Comprehensive human membrane protein database

Min-sung Kim
Dept. of Information and Communications Engineering
KAIST
Daejeon, Korea
kmsid2@kaist.ac.kr

Gwan-Su Yi
Dept. of Bio and Brain Engineering
KAIST
Daejeon, Korea
gsyi@kaist.ac.kr

Abstract— **Membrane proteins are an important protein group involved in cell signaling, molecule transport, and cell-cell communication and regarded as very important pharmaceutical targets. However, most membrane protein resources are built for specific function groups and it is hard to find comprehensive information about membrane proteins. This circumstance hampers the search for both the general and specific characteristics of membrane proteins, which characteristics may provide critical information for the investigation of novel functions or drug targets. We collected 28,701 human membrane proteins from various sources and suggest 9,616 novel candidates using the information in collected proteins and orthologs from 51 other species. We also annotated various characteristics of membrane proteins including their conserved motifs and domains, functions, related cellular processes, and associated diseases. This database is freely available at http://fcode.kaist.ac.kr/chmpd.**

Keywords-component; Human Membrane Protein, Novel Membrane Protein Prediction, Interaction Type with Membrane, Pharamacetucial Research

I. INTRODUCTION

Membrane proteins are essential in diverse cellular functions such as cell signaling, molecule transport and cell-cell communications. They have been considered major therapeutic targets because of their biological importance. In fact, 60% of FDA approved drug targets are known as membrane proteins [1]. The importance of membrane proteins has urged the development of several organized databases for well-known membrane protein groups including transporters [2, 3] and specific receptors [4-6]. However, there is no comprehensive resource that focuses on all human membrane proteins to support pharmaceutical research in the membrane proteome [7].

The scope of such pharmaceutically important membrane proteins can be delineated based on how they interact with the membrane. An Integral Membrane Protein (IMP) is a protein in which all or part of its sequence is buried in the membrane. A Lipid-Anchored Protein (LAP) is bound to the membrane using a covalently attached anchor instead of any embedded peptide sequences in the membrane. The last type is a Peripheral Membrane Protein (PMP), which has two major types of interaction with the membrane: proteins interacting with lipid head groups rather than the hydrophobic core of the lipid bilayer and proteins indirectly localized at the membrane by binding to IMPs. PMP and LAP are as important as IMP. They behave as regulatory subunits of many receptors and ion channels. Recent G Protein-Coupled Receptor (GPCR) databases additionally provide information on heterotrimeric G proteins because the GPCRs cannot transmit extracellular signals into the intracellular region without the G proteins. These proteins are also correlated with various types of diseases. Ras proteins are known as proto-oncogenes by containing mutations found in approximately 30% of all human tumors [8].

The primary source of membrane protein is subcellular localization resources. Novel membrane protein prediction is also needed to complement the unidentified portion of membrane proteins among resources. There were three types of approach related with the prediction: subcellular localization prediction, membrane protein type prediction, and membrane topology prediction. Most subcellular localization prediction tools only predict proteins localized at the plasma membrane, except organelle membranes. They also predict membrane proteins without considering how those proteins interact with the membrane, even though the localization mechanisms are completely different from each other. Research related to the membrane protein type prediction attempt to classify membrane proteins into sub-classes that reflect the concept of interaction type with the membrane. However, such studies train their classifiers by using the pre-classified membrane proteins without the Non-Membrane Proteins (NMPs). Some of those studies implement an additional classifier to distinguish membrane proteins from NMPs [9]. However, they only compare all the collected membrane proteins with NMPs, rather than directly comparing each type of membrane protein with NMPs. Membrane topology prediction tools can be used to predict only IMP. Therefore, a different approach is required to identify novel membrane proteins.

In this paper, we introduce CHMPD, a comprehensive resource that aims to cover all the human membrane proteins with various characterizations. Membrane proteins were collected from 19 different public resources. The collected membrane proteins were characterized with sequence features. The database contains 9,616 novel membrane protein candidates from membrane protein orthologs of 51 different model organisms and novel membrane protein prediction by considering interaction type with the membrane. The collected proteins were classified into 1,409 hierarchical molecular function groups and annotated with

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

biological process and pharmaceutical information. CHMPD will enable researchers to identify the membrane proteins and their characteristics comprehensively in one place and to explore the novel membrane protein candidates.

II. CONSTRUCTION AND CONTENT

A. Collection of membrane proteins

We collected membrane proteins from 19 different publicly available resources to construct the membrane protein dataset, illustrated in Fig 1. The collected membrane proteins have their unique UniProt accession numbers. Membrane proteins were collected from the GO Cellular Component (CC) [10], UniProt Subcellular Locations (SL), and UniProt Keywords. To supplement the limited coverage of these resources, we additionally gathered membrane proteins from other eukaryotic subcellular localization databases without prediction results; LOCATE [11], eSLDB [12], Organelle DB [13], DBMLoc [14], and DBSubLoc [15]. There are annotation terms that can be included for membrane proteins like "GO:0004888 : transmembrane receptor activity" and "GO:0022857 : transmembrane transporter activity" in molecular function ontology resources. The member proteins of the manually selected terms from GO Molecular Function (MF) and UniProt Keywords were merged into the CHMPD. Curated membrane proteins were also accumulated from 7 specific membrane protein group databases: KEGG BRITE [16], UniProt 7-transmembrane G-linked receptors, gpDB, GPCRDB, IUPHAR-DB, VKCDB, and TCDB. Membrane topology resources provide membrane proteins by designating which regions are buried in the membrane. The proteins were retrieved from TOPDB [17], ExTopoDB [18], and the UniProt sequence section.

The collected membrane proteins were allocated with evidence codes depending on their annotated evidence in the integrated resources: from experiment, from the literature, from manual curation, from computational prediction, and non-traceable. The constructed dataset contains redundant sequences including identical sequences and sub-fragments. UniRef100 [19] was adapted to construct the non-redundant membrane protein dataset because this resource provides clustered sets of redundant sequences with sequence identity 100% and a representative protein in each clustered set.

Novel membrane proteins were predicted based on the membrane protein orthologs in other organisms. To collect the orthologs from other species, we used 51 eukaryotic species that are currently supported model organisms in Ensembl [20]. Membrane proteins were collected with the same procedure in the target species. The proteins with reliable evidence codes were employed for searching the orthologs. Human ortholog information for the organisms was retrieved from Ensembl Compara. The 987 novel membrane proteins were predicted by excluding proteins already annotated as membrane proteins.

B. Sequential characterization of membrane proteins

We integrated 4 resources and 8 prediction tools to characterize the constructed membrane proteins at sequence level. The membrane topology reveals which regions of the sequence span the membrane and which are located in the inside or the outside regions. Known membrane topology region information was gathered from the UniProt sequence section and PDBTM [21]. We also integrated 5 membrane topology prediction programs to predict uncharacterized proteins among the resources and to analyze user's input sequences: TMHMM [22], STHMHMM [23], SCAMPI [24], HMMTOP [25], and PHOBIUS [26]. Signal peptide is a protein sorting signal that targets a protein that is translocated across the ER membrane and exported by secretory vesicles. Identification of the signal peptide is important because the membrane topology prediction tools often predict it as N-terminal transmembrane helix. These were aggregated from the UniProt Sequence Section and the prediction results of SignalP 4.0 [27] and PHOBIUS. A lipid-anchor is a fatty acid covalently attached to a protein. It may anchor proteins to the plasma membrane. Myristoylation, palmitoylation, prenylation, and GPI-anchor sites were retrieved from dbPTM [28]. The predicted GPI-anchor and myristoylation sites were integrated from two programs: FragAnchor [29] and Myristoylator [30]. Protein domains, motifs, and families are conserved sequences in certain protein groups and are highly correlated with their function. We gathered the information using InterPro [31].

C. Novel membrane protein prediction based on interaction type with membrane

The collected membrane proteins were categorized based on their interaction type with membrane: IMP, PMP, and LAP. The method for the prediction of novel membrane protein candidates is hierarchically organized with three random forest classifiers, as depicted in Fig 2.

The NMP dataset is prerequisite for the control group of the classifiers. It was created from the subcellular localization resources by searching proteins from non-membrane regions covering extracellular region, cytoplasm, and organelle lumen. To construct a reliable training dataset, the following steps were performed: (1) Computationally predicted membrane proteins were not used. (2) The sequences that have fragment status or contain ambiguous residues (such as "X", "B", "J", and "Z") were excluded. (3) The proteins designated "putative", "probable", "uncharacterized", and "predicted" were removed. (4) The CD-HIT [32] program was utilized to get rid of highly homologous sequences with 90% sequence identity threshold. The 4801 IMPs, 581 PMPs, 219 LAPs, and 5277 NMPs were finally prepared for the training dataset.

Sequence feature vectors for the classifiers are listed in Table 1. EMBOSS PEPSTAT [33] and PseAAC-Builder [34] were utilized to generate the amino acid properties. Pseudo amino acid composition is analyzed for the sequence features because it can incorporate sequence order effects in addition to the conventional amino acid composition [35]. The predicted number of transmembrane regions was the other feature that can reflect the properties of IMP. The classifiers were implemented based on the R randomForest package [36].

The three classifiers were organized in a hierarchical

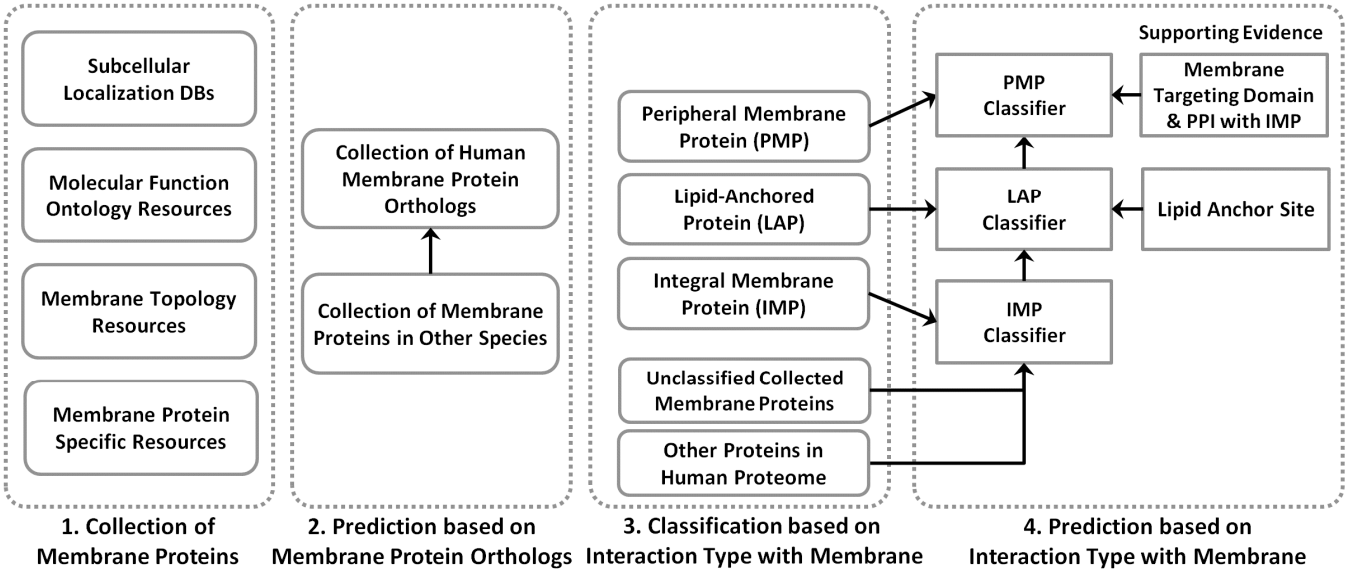

Figure 1. Whole procedure for the collection of membrane proteins and the prediction of novel membrane protein candidates depending on their interaction type with membrane in CHMPD

structure to accommodate the nature of each type of membrane protein. The size of the reliable IMP dataset can be comparable to those of other types of proteins. However, the number of PMPs and LAPs is too small to directly compare with the number of IMPs and NMPs. This circumstance causes imbalanced dataset problems. Therefore, we randomly sampled equal amounts of proteins from the negative dataset to train the PMP classifier and the LAP classifier. Although the IMPs have membrane embedded sequence regions that can fairly well discriminate them from the others, the PMPs don't have any distinctive sequence properties that can cover most of them. Protein families belonging to the PMP group are also very diverse compared to the size of the currently annotated member proteins. These make it difficult to classify PMPs and NMPs. LAPs are relatively less diverse compared to PMPs. Therefore, we prioritized the classifiers in the hierarchical structure.

The performance of the classifiers at each stage was measured with 5-fold cross validation. The performance of the IMP classifier was compared using the integrated membrane topology prediction tools that are commonly used to predict IMPs, as shown in Table 2. The performances of the other classifiers were also calculated by averaging the results from 1,000 randomly sampled negative datasets. The results are listed in Table 3. The high accuracy of the IMP classifier enables us to apply it to analyzing the whole human membrane proteome. However, the prediction results of the two classifiers are not so reliable compare to those of the IMP classifier. Therefore, we integrated additional evidence for the classifiers in order to increase the reliability of the predicted membrane proteins as shown in the 4th step of Fig 1. The input protein sequence is examined using the IMP classifier to evaluate whether it is IMP or not. If it is not predicted to be IMP, it is tested for LAP. Only proteins with positive prediction results from the LAP classifier and predicted lipid-anchor sites are allocated as LAP. The integrated lipid-anchor site and prediction results from FragAnchor, Myristoylator, and dbPTM were used for evidence. If the target protein is not IMP or LAP, it is checked with the PMP classifier and the existence of two possible types of evidences for PMP. Known membrane targeting domains can be used to identify candidate proteins bound to the lipid head groups. Protein-protein interaction (PPI) can also be used to retrieve proteins that interact with IMPs. The list of 9 known membrane targeting domains is made by referencing MeTaDoR (Membrane Targeting Domain Resource) [37]. The interaction information for the membrane proteins was assembled from our comprehensive protein interaction database, COMBICOM [38].

Table 3 summarizes the current statistics of membrane proteins in CHMPD. It contains 38,317 human membrane proteins, including our novel membrane protein candidates. The size is about 36.5% of human UniRef100 proteins and 21.0% of human Entrez RefSeq genes.

D. Classification of membrane proteins based on their subcellular localization groups

Cell has membrane to separate the interior part from the extracellular environment. Although this plasma membrane is a representative type of membrane, there are other types of organelle membranes such as endoplasmic reticulum membrane and nucleus membrane. Therefore, the collected membrane proteins can be alternatively categorized based on which types of membrane they are associated with. Whole structure of the classification is composed of 45 hierarchical classes by extracting common child terms of "membrane" in GO CC and UniProt SL.

E. Classification of membrane proteins based on their molecular function groups

Classification of membrane proteins into smaller functional groups is a key step in analyzing commonality and

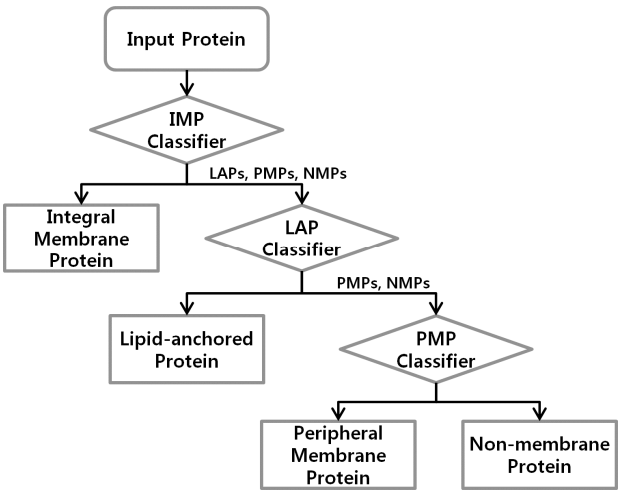

Figure 2. Hierarchical structure of classifiers to predict novel membrane proteins based on their interaction type with membrane

specificity among membrane proteins. It is imperative to identify the major molecular function categories of membrane proteins before the classification. The identification was performed with the GO MF term set because this set supports the largest number of terms and has the largest variety of aspects of molecular activity compared to those of other controlled term sets. We prioritized the GO MF terms based on their proportion in membrane proteins and manually selected abstract terms by traversing GO MF hierarchies. Four major functional categories were identified; "Receptor", "Transporter", "Enzyme", and "Others", which is further classified to "Structural molecule" and "Ligand". This categorization is matched with functional classification results for membrane proteins in previous research [39]. We added "Cell adhesion molecule" that is not found in the GO MF hierarchies. The identified major categories were further classified based on the term sets from GO MF and related protein databases. There are classification structures for certain types of proteins that are widely accepted. These are International Union of Biochemistry and Molecular Biology (IUBMB) enzyme nomenclature, IUBMB membrane transport proteins nomenclature, phylogenetic classification of the GPCR family, etc. We outlined the whole hierarchical structure with the GO MF terms; and detailed classifications were constructed by adapting the classification terms in other resources. CHMPD currently uses 11 different resources to make its classification structure: GO MF, UniProt Keywords, KEGG BRITE, PRRDB [40], IUPHAR-DB, UniProt 7-transmembrane G-linked receptors, gpDB, GPCRDB, VKCDB, TCDB, and ExPASy – ENZYME [41]. Mapping information of the controlled vocabularies to GO terms was downloaded from the GO database. We used the mapping information as one of reference guides for the manual integration. The classification terms from the sources were manually integrated into 1,409 hierarchical classes.

F. *Characterization of biological process information of membrane proteins*

The molecular function classification of membrane proteins only covers their individual functions. It doesn't describe functional characteristics by interacting with other proteins or molecules. GO Biological Process (BP) and UniProt Keywords were selected for the characterization. Signaling pathway resources provide more comprehensive information of signaling mechanism and curated member proteins compared to the biological process resources. Therefore, we also integrated signaling pathway information from 4 pathway databases; KEGG, PID [42], WikiPathways [43], and Reactome [44].

TABLE I. SEQUENCE FEATURES USED IN THE CLASSIFIERS

Feature type	***Feature***
Amino acid properties from EMBOSS PEPSTAT and PseAAC-Builder	Number of residues Molecular weight Average molecular weight Isoelectric point Charge Frequency of tiny (A+C+G+S+T) Frequency of small (A+B+C+D+G+N+P+S+T+V) Frequency of aliphatic (A+I+L+V) Frequency of aromatic (F+H+W+Y) Frequency of charged (B+D+E+H+K+R+Z) Frequency of polar (D+E+H+K+N+Q+R+S+T+Z) Frequency of non-polar (A+C+F+G+I+L+M+P+V+W+Y) Frequency of basic (H+K+R) Frequency of acidic (B+D+E+Z) Pseudo amino acid composition
Membrane topology	Predicted # of TM from TMHMM Predicted # of TM from STMHMM Predicted # of TM from SCAMPI Predicted # of TM from HMMTOP Predicted # of TM from PHOBIUS

G. *Characterization of pharmaceutical information of membrane proteins*

We collected known drug target proteins and disease associated proteins to construct pharmaceutical information of membrane proteins. Drugbank [45] and KEGG DRUG were used to extract known drug-target protein relationships and drug information covering their name, indication, and drug classification based on Anatomical Therapeutic Chemical (ATC) code. Disease associated proteins were gathered from OMIM [46], Genetic Association Database [47], and KEGG DISEASE.

H. *Identification of pharmaceutical association of the classified membrane protein groups*

The integration of classified membrane proteins and their pharmaceutical information makes it possible to derive the pharmaceutical association of each classified protein group. We calculated the total number of drug target proteins and disease associated proteins and the proportion of these in member proteins of each class. To measure the specific association in comparison with with other protein groups, we also calculated the p-value using hypergeometric test with false discovery rate (FDR) multiple testing correction, which was implemented in our previous research for enrichment analysis [48].

TABLE II. PERFORMANCE COMPARISON OF INTEGRAL MEMBRANE PROTEIN PREIDCTION METHODS

Prediction Tool	*Sensitivity*	*Accuracy*	*Specificity*	*F-score*
IMP Classifier in CHMPD	0.9307	0.9531	0.9707	0.9459
PHOBIUS	0.9175	0.9402	0.9582	0.9313
TMHMM	0.8961	0.9292	0.9554	0.9179
S-TMHMM	0.9311	0.9011	0.8774	0.8926
SCAMPI	0.9536	0.8881	0.8364	0.8827
HMMTOP	0.9517	0.7943	0.6699	0.8033

TABLE III. PERFORMANCE OF CLASSIFIERS FOR LIPID-ANCHORED PROTEIN AND PERIPHERAL MEMBRANE PROTEIN

Prediction Tool	*Sensitivity*	*Accuracy*	*Specificity*	*F-score*
LAP Classifier	0.7552	0.7445	0.7338	0.7467
PMP Classifier	0.6379	0.6471	0.6563	0.6434

III. CONCLUSION

We proposed a comprehensive database that focused on human membrane proteins to support novel pharmaceutical discoveries in the membrane proteome. Compared to other integral membrane protein database, this database covers biologically important non-integral membrane proteins and provides novel membrane protein candidates by considering their interaction type with membrane.

To characterize the collected membrane proteins, we hierarchically classified the membrane proteins into subcellular localization and molecular function groups and annotated them with various features including pharmaceutical information. We also measured the pharmaceutical associations of each classified group. Therefore, researchers can infer what kind of membrane protein group is more pharmaceutically targeted and can investigate what kinds of features the target proteins have in various aspects. Researches can also identify common and specific features of certain membrane proteins in one place, which is not possible in other specific membrane protein group databases. CHMPD will be a valuable resource for the in-depth study of the various features and mechanisms of membrane proteins and of novel disease associated protein discovery in the human membrane proteome.

TABLE IV. CURRENT STATISTICS OF MEMBRANE PROTEINS IN CHMPD

Resource Type	*Protein Count*
Collected membrane proteins from other resources	28,701
Membrane protein orthologs from 51 other species	987
Predicted integral membrane proteins	6,376
Predicted lipid-anchored proteins	590
Predicted peripheral membrane proteins	1,663
Total	38,317

ACKNOWLEDGMENT

This work was partially supported by the Korea Institute of Science and Technology Information (KISTI), partially supported by the Converging Research Center Program funded by the Ministry of Education, Science and Technology (Project No. 2011K000864), partially supported by the KAIST Future Systems Healthcare Project from the Ministry of Education, Science and Technology, and partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST)(No. 2012-0001001)

REFERENCES

[1] M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, and M. Vidal, "Drug-target network," *Nat Biotechnol,* vol. 25, pp. 1119-26, Oct 2007.

[2] M. H. Saier, Jr., M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan, "The Transporter Classification Database: recent advances," *Nucleic Acids Res,* vol. 37, pp. D274-8, Jan 2009.

[3] W. J. Gallin and P. A. Boutet, "VKCDB: voltage-gated K+ channel database updated and upgraded," *Nucleic Acids Res,* vol. 39, pp. D362-6, Jan 2011.

[4] M. C. Theodoropoulou, P. G. Bagos, I. C. Spyropoulos, and S. J. Hamodrakas, "gpDB: a database of GPCRs, G-proteins, effectors and their interactions," *Bioinformatics,* vol. 24, pp. 1471-2, Jun 15 2008.

[5] J. L. Sharman*, et al.*, "IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data," *Nucleic Acids Res,* vol. 39, pp. D534-8, Jan 2011.

[6] B. Vroling*, et al.*, "GPCRDB: information system for G protein-coupled receptors," *Nucleic Acids Res,* vol. 39, pp. D309-19, Jan 2011.

[7] R. Schwacke*, et al.*, "ARAMEMNON, a novel database for Arabidopsis integral membrane proteins," *Plant Physiol,* vol. 131, pp. 16-26, Jan 2003.

[8] A. A. Adjei, "Blocking oncogenic Ras signaling for cancer therapy," *J Natl Cancer Inst,* vol. 93, pp. 1062-74, Jul 18 2001.

[9] K. C. Chou and H. B. Shen, "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochem Biophys Res Commun,* vol. 360, pp. 339-45, Aug 24 2007.

[10] M. Ashburner*, et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet,* vol. 25, pp. 25-9, May 2000.

[11] J. Sprenger*, et al.*, "LOCATE: a mammalian protein subcellular localization database," *Nucleic Acids Res,* vol. 36, pp. D230-3, Jan 2008.

[12] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio, "eSLDB: eukaryotic subcellular localization database," *Nucleic Acids Res,* vol. 35, pp. D208-12, Jan 2007.

[13] N. Wiwatwattana, C. M. Landau, G. J. Cope, G. A. Harp, and A. Kumar, "Organelle DB: an updated resource of eukaryotic protein localization and function," *Nucleic Acids Res,* vol. 35, pp. D810-4, Jan 2007.

[14] S. Zhang, X. Xia, J. Shen, Y. Zhou, and Z. Sun, "DBMLoc: a Database of proteins with multiple subcellular localizations," *BMC Bioinformatics,* vol. 9, p. 127, 2008.

[15] T. Guo, S. Hua, X. Ji, and Z. Sun, "DBSubLoc: database of protein subcellular localization," *Nucleic Acids Res,* vol. 32, pp. D122-4, Jan 1 2004.

[16] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res,* vol. 38, pp. D355-60, Jan 2010.

[17] G. E. Tusnady, L. Kalmar, and I. Simon, "TOPDB: topology data bank of transmembrane proteins," *Nucleic Acids Res,* vol. 36, pp. D234-9, Jan 2008.

[18] G. N. Tsaousis*, et al.*, "ExTopoDB: a database of experimentally derived topological models of transmembrane proteins," *Bioinformatics,* vol. 26, pp. 2490-2, Oct 1 2010.

[19] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics,* vol. 23, pp. 1282-8, May 15 2007.

[20] P. Flicek*, et al.*, "Ensembl 2012," *Nucleic Acids Res,* vol. 40, pp. D84-90, Jan 2012.

[21] G. E. Tusnady, Z. Dosztanyi, and I. Simon, "PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank," *Nucleic Acids Res,* vol. 33, pp. D275-8, Jan 1 2005.

[22] S. Moller, M. D. Croning, and R. Apweiler, "Evaluation of methods for the prediction of membrane spanning regions," *Bioinformatics,* vol. 17, pp. 646-53, Jul 2001.

[23] H. Viklund and A. Elofsson, "Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information," *Protein Sci,* vol. 13, pp. 1908-17, Jul 2004.

[24] A. Bernsel*, et al.*, "Prediction of membrane-protein topology from first principles," *Proc Natl Acad Sci U S A,* vol. 105, pp. 7177-81, May 20 2008.

[25] G. E. Tusnady and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics,* vol. 17, pp. 849-50, Sep 2001.

[26] L. Kall, A. Krogh, and E. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method," *J Mol Biol,* vol. 338, pp. 1027-36, May 14 2004.

[27] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions," *Nat Methods,* vol. 8, pp. 785-6, 2011.

[28] T. Y. Lee*, et al.*, "dbPTM: an information repository of protein post-translational modification," *Nucleic Acids Res,* vol. 34, pp. D622-7, Jan 1 2006.

[29] G. Poisson, C. Chauve, X. Chen, and A. Bergeron, "FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring," *Genomics Proteomics Bioinformatics,* vol. 5, pp. 121-30, May 2007.

[30] G. Bologna, C. Yvon, S. Duvaud, and A. L. Veuthey, "N-Terminal myristoylation predictions by ensembles of neural networks," *Proteomics,* vol. 4, pp. 1626-32, Jun 2004.

[31] S. Hunter*, et al.*, "InterPro in 2011: new developments in the family and domain prediction database," *Nucleic Acids Res,* vol. 40, pp. D306-12, Jan 2012.

[32] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics,* vol. 22, pp. 1658-9, Jul 1 2006.

[33] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite," *Trends Genet,* vol. 16, pp. 276-7, Jun 2000.

[34] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Anal Biochem,* vol. 425, pp. 117-9, Jun 15 2012.

[35] K. C. Chou and Y. D. Cai, "Prediction of membrane protein types by incorporating amphipathic effects," *J Chem Inf Model,* vol. 45, pp. 407-13, Mar-Apr 2005.

[36] L. Breiman, *Random Forests*: Machine Learning, 2001.

[37] N. Bhardwaj, R. V. Stahelin, G. Zhao, W. Cho, and H. Lu, "MeTaDoR: a comprehensive resource for membrane targeting domains and their host proteins," *Bioinformatics,* vol. 23, pp. 3110-2, Nov 15 2007.

[38] H. Youngwoong, S. Choong-Hyun, K. Min-Sung, and Y. Gwan-Su, "Combined Database System for Binary Protein Interaction and Co-complex Association," in *Computer Science and Information Technology - Spring Conference*, pp. 538-542.

[39] M. S. Almen, K. J. Nordstrom, R. Fredriksson, and H. B. Schioth, "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin," *BMC Biol,* vol. 7, p. 50, 2009.

[40] S. Lata and G. P. Raghava, "PRRDB: a comprehensive database of pattern-recognition receptors and their ligands," *BMC Genomics,* vol. 9, p. 180, 2008.

[41] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Res,* vol. 28, pp. 304-5, Jan 1 2000.

[42] C. F. Schaefer*, et al.*, "PID: the Pathway Interaction Database," *Nucleic Acids Res,* vol. 37, pp. D674-9, Jan 2009.

[43] A. R. Pico*, et al.*, "WikiPathways: pathway editing for the people," *PLoS Biol,* vol. 6, p. e184, Jul 22 2008.

[44] L. Matthews*, et al.*, "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Res,* vol. 37, pp. D619-22, Jan 2009.

[45] C. Knox*, et al.*, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res,* vol. 39, pp. D1035-41, Jan 2011.

[46] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res,* vol. 33, pp. D514-7, Jan 1 2005.

[47] Y. Zhang*, et al.*, "Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information," *BMC Med Genomics,* vol. 3, p. 1, 2010.

[48] C. H. Sun, M. S. Kim, Y. Han, and G. S. Yi, "COFECO: composite function annotation enriched by protein complex data," *Nucleic Acids Res,* vol. 37, pp. W350-5, Jul 2009.