SHORT COMMUNICATION

Prediction of Buried Helices in Multispan Alpha helical Membrane Proteins

Larisa Adamian and Jie Liang*
Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois

ABSTRACT **Analysis of a database of structures of membrane proteins shows that membrane proteins composed of 10 or more transmembrane (TM) helices often contain buried helices that are inaccessible to phospholipids. We introduce a method for identifying TM helices that are least phospholipid accessible and for prediction of fully buried TM helices in membrane proteins from sequence information alone. Our method is based on the calculation of residue lipophilicity and evolutionary conservation. Given that the number of buried helices in a membrane protein is known, our method achieves an accuracy of 78% and a Matthew's correlation coefficient of 0.68. A server for this tool (RANTS) is available online at http://gila.bioengr.uic.edu/lab/. Proteins 2006;63:1–5.**

Key words: membrane protein; transmembrane helix; entropy; lipophilicity; TMLIP

INTRODUCTION

Integral membrane proteins (MP) fulfill important cellular roles and constitute up to 25–30% of a typical genome.[1,2] However, there is only a handful of structures of membrane proteins available at the present time because of the experimental difficulties in obtaining high-quality crystals. Accurate modeling of MP structures is an important task that can help to fill the gap between sequence, structure, and function of the membrane proteins. Prediction of helix orientation is an important first step in the modeling of the structures of multispan membrane proteins.[3,4] There are several highly effective methods to predict the orientation of helices (i.e., to identify the regions on a helix that face phospholipids).[5] However, membrane proteins with 10 or more transmembrane (TM) helices often contain helices that are completely buried within the helical bundle and are not accessible to lipids. Global topology analysis of the *Escherichia coli* inner membrane proteome estimates that there are 20–25% of membrane proteins with 10+ TM helices, which are often involved in transport of small molecules across the membrane.[6] A significant number of these proteins are likely to contain buried helices. Thus, modeling of the larger membrane proteins would require a method to identify the fully buried helices. To the best of our knowledge, no methods currently exist that can predict buried helices in integral α-helical membrane proteins. In this study, we present a method that uses sequence information alone to rank transmembrane helices by their lipid accessibility. Given that the number of buried helices in a membrane protein is known, our method can identify these buried helices with an accuracy of 78% and a Matthew's correlation coefficient (MCC) of 0.68. A server for this tool (RANTS) is available online at http://gila.bioengr.uic.edu/lab/.

MATERIALS AND METHODS

Empirical Burial Function to Identify the Most Buried Helices

We develop an empirical helix burial function f based on a few assumptions. First, the residues in the buried helices are more conserved because of structural and functional constraints. Second, the residue composition of the buried helices is different from the composition of helices facing the lipid environment. Finally, the difference between minimal and maximal values of conservation entropy for every position in the multiple sequence alignment of the TM helix should be smaller in buried helices than in lipid-exposed helices because of the homogeneity of the environment. We design the helix burial function f as the product of the average entropy $\bar{s}$ of all residue positions of the TM helix,[7] the average lipophilicity $\bar{l}$ as calculated using lipophilicity scales TMLIP2,[8] and the slope k of the sorted entropy values of all residue positions in a helix of length d for helices $1 \ldots n$ of the membrane protein

$$f = k \cdot \bar{s} \cdot \bar{l}$$

Grant sponsor: National Science Foundation; grant number: CAREER DBI0133856; Grant sponsor: National Institutes of Health; grant number: GM68958; Grant sponsor: Office of Naval Research; grant number: N000140310329; Grant sponsor: Whitaker Foundation; grant number: TF-04-0023

*Correspondence to: Jie Liang, Department of Bioengineering, M/C 53, University of Illinois at Chicago, 835 South Wolcott, Room 103, Chicago, IL 60612-7340. E-mail: jliang@uic.edu

Received 30 June 2005; Revised 1 November 2005; Accepted 4 November 2005

Published online 17 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20874

where $\bar{s} = (s_1 + \ldots + s_d)/d$ and $\bar{l} = (l_1 + \ldots + l_d)/d$. Here, the most buried helices are the ones with the minimal value of burial function *f*.

The *f* values of the buried as well as the exposed helices may differ significantly from protein to protein. An important reason is that there are usually different numbers of sequences in the multiple sequence alignments for two different proteins, which reflects the different evolutionary history of the two genes. This affects the values of the average entropy $\bar{s}$ of the TM helices. For multisubunit membrane proteins, the *f* values will only be comparable for the TM helices within the same subunit. To compare *f* values for TM helices from different subunits, one needs to build multiple sequence alignments using exactly the same set of sequences from the same set of species for all subunits, which is not always feasible. In addition, the stability of different membrane proteins in the lipid environment as reflected by $\bar{l}$ may also be different. For example, a membrane protein from a thermophilic archaea may have very different stability in lipid environment than a membrane protein from a eukaryote.

To account for the ambiguity in the definition of the ends of TM helices, we calculate the *f* values for variant sequences of the same helix by changing the putative positions of both N- and C-termini up and down within a range of 5 residues. We generate a total of $5 \times 5 = 25$ candidate sequences for each of the *n* TM helices in the helical bundle. We then randomly sample $100 \cdot n$ combinations of the candidate helical sequences from the pool of 25^n possible combinations and rank helices by their *f* values for every combination. We calculate the expected ranking $\bar{r}_i$ for every helix *i* as

$$\bar{r}_i = \frac{\sum_{r_i=1}^{n} c(r_i) \cdot r_i}{100 \cdot n},$$

where $c(r_i)$ is the number count of termini combinations where helix *i* had ranking r_i. For example, helix 1 in a helical bundle of $n = 10$ helices had rank $r_1 = 3$ in 800 combinations and rank $r_1 = 4$ in 200 combinations of helices. Then the expected rank $\bar{r}_1$ for helix 1 would be

$$\bar{r}_1 = \frac{800 \cdot 3 + 200 \cdot 2}{1000} = 2.8.$$

An important factor for our prediction is the quality of the multiple sequence alignment, which is crucial for the computation of the average entropy and average lipophilicity. We manually inspect all sequences obtained from BLAST searches and include only those that represent the same protein (with identical annotation) and are at least 35–40% identical to the query sequence. We manually close all gaps found in the MSA of TM segments using the Pfaat[9] program.

We compare our results with calculated lipid accessibility of the residues in the TM helix. The lipid-accessible surface for the whole TM region is calculated with a 1.9 Å probe using the VOLBL algorithm (http://www.cs.ust.hk/faculty/edels/alpha3d.html) as described previously.[8] The fraction $f(SA)$ of the total lipid-accessible surface of the TM region was then calculated for every TM helix.

RESULTS AND DISCUSSION

Ranking of Transmembrane Helices by Helix Burial Function *F* and Robustness Coefficient α

We have tested our method using six different polytopic membrane proteins (1EUL, 1IWG, 1KPL, 1OCR, 1Q90, 1RH5), all of which contain completely or partially buried helices. Figure 1 shows the sorted expected rankings $\bar{r}_i$ by the helix burial function *f* and the corresponding fraction of lipid-accessible surface $f(SA)$ for TM helices from these proteins. The best prediction results were obtained for some of these proteins: Ca^{2+}-ATPase (1EUL), multidrug efflux transporter AcrB (1IWG), b_6 subunit of cytochrome b_6f complex (1Q90), and for the protein-conducting channel (1RH5).

In Ca^{2+}-ATPase (1EUL),[10] helices M4, M5, M6, and M8 form the internal layer of helices in the helical bundle [Fig. 2(a)]. They are the least accessible to the 1.9 Å probe, have the lowest values of *f,* and are consistently the four lowest ranked helices as shown on Figure 1(a). In bacterial multidrug efflux transporter AcrB (1IWG),[11] helices TM4 and TM10 [Fig. 2(b)] are the most buried and consistently have the lowest *f* values as shown on Figure 1(b). Our method also correctly identifies the buried helix B of cytochtome b_6 subunit of cytochrome b_6f complex from *Chlamydomonas reinhardtii* (1Q90)[12] as shown on Figure 1(c).

There are no completely buried helices in the protein conducting channel (1RH5),[13] but the *f* values for helices from α-subunit are in a good agreement with the lipid accessibility of the helices [Fig. 1(d)], with only one incorrectly ranked helix (helix 7).

The ranking of helices in subunit I of cytochrome c oxidase (1OCR)[14] is presented in Figure 1(e). Here, buried helices 2, 6, and 10 are correctly predicted. Helix 8, which has a lipid-facing surface and interacts with TM helices 1 and 2 from subunit II, is also predicted to be buried. The latest structure of cytochrome c oxidase at 1.8 Å resolution (1V54)[15] shows a cardiolipin molecule that is tightly bound to helix 8 of subunit I and helix 2 of subunit II. Robinson et al[16] showed that two molecules of tightly bound cardiolipin are required for full functional activity of each monomer of detergent-solubilized bovine heart cytochrome c oxidase. It is clear that cardiolipin serves as one of the cofactors for cytochrome c oxidase. This interaction requires conservation of amino acid residues in the binding site to ensure a stable and specific association between a protein and a phospholipid. Figure 2(c) shows the subunits I (blue atoms) and II (green atoms) of cytochrome c oxidase and a bound cardiolipin molecule (yellow atoms). Residues from helix 8 are shown in magenta. This figure demonstrates that interacting acyl chains of cardiolipin cover almost all lipid accessible surface of helix 8. Thus, helix 8 is structurally constrained due to the packing interactions with neighboring helices from subunits I, II, and a cardiolipin molecule. If these considerations are taken into account,

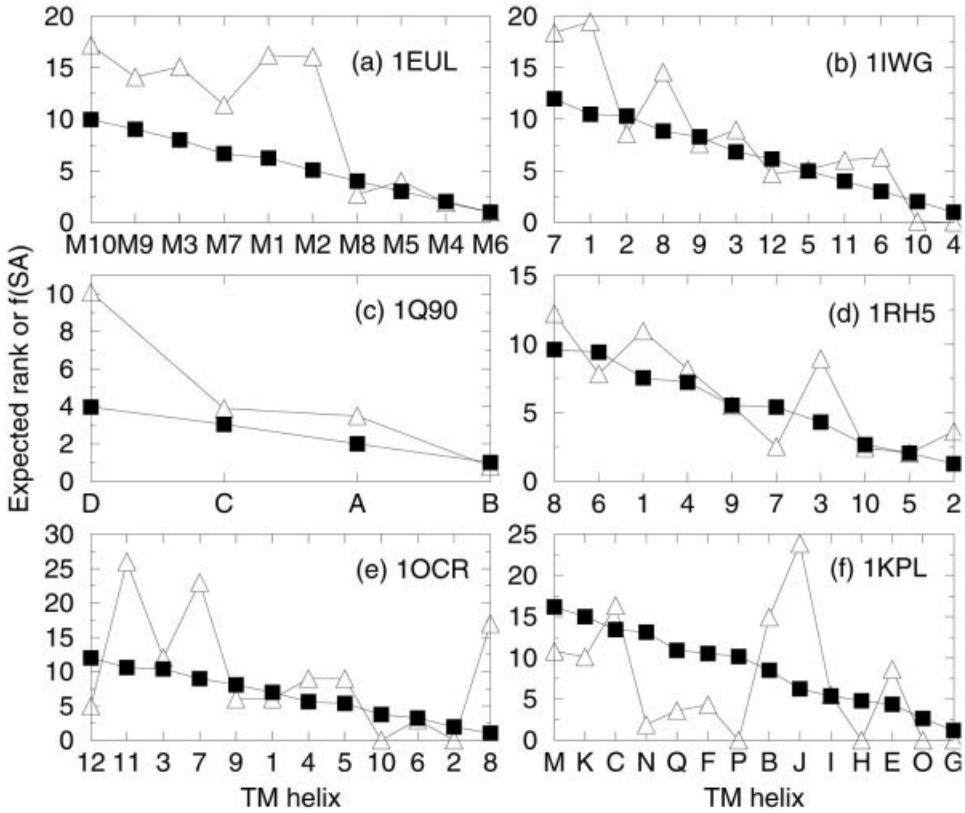

Fig. 1. Comparison of expected ranking $\bar{r}_i$ by helix burial function (■) and the fraction of lipid-accessible surface area (*f*(SA), in symbol △) for every TM helix. *f*(SA) for a helix is the percentage of its lipid-accessible surface area for the whole TM region of the protein. (**a**) Transmembrane helices of Ca^{2+}-ATPase (1EUL). TM helices are denoted as in Ref. [10]. (**b**) TM helices of multidrug efflux transporter AcrB (1IWG).[11] (**c**) TM helices of cytochrome b_6 subunit from cytochrome b_6f complex 1Q90.[12] (**d**) TM helices of α-subunit of a protein-conducting channel (1RH5).[13] (**e**) TM helices of subunit I of cytochrome c oxidase (1OCR).[14] (**f**) Transmembrane helices of ClC chloride channel (1KPL).[17] Overall, among the 19 buried helices and 43 lipid-exposed helices in the set of 6 multispan helical membrane proteins, 15 buried helices are predicted correctly (78%), and 39 exposed helices are predicted correctly (90%).

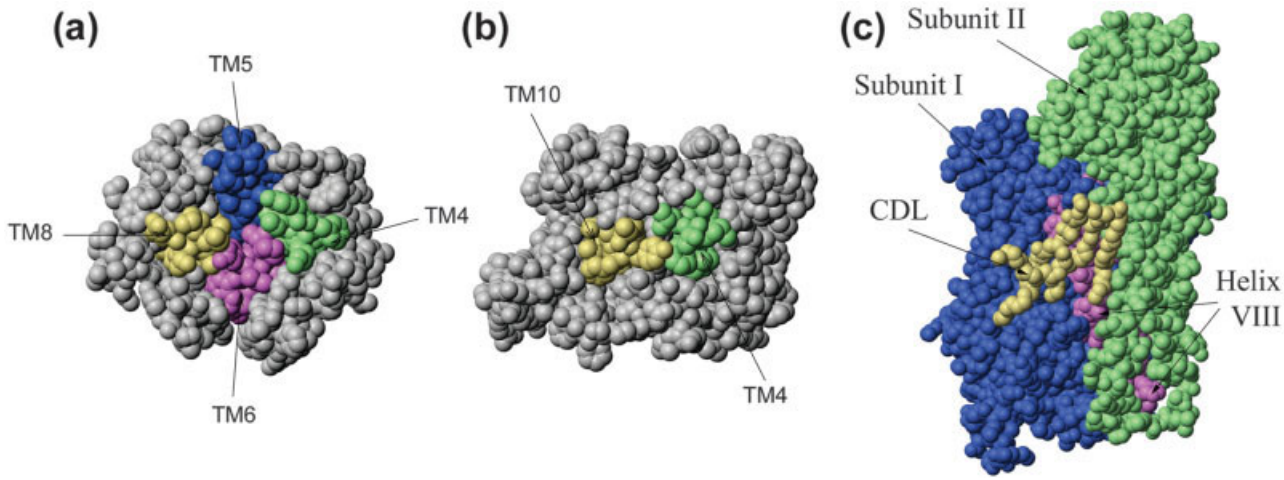

Fig. 2. Examples of buried TM helices. (**a**) Top view of the TM region of Ca^{2+}-transporting ATPase (1EUL). Buried helices TM4, TM5, TM6, and TM8 are shown in color. (**b**) Top view of the TM region of multidrug efflux transporter (1IWG). Buried helices TM4 and TM10 are shown in color. (**c**) Side view of subunits I (blue) and II (green) of *Bos taurus* cytochrome c oxidase (PDB: 1V54) with bound cardiolipin (CDL) molecule (yellow). Helix 8 (magenta) from subunit I is almost completely shielded from phospholipids by CDL and subunit II. It is predicted as buried.

the fact that helix 8 is not lipid-exposed is correctly predicted.

The ClC chloride channel (1KPL)[17] has a complex topology for its TM helices, where helices G, N, and O are significantly buried within a TM bundle. In addition, helices H and P are buried at the oligomerization interface of the functionally important dimer. Figure 1(f) shows that helices G, H, and O are correctly predicted as buried. However, the ranking of helices N and P places them among the helices with much higher lipid-accessibility [Fig. 1(f)].

To summarize, among the 19 buried helices and 43 lipid-exposed helices in the set of 6 multispan helical membrane proteins, 15 buried helices are predicted correctly (78%), and 39 exposed helices are predicted correctly (90%).

Comparison of Ranking Results for Computationally Predicted versus Structure-based Transmembrane Helices

We compare the ranking results of lipid exposure of predicted TM helices and helices obtained from the X-ray structure. Our goal is to estimate the applicability of the proposed method to membrane proteins for which only sequence data are available. We chose a structure of Leu transporter $LeuT_{Aa}$ from *Aquifex aeolicus* (PDB ID: 2A65),[18] which is a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. The hidden Markov model topology predictor TMHMM[19] predicted 12 transmembrane helices, which had a good overlap with TM helices from the structure. Unfortunately, a BLAST search of nonredundant database of protein sequences found no additional sequences annotated as Na^+-dependent Leu transporters. The majority of the retrieved homologues of $LeuT_{Aa}$ were annotated as "Na^+-dependent transporters of the SNF family," for which the exact transported substrates were unknown. These sequences were used to build a multiple sequence alignment for predicted and structure-based TM helices. The f values calculated for predicted and structure-derived TM helices are very similar, with a correlation coefficient R = 0.94. The results of ranking of the burial functions f are shown on Figure 3. There are three significantly buried helices in $LeuT_{Aa}$: helices 1, 6, and 8. Helices 1, 2, and 6 are consistently given the lowest f values for both predicted and structure-based sets of transmembrane helices. Of these, helices 1 and 6 are true positives, whereas helix 2 is a false positive. Buried helix 8 was ranked among the lipid-exposed helices and is a false negative. Buried helices 1 and 6 comprise the sodium-binding site and are the mostly conserved helices in all Na^+-dependent transporters. Buried helix 8 forms a substrate-binding site together with helix 3. The failure to correctly predict the solvent accessibility for helix 8 is likely due to the quality of the multiple sequence alignment, which contained sequences of different transporters.

Performance of the Method

The overall accuracy of the prediction is difficult to estimate without a priori knowledge of the total number of buried helices. Unfortunately, the number of structures of large multispan membrane proteins is still too small to develop a statistically sound relationship between the total number of TM helices and the number of buried helices. However, if the number of buried helices in the membrane protein is known a priori (e.g., 4 in 1EUL, 2 in 1IWG, 5 in 1KPL, 3 in 1OCR (Subunit I), 1 in 1Q90 (cytochrome b_6 subunit), 4 in 1RH5 (α-subunit), and 3 in 2A65), then the performance of our prediction, which produces over- and under-predictions, can be measured using the Matthews correlation coefficient (MCC)[20,21] as follows

$$MCC = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}.$$

Here, p is the number of correctly predicted buried helices (17 for this set of 7 multispan membrane proteins), n is the number of correctly predicted lipid-accessible helices (47), o is the number of incorrectly predicted buried helices (5), and u is the number of incorrectly predicted lipid-accessible helices (5). The Matthew's correlation coefficient ranges from $-1 \leq MCC \leq 1$, where $MCC = 1$ indicates the best possible prediction. Our method of ranking of solvent accessibility of transmembrane helices gives a good Matthew's correlation coefficient of 0.68.

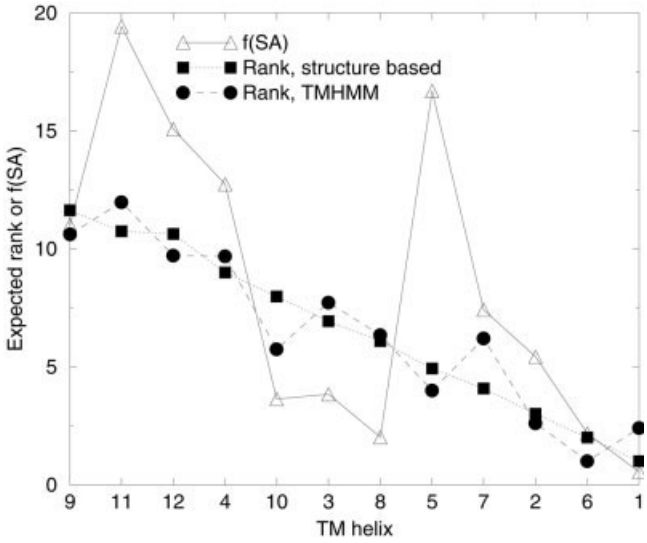

Fig. 3. Ranking results of predicted and structure-derived transmembrane helices from $LeuT_{Aa}$, a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporter. Helices 1 and 6 are correctly predicted for both sets of TM helices, whereas helices 2 and 8 were ranked incorrectly and are false positive and false negative, respectively.

CONCLUSIONS

We describe here for the first time an empirical helix burial function for ranking TM helices by their phospholipid accessibility and for predicting buried helices in polytopic membrane proteins, if the number of such helices is known a priori. Our method can provide informative predictions with accuracy of 78% and a Matthew's correlation coefficient of 0.68.

ACKNOWLEDGMENTS

We thank Dr. E. A. Berry for critical reading of the manuscript. We thank Matthew Dabrowski for technical assistance in the development of the online RANTS server.

REFERENCES

1. Boyd D, Schierle C, Beckwith J. How many membrane proteins are there? Protein Sci 1998;7:201–205.
2. Wallin E, Von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7:1029–1038.
3. Trabanino RJ, Hall SE, Vaidehi N, Floriano WB, Kam VWT, Goddard WA. First principles predictions of the structure and function of G-protein-coupled receptors: validation for bovine rhodopsin. Biophys J 2004;86:1904–1921.
4. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M, Naor Z, Noiman S, Becker OM. PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins 2004;57:51–86.
5. Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics 2004;20(12):1822–1835.
6. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G. Global topology analysis of the Escherichia coli inner membrane proteome. Science 2005;308:1321–1323.
7. Larson SM, Davidson AR. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. Protein Sci 2000;9:2170–2180.
8. Adamian L, Vikas N, DeGrado W, Liang J. Empirical lipid potentials of amino acid residues in multispan alpha helical membrane proteins. Proteins 2005;59(3):496–509.
9. Johnson JM, Mason K, Moallemi C, Xi H, Somarro S, Huang ES. Protein family annotation in a multiple alignment viewer. Bioinformatics 2003;19(4):544–545.
10. Toyoshima C, Nakasako M, Nomura H, Ogawa H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. Nature 2000;405:647–655.
11. Murakami S, R. N, Yamashita E, Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. Nature 2002;419:587–593.
12. Stroebel D, Choquet Y, Popot J-L, Picot D. An atypical haem in the cytochrome b_6f complex. Nature 2003;426:413–418.
13. van den Berg B, Clemons WM, Collinson I, Modis Y, Hartmann E, Harrison SC, Rapoport TA. X-ray structure of a protein-conducting channel. Nature 2004;427:36–44.
14. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi A, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. The whole structure of the 13-subunit oxidized cytochrome coxidase at 2.8 A. Science 1996;272:1136–1144.
15. Tsukihara T, Shimokata K, Katayama Y, Shimada H, Muramoto K, Aoyama H, Mochizuki M, Shinzawa-Itoh K, Yamashita E, Yao M, Ishimura Y, Yoshikawa S. The low-spin heme of cytochrome c oxidase as the driving element of the proton-pumping process. Proc Natl Acad Sci USA 2003;100(26):15304–15309.
16. Robinson NC, Zborowski J, Talbert LH. Cardiolipin-depleted bovine heart cytochrome c oxidase: binding stoichiometry and affinity for cardiolipin derivatives. Biochemistry 1990;29:8962–8969.
17. Dutzler R, Campbell EB, Cadene M, Chait BT, MacKinnon R. X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. Nature 2002;415(6869):287–294.
18. Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. Nature 2005;437:215–223.
19. Krough A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305(3):567–580.
20. Kaur H, Raghava GPS. A neural-network based method for prediction of g-turns in proteins from multiple sequence alignment. Protein Sci 2003;12:923–929.
21. Shepherd AJ, Gorse D, Thornthon JM. Prediction of the location and type of b-turns in proteins using neural networks. Protein Sci 1999;8:1045–1055.