ISSN 0006-3509, Biophysics, 2008, Vol. 53, No. 2, pp. 134–139. © Pleiades Publishing, Inc., 2008.

MOLECULAR BIOPHYSICS

Benchmarking of Programs that Predict the Position of Transmembrane Segments in Beta-Barrel Proteins

N. S. Sadovskaya[a] and M. S. Gelfand[b, c]

[a] *State Research Center GosNIIGenetika, Moscow, 113545 Russia*
[b] *Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992 Russia*
[c] *Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia*
e-mail: nathalie@iitp.ru

Received December 24, 2007

Abstract—We propose a method for comparative analysis of the programs that locate transmembrane segments in proteins. On this basis, we have compiled a sample of beta-barrel protein that allows unequivocal assessment of prediction quality. Upon testing of several Internet servers, B2TMR proves the best, with B2TMPRED, HMM-B2TMR, and PROFtmb following.

Key words: beta-barrel proteins, testing sample, homologous proteins, transmembrane segment prediction, benchmarking of Internet servers

DOI: 10.1134/S0006350908020036

INTRODUCTION

Beta-barrel (BB) proteins have been found in the outer membranes of bacteria, mitochondria, and chloroplasts [1]. They play an important part in cell activities, providing for metabolite transport [2], operating as enzymes [3] and receptors [4], and performing defense functions [5].

To date, the 3D structures are known for a limited number of BB proteins. In December 2007, the PDB_TM database contained [6] 126 such entries; less then half would remain upon removing homologs. The situation is not surprising, as these proteins are present in small amounts insufficient for detailed analysis. On the other hand, their production by gene engineering encounters difficulties in attaining high expression [7] as well as in crystallization [8, 9]. Therefore, quite important are *in silico* studies as a starting point for laboratory research.

The BB protein structure is an antiparallel β-sheet closed into a cylindrical shape (barrel). Most often it is composed of an even number of segments inclined about 45° relative to the barrel axis, which corresponds to the overall turn of the β-sheet. Only one of the two possible structures is realized in proteins, the mirror structure being unfavorable in energy [1]. The properties of these proteins depend on the number of segments and the shear number [11] which correlates with the segment inclination. For a protein of n segments, the shear number is about $n + 2$. The number of segments in BB proteins mostly ranges from 8 (NspA [12], OmpA [13, 14], OmpX [5]) to 22 (FepA [15], FhuA [16, 17]), the shear number ranges from 8 to 24.

In the periplasmic part of the barrel, the transmembrane (TM) segments are connected by short loops (several residues), whereas in the outr part the loops are long. The termini are usually periplasmic. The barrel surface in contact with the nonpolar membrane is built of the side chains of aliphatic amino acids, making a nonpolar ribbon ~22 Å wide. The edges are formed by the side groups of weakly polar aromatic residues on the membrane surface. It should be noted that BBs often form oligomers where each monomer is an individual polypeptide.

Notwithstanding, prediction of TM segments in BBs from their amino acid sequence remains a difficult task because the segments are quite short; seven residues suffice for spanning the membrane [1]. To add, the motif of alternating polar residues inside the barrel and nonpolar ones outside is broken by numerous nonpolar residues, especially in porins.

In recent years, a number of programs predicting the position of TM segments in beta-barrels have been designed and implemented as Internet servers; these are listed in Table 1.

As one attempts comparative statistical testing, it turns out that there is not enough experimental data on the secondary structure of TM proteins. The available data most probably have already been used in building the servers. Hence, it is not clear what data can be used for independent benchmarking.

This work was aimed to analyze the Internet servers for prediction quality using a specially created unique sample and the new method of testing based on the similarity of predictions for homologous proteins.

Table 1. Servers predicting TM segment position in proteins; those assessed here are in boldface

ConBBPRED	[22]	http://bioinformatics.biol.uoa.gr/ConBBPRED
B2TMPRED	[23]	http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi
B2TMR	[23]	http://gpcr.biocomp.unibo.it/predictors/
HMM-B2TMR	[24]	http://gpcr.biocomp.unibo.it/predictors/
PRED-TMBB (N-best method)	[25], [26]	http://bioinformatics.biol.uoa.gr/PRED-TMBB
PRED-TMBB (Posterior decoding method)	[25], [26]	http://bioinformatics.biol.uoa.gr/PRED-TMBB
PRED-TMBB (Viterbi method)	[25], [26]	http://bioinformatics.biol.uoa.gr/PRED-TMBB
PROFtmb	[27]	http://rostlab.org/cgi-bin/var/bigelow/proftmb/query
Tbbpred	[28]	http://www.imtech.res.in/raghava/tbbpred/
TMBETA-NET	[29], [30]	http://psfs.cbrc.jp/tmbeta-net/

METHODS AND DATA

The study was based on the following evolutionary premises:

—a group of closely related proteins should retain the overall structure,

—the position of TM segments should also be conserved in such proteins,

—in a pair of aligned closely related proteins the TM segments should be quite accurately projected upon each other.

The assessment procedure consisted of three parts:

(1) choice of evaluation criteria based on evolutionary considerations;

(2) compilation of a unique sample not used theretofore in building or testing any server;

(3) evaluation of the servers to determine the most reliable one.

The two indices chosen to evaluate the prediction quality were the Jaccard coefficient (Q) and the segment overlap coefficient (C) [18].

The predictions were compared residue by residue with the Q score. For each pair of aligned proteins, it is determined as the size of overlap of the segments divided by the size of their aggregate. That is, let S be the number of aligned residues predicted to enter TM segments in both proteins, U be the number of residues found in at least one of the segments; then

$$Q = S/U.$$

The C value was determined as the share of TM segments present in the compared pair of proteins. Let n_1 and n_2 be the number of segments predicted for the respective protein, and $i = 1, \ldots, n_1, j = 1, \ldots, n_2$ be the ordinal number of the segment in the protein. Consider all pairs of segments ij for which the projections overlap by at least one residue. Introducing V_{ij} to represent partial overlap of segment i with segment j, take $V_{ij} = 1$ if at least half of i overlaps j, and $V_{ij} = 0$ otherwise. That is, if L_i is segment i length, M_j is segment j length, and K_{ij} is the extent of i–j overlap, then

$$V_{ij} = 1 \quad \text{if} \quad K_{ij}/L_i \geq 0.5,$$
$$V_{ij} = 0 \quad \text{if} \quad K_{ij}/L_i < 0.5.$$

In the same way, the local overlap of j with segment i is

$$W_{ji} = 1 \quad \text{if} \quad K_{ij}/M_j \geq 0.5,$$
$$W_{ji} = 0 \quad \text{if} \quad K_{ij}/M_j < 0.5.$$

Now C for a protein pair is calculated as the sum of local overlaps for all TM segment pairs divided by the total number of predicted segments:

$$C = \sum_{ij} (V_{ij} + W_{ji})/(n_1 + n_2).$$

If the predictions for two related proteins are similar, the Q and C values must be close to unity. If the TM segments predicted for the two aligned proteins overlap at least by half, so that the number of segments is retained but the positions of edge residues differ, then $Q < 1$ and $C = 1$. On the other hand, if a TM segment in one protein is predicted as two close segments (interrupted) while in the other protein the corresponding segment is predicted to be whole, then $Q \approx 1$ and $C < 1$.

In this way, the two indices describe two different aspects of similarity in TM segment prediction for related sequences.

To make a testing sample of BBs, we started with TCDB http://www.tcdb.org/ [19], taking one representative of bacterial transporters from the TC.1B class for each family. Initially we considered only *E. coli* proteins. For the chosen proteins, we found clusters of orthologous genes (COG) http://www.ncbi.nlm.nih.gov/COG/ [20] from Gram-negative bacteria. Proteins shorter than 80% of the protein used to find the cluster were rejected. The resulting sample comprised 5673 pairs of proteins from 14 COGs.

Comparisons were made for all pairs of proteins from the same class. Sequences were aligned with

Table 2. Server performance ranked as arithmetic means of Q and C with standard deviations σ for different similarity intervals *ID*. Italics marks PROFtmb and HMM-B2TMR upon removal of the proteins they classed as non-TM

ID	0–50%		51–100%		All	
	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$
B2TMR	0.67 ± 0.15	0.84 ± 0.11	0.85 ± 0.18	0.93 ± 0.11	0.68 ± 0.15	0.84 ± 0.11
PROFtmb	*0.64 ± 0.15*	*0.82 ± 0.13*	*0.83 ± 0.19*	*0.92 ± 0.13*	*0.64 ± 0.16*	*0.82 ± 0.13*
HMM-B2TMR	*0.64 ± 0.16*	*0.82 ± 0.13*	*0.83 ± 0.19*	*0.93 ± 0.12*	*0.64 ± 0.16*	*0.83 ± 0.13*
B2TMPRED	0.49 ± 0.15	0.68 ± 0.14	0.79 ± 0.23	0.88 ± 0.17	0.50 ± 0.16	0.68 ± 0.14
HMM-B2TMR	0.55 ± 0.26	0.71 ± 0.31	0.70 ± 0.23	0.78 ± 0.35	0.55 ± 0.27	0.71 ± 0.31
PROFtmb	0.54 ± 0.26	0.70 ± 0.31	0.70 ± 0.35	0.78 ± 0.35	0.55 ± 0.27	0.71 ± 0.31
PRED-TMBB (N-best method)	0.37 ± 0.18	0.56 ± 0.24	0.67 ± 0.27	0.78 ± 0.25	0.38 ± 0.19	0.57 ± 0.25
PRED-TMBB (Viterbi method)	0.37 ± 0.18	0.56 ± 0.24	0.67 ± 0.27	0.78 ± 0.25	0.38 ± 0.19	0.57 ± 0.25
PRED-TMBB (Posterior decoding method)	0.37 ± 0.17	0.56 ± 0.23	0.66 ± 0.28	0.78 ± 0.24	0.37 ± 0.18	0.57 ± 0.24
TMBETA-NET	0.36 ± 0.08	0.54 ± 0.10	0.66 ± 0.24	0.79 ± 0.20	0.37 ± 0.10	0.54 ± 0.11

ClustalW [21]. The positions of TM segments were predicted with the servers specified above, using their default settings. Intersecting and contiguous segments were treated as a single whole segment. In the case when the PRED-TMBB output indicated that a query protein is not a BB, the request was repeated to class it with the BB type.

Then the Q and C values were determined for each server.

RESULTS AND DISCUSSION

Here we examined the performance of the following eight servers: B2TMPRED, B2TMR, HMM-B2TMR, PRED-TMBB (N-best method), PRED-TMBB (Posterior decoding method), PRED-TMBB (Viterbi method), PROFtmb, and TMBETA-NET. Our choice was based on the availability through Internet, the possibility of obtaining multiple predictions, and independence of other servers: the input data should have been the amino acid sequence as such, rather than predictions of other servers.

The arithmetic means of Q and C with standard deviations σ for each server are listed in decreasing order in Table 2. For the four best servers we show diagrams 1–16 plotting the number of proteins N in the given Q and C range.

The most consistent predictions were obtained from B2TMR, followed with a sizable lag by B2TMPRED. With HMM-B2TMR and PROFtmb, the test BB proteins were quite often classed as "non-TM." These servers are ranked medium, though their results were comparable to those of B2TMPRED. When the "non-TM" proteins were removed from the sample, 4997 pairs remained for PROFtmb and 5018 for HMM-B2TMR. Thereupon the mean Q and C improved to become comparable to B2TMR (italicized in Table 2).

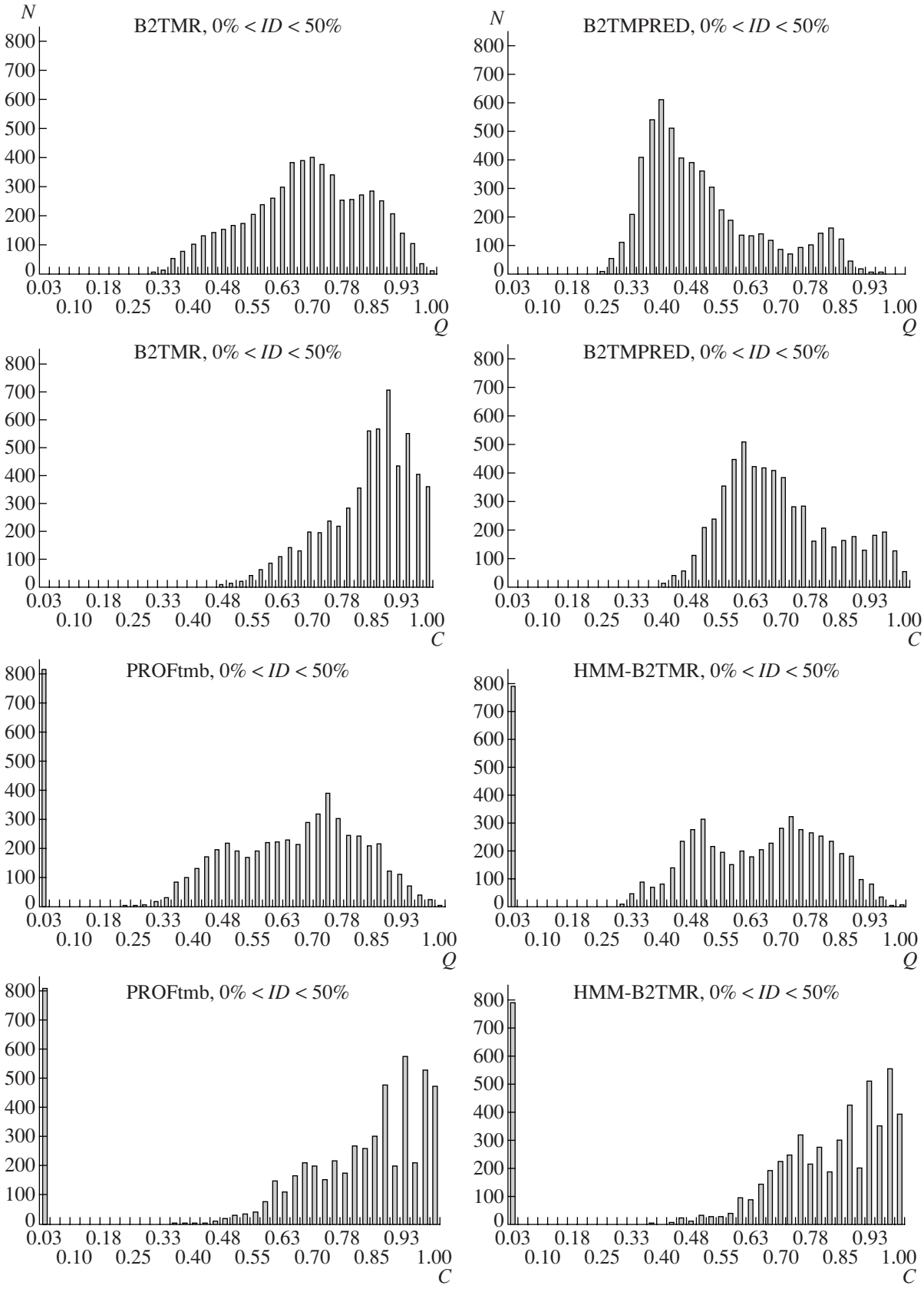
The PRED-TMBB group and TMBETA-NET are at the end of the list. The results within the group (N-best method, Posterior decoding method, Viterbi method) were rather similar; however, PRED-TMBB often related the query proteins to the non-TM class.

Bagos et al. [22] assert that the best predictions are made by HMM-B2TMR, PRED-TMBB, and ProfTMB; somewhat worse are those of B2TMPRED and TMBETA-NET. Note that these authors did not consider B2TMR.

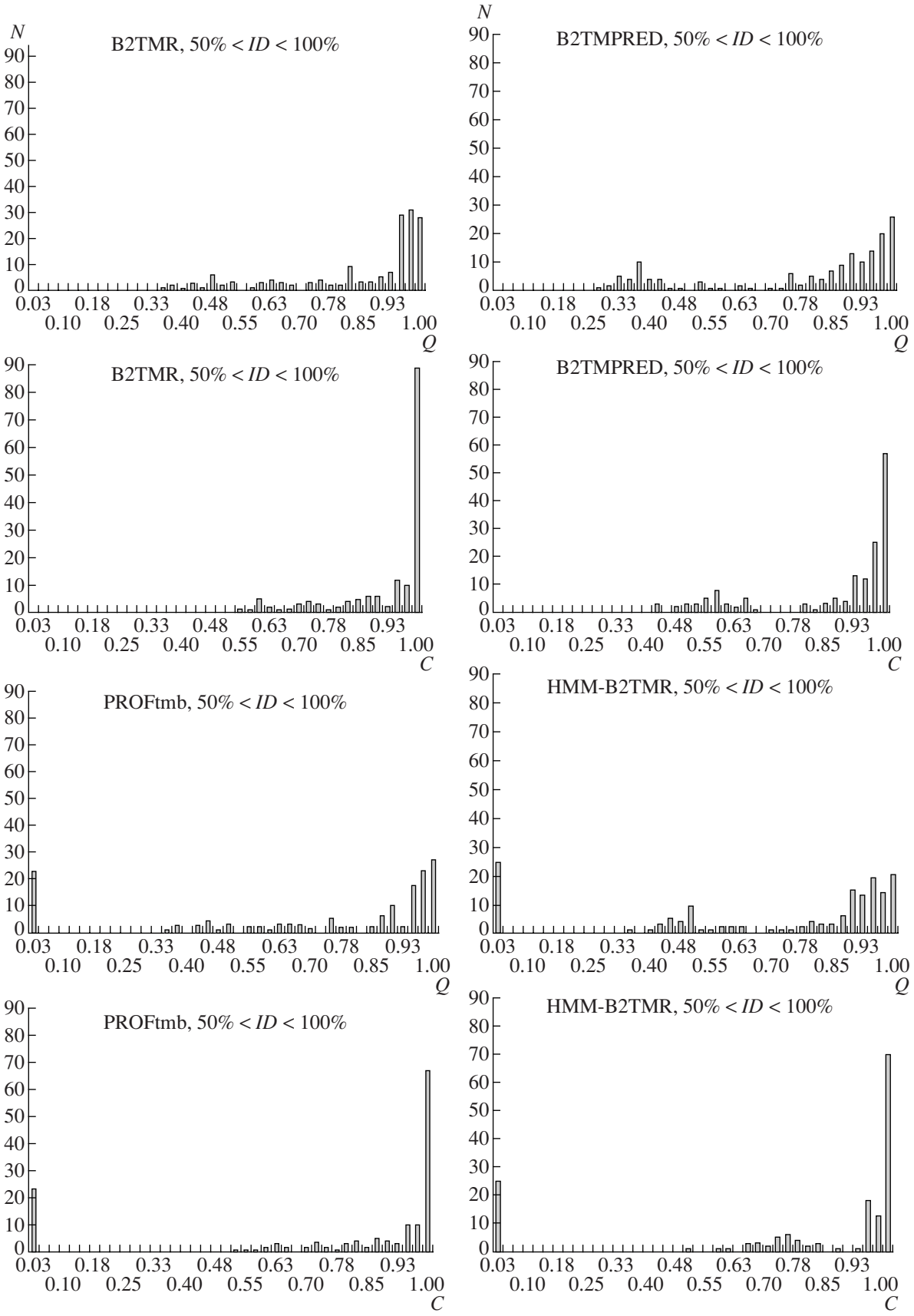
Thus, the results of two independent assessments are quite similar despite the different criteria. The discrepancy concerning B2TMPRED and PRED-TMBB (N-best method, Posterior decoding method, Viterbi method) may be explained by the large difference in the testing sample size: Bagos et al. used only 20 proteins of known structure, whereas our sample comprised 5673 pairs.

Currently, there is a shortage of TM proteins with established membrane marking. Here we propose an evolutionarily based method that can partly make up for this shortage. Thereby one can compile a protein sample that has not been used before. Such samples of orthologous proteins expand the opportunities for testing and improving the servers that predict the positions of TM segments, and can also be used to choose the server most suitable for a particular task.

Our assessments show that the best server for membrane marking in beta-barrel proteins is B2TMR, which gives the most self-consistent predictions; the next best choice are B2TMPRED, HMM-B2TMR, and PROFtmb. When BB proteins are placed in the non-TM class as



Diagrams 1–16. The numbers N of proteins in specified Q and C intervals. Data for the best servers in two series of testing protein pairs with similarities $ID \le 50\%$ and $ID > 50\%$.



Diagrams 1–16. (Contd.)

often done by the latter two servers, another one should be asked.

In important cases it is expedient to get results from several servers; to obtain a higher-quality prediction, it is recommended to examine not one protein but the whole family of homologs.

ACKNOWLEDGMENTS

The work was partly supported by the Molecular and Cell Biology program of RAS and the Russian Foundation for Basic Research (07-04-91555).

REFERENCES

1. G. E. Schulz, Curr. Opin. Struct. Biol. **10**, 443 (2000).
2. T. Schirmer, T. A. Keller, Y. F. Wang, and J. P. Rosenbusch, Science **267**, 512 (1995).
3. R. E. Bishop, Biochim. Biophys. Acta (2007).
4. H. A. Eisenhauer, S. Shames, P. D. Pawelek, and J. W. Coulton, J. Biol. Chem. **280**, 30574 (2005).
5. J. Vogt and G. E. Schulz, Structure **7**, 1301 (1999).
6. G. E. Tusnady, Z. Dosztanyi, and I. Simon, Nucleic Acids Res. **33**, D275 (2005).
7. R. Grisshammer and C. G. Tate, Q. Rev. Biophys. **28**, 315 (1995).
8. Y. Qutub, I. Reviakine, C. Maxwell, et al., J. Mol. Biol. **343**, 1243 (2004).
9. H. Zhang and W. A. Cramer, J. Struct Funct Genomics **6**, 219 (2005).
10. G. E. Schulz, Biochim. Biophys. Acta **1565**, 308 (2002).
11. W. M. Liu, J. Mol. Biol. **275**, 541 (1998).
12. L. Vandeputte-Rutten, M. P. Bos, J. Tommassen, and P. Gros, J. Biol. Chem. **278**, 24825 (2003).
13. A. Arora, F. Abildgaard, J. H. Bushweller, and L. K. Tamm, Nat. Struct. Biol. **8**, 334 (2001).
14. A. Pautsch and G. E. Schulz, J. Mol. Biol. **298**, 273 (2000).
15. S. K. Buchanan, B. S. Smith, L. Venkatramani, et al., Nat. Struct. Biol. **6**, 56(1999).
16. A. D. Ferguson, E. Hofmann, J. W. Coulton, et al., Science **282**, 2215 (1998).
17. K. P. Locher, B. Rees, R. Koebnik, et al., Cell **95**, 771 (1998).
18. N. S. Sadovskaya, R. A. Sutormin, and M. S. Gelfand, J. Bioinform. Comput. Biol. **4**, 1033 (2006).
19. W. Busch and M. H. Saier, Jr., Mol. Biotechnol. **27**, 253 (2004).
20. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, et al., Nucleic Acids Res. **29**, 22 (2001).
21. J. D. Thompson, D. G. Higgins, and T. J. Gibson, Nucleic Acids Res. **22**, 4673 (1994).
22. P. G. Bagos, T. D. Liakopoulos, and S. J. Hamodrakas, BMC Bioinformatics **6**, 7 (2005).
23. I. Jacoboni, P. L. Martelli, P. Fariselli, et al., Protein Sci. **10**, 779 (2001).
24. P. L. Martelli, P. Fariselli, A. Krogh, and R. Casadio, Bioinformatics **18**, S46 (2002).
25. P. G . Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, Nucleic Acids Res. **32**, W400 (2004).
26. P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, BMC Bioinformatics **5**, 29 (2004).
27. H. R. Bigelow, D. S. Petrey, J. Liu, et al., Nucleic Acids Res. **32**, 2566 (2004).
28. N. K. Natt, H. Kaur, and G. P. Raghava, Proteins **56**, 11 (2004).
29. M. M. Gromiha, S. Ahmad, and M. Suwa, J. Comput. Chem. **25**, 762 (2004).
30. M. M. Gromiha and M. Suwa, Bioinformatics **21**, 961 (2005).