungarian Academy of Sciences [LP2012-35].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
2. Overington,J.P., Al-Lazikani,B. and Hopkins,A.L. (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.
3. Käll,L. and Sonnhammer,E.L.L. (2002) Reliability of transmembrane predictions in whole-genome data. *FEBS Lett.*, **532**, 415–418.
4. Dobson,L., Reményi,I. and Tusnády,G.E. (2015) The human transmembrane proteome. *Biol. Direct*, **10**, 31.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Kozma,D., Simon,I. and Tusnády,G.E. (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524–D529.
7. Büssow,K., Scheich,C., Sievert,V., Harttig,U., Schultz,J., Simon,B., Bork,P., Lehrach,H. and Heinemann,U. (2005) Structural genomics of human proteins–target selection and generation of a public catalogue of expression clones. *Microb. Cell Fact.*, **4**, 21.
8. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
9. Dobrovetsky,E., Lu,M.L., Andorn-Broza,R., Khutoreskaya,G., Bray,J.E., Savchenko,A., Arrowsmith,C.H., Edwards,A.M. and Koth,C.M. (2005) High-throughput production of prokaryotic membrane proteins. *J. Struct. Funct. Genomics*, **6**, 33–50.
10. Kelly,L., Pieper,U., Eswar,N., Hays,F.A., Li,M., Roe-Zurz,Z., Kroetz,D.L., Giacomini,K.M., Stroud,R.M. and Sali,A. (2009) A survey of integral alpha-helical membrane proteins. *J. Struct. Funct. Genomics*, **10**, 269–280.
11. Punta,M., Love,J., Handelman,S., Hunt,J.F., Shapiro,L., Hendrickson,W.A. and Rost,B. (2009) Structural genomics target selection for the New York consortium on membrane protein structure. *J. Struct. Funct. Genomics*, **10**, 255–268.
12. Slabinski,L., Jaroszewski,L., Rychlewski,L., Wilson,I.A., Lesley,S.A. and Godzik,A. (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*, **23**, 3403–3405.
13. Chen,L., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860–2862.
14. Mizianty,M.J. and Kurgan,L. (2011) Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, **27**, i24–i33.
15. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
16. Wang,H., Wang,M., Tan,H., Li,Y., Zhang,Z. and Song,J. (2014) PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PLoS One*, **9**, e105902.
17. Wang,H., Feng,L., Zhang,Z., Webb,G.I., Lin,D. and Song,J. (2016) Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep.*, **6**, 21383.
18. Mizianty,M.J., Fan,X., Yan,J., Chalmers,E., Woloschuk,C., Joachimiak,A. and Kurgan,L. (2014) Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr. D. Biol. Crystallogr.*, **70**, 2781–2793.
19. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
20. Dosztányi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
21. Tusnády,G.E., Dobson,L. and Tompa,P. (2015) Disordered regions in transmembrane proteins. *Biochim. Biophys. Acta*, **1848**, 2839–2848.
22. Siegel,J.B., Zanghellini,A., Lovick,H.M., Kiss,G., Lambert,A.R., St Clair,J.L., Gallaher,J.L., Hilvert,D., Gelb,M.H., Stoddard,B.L. *et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, **329**, 309–313.
23. Khare,S.D., Kipnis,Y., Greisen,P., Takeuchi,R., Ashani,Y., Goldsmith,M., Song,Y., Gallaher,J.L., Silman,I., Leader,H. *et al.* (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol.*, **8**, 294–300.
24. Marks,D.S., Hopf,T.A. and Sander,C. (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.
25. Dobson,L., Reményi,I. and Tusnády,G.E. (2015) CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.*, **43**, W408–W412.
26. Kozma,D., Simon,I. and Tusnady,G.E. (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.*, **41**, D524–D529.

27. Gabanyi,M.J., Adams,P.D., Arnold,K., Bordoli,L., Carter,L.G., Flippen-Andersen,J., Gifford,L., Haas,J., Kouranov,A., McLaughlin,W.A. *et al.* (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics*, **12**, 45–54.
28. Dobson,L., Lango,T., Remenyi,I. and Tusnady,G.E. (2014) Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.*, **43**, D283–D289.
29. Lomize,A.L. and Pogozheva,I. (2015) Membranome: a database of single-spanning transmembrane proteins. *Biophys. J.*, **108**, 249a–250a.
30. Deisenhofer,J., Epp,O., Miki,K., Huber,R. and Michel,H. Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution. *Nature*, **318**, 618–624.
31. Oberai,A., Ihm,Y., Kim,S. and Bowie,J.U. (2006) A limited universe of membrane protein families and folds. *Protein Sci.*, **15**, 1723–1734.
32. Kozma,D. and Tusnády,G.E. (2015) TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool. *BMC Bioinformatics*, **16**, 201.
33. Kozma,D. and Tusnády,G.E. (2015) TMFoldWeb: a web server for predicting transmembrane protein fold class. *Biol. Direct*, **10**, 54.
34. Kloppmann,E., Punta,M. and Rost,B. (2012) Structural genomics plucks high-hanging membrane proteins. *Curr. Opin. Struct. Biol.*, **22**, 326–332.
35. White,S.H. (2004) The progress of membrane protein structure determination. *Protein Sci.*, **13**, 1948–1949.
36. Pieper,U., Schlessinger,A., Kloppmann,E., Chang,G.A., Chou,J.J., Dumont,M.E., Fox,B.G., Fromme,P., Hendrickson,W.A., Malkowski,M.G. *et al.* (2013) Coordinating the impact of structural genomics on the human α-helical transmembrane proteome. *Nat. Struct. Mol. Biol.*, **20**, 135–138.
37. Bray,J.E. (2012) Target selection for structural genomics based on combining fold recognition and crystallisation prediction methods: application to the human proteome. *J. Struct. Funct. Genomics*, **13**, 37–46.
38. Oliveros,J.C. (2007) VENNY. An interactive tool for comparing lists with Venn Diagrams. *VENNY. An Interact. tool Comp. List. with Venn Diagrams*. http://bioinfogp.cnb.csic.es/tools/venny/index.html.