RARE RESIDUES FORM THE CHANNEL IN TRANSMEMBRANE TRANSPORTERS

Kalina O.V., Makeev V.Ju., Sutormin R.A., Gelfand M.S., Rakhmaninova A.B.

State Scientific Center GosNILGenetica, Moscow, 113545, Russia
Integrated Genomics, P.O. Box 348, Moscow, 117333, Russia
e-mail: linel@mail.ru

Motivation. Transmembrane transport is of extremely importance for the cell life. Many genes encoding real or putative transport proteins are found in bacterial genomes. But in most cases their substrate specificity is not experimentally determined and only approximately predicted from genomic studies. Even less is known about the 3D-structure of transporters. Nevertheless the published experimental data lets us assume that determination of the channel-forming residues would allow make suppositions about substrate specificity of secondary transporters.

Results. We have developed a simple computational method for identification of channel-forming residues in transporter sequence, based on our original data about amino acids frequencies in bacterial secondary transporters. We have applied it to transmembrane proteins with resolved 3D structure and the prediction showed a sufficiently good agreement with the real protein structure.

Availability. All results are available from the authors (linel@mail.ru).

Introduction

Transmembrane (TM) transporter proteins are the major mechanism of the flow of compounds in and out the bacterial cell. Up to eleven percent of a prokaryotic proteome are the membrane transporter systems, and thus prediction of their substrate specificity not only is important for the genome annotation, but also is of major practical interest. 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. Therefore, prediction of the substrate specificity requires determination of the channel-forming residues.

Although only few resolved 3D structures of transporters are known, there are many structural models based as on various indirect experimental data. 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. In an accompanying abstract (Sutormin et al.) we introduce the concept of TM-kernels as a protein fragment consistently predicted to be a transmembrane segment.

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

Statistical analysis of TM-kernels

We have analyzed 18908 kernels from 2172 proteins (bacterial secondary transporters, class 2A according to the Saier-Paulsen classification). The TM-kernels retain the periodic distribution of residues described for the whole TM-helices.

To reveal amino acid residues propensity 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, \( \alpha \) and \( \beta \). The positional correlation for each distance \( n \) was calculated as follows. Let \( N_{n}^{\alpha \beta} \) be the number of such residue pairs, where the first residue belongs to group \( \alpha \), the second residue belongs to group \( \beta \) and the distance between the residues is \((n-1)\): Let \( \ell^\alpha(x) \) be such a function: \( \ell^\alpha(x) = 1 \), if \( x \in \alpha \), \( \ell^\alpha(x) = 0 \), else.

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 $p_\alpha$ be the frequency of residues from group $\alpha$ in the sample of TM-kernels: 

$$p_\alpha = \frac{\sum_{k=1}^{M} l_k \cdot \delta_i (x_i)}{\sum_{k=1}^{M} l_k}.$$ 

Then the positional correlation coefficient in point $n$ is 

$$\text{corr}(n) = \frac{N_{n}^{\alpha \beta} - N_{n} p_\alpha p_\beta}{\sqrt{p_\alpha (1 - p_\alpha) p_\beta (1 - p_\beta) \cdot \sum_{k=1}^{M} l_k}}.$$ 

We have observed that tryptophan and tyrosine tend to lie at the same side of the helix as charged and polar residues (Fig. 1). We assume that this is the channel side and therefore call them channel residues. The common property of these residues is that according to our data (Sutormin et al.) their frequency in TM-kernels is significantly less than in protein in general. Still, the average number of channel residues per kernel is 2.6 (Fig. 2), which might be sufficient for determination of the channel side of a helix.

![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.](image1)

![Fig. 2. Distribution of the number of channel amino acid residues in kernels.](image2)

**Calculation of the channel moment**

Two scales of channel propensity were constructed as follows:

$$P^{(1)}_\alpha = \log \frac{f_{a}^{\text{tm}}}{f_{a}^{\text{sv}}} \quad \quad P^{(2)}_\alpha = \log \frac{f_{a}^{\text{tm}}}{1/20},$$

where $P^{(\nu)}_\alpha$ is the channel propensity of residue $\alpha$, $f_{a}^{\text{tm}}$ is the frequency of $a$ in TM-kernels, $f_{a}^{\text{sv}}$ is the frequency of $a$ in all proteins.

Correlation of these scales with about 90 different scales of amino acid attributes [http://pref.etfos.hr/split/] used for prediction of TM helices 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 (Fig. 3a, b). It must be noted that both scales showed only poor correlation with most popular scales, such as the Kyte-Doolittle scale (Kyte, Doolittle, 1982) ($P^{(1)}$ and $P^{(2)}$: 0.84), Eisenberg scale (Eisenberg et al., 1984) ($P^{(2)}$: 0.79) and kPROT (Pilpel et al., 1999) ($P^{(1)}$: 0.46, $P^{(2)}$: 0.48).

![Fig. 3a. Correlation of $P^{(1)}$ (horizontal) with Engelman scale (Engelman et al., 1986) (vertical) (correlation coefficient = 0.93).](image3a)

![Fig. 3b. Correlation of $P^{(2)}$ (horizontal) with the Kuhn-Leigh scale (Kuhl, Leigh, 1985) (vertical) (correlation coefficient = 0.90).](image3b)
These scales were used for identification of channel residues. The channel moment $C$ was defined analogously to the hydrophobic moment (Eisenberg et al., 1984):

$$\vec{C} = \sum_i \vec{c}_i,$$

where $\vec{c}_i = \vec{r}_i \cdot P^{(v)}$, $\vec{r}_i$ is the radius-vector of residue at position $i$, $P^{(v)}$ is the channel propensity scale $\#v$ ($v = 1, 2$).

Several bacterial and archaeal TM proteins with resolved 3D structure were used to test the reliability of rotational orientation of TM-segments by the channel moment.

To determine the “true” orientation of the channel vector we calculated the vector pointing to the most exposed side of the helix and assumed that it points to the membrane, that is, out of the channel. That was done using the solvent accessibility surfaces from the DSSP database [http://www.sander.ebi.ac.uk/dssp/] or calculated using the program SPDBV [http://cn.expasy.org/spdbv/]. We considered only proteins which had an inner cavity or channel and an easily detectable single layer of helices surrounding this cavity: 1FBB (bacteriorhodopsin, Halobacterium salinarum), 1E12 (light-driven chloride pump, Halobacterium salinarum), 1H68 (sensory rhodopsin II, Natronomonas pharaonis), 1FX8 (glycerol-conducting channel, Escherichia coli), 1MSL (mechanosensitive ion channel MSCL homolog, chain A, Mycobacterium tuberculosis), 1BL8 (KCSA, potassium channel, chain A, Streptomyces lividans). 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. Total number of TM helices in the study was 32.

**Results**

The angle differences between the calculated channel moments and the directions of “true” channel vectors for all 32 studied TM helices are shown in Fig. 4. One can see that the obtained predictions are comparable to the ones obtained using the Kyte-Doolittle scale: in approximately 2/3 cases the channel side is predicted with deviation less than 50° from the “true” direction and in about 1/3 cases the channel side is predicted badly. The latter phenomenon is coupled with the objective limit of accuracy for such predictions: some helices contain charged residues that face the membrane, possibly establishing interactions between protein subunits.

Additionally, we have analyzed MsbA (Chang, Roth, 2001), which is the only bacterial transporter with resolved 3D structure. The numerical analysis is impossible since the X-ray structure of MSBA is still incomplete: only coordinates of Cz-atoms are published. The visual analysis reveals good accuracy of our predictions: We have analyzed the residues that lie within the sector of 90° facing the predicted channel direction ($\pm45^\circ$ from the channel moment) and all but three of them indeed face the channel. Moreover, all six residues, shown in (Chang, Roth, 2001) to face the channel, lie in the predicted sector.

The most intriguing result of this study is that predictions done using $P^{(2)}$ are not worse than those obtained by any other scale. It means that one can predict the channel side of a TM-helix without prior assumptions about the amino acids properties and using only amino acid frequencies in TM-kernels.
We plan to use this method for modeling secondary transporters. We expect a good accuracy of prediction, especially for monomeric proteins.

Acknowledgements

This study was partially supported by grants from RFBR, INTAS (99-1476), HHMI, and LICR/CRDF. We are grateful to A.A. Mironov for useful discussion.

References

7. Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. BATMAS30 - the amino acid substitution matrix for alignment of bacterial transporters. This volume.