In Silico Biology 3 (2003) 197–204
IOS Press

The Channel in Transporters is Formed by Residues That Are Rare in Transmembrane Helices

Olga V. Kalinina[1,*], Vsevolod J. Makeev[1], Roman A. Sutormin[1], Mikhail S. Gelfand[1,2] and Aleksandra B. Rakhmaninova[2]

[1]*State Scientific Center GosNIIGenetika, Moscow, 113545, Russia*
[2]*Integrated Genomics, P.O. Box 348, Moscow, 117333, Russia*

Edited by H. Michael; received 27 September 2002; revised and accepted 29 November 2002; published 19 December 2002

ABSTRACT: Transmembrane transport is an essential component of the cell life. Many genes encoding known or putative transport proteins are found in bacterial genomes. In most cases their substrate specificity is not experimentally determined and only approximately predicted by comparative genomic analysis. Even less is known about the 3D structure of transporters. Nevertheless, the published experimental data demonstrate that channel-forming residues determine the substrate specificity of secondary transporters and analysis of these residues would provide better understanding of the transport mechanism.
We developed a simple computational method for identification of channel-forming residues in transporter sequences. It is based on the analysis of amino acids frequencies in bacterial secondary transporters. We applied this method to a variety of transmembrane proteins with resolved 3D structure. The predictions are in sufficiently good agreement with the real protein structure.

KEYWORDS: membrane proteins, bacteria, transporters, statistical analysis, functional sites

INTRODUCTION

Transmembrane (TM) transporter proteins are a major mechanism of the flow of compounds in and out the bacterial cell. The membrane transporter systems constitute up to eleven per cent of a prokaryotic proteome, and thus prediction of their substrate specificity not only adds to the genome annotation, but also is of major practical interest [1]. The experimental data, though scarce, indicates that in the case of secondary transporters, the substrate specificity is determined by the general structure of the TM channel [2–4]. This means that identification of channel-forming residues would improve our understanding of the general properties of this structure and hence to the determination of the substrate specificity.

Although only few resolved 3D structures of transporters are known [5], there are many structural models based both on computer predictions of TM-segments and various indirect experimental data [2–4]. However, different prediction algorithms yield contradictory results when applied to the same sequence, and the same algorithm may yield contradictory results when applied to orthologous proteins [6].

* Corresponding author. E-mail: ok81@yandex.ru.

Electronic publication can be found in *In Silico Biol.* **3**, 0017 <http://www.bioinfo.de/isb/2003/03/0017/>, 19 December 2002.

In [7} we introduced the concept of TM-kernels defined as protein fragments consistently predicted to be transmembrane segments.

The aim of this study was to develop a method for identification of channel-forming residues using statistical analysis of TM-kernels.

METHODS

TM-kernels

TM-kernel is defined as a protein segment consistently predicted to be a transmembrane segment, i.e. satisfying two conditions: agreement of several prediction algorithms and consistency of prediction for homologous proteins (for details see [7]). We have analyzed 18908 kernels from 2172 proteins (bacterial secondary transporters, class 2.A according to the Saier–Paulsen classification [1,8]).

Positional correlation for groups of amino acid residues

To reveal the propensity of amino acid residues to lie on the same or on the opposite sides of a TM-helix we calculate positional correlation for groups of amino acid residues.

Let M be the number of TM-kernels in the sample. Let l_k be the number of residues (length) of k-th kernel. Consider two disjoint groups of residues, α and β. Positional correlation for each distance n was calculated as follows. Let $N_n^{\alpha\beta}$ be the number of residue pairs, where the first residue belongs to group α, the second residue belongs to group β and the distance between the residues is $(n-1)$:

$$N_n^{\alpha\beta} = \sum_{k=1}^{M}\sum_{i=1}^{l_i} I^{\alpha}(x_i) I^{\beta}(x_{i+n}), \text{ where } I^{\xi}(x) = \begin{cases} 1, x \in \xi \\ 0, x \notin \xi \end{cases}.$$

Let N_n be the number of all pairs at the distance $(n-1)$. $N_n = \sum_{k=1}^{M}(l_k - n)$.

Finally, let p_ξ be the frequency of residues from group ξ in the sample of TM-kernels:

$$p_\xi = \sum_{k=1}^{M}\sum_{i=1}^{l_i} I^{\alpha}(x_i) \Big/ \sum_{k=1}^{M} l_k .$$

Then *the positional correlation coefficient* in point n is $\mathrm{corr}(n) = \dfrac{N_n^{\alpha\beta} - N_n p_\alpha p_\beta}{\sqrt{p_\alpha(1-p_\alpha)p_\beta(1-p_\beta)} \cdot \sum_{k=1}^{M} l_k}$.

The channel moment of a TM-segment

Two scales of channel propensity are constructed as follows:

$$P^{(1)}{}_a = \log \frac{f_a^{tm}}{f_a^{av}},$$

$$P^{(2)}{}_a = \log \frac{f_a^{tm}}{1/20},$$

where $P^{(v)}{}_a$ is the channel propensity of residue a, f_a^{tm} is the frequency of a in TM-kernels, f_a^{av} is the frequency of a in all proteins.

The channel moment C of a TM-segment is defined analogously to the hydrophobic moment [9]:

$$C = \sum c_i$$

where $c_i = r_i \cdot P^{(v)}{}_a$, r_i is the radius-vector of residue at position i, $P^{(v)}{}_a$ is the channel propensity scale v (v = 1, 2).

Testing

To compare the calculated channel moment with the real orientation of TM-helices, several eubacterial and archaeal alpha-helical TM proteins with resolved 3D structure were used [10–15].

To determine the orientation of the channel vector, we calculated the vector pointing to the most exposed surface side of the helix and assumed that it points to the membrane, that is, out of the channel. That was done using the amino acids solvent accessibility surfaces given in the DSSP database [16] or calculated using the program SPDBV [17]. We considered only proteins which had an inner cavity or a channel and an easily detectable single layer of helices surrounding this cavity: 1FBB (bacteriorhodopsin, *Halobacterium salinarum*) [10], 1E12 (light-driven chloride pump, *Halobacterium salinarum*) [11], 1H68 (sensory rhodopsin II, *Natronomonas pharaonis*) [12], 1FX8 (glycerol-conducting channel, *Escherichia coli*) [13], 1MSL (mechanosensitive ion channel MSCL homolog, chain A, *Mycobacterium tuberculosis*) [14], 1BL8 (KCSA, potassium channel, chain A, *Streptomyces lividans*) [15]. In the latter case the outer helices were removed from the PDB file. Visual control and analysis of positions of functionally important residues showed that this procedure adequately describes the channel. The total number of TM-helices in this study was 32. The test sample did not contain any secondary transporters, as no such structures were available.

RESULTS

Properties of TM-kernel

We have observed that TM-kernels retain the periodic distribution of residues described for complete TM-helices. In particular, Figure 1 demonstrates that aromatic amino acid residues tend to be separated from charged and polar residues by 3–4 positions, which agrees with the period of the alpha-helix. Thus aromatic and charged and polar residues lie at the same side of the helix. We assume that this is the channel side and therefore call all these residues (K, R, H, Q, D, E, N, F, W, Y) the *channel residues*. The common property of these residues is that according to our data [7] their frequency in TM-kernels is significantly lower than in proteins in general. Still, the average number of channel residues per kernel is 2.6 (Figure 2), which is sufficient for determination of the channel side of a helix.

Comparison of the channel propensity scales

Correlation of two channel propensity scales with about 90 various scales of amino acid attributes used

for prediction of TM-helices [18] was computed. As expected, $P^{(1)}$ turned out to be similar (correlation coefficient >0.85) to several scales but there still are some numerical differences. The other scale, $P^{(2)}$, correlates with only one scale (Figure 3a and b). It must be noted that both scales showed relatively weak correlation with the most popular scales, such as the Kyte–Doolittle scale [19] ($P^{(1)}$ and $P^{(2)}$: 0.84), the Eisenberg scale [9] ($P^{(2)}$: 0.79) and kPROT [20] ($P^{(1)}$: 0.46, $P^{(2)}$: 0.48).

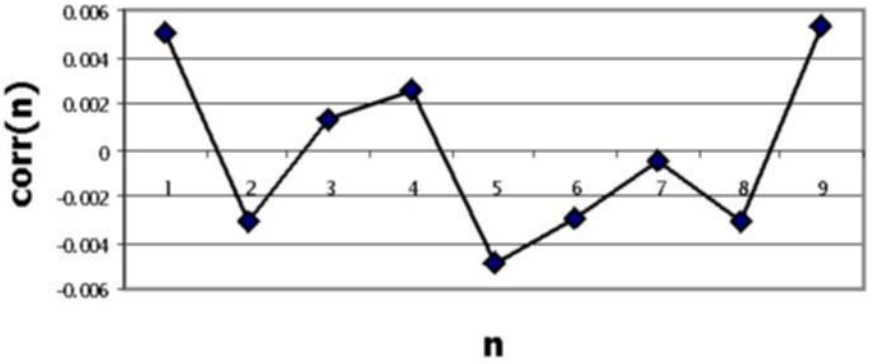

Fig. 1. Positional correlation between two groups of amino acids: charged (K, R, H, Q, D, E, N) and aromatic (F, W, Y) amino acids. Horizontal axis (n): the distance between residues (positions).

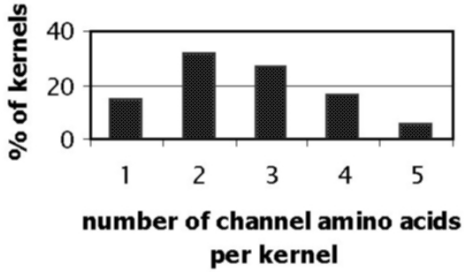

Fig. 2. Distribution of the number of channel amino acid residues in kernels.

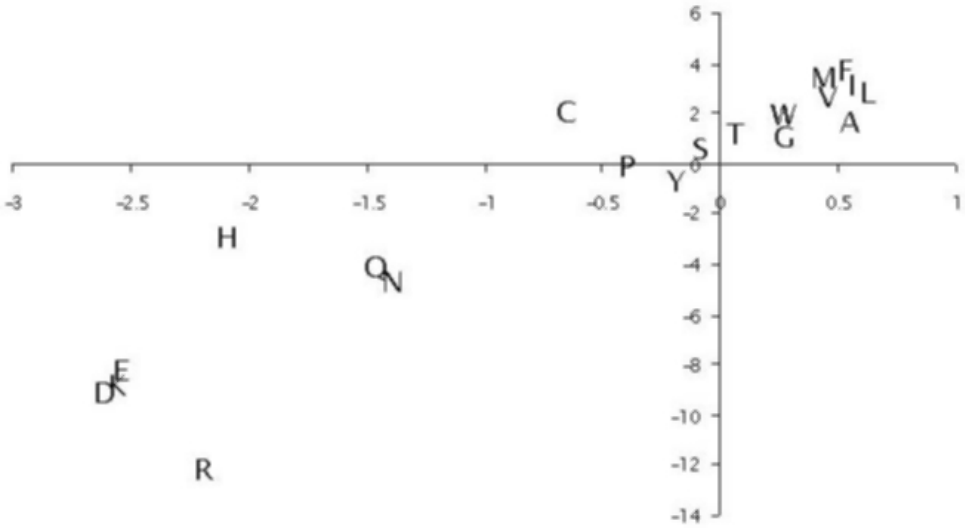

Fig. 3. The published scales having the highest correlation with the channel propensity scales.
Fig. 3a. Correlation of $P^{(1)}$ (horizontal axis) with the Engelman scale [21] (vertical axis). Correlation coefficient = 0.93.

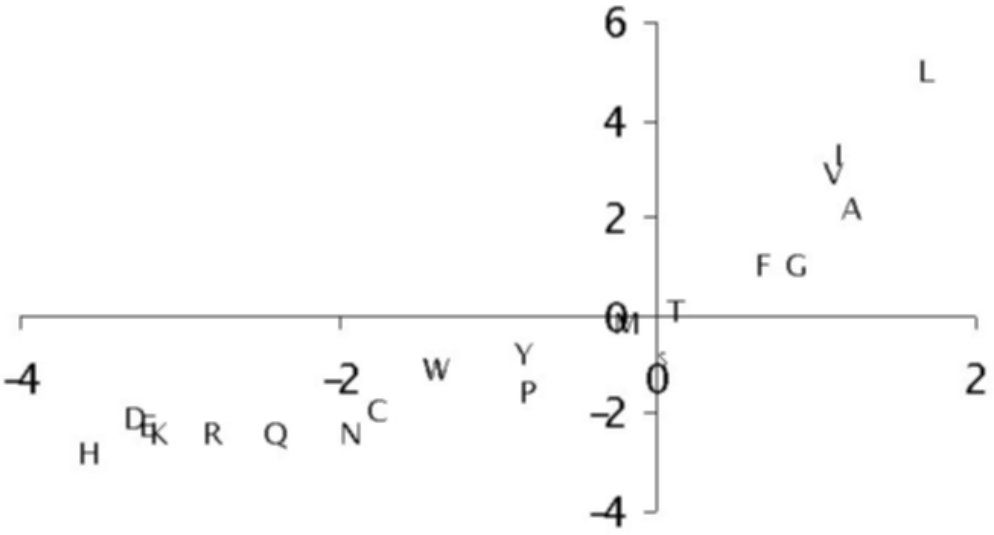

Fig. 3b. Correlation of $P^{(2)}$ (horizontal axis) with the Kuhn–Leigh scale [22] (vertical axis). Correlation coefficient = 0.90.

Evaluation of the prediction quality

The angle differences between the calculated channel moments and the directions of the channel vectors for all 32 studied TM-helices are shown in Figure 4. One can see that the obtained predictions are comparable to the ones obtained using the most popular Kyte–Doolittle scale. In approximately two thirds of all cases the channel side is predicted with a deviation less than 60° from the true direction, whereas in the remaining cases the channel side is predicted badly by all scales. This seems to be caused by the objective limit of accuracy for such predictions. Indeed, some helices contain charged residues that face the membrane, possibly establishing interactions between protein subunits.

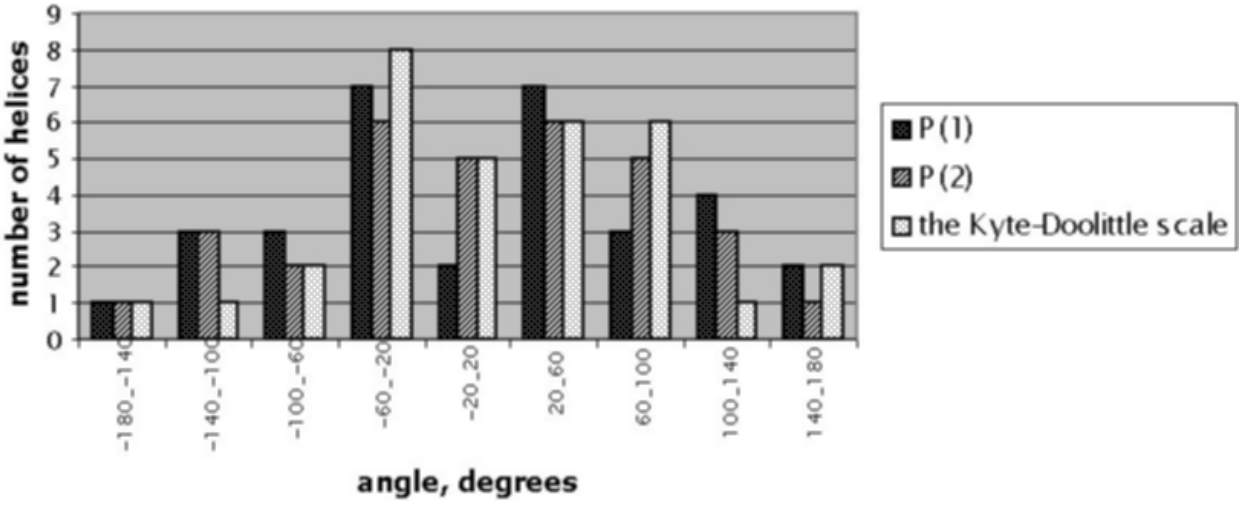

Fig. 4. Comparison of different scales for orientation of TM-helices relative to the channel. Horizontal axis: the angle between the channel moment and the true channel direction.

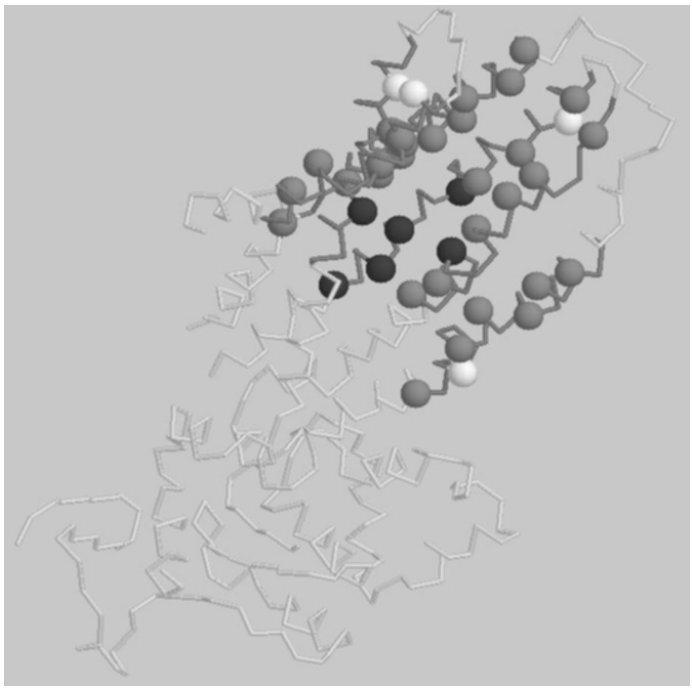

Fig. 5. X-ray-structure of MsbA according to [23] (only Cα-atoms are shown) with predicted channel residues highlighted.
Fig. 5a. All predicted residues are showed by spheres. Among them: residues that clearly face the channel are colored *gray*, residues that seem not to face the channel are colored *white*, residues that have been experimentally shown to face the channel are colored *dark gray*.

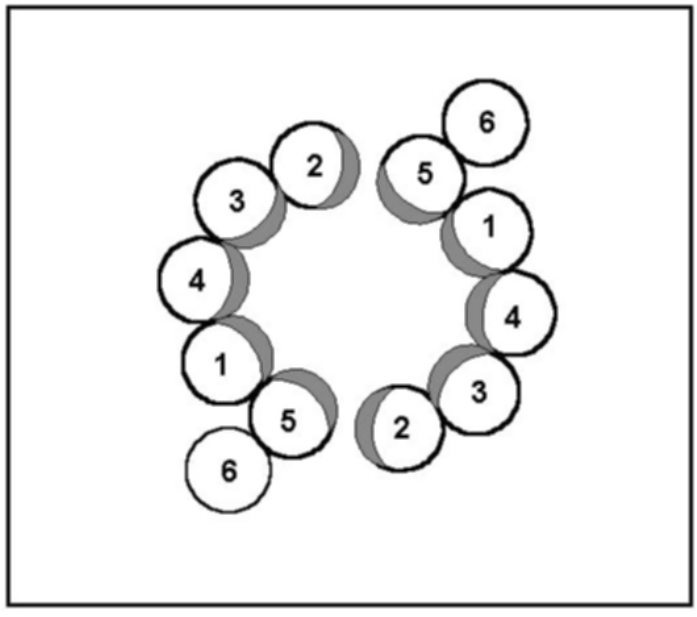

Fig. 5b. Model of the MsbA channel (view perpendicular to the membrane) with predicted channel residues highlighted. The side of the TM-helix that contains predicted residues is colored *dark gray*. Helices are numbered according to [23]. TM6 is ignored, as it clearly does not face the channel.

Additionally, we analyzed MsbA [23], which is the only bacterial transporter with resolved 3D structure (Figure 5). The numerical analysis is impossible since the X-ray structure of MsbA is still incomplete (only coordinates of Cα-atoms are published). Visual analysis revealed good accuracy of our predictions: Among the residues that lie within the sector of 90° facing the predicted channel direction (±45º from the channel moment) all but three indeed face the channel. Moreover, all six residues, shown in [23] to face the channel, lie in the predicted sector.

DISCUSSION

The growth of genomic data overwhelms the capacity of experimentalists to test the proteins functions. Therefore prediction of the substrate specificity of transporters could be very useful for genetic engineering, e.g. for creation of strains producing various bioactive compounds. The standard approach to determination of the substrate specificity based on protein similarity can easily lead to mistakes because of skewed amino acid composition of TM-proteins.

It is clear that in order to identify the amino acid residues interacting with the substrate one has to examine the protein's 3D structure. Unfortunately, due to the fact that TM-proteins crystallize poorly, there are very little data about 3D structures of TM-proteins. The total number of resolved 3D structures is 62, including two bacterial ABC-transporters (one of which was published after completion of this project), four bacterial ion conducting channels, and no secondary transporters [5]. Furthermore, current methods of secondary structure prediction for TM-proteins are hardly reliable [6]. One possible explanation is that prediction programs merge statistics of TM-proteins from both prokaryotic and eukaryotic organisms [24,25], although even the amino acid composition of eukaryotic transporters significantly differs from the amino acid composition of prokaryotic transporters [7].

As indicated by the experimental data [2–4], in the case of secondary transporters the substrate specificity is determined by the general structure of the TM-channel. Hence, the determination of the channel-forming residues is a prerequisite for the determination of specificity.

In this study we develop statistics specially designed for TM-kernels of secondary transporters. It turned out that for identification of channel-forming residues, which are most likely to determine the substrate specificity of a transporter, it is sufficient to consider a rather short segment consistently predicted to be transmembrane, that is the TM-kernel.

Despite the fact that the test dataset contained no secondary transporters, as no 3D structures of these proteins were available, and many proteins in the test dataset were oligomeric, our method, designed for secondary transporters, showed good results, especially for MsbA, the closest relative to the secondary transporters among the proteins from the test dataset.

Our results were obtained using two newly designed scales that differ from any other known hydrophobicity scale. Although these scales rely only on the distribution of amino acid residues in TM-kernels, they produce predictions not worse than those obtained by any other scale. This means that we can make reasonable predictions without any prior assumptions about physical and chemical properties of amino acid residues and of their environment.

ACKNOWLEDGMENTS

This study was partially supported by grants from the Howard Hughes Medical Institute (5500309) and the Ludwig Institute of Cancer Research (CRDF RB0-1268). We are grateful to A. A. Mironov for useful discussions.

REFERENCES

[1] Paulsen, I. T., Sliwinski, M. K. and Saier Jr., M. H. (1998). Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* **277**, 573–592.

[2] Kaback, H. R., Voss, J. and Wu, J. (1997). Helix packing in polytopic membrane proteins: the lactose permease of Escherichia coli. *Curr. Opin. Struct. Biol.* **7**, 537–542.

[3] Cosgriff, A. J., Brasier, G., Pi, J., Dogovski, C., Sarsero, J. P. and Pittard, A. J. (2000). A study of AroP-PheP chimeric proteins and identification of a residue involved in tryptophan transport. *J. Bacteriol.* **182**, 2207–2217.

[4] Hastings Wilson, T. and Wilson, D. M. (1998). Evidence for a close association between helix IV and helix XI in the melibiose carrier of Escherichia coli. *Biochim. Biophys. Acta* **1374**, 77–82.

[5] http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html

[6] Sadovskaya, N. S., Sutormin, R. A., Rakhmaninova, A. B. and Gelfand, M. S. (2002). Benchmarking of programs for recognition of transmembrane segments in transporter proteins. *In:* Proc. 3rd Int. Conf. On Bioinformatics of Genome Regulation and Structure BGRS'2002 (Novosibirsk, Russia, July 2002) 3, 115–116.

[7] Sutormin, R. A., Rakhmaninova, A. B. and Gelfand, M. S. (2003). BATMAS30 — the amino acid substitution matrix for alignment of bacterial transporters. Proteins **51**, 85–95.

[8] Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. and Saier Jr., M. H. (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301**, 75–100.

[9] Eisenberg, D., Schwarz, E., Komarory, M. and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142.

[10] Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J.-P. and Lanyi, J. K. (1999). Structural changes in bacteriorhidopsin during ion transport at 2 angstrom resolution. *Science* **286**, 255–261.

[11] Kolbe, M., Besir ,H., Essen, L.-O. and Oesterhelt, D. (2000). Structure of the light-griven chloride pump halorhodopsin at 1.8 Å resolution. *Science* **288**, 1390–1396.

[12] Royant, A., Nollert ,P., Edman, K., Neutze ,R., Landau ,E. M., Pebay-Peyroula, E. and Navarro, J. (2001). X-ray structure of sensory rhodopsin II at 2.1-A resolution. *Proc. Natl. Acad. Sci. USA* **98**, 10131–10136.

[13] Fu, D., Libson, A., Miercke, L. J., Weitzman, C., Nollert, P., Krucinski, J. and Stroud, R. M. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **290**, 481–486.

[14] Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. and Rees, D.C. (1998). Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensetive ion channel. *Science* **282**, 2220–2226.

[15] Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. and MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* **280**, 69–77.

[16] http://www.sander.ebi.ac.uk/dssp/

[17] http://cn.expasy.org/spdbv/

[18] http://pref.etfos.hr/split/

[19] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

[20] Pilpel, Y., Ben-Tal, N. and Lancet D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **294**, 921–935.

[21] Engelman, D. M., Steitz, T. A. and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Ann. Rev. Biophys. Chem.* **15**, 321–353.

[22] Kuhn, L. A. and Leigh jr., J. S. (1985). A statistical technique for predicting membrane protein structure. *Biochim. Biophys. Acta* **828**, 351–361.

[23] Chang, G. and Roth, C. B. (2001). Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**, 1793–1800.

[24] Ng, P. C., Henikoff, J. G. and Henikoff, S. (2000). PHAT: A transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics* **16**, 760–766.

[25] Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994). A mutation data matrix for transmembrane proteins. *FEBS Lett.* **339**, 269–275.