

Database Development and Discrimination Algorithms for Membrane Protein Functions

M. Michael Gromiha, Y. Yabuki, K. Imai, P. Horton, and K. Fukui

Abstract—We have developed a database for membrane protein functions, which has more than 3000 experimental data on functionally important amino acid residues in membrane proteins along with sequence, structure and literature information. Further, we have proposed different methods for identifying membrane proteins based on their functions: (i) discrimination of membrane transport proteins from other globular and membrane proteins and classifying them into channels/pores, electrochemical and active transporters, and (ii) β -signal for the insertion of mitochondrial β -barrel outer membrane proteins and potential targets. Our method showed an accuracy of 82% in discriminating transport proteins and 68% to classify them into three different transporters. In addition, we have identified a motif for targeting β -signal and potential candidates for mitochondrial β -barrel membrane proteins. Our methods can be used as effective tools for genome-wide annotations.

Keywords—Membrane proteins; database; transporters; discrimination; β -signal.

I. INTRODUCTION

MEMBRANE proteins perform a diverse variety of functions and are used as main drug targets of pharmaceutical agents. The collection of information on potential amino acid residues for the function of membrane proteins is important for understanding the sequence-structure-function relationship of membrane proteins as well as predicting the functional residues from sequence/structure. The discrimination algorithms for membrane protein structure and function would be valuable tools in the advancement of structural and functional genomics.

On the structural aspect, several methods have been proposed for discriminating α -helical and β -barrel membrane proteins and predicting their membrane spanning segments. These methods are mainly based on statistical analysis [1-3], hidden Markov models [4,5] and machine learning techniques [6-9]. The prediction algorithms have also been used to annotate membrane proteins in genomic sequences [10].

On the other hand, the functional aspects of membrane proteins have been studied with few approaches such as the development of databases [11-14], characterization of transport families [15], dissecting the sorting signal of mitochondrial β -barrel membrane proteins [16] etc. There is

no database for wide spectrum of functionally important residues in membrane proteins and the knowledge about functional discrimination of membrane proteins is still limited.

In this work, we have collected the information on functionally important amino acid residues in membrane proteins from the experimental data available in the literature. The functional data has been integrated with structural and sequence information, and other membrane protein databases. The database has several proteins that perform diverse functions and membrane transporters represent a large and diverse group of proteins. They play indispensable roles in the fundamental cellular processes of all organisms [17]. We have devised a method for discriminating membrane transport proteins from other globular and membrane proteins and classifying them into channels/pores, electrochemical and active transporters. In addition, we have analyzed the sequences of mitochondrial β -barrel outer membrane proteins and identified the motif for β -signal as well as probable targets. The salient features of the results will be discussed.

II. DATABASE FOR FUNCTIONAL RESIDUES IN MEMBRANE PROTEINS

A. Organization of Database

The organization of membrane protein function database is illustrated in Fig. 1. Each entry in the database contains the following information [18]: (i) protein name and source, (ii) main function of the protein, (iii) experimental data, (iv) methods and conditions and (v) literature.

We have provided the sequence and structure information in the form of Uniprot [19] and Protein Data Bank, PDB [20] codes. The functional information includes relative activity of mutants with respect to wild type protein, affinity for binding, channel, drug, glycosylation, membrane insertion, cellular signaling, membrane translocation, transport etc. The experimental data has numerical values for binding affinity, V_{\max} (maximal velocity of transport), IC50 (measure of the effectiveness of a compound in inhibiting biological function), drug sensitivity, dissociation constant, uptake etc.

B. Features of Database

The database has several features such as the retrieval of data using various conditions and displaying the results. We have provided direct links to Uniprot, PDB and PUBMED literature database. In addition links are given to related

MMG, YY, KI, PH and KF are with the Computational Biology Research Center of the National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, Japan (phone: +81-3-3599-8046; fax: +81-3-3599-8081; e-mail: michael-gromiha@aist.go.jp).

structural, functional and genomic databases as well as to prediction methods.

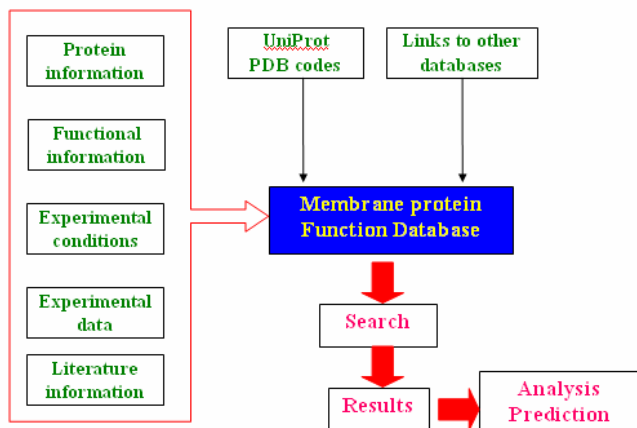


Fig. 1 Flowchart showing the organization of membrane protein function database

C. Example

The utility of the database is explained with an example. The available search options are shown in Fig. 2.

The database accepts and performs the queries in the search option (protein name, Uniprot ID, source, α -helical/ β -strand; function, parameter, mutation, keyword, authors and year) and displays the results.

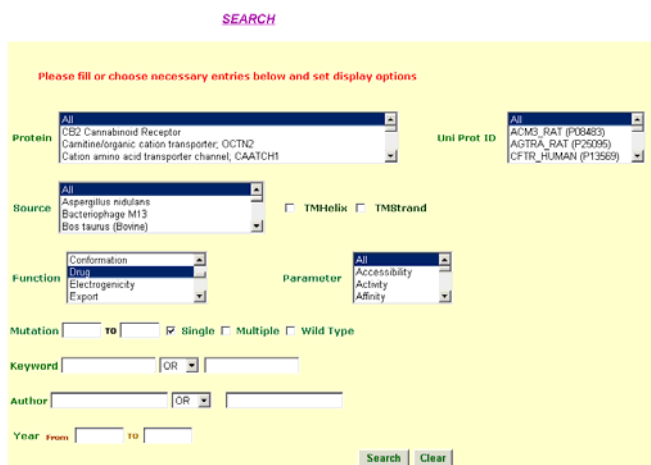


Fig. 2 Search options in membrane protein function database

III. DISCRIMINATION OF MEMBRANE PROTEINS BASED ON FUNCTIONS

We have developed different sets of data and utilized machine learning techniques for discriminating membrane proteins based on their functions.

A. Construction of Datasets

We have constructed different datasets for the present study: (i) channels/pores, electrochemical transporters and

active transporters from the information available in Transport Classification Database, TCDB [11]. We have removed the redundant sequences using blastclust program [21] so that no two proteins have the sequence identity of more than 20%. The final dataset contains 1718 proteins, which has 510 channels/pores, 502 electrochemical and 706 active transporters; (ii) a dataset of 3336 non-transport membrane proteins from SWISS-PROT and (iii) 1712 globular proteins from Protein Data Bank [20].

B. Machine Learning Algorithms

We have analyzed several machine learning techniques implemented in WEKA program [22] for discriminating membrane transporters from other proteins and classifying them into channels/pores, electrochemical and active transporters. This program includes several methods based on Bayes function, Neural network, Logistic function, Support vector machine, Regression analysis, Nearest neighbor, Meta learning, Decision tree and Rules.

C. Assessment

We have performed a 5-fold cross-validation test for assessing the validity of the present work. In this method, the data set is divided into five groups, four of them are used for training and the rest is used for testing the method. The same procedure is repeated for five times and the average is computed for obtaining the accuracy of the method.

We have used different measures, such as specificity, precision, F-measure and accuracy to assess the performance of discriminating channels/pores, electrochemical and active transporters. The term sensitivity shows the correct prediction of specific transporters and accuracy indicates the overall assessment. F-measure is the balance between sensitivity and precision. These terms are defined as follows:

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$F\text{-measure} = 2TP / (2TP + FP + FN)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN),$$

where, TP, FP, TN and FN refer to the number of true positives, false positives, true negatives and false negatives, respectively.

D. Sequence and Structural Analysis

We have computed the amino acid occurrence for the 20 amino acid residues in channels/pores, electrochemical and active transporters and analyzed the results. We noticed that the residues Asn and Gln are dominant in channels/pores among all the transporters [23]. Interestingly, these residues play important roles to the stability and function of β -barrel membrane proteins [24].

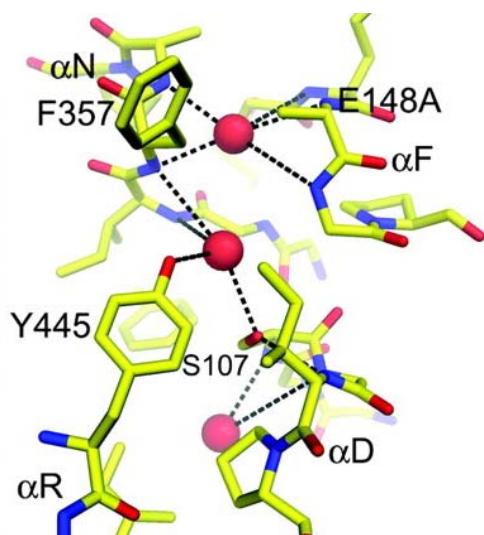


Fig. 3 Ion binding sites of E148A in CIC chloride channel. The hydrogen bonds are shown as black dashed lines

The structural analysis on cobalamin transporter protein (BtuB) that transports substrates across the outer membrane, showed that the residues, Asn185 and Asn276 are important for the stability of the upper surface of cyanocobalamin (vitamin B₁₂; CN-Cbl) binding pocket [25,26], which is important for its function. Further, the residues Glu166 and Glu 148 are important for the channel function in CIC chloride channel proteins as seen in Fig. 3 [27].

The residues Phe and Leu are dominant in electrochemical transporters. In addition, the composition of Ala, Ile, Val and Trp are higher in this class of proteins compared with other two transporters. Interestingly, in glycerol-3-phosphate transporter the space between helices 1 and 7 is filled by nine aromatic side chains and the occurrence of bulky aromatic residues helps to close the pore completely [28]. In lactose permease the substrate binding site is composed of residues that include Trp151 [29].

E. Discrimination of Membrane Transport Proteins

We have utilized a dataset of 5048 non-transporters (α -helical and β -barrel membrane proteins as well as globular proteins) and 1718 membrane transporters to discriminate the transporters. We have used “amino acid occurrence” as features and several machine learning techniques for discrimination. It has been shown that amino acid occurrence is one of the best parameters for discriminating proteins of different folds [30]. Our method showed the 5-fold cross-validation accuracy of 78.7% in discriminating transporters and non-transporters. Further, we have used the same number of proteins in transporters and non-transporters and repeated the calculations. We obtained the accuracy of 81.5% in distinguishing them.

F. Classification of Membrane Transport Proteins

We have analyzed different machine learning methods for classifying channels/pores, electrochemical and active transporters with amino acid occurrence as features. The results showed that the neural network is one of the best

methods and its performance is presented in Table I. It has been also shown that neural network is an efficient method for discriminating β -barrel membrane proteins [7,8].

The sensitivity is 0.55, 0.70 and 0.76 for channels/pores, electrochemical and active transporters, respectively. The precision is 0.70, 0.78 and 0.62, and F-measure is 0.61, 0.74 and 0.68. The average accuracy in classifying channels/pores, electrochemical and active transporters is 68% using 5-fold cross-validation.

We have analyzed the capability of BLAST to discriminate the three different types of transporters based on homology search. For each protein we have computed the sequence identity with all proteins in the three transporters and assigned the group, which has the highest sequence identity or best e-value. The calculations have been repeated for all the 1708 proteins and computed the overall accuracy. This method showed an accuracy of 51.6% in discriminating channels/pores, electrochemical and active transporters. Our method showed the accuracy of 68%, which is superior to simple BLAST search and the analysis revealed the better performance of the present method.

TABLE I
DISCRIMINATION OF CHANNELS/PORES, ELECTROCHEMICAL AND ACTIVE TRANSPORTERS USING K-NEAREST NEIGHBOR METHOD

Measure	Transporter	Performance
Sensitivity	Channels/pores	0.549
	Electrochemical	0.701
	Active	0.761
Precision	Channels/pores	0.695
	Electrochemical	0.780
	Active	0.622
F-measure	Channels/pores	0.613
	Electrochemical	0.739
	Active	0.684
Accuracy (%)	Overall	68.1

G. Genome-Wide Applications

The method developed for identifying different classes of transporters can be used to detect transporters in genomic sequences and annotate their functions. The protocol is shown in Fig. 4. For a new sequence, first it can be discriminated as a transporter or non-transporter using the discrimination method to classify the transporters (section E). This can be done with the highest accuracy of 82%. Further, for a transporter, it can be identified into channels/pores, electrochemical and active transporters with an accuracy of 68%. Hence, the two-way prediction system can be used to detect different types of transporters in genomic sequences.

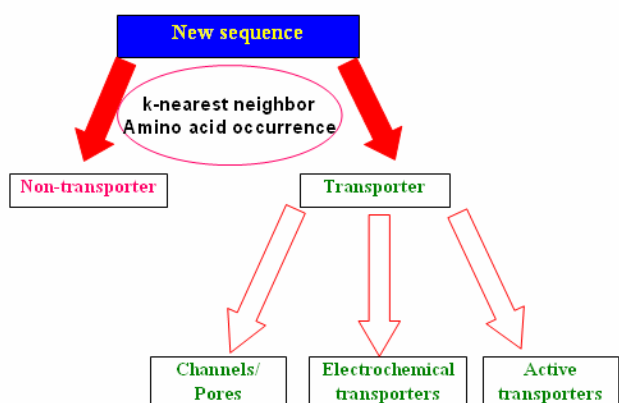


Fig. 4 Protocol to detect transporters in genomic sequences

IV. DETECTING BETA SIGNAL AND TARGET IN MITOCHONDRIAL BETA-BARREL MEMBRANE PROTEINS

Recently Kutik et al. [16] identified a sorting signal for mitochondrial β -barrel outer membrane proteins (MBOMPs), which has the motif PxGlyxxHxH (P: polar; Gly: glycine; H: hydrophobic; x: any amino acid) in the last β -strand based on their work on SAM complex. We have systematically carried out bioinformatics approaches to refine the signal, detect novel potential targets and the survey on available and probable MBOMPs.

A. Potential new MBOMPs

We have devised a procedure to detect novel MBOMPs using Gene Ontology annotation in Uniprot database, β -signal motif, evolutionary conservation, and predicted secondary structure [31]. The steps to identify MBOMPs are shown in Table II.

Starting from over 9,000 Eukaryotic proteins annotated as being mitochondrial in either Uniprot or Gene Ontology, we applied an automated procedure which reduced the number to 60 by requiring the refined β -signal motif to match within 40 residues of the C-terminus of each homolog cluster, and finally to 12 by consideration of predicted secondary structure and available annotation. Of the 12 clusters which remained after manual inspection, 11 were members of the known MBOMP families: porin, Tom40, Sam50 and Mdm10. The remaining cluster contained SUN family proteins with dual cell wall and mitochondrial localization [32]. One of the proteins in the cluster, yeast UTH1, is promising because it is a mitochondrial integral outer membrane protein [32].

TABLE II
STEPS USED TO IDENTIFY POTENTIAL MBOMPs

step/filter	clusters
Blastclust (40% identity)	2133
Motif match near C-terminus	60
Not α -helical MP (TMHMM)	44
Manual inspection	12
Remove known	1

B. Probable Number of MBOMPs

In earlier studies it has been reported that the yeast proteome would contain more than 100 MBOMPs [33]. Yet, five years and many completed genome sequences, we still only know of five MBOMPs. Our analysis also did not show the presence of many MBOMPs. Hence, we suggest that there may not be many unknown MBOMPs remaining.

C. Refined β -signal for Membrane Protein Insertion

We have carried out multiple sequence alignment analysis with the homologs of known five families of MBOMPs. The result reveals a set of 54 distinct octomers which presumably can function as β -signals. Analysis of these octomers shows that in 53/54 cases the residue following Glycine is hydrophobic {L:21, I:12, V:8, F:4, A:4, C:2, W:1, M:1}, with the single exception being threonine found in a fungal porin (VDAC_NEUCR). These observations yield the motif: PxGlyHxHxH, whose alternating hydrophobic residues probably reflect the dyad repeat structure of β -strands. Recently, Hiller et al. [34] reported the solution structure of human VDAC-1 in which the last β -strand contains the β -signal (KLGLGLEF) and it satisfies the refined motif. The residues constituting the β -signal form hydrogen bonds with the penultimate strand in anti-parallel orientation, and the first strand in parallel orientation, which is in contrast to known bacterial OMP structures that have an even number of strands connected exclusively in anti-parallel orientation [35].

The database and prediction algorithms can be used for understanding the sequence-structure-function relationship of membrane proteins.

ACKNOWLEDGMENT

The authors wish to thank Dr. M. Suwa for useful comments.

REFERENCES

- [1] T. Hirokawa, S. Boon-Chieng, S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics* 14 (1998) 378-379.
- [2] M.M. Gromiha, M. Suwa. A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics* 21 (2005) 961-968.
- [3] Y.D. Cai, K.C. Chou. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. *J Theor Biol.* (2006) 238: 395-400.
- [4] P.L. Martelli, P. Fariselli, A. Krogh, R. Casadio, A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins, *Bioinformatics* 18 (2002) S46-S53.
- [5] P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, S.J. Hamodrakas, A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins, *BMC Bioinformatics* 5 (2004) 29.
- [6] N.K. Natt, H. Kaur, G.P. Raghava. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins* 56 (2004) 11-18.
- [7] M.M. Gromiha, M. Suwa. Discrimination of outer membrane proteins using machine learning algorithms. *Proteins* 63 (2006) 1031-1037.
- [8] M.M. Gromiha, M. Suwa. Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. *Biochim Biophys Acta.* 2006 Sep;1764(9):1493-7.

- [9] Gromiha MM, Ahmad S, Suwa M. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J Comput Chem.* 2004 Apr 15;25(5):762-7.
- [10] Gromiha MM, Yabuki Y, Kundu S, Suharnan S, Suwa M. (2007) TMBETA-GENOME: database for annotated beta-barrel membrane proteins in genomic sequences. *Nucleic Acids Res.* 35: D314-6
- [11] M.H. Saier, Jr, C.V. Tran, R.D. Barabote. (2006) TCDB: the transporter classification database for membrane transport protein analyses and information *Nucleic Acids Res.* 34, D181–D186
- [12] Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* 31:294-297.
- [13] Schwacke R, Schneider A, Van Der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugge UI, Kunze R. (2003) ARAMEMNON, a Novel Database for Arabidopsis Integral Membrane Proteins. *Plant Physiol.* 131: 16-26.
- [14] Edvardsen O, Reiersen AL, Beukers MW, Kristiansen K. (2002) tGRAP, the G-protein coupled receptors mutant database. *Nucleic Acids Res.* 30: 361-3
- [15] Li H, Dai X, Zhao X. (2008) A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. *Bioinformatics.* 24:1129-36.
- [16] Kutik S, Stojanovski D, Becker L, Becker T, Meinecke M, Krüger V, Prinz C, Meisinger C, Guiard B, Wagner R, Pfanner N, Wiedemann N. Dissecting membrane insertion of mitochondrial beta-barrel proteins. *Cell.* 2008 Mar 21;132(6):1011-24.
- [17] M.H. Saier, Jr. A functional-phylogenetic classification system for transmembrane solute transporters *Microbiol. Mol. Biol. Rev.*, (2000) 64, 354–411
- [18] Gromiha MM, Yabuki Y, Suresh MX, Thangakani AM, Suwa M, Fukui K. TMFunction: database for functional residues in membrane proteins. *Nucleic Acids Res.* 2008; doi:10.1093/nar/gkn672.
- [19] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34: D187-91.
- [20] Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. (2007) *Nucleic Acids Res.* 35: D301-3
- [21] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25: 3389-3402.
- [22] Witten IH, Frank E: *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
- [23] Gromiha MM, Yabuki Y. (2008) Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics.* 9:135.
- [24] Gromiha MM, Suwa M. (2007) Current developments on beta-barrel membrane proteins: sequence and structure analysis, discrimination and prediction. *Curr. Protein Pept Sci.* 8: 580-99.
- [25] Chimento DP, Mohanty AK, Kadner RJ, Wiener MC. Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. *Nat Struct Biol.* 2003, 10: 394-401.
- [26] Chimento DP, Kadner RJ, Wiener MC. The Escherichia coli outer membrane cobalamin transporter BtuB: structural analysis of calcium and substrate binding, and identification of orthologous transporters by sequence/ structure conservation. *J Mol Biol.* 2003, 332: 999-1014.
- [27] Dutzler R, Campbell EB, MacKinnon R. Gating the selectivity filter in CIC chloride channels. *Science.* 2003; 300:108-12.
- [28] Huang Y, Lemieux MJ, Song J, Auer M, Wang DN. Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli. *Science.* 2003; 301:616-20.
- [29] Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S. Structure and mechanism of the lactose permease of Escherichia coli. *Science.* 2003; 301:610-5.
- [30] Taguchi YH, Gromiha MM. Application of amino acid occurrence for discriminating different folding types of globular proteins. *BMC Bioinformatics* 2007, 8, 404.
- [31] Imai K, Gromiha MM, Horton P. Mitochondrial β -Signal; The End of the Story? *Cell* (in press).
- [32] Velours G, Boucheron C, Manon S, Camougrand N. Dual cell wall/mitochondria localization of the 'SUN' family proteins. *FEMS Microbiol Lett.* 2002 Feb 5;207(2):165-72
- [33] Wimley WC. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol.* 2003 Aug;13(4):404-11
- [34] Hiller S, Garces RG, Malia TJ, Orekhov VY, Colombini M, Wagner G. Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science.* 2008 Aug 29;321(5893):1206-10
- [35] Tusnády GE, Dosztányi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D275-8.