D314–D316 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 6 November 2006*
doi:10.1093/nar/gkl805

TMBETA-GENOME: database for annotated β-barrel membrane proteins in genomic sequences

M. Michael Gromiha[1,*], Yukimitsu Yabuki[1], Srinesh Kundu[2], Sivasundaram Suharnan[2] and Makiko Suwa[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan and [2]Axiohelix Pvt Limited, Tokyo, Japan

Received August 7, 2006; Revised September 20, 2006; Accepted September 29, 2006

ABSTRACT

We have developed the database, TMBETA-GENOME, for annotated β-barrel membrane proteins in genomic sequences using statistical methods and machine learning algorithms. The statistical methods are based on amino acid composition, reside pair preference and motifs. In machine learning techniques, the combination of amino acid and dipeptide compositions has been used as main attributes. In addition, annotations have been made using the criterion based on the identification of β-barrel membrane proteins and exclusion of globular and transmembrane helical proteins. A web interface has been developed for identifying the annotated β-barrel membrane proteins in all known genomes. The users have the feasibility of selecting the genome from the three kingdoms of life, archaea, bacteria and eukaryote, and five different methods. Further, the statistics for all genomes have been provided along with the links to different algorithms and related databases. It is freely available at http://tmbeta-genome.cbrc.jp/annotation/.

INTRODUCTION

The β-barrel membrane proteins perform a variety of functions, such as pore formation, membrane anchoring, enzyme activity, bacterial virulence, mediating non-specific, passive transport of ions and small molecules, selectively passing the molecules such as maltose and sucrose and are involved in voltage-dependent anion channels (1). The annotation of β-barrel membrane proteins in genomic sequences will be helpful for understanding their functions. In our earlier works, we have developed statistical methods and machine learning techniques for discriminating transmembrane β-barrel proteins (TMBs) from globular and transmembrane helical (TMH) proteins (2–5) and we obtained the accuracy in the range of 89–95% in discriminating TMBs. These methods are mainly based on amino acid composition, residue pair preference/dipeptide composition, motifs and the combinations of them. On the other hand, different methods, such as hidden Markov models, neural networks and nearest neighbor algorithms, have been proposed for discriminating TMBs and screening them in genome sequences (6–12). However, there is no electronically available database for the annotated β-barrel membrane proteins in genomes.

In this work, we have developed a database, TMBETA-GENOME, which has the annotated β-barrel membrane proteins for all the completed genomes. The annotation has been carried out with several statistical methods and machine learning techniques along with a new approach based on detecting β-barrel membrane proteins and eliminating other folding types of globular and membrane proteins. The users have the feasibility of selecting the method and the genome. The database is freely available at http://tmbeta-genome.cbrc.jp/annotation/.

CONTENTS OF THE DATABASE

TMBETA-GENOME contains the annotated β-barrel membrane proteins for 275 completed genomes, including 23 genomes from archaea, 237 from bacteria and 15 from eukaryote. It may be noted that very few TMBs are annotated experimentally in the eukaryote proteomes. Further, for a genome that have different chromosomes (e.g. human genome has 24 chromosomes) the data for each chormosome is given individually. This increased the total number of entries into 24, 254 and 149, respectively, for archaea, bacteria and eukaryote. The total number of proteins in these three kingdoms of life is 52 241, 686 562 and 165 186, respectively, with the total of 903 989 sequences. The amino acid sequences of all the genomes have been taken from the NCBI database (http://www.ncbi.nih.gov/). In addition, we have provided the statistics for the annotated β-barrel membrane proteins in all the genomes by different discrimination methods (see below).

*To whom correspondence should be addressed. Tel: +81 3 3599 8046; Fax: +81 3 3599 8081; Email: michael-gromiha@aist.go.jp

DISCRIMINATION METHODS

The annotation results are accumulated for different statistical methods and machine learning techniques. The statistical methods include the composition of amino acid residues (2), residue pair preference (3) and motifs (4). In these methods the compositions of amino acid residues/residue pairs/residue pairs with a gap (motif) have been computed for a training set of 674 globular and 377 TMB proteins obtained from Protein Data Bank (PDB) and PSORT database, respectively (2–4,13,14). For a new protein X, we have calculated the deviations of the amino acid composition between protein X and globular/TMBs. The protein is said to be a TMB if the deviation is the lowest for TMB and vice versa (2). For residue pair preference and motif, the compositional difference between globular and TMBs have been calculated ($\sigma_{\text{TMB-glob}}$). The weighted average of $\sigma_{\text{TMB-glob}}$ with the dipeptide/motif composition of protein X discriminates the TMB and globular protein (3,4). We have also applied support vector machines (5) for discriminating TMBs, which uses the combinations of amino acid composition and residue pair preference. Further, the program, SOSUI (15) has been used for finding the TMH in genomic sequences. In the new approach we have used the following criteria: (i) identification of TMBs using the preference of residue pairs in globular, TMH and TMBs, (ii) elimination of globular/TMH proteins that show the sequence identity of >70% for the coverage of 80% residues with known structures in PDB, (iii) elimination of globular/TMH proteins that have the sequence identity of >60% with known sequences in SWISS-PROT, and (iv) exclusion of TMH proteins using SOSUI, a prediction system for TMH proteins. This method also showed good agreement with experimental observations.

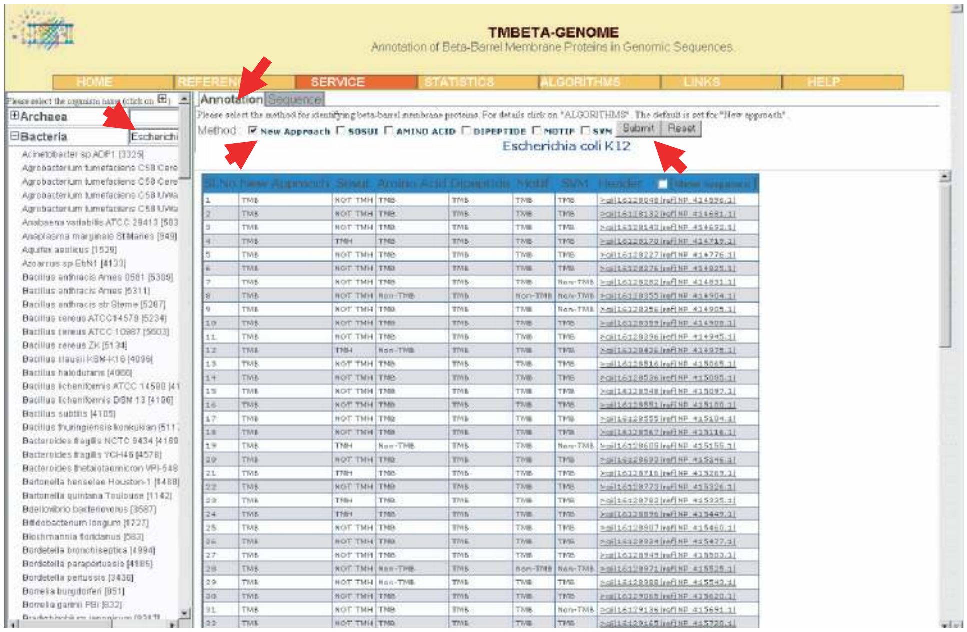
FEATURES OF TMBETA-GENOME

TMBETA-GENOME includes several features, such as, the service for detecting TMBs in genomic sequences using various methods, related references, statistics for the detected TMBs by different methods for each genome, details about all algorithms used to detect TMBs, relative links to other databases and a help page. The 'help' section illustrates the details to perform the search and to obtain the results.

ACCESS TO TMBETA-GENOME

TMBETA-GENOME can be directly accessed through the web at http://tmbeta-genome.cbrc.jp/annotation/. The users have the feasibility to select the method of annotation and name of the genome. This server provides the annotated TMBs using our previous methods, such as, statistical, dipeptide, motif and SVM as well as the 'New Approach'. The new approach considerably reduced the number of false positives and it has the ability of picking up most of the real TMBs. An example is shown in Figure 1. In this figure the results are shown for *Escherichia coli* K12 genome. This can be obtained by clicking on the button + Bacteria and selecting the name of the genome. It is also possible to get the data by entering the name of the genome.

We have selected the method, 'New Approach' for obtaining the annotated TMBs. This search picked up 87 entries and the TMBs identified by the new approach are shown with the identification number. In addition, the results obtained with other methods are also given for comparison. The method SVM yielded 337 TMBs and the combination of 'Amino acid' and 'Dipeptide' showed 501 TMBs. It is noteworthy that several discrimination methods including statistical,



Please select the organism name (click on ⊞)

⊞Archaea

⊟Bacteria Escherichi

Acinetobacter sp.ADP1 [3325]
Agrobacterium tumefaciens C58 Cere
Agrobacterium tumefaciens C58 Cere
Agrobacterium tumefaciens C58 UWa
Agrobacterium tumefaciens C58 UWa
Anabaena variabilis ATCC 29413 [503
Anaplasma marginale St Maries [949]
Aquifex aeolicus [1529]
Azoarcus sp EbN1 [4133]
Bacillus anthracis Ames 0581 [5309]
Bacillus anthracis Ames [5311]
Bacillus anthracis str Sterne [5287]
Bacillus cereus ATCC14579 [5234]
Bacillus cereus ATCC 10987 [5603]
Bacillus cereus ZK [5134]
Bacillus clausii KSM-K16 [4096]
Bacillus halodurans [4066]
Bacillus licheniformis ATCC 14580 [41
Bacillus licheniformis DSM 13 [4196]
Bacillus subtilis [4105]
Bacillus thuringiensis konkukian [511
Bacteroides fragilis NCTC 9434 [4189
Bacteroides fragilis YCH46 [4578]
Bacteroides thetaiotaomicron VPI-548
Bartonella henselae Houston-1 [1488]
Bartonella quintana Toulouse [1142]
Bdellovibrio bacteriovorus [3587]
Bifidobacterium longum [1727]
Blochmannia floridanus [583]
Bordetella bronchiseptica [4994]
Bordetella parapertussis [4185]
Bordetella pertussis [3436]
Borrelia burgdorferi [851]
Borrelia garinii PBi [832]

Please select the method for identifying beta-barrel membrane proteins. For details click on "ALGORITHMS". The default is set for "New approach".

Method: ☑ New Approach ☐ SOSUI ☐ AMINO ACID ☐ DIPEPTIDE ☐ MOTIF ☐ SVM

Escherichia coli K12

Sl No	New Approach	Sosui	Amino Acid	Dipeptide	Motif	SVM	Header	☐ Show sequence
1	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128048|ref|NP_414596.1|	
2	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128132|ref|NP_414681.1|	
3	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128143|ref|NP_414692.1|	
4	TMB	TMH	TMB	TMB	TMB	TMB	>gi|16128170|ref|NP_414719.1|	
5	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128227|ref|NP_414776.1|	
6	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128276|ref|NP_414825.1|	
7	TMB	NOT TMH	TMB	TMB	TMB	Non-TMB	>gi|16128282|ref|NP_414831.1|	
8	TMB	NOT TMH	Non-TMB	TMB	Non-TMB	Non-TMB	>gi|16128355|ref|NP_414904.1|	
9	TMB	NOT TMH	TMB	TMB	TMB	Non-TMB	>gi|16128356|ref|NP_414905.1|	
10	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128359|ref|NP_414908.1|	
11	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128396|ref|NP_414945.1|	
12	TMB	TMH	Non-TMB	TMB	TMB	TMB	>gi|16128426|ref|NP_414975.1|	
13	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128516|ref|NP_415065.1|	
14	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128536|ref|NP_415085.1|	
15	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128548|ref|NP_415097.1|	
16	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128551|ref|NP_415100.1|	
17	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128555|ref|NP_415104.1|	
18	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128567|ref|NP_415116.1|	
19	TMB	TMH	Non-TMB	TMB	TMB	Non-TMB	>gi|16128605|ref|NP_415155.1|	
20	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128693|ref|NP_415246.1|	
21	TMB	TMH	TMB	TMB	TMB	TMB	>gi|16128716|ref|NP_415263.1|	
22	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128773|ref|NP_415326.1|	
23	TMB	TMH	TMB	TMB	TMB	TMB	>gi|16128782|ref|NP_415335.1|	
24	TMB	TMH	TMB	TMB	TMB	TMB	>gi|16128896|ref|NP_415447.1|	
25	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128907|ref|NP_415460.1|	
26	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128924|ref|NP_415477.1|	
27	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16128949|ref|NP_415503.1|	
28	TMB	NOT TMH	Non-TMB	TMB	Non-TMB	Non-TMB	>gi|16128971|ref|NP_415525.1|	
29	TMB	NOT TMH	Non-TMB	TMB	TMB	TMB	>gi|16128988|ref|NP_415542.1|	
30	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16129065|ref|NP_415620.1|	
31	TMB	NOT TMH	TMB	TMB	TMB	Non-TMB	>gi|16129136|ref|NP_415691.1|	
32	TMB	NOT TMH	TMB	TMB	TMB	TMB	>gi|16129165|ref|NP_415720.1|	

Figure 1. Illustration for searching the database. The arrows indicate the selected items, (i) name of the genome (*Escherichia coli* K12), (ii) method (New Approach) and (iii) annotation (set as default). Each change can be taken into account by clicking on the 'Submit' button. The displayed results are also shown in this figure.

dipeptide and motifs have a tendency of providing many false positives. The new approach results are reasonable that it identified just 2.05% of the proteins as TMBs. Further, the comparison between identified TMBs and experimentally known TMBs revealed that the new approach could correctly pick up all the 11 TMBs of known three-dimensional structures obtained from *E.coli* and representative proteins in all the families of Transport Classification Database (TCDB, 16). Hence, we suggest to use the new approach for obtaining TMBs in genomic sequences.

Few other methods have also been proposed for detecting TMBs and these methods have their own advantages and shortcomings (6–12). Zhai and Saier (10) developed a program, β-barrel finder, based on secondary structure, hydropathy and amphipathicity for identifying β-barrel membrane proteins in prokaryotic genomes. This method detected 118 TMBs in *E.coli* genome and it could identify representative proteins from nine among 15 families available in TCDB. The method based on *k*-nearest neighbor missed few TMBs with the *E*-value > 3 although these proteins have been used in the training set to develop the method (9). The BOMB server based on (i) C-terminal pattern typical of many integral β-barrel proteins and (ii) integral β-barrel score based on the extent to which the sequence contains stretches of amino acids typical of transmembrane β-strands missed eight proteins (11). The profile based hidden Markov model (12) identified TMBs belonging to 12 families in TCDB. This analysis reveals that the results obtained with new approach are better than other methods in the literature.

For more details about the protein the appropriate links to NCBI protein sequences have been provided for each annotated protein. Further, the users have the feasibility of downloading the annotated TMBs in FASTA format, which can be used for further analysis. The complete sequences of the genome have been obtained by selecting the 'Sequence' button. It is also possible to download all the protein sequences of the specific genome in FASTA format.

LINKS TO OTHER DATABASES

Each protein in the specified genome as well as the annotated TMBs are directly linked with NCBI protein sequences. Further, TMBETA-GENOME is linked with several related genome, structure and sequence databases, such as, Genome Online Database [GOLD; (17)], NCBI (http://www.ncbi.nlm.nih.gov/Genomes/), KEGG (http://www.genome.jp/), PDB (13), SWISS-PROT (http://www.expasy.org/sprot/), Protein Information Resource (PIR; http://pir.georgetown.edu), Uniprot (18), Protein Data Bank of Transmembrane proteins [PDBTM; (19)], Transport Classification Database [TCDB; (16)] etc.

AVAILABILITY AND CITATION OF TMBETA-GENOME

The database can be freely accessible at http://tmbeta-genome.cbrc.jp/annotation/. If this database is used as a tool in your published research work, please cite this article including the URL. Suggestions and comments are welcome and should be sent to michael-gromiha@aist.go.jp.

ACKNOWLEDGEMENTS

We thank Dr Yutaka Akiyama for encouragement. Funding to pay the Open Access publication charges for this article was provided by the National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Koebnik,R., Locher,K.P. and Van Gelder,P. (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol*., **37**, 239–253.
2. Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics*, **21**, 961–968.
3. Gromiha,M.M., Ahmad,S. and Suwa,M. (2005) Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput. Biol. Chem*., **29**, 135–142.
4. Gromiha,M.M. (2005) Motifs in outer membrane protein sequences: applications for discrimination. *Biophys. Chem*., **117**, 65–71.
5. Park,K.J., Gromiha,M.M., Horton,P. and Suwa,M. (2005) Discrimination of outer membrane proteins using support vector machines. *Bioinformatics*, **21**, 4223–4229.
6. Martelli,P.L., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**, S46–S53.
7. Bagos,P.G., Liakopoulos,T.D., Spyropoulos,I.C. and Hamodrakas,S.J. (2004) A hidden Markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.
8. Natt,N.K., Kaur,H. and Raghava,G.P. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, **56**, 11–18.
9. Garrow,A.G., Agnew,A. and Westhead,D.R. (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. *Nucleic Acids Res*., **33**, W188–W192.
10. Zhai,Y. and Saier,M.H., Jr (2002) The β-barrel finder (BBF) program, allowing identification of outer membrane β-barrel proteins encoded within prokaryotic genomes. *Protein Sci*., **11**, 2196–2207.
11. Berven,F.S., Flikka,K., Jensen,H.B. and Eidhammer,I. (2004) BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res*., **32**, W394–W399.
12. Bigelow,H.R., Petrey,D.S., Liu,J., Przybylski,D. and Rost,B. (2004) Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res*., **32**, 2566–2577.
13. 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.
14. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., de Fays,K., Lambert,C., Nakai,K. and Brinkman,F.S. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.
15. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
16. Busch,W. and Saier,M.H., Jr (2002) The transporter classification (TC) system. *Crit Rev Biochem Mol Biol*., **37**, 287–337.
17. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*., **34**, D332–D334.
18. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The universal protein resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
19. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*., **33**, D275–D278.