W188–W192 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki384

TMB-Hunt: a web server to screen sequence sets for transmembrane β-barrel proteins

Andrew G. Garrow, Alison Agnew and David R. Westhead*

School of Biochemistry and Microbiology, University of Leeds, Leeds LS2 9JT, UK

Received February 14, 2005; Revised and Accepted March 9, 2005

ABSTRACT

TMB-Hunt is a program that uses a modified *k*-nearest neighbour (*k*-NN) algorithm to classify protein sequences as transmembrane β-barrel (TMB) or non-TMB on the basis of whole sequence amino acid composition. By including differentially weighted amino acids, evolutionary information and by calibrating the scoring, a discrimination accuracy of 92.5% was achieved, as tested using a rigorous cross-validation procedure. The TMB-Hunt web server, available at www.bioinformatics.leeds.ac.uk/betaBarrel, allows screening of up to 10 000 sequences in a single query and provides results and key statistics in a simple colour coded format.

INTRODUCTION

Integral membrane proteins can be grouped into two distinct structural classes: α-helical and β-barrel. Of these the transmembrane β-barrel (TMB) proteins are the least well characterized and have proven with high-resolution structures only been for proteins spanning the outer membranes of Gram −ve and acid-fast Gram +ve bacteria. It is also widely believed that TMB proteins are present in the outer membranes of chloroplasts and mitochondria, presumably owing to the endosymbiotic theory. As with α-helical transmembrane (AHTM) proteins, TMB proteins play both functionally important and diverse roles. Currently >90 TMB protein structures can be found in the Protein Data Bank (PDB) (1), fitting 23 families in the transmembrane protein structure database (2) and several folds of the SCOP hierarchy (3). From this diversity, it seems likely that TMB proteins have multiple evolutionary origins.

Unlike with the AHTM proteins, which can be easily distinguished on the basis of long stretches (>20 amino acids) of hydrophobic residues, development of TMB protein discriminators has proven difficult. This is due to a short and cryptic inside–outside dyad repeat motif in which only alternate residues are lipid facing and thus hydrophobic (4). Despite these difficulties, recently published algorithms have led to discrimination accuracies ranging from 80 to 90% (5–11) – levels acceptable if analysing a particular sequence of interest. Unfortunately, with these accuracy levels, problems will still occur when screening entire genomes owing to the large numbers of sequences tested, of which TMB proteins constitute only a small fraction. There is, therefore, still a need for improved algorithms.

TMB-Hunt uses whole sequence amino acid composition to discriminate between TMB and non-TMB proteins. Whole sequence amino acid composition has been applied to a number of other protein classification problems, including discrimination between intra- and extra-cellular proteins (12), membrane protein type (13), subcellular location (14) and structural class (15). However, although studies of TMB protein composition have been made, whole sequence amino acid composition has not yet been applied to the discrimination problem.

Because TMB-Hunt puts no emphasis on identification of TM β-strands, we were not dependent on sequences with resolved structures, thus allowing the use of training sets that were larger and more representative than those used for other predictors. TMB-Hunt is at least as accurate as other predictors. However, we believe that its major advantage is that, because it adopts a completely different approach from those of other methods, it will prove valuable to the development of consensus approaches that can more accurately be applied to searching diverse proteomes for novel families of candidate TMB proteins.

TRAINING AND TESTING SETS

Training sets for TMB, AHTM and non-transmembrane (NTM) proteins were gathered from a number of manually curated and published sources. The PDB accessions of 3159 NTM proteins were acquired from PDB-REPRDB via the Papia database (16), and respective sequences were extracted. Sequences of 189 AHTM proteins were obtained from datasets available from the Sanger Centre (17). TMB proteins came

*To whom correspondence should be addressed. Tel: +44 113 2333116; Fax: +44 113 2333167; Email: westhead@bmb.leeds.ac.uk

from a number of sources, including 957 from Uniprot (18), 134 from the transporter classification (TC) database (19) and 35 from the PDB files of TMB proteins found in SCOP. The TMB protein training set included a diverse range of proteins, including atypical TMB-forming proteins, TolC (20), α-hemolysin (21) and the mycobacterial porin MspA (22) in addition to a number of eukaryotic proteins expressed in the mitochondria and chloroplast.

Sequences of <120 residues were removed from training sets. The remaining sequences were then grouped into clusters using BLASTclust (ftp://ftp.ncbi.nih.gov/blast/) with a sequence similarity threshold of 23%. Amino acid composition frequency profiles were calculated for each of these clusters using evolutionary information as described below. In total, 1763 composition profiles were calculated for NTM proteins, 132 for AHTM proteins and 196 for TMB proteins.

ALGORITHM

The *k*-NN algorithm

TMB-Hunt classifies sequences using a *k*-nearest neighbour (*k*-NN) algorithm, which is a simple instance-based learning algorithm in which the class (i.e. TMB, AHTM or non-TMB) of a query instance (a protein in this case) is predicted using the class of its *k*-nearest neighbours within the training set. The *k*-NN algorithm is thus a local approximation, focusing on the neighbourhood of the query instance. One of its major advantages is that it is robust to noisy data (provided a large dataset), as taking the weighted average of the nearest neighbours smoothes out isolated training examples.

Here the difference between two proteins, $d^2(x_i, x_j)$, is measured using the standard Euclidean metric

$$d^2(x_i, x_j) = \sum_{r=1}^{n} \left[a_r(x_i) - a_r(x_j)\right]^2,$$

where $a_r(x)$ is the relative frequency occurrence of amino acid r in protein x.

In the standard *k*-NN algorithm, a score $S(x_q, c)$ is assigned to each possible class c using

$$S(x_q, c) = \sum_{i=1}^{k} \delta[c, c(x_i)] / d^2(x_q, x_i),$$

where $\delta(c_1, c_2) = 1$ if the classes c_1 and c_2 are equal and zero otherwise. Thus the score for each class is a sum of positive contributions from each of the nearest neighbours from that class, where the contribution is weighted according to the reciprocal square distance between query instance and neighbour, with closer neighbours contributing more strongly. However, because we are interested in a binary classification problem (i.e. TMB or non-TMB), we define a discrimination score

$$D(x_q, c) = S(x_q, c) - \sum_{c' \neq c} S(x_q, c'),$$

which is calculated as the score for the TMB class minus the scores for other classes.

Calibration and scoring

It is possible to convert discrimination scores into a convenient log likelihood ratio (LLR),

$$R(D) = \log(P(\text{TMB} \mid D) / P(\text{other} \mid D)),$$

where $P(\text{TMB} \mid D)$ denotes the probability of a TMB protein obtaining a score of *at least D* and $P(\text{other} \mid D)$ denotes the probability of a protein from the other class obtaining a score of *D or greater*. Negative values of *R* indicate a query protein more likely to come from the other class, and positive values indicate a protein more likely to come from the TMB class.

To take into account the multiple testing involved in screening large sequence sets, an expectancy value is also calculated using $E(D) = N\,P(\text{other} \mid D)$, where N indicates the number of query sequences tested. This measure is related to the standard Bonferroni correction and is directly analogous to the *E*-values reported by the popular sequence search programs FASTA (23) and BLAST (24).

Differential amino acid weightings

To account for the fact that some dimensions contribute information more valuable to classification than others, weights were applied to each of the dimensions used in calculating Euclidean distances. Optimal weightings were calculated with a genetic algorithm and were applied using a modified Euclidean distance

$$d^2(x_i, x_j) = \sum_{r=1}^{n} g_r \left[a_r(x_i) - a_r(x_j)\right]^2,$$

where g_r is the weight applied to the rth dimension.

Inclusion of evolutionary information

Random noise in amino acid composition was reduced by the inclusion of evolutionary information. Evolutionary information was included by building a feature vector using both the query sequence and a number of close homologues (as determined by a BLAST query against SwissProt with an *E*-value threshold of 0.0001 and a maximum of 25 homologues) to calculate an average amino acid composition vector for the sequence and its close evolutionary relatives. A weighted average composition was used, with more distant homologues contributing more to the average (since the more distant sequences contain more new information).

Table 1. Program performance using a variety of settings

	Sensitivity (%)	Positive predicative value (%)	Accuracy (%)
Plain	83	86.5	85
Weighted amino acids	84	90.3	87.5
Evolutionary information	89	93.7	91.5
Evolutionary information + weighted amino acids	91	93.8	92.5

The ability of TMB-Hunt to discriminate between BTM and NTM proteins, tested using the 'leave homologues out' cross-validation method and with a range of different features.

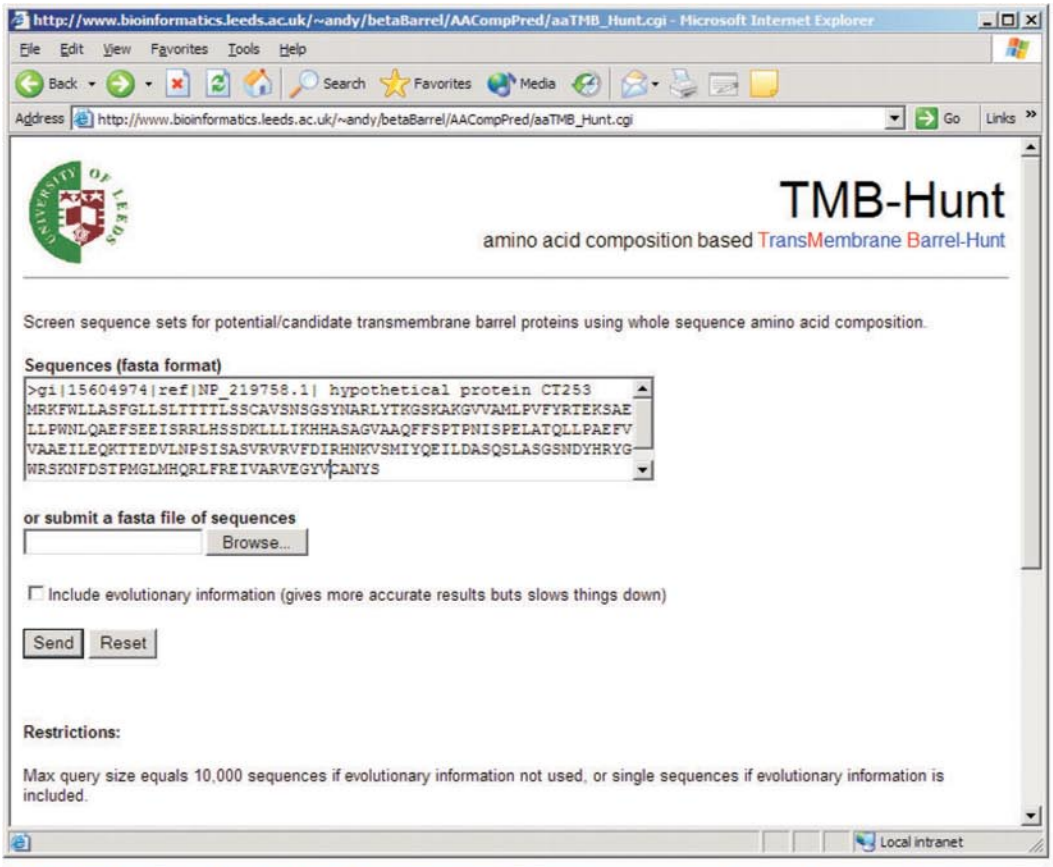

Figure 1. Images of the TMB-Hunt web server query form and output. Results provide key statistics in a simple-to-understand colour-coded format, with a more detailed output format available. Links are also provided to a number of prescreened genomes.

PERFORMANCE

Performance was measured using a 'leave homologues out' cross-validation. This involved precomputing sequences similar (with a BLAST *E*-value threshold <1) to each query sequence and then removing these in turn from the training set and seeing if whether algorithm could correctly reassign them. We found that best performance occurred with $k = 5$, although performance was generally insensitive to the precise value of k, with similar performance shown for moderate values $\geqslant 5$. Table 1 summarizes performance with a variety of settings. Using the 'leave homologues out' cross-validation, without inclusion of evolutionary information and without differential amino acid weightings, the program was able to discriminate between TMB and NTM proteins with 85% accuracy, 83% sensitivity and 86.5% positive predictive value (PPV). With inclusion of query sequence evolutionary information and weighted amino acids, discrimination accuracy increased to 92.5%, with 91% sensitivity and 93.8% PPV. Similar levels of accuracy were seen for discriminating between TMB and AHTM proteins, although clearly there are more specialized predictors available for this. These results suggest that TMB-Hunt is more accurate than other predictors (5–11), although direct comparison is difficult owing to differences in algorithms, test methods and test/training set size and compositions.

Using the cross-validation, TMB-Hunt correctly classified all TMB proteins tested that have structures resolved, except for OmpA (P02934) (25); however, OmpA is correctly classified in a self-consistency test. Among these correctly classified proteins were a number of TMB proteins with structures recently resolved, including NalP (Q8GKS5) (26), Tsx (P22786) (27), FadL (P10384) (28) and BtuB (P06129) (29). TMB-Hunt was also able to correctly classify a number of atypical TMB-forming proteins including the mycobacterial porin MspA (Q9RLP7) (22), TolC (P02930) (20) and α-hemolysin (O68404) (21), as well as TMB proteins from locations other than the Gram −ve bacterial outer membrane, for example, the mitochondrial porin VDAC (Q60931) and plastid porins Toc75 (Q43715) and OEP24 (O49929). TMB-Hunt also positively classifies a number of controversial TMB proteins, including secretin (P31700) and usher proteins (P30130). The lowest-scoring protein in the TMB test set was a 60 kDa cysteine-rich outer-membrane protein (P26758) (30). However, the experimental evidence that this is a genuine TMB protein is weak, and it has been suggested that it is falsely annotated (9).

SCOP structural classes were compared with non-TMB protein cross-validation results. TMB-Hunt was able to correctly reject 92.9% of the all-α proteins, 85.6% of all-β, 92% of α/β, 91.8% of α + β, 96.5% of other classified and 95% of those not classified within SCOP. Further analysis of the all-β results revealed that the main weakness was with the all-β proteins annotated as being secreted [e.g. sialidase (P37060), anhydrosialidase (Q27701), candidapepsin (Q00663) and galactose oxidase (Q01745)].

THE WEB SERVER

The web server (Figure 1) takes FASTA format sequences as input and has the option for inclusion of evolutionary information. The algorithm is very quick, capable of screening >400 sequences in <1 min using a 2 Ghz Pentium processor. On this basis, the web server allows up to 10 000 sequences in a single query. Results are reported in a simple, user-friendly colour-coded output with protein description line, log-likelihood ratio and *E*-value statistics. A warning is given if query sequence compositions are unusually distant to from of the training instances. Such problems may occur when screening short peptides (e.g. <50 amino acids) or open reading frames automatically predicted from a genome sequence. A link is also available to a 'full format' result that includes information on query sequence composition and individual-neighbour Euclidean distances. The web server comes with detailed instructions and also provides links to a number of prescreened genomes.

CONCLUSION

TMB-Hunt is a program that uses whole sequence amino acid composition to discriminate between TMB and non-TMB proteins. Using rigorous cross-validation procedures, accuracy levels were achieved that were higher than those previously reported. To our knowledge, this is the first time such an algorithm has been applied to the TMB protein discrimination problem, and it is thus hoped that it will prove a valuable step towards the development of consensus approaches. TMB-Hunt is extremely quick, and so the web server allows screening of up to 10 000 sequences in a single query. Results and key statistics are reported using a simple colour coding system.

ACKNOWLEDGEMENTS

The authors would like to thank the Medical Research Council for funding and reviewers for their constructive criticism. Funding to pay the Open Access publication charges for this article was jointly provided by the University of Leeds and JISC.

Conflict of interest statement. None declared.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F.J., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
2. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, **20**, 2964–2972.
3. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
4. Schulz,G.E. (2000) beta-Barrel membrane proteins. *Curr. Opin. Struct. Biol.*, **10**, 443–447.
5. 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.
6. Bagos,P.G., Liakopoulos,T.D., Spyropoulos,I.C. and Hamodrakas,S.J. (2004) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res.*, **32**, W400–W404.
7. Berven,F.S., Flikka,K., Jensen,H.B. and Eidhammer,I. (2004) BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.*, **32**, W394–W399.

8. 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.
9. Liu,Q., Zhu,Y., Wang,B. and Li,Y. (2003) Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput. Biol. Chem*., **27**, 355–361.
10. 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.
11. 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.
12. Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol*., **238**, 54–61.
13. Chou,K.C. and Elrod,D.W. (1999) Prediction of membrane protein types and subcellular locations. *Proteins*, **34**, 137–153.
14. Cedano,J., Aloy,P., Perez-Pons,J.A. and Querol,E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol*., **266**, 594–600.
15. Zhang,C.T. and Chou,K.C. (1992) An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci*., **1**, 401–408.
16. Noguchi,T., Matsuda,H. and Akiyama,Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res*., **29**, 219–220.
17. Moller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterised integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.
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. Busch,W. and Saier,M.H. (2002) The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol*., **37**, 287–337.
20. Koronakis,V., Sharff,A., Koronakis,E., Luisi,B. and Hughes,C. (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature*, **405**, 914–919.
21. Song,L., Hobaugh,M.R., Shustak,C., Cheley,S., Bayley,H. and Gouaux,J.E. (1996) Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science*, **274**, 1859–1866.
22. Faller,M., Niederweis,M. and Schulz,G.E. (2004) The structure of a mycobacterial outer-membrane channel. *Science*, **303**, 1189–1192.
23. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
24. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
25. Pautsch,A. and Schulz,G.E. (2000) High-resolution structure of the OmpA membrane domain. *J. Mol. Biol*., **298**, 273–282.
26. Oomen,C.J., Van Ulsen,P., Van Gelder,P., Feijen,M., Tommassen,J. and Gros,P. (2004) Structure of the translocator domain of a bacterial autotransporter. *EMBO J*., **23**, 1257–1266.
27. Ye,J. and Van Den Berg,B. (2004) Crystal structure of the bacterial nucleoside transporter Tsx. *EMBO J*., **23**, 3187–3195.
28. van den Berg,B., Black,P.N., Clemons,W.M.,Jr and Rapoport,T.A. (2004) Crystal structure of the long-chain fatty acid transporter FadL. *Science*, **304**, 1506–1509.
29. Chimento,D.P., Mohanty,A.K., Kadner,R.J. and Wiener,M.C. (2003) Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. *Nature Struct. Biol*., **10**, 394–401.
30. Everett,K.D. and Hatch,T.P. (1995) Architecture of the cell envelope of *Chlamydia psittaci* 6BC. *J. Bacteriol*., **177**, 877–882.