BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 608–616

Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces

Shigeki Mitaku [1,], Takatsugu Hirokawa [1, 2] and Toshiyuki Tsuji [1]*

[1]Tokyo University of Agriculture and Technology, Faculty of Technology, Department of Biotechnology, Nakacho, Koganei, Tokyo 184-8588, Japan and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koutou-ku, Tokyo 135-0064, Japan

Received on December 12, 2000; revised on May 10, 2001; accepted on September 20, 2001

ABSTRACT

Motivation: An amphiphilicity index of amino acid residues was developed for improving the method of transmembrane helix prediction.

Results: The transfer energy of a hydrocarbon stem group beyond the γ-carbon was calculated from the accessible surface area, and used to index the amphiphilicity of the residue. Non-zero amphiphilicity index values were obtained for lysine, arginine, histidine, glutamic acid, glutamine, tyrosine and tryptophan. Those residues were found to be abundant in the end regions of transmembrane helices, indicating their preference for the membrane–water interface. The moving average of the amphiphilicity index actually showed significant peaks in the end regions of most transmembrane helices. A dispersion diagram of average amphiphilicity index versus average hydrophobicity index was devised to facilitate discrimination of transmembrane helices.

Availability: The amphiphilicity index has been incorporated into a system, SOSUI, for the discrimination of membrane proteins and the prdiction of tranmembrane helical regions (http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html).

Contact: mitaku@cc.tuat.ac.jp

INTRODUCTION

Total genomes of many biological organisms are currently being analysed, but a considerably large fraction of total ORFs code for orphan proteins whose structure or function cannot be annotated (Brown, 1999). Analyses of amino acid sequences from total genomes have shown that approximately a quarter of ORFs code for membrane proteins (Wallin and von Heijne, 1998; Mitaku *et al.*, 1999). However, experimental analyses, (particularly structural analyses) of membrane proteins, while biologically important, are very difficult to perform. Therefore, high-performance software systems designed for membrane protein prediction are invaluable aids in the investigation of membrane protein structure.

Amino acid indices are generally very useful for making a software system of the classification and prediction of protein structure, and many kinds of amino acid indices have been developed for this purpose (Nakai *et al.*, 1988; Tomii and Kanehisa, 1996). A good example is the use of a hydrophobicity (hydropathy) index for visualizing hydrophobic segments in membrane proteins (Kyte and Doolittle, 1982; Engelman *et al.*, 1982; Eisenberg *et al.*, 1984; Eisenberg and McLachlan, 1986; Mitaku *et al.*, 1985). However, the hydrophobicity index alone is not enough for accurate prediction of membrane proteins, leading to the necessity of other kinds of indices for further improvement (Mitaku and Hirokawa, 1999).

Propensity studies of amino acids in membrane proteins indicate that the polar residues lysine, arginine and tryptophan are preferable at the end regions of transmembrane helices (Schiffer *et al.*, 1992; von Heijne, 1992; Reithmeier, 1995; Braun and von Heijne, 1999; Ridder *et al.*, 2000). The reason why the polar residues are distributed at the end regions has also been discussed, and a snorkel hypothesis was proposed suggesting that polar groups of lipids interact with positively charged residues and polar-aromatic residues (Segrest *et al.*, 1990; Killian and von Heijne, 2000). However, it seems that this kind of amino acid preference does not indicate a definite motif, and preferable amino acids appear in a statistical manner at the membrane surface. When certain kinds of amino acids are statistically preferred, local averaging of an appropriate index is useful for determining the statistical preference of occurrence of amino acids. Statistical noise is reduced by the local averaging procedure, and peaks in the local average of the index coincide with clusters of preferable amino acids. Because the occurrence of lysine,

*To whom correspondence should be addressed.

arginine, tryptophan and tyrosine at the end regions of transmembrane helices appears to be statistical in nature, the indexing of those amino acids will be useful in a quantitative discussion of the stability of transmembrane helices at the interfacial region of membranes.

In this study, we developed a novel index of amino acids that represents the amphiphilicity of each polar side chain. Amphiphilicity values are positive for polar residues with large hydrophobic stems beyond the γ-carbon (lysine, arginine, histidine, glutamic acid, glutamine, tryptophan and tyrosine), and small polar residues and hydrophobic residues have an amphiphilicity value of zero. Propensity analyses of membrane proteins of known 3D structure showed that amino acids with positive amphiphilicity values were preferable at the end regions of transmembrane helices. In addition to the already reported preference of positively charged residues and polar-aromatic residues to the membrane–water interface, we found that the propensity of glutamic residues (glutamic acid and glutamine) was larger than aspartic residues (aspartic acid and aspargine) at the end regions of transmembrane helices. The moving average of the amphiphilicity index for a window of seven residues actually showed significant peaks at the end regions of transmembrane helices, a fact which will be useful in the prediction of transmembrane helices. In fact, a dispersion diagram of average amphiphilicity index versus average hydrophobicity could aid in discriminating α-helices in membrane proteins from those in soluble proteins.

METHODS

Calculation of amphiphilicity values of polar amino acids

We indexed the amphiphilicity of polar amino acids according to the transfer energy calculated from the accessible surface area of the stem groups of their polar side chains. The accessible surface area of the stem group beyond the γ-carbon was calculated using Insight II (Molecular Simulations Inc; Figure 1). The transfer energy, $\Delta G_{\text{transfer}}$, was then calculated from the accessible surface area of the hydrophobic stem, ΔA_{stem}, using (1):

$$\Delta G_{\text{transfer}} = \sigma \cdot \Delta A_{\text{stem}}. \qquad (1)$$

In (1), σ represents surface tension at the molecular level. We assumed a surface tension of 40 dyn cm^{-1} (Mitaku, 1993).

The amphiphilicity index of amino acids was defined by the transfer energy of stem groups of side chains. Since polar groups of aspartic acid, aspargine, serine and threonine are connected to the main chain through β-carbon, amphiphilicity index values of those amino acids are zero. Amphiphilicity indices for aromatic and aliphatic hydrophobic stems were calculated using the same procedure. Positive values were obtained for seven amino acids (arginine, lysine, histidine, glutamic acids, glutamine, tyrosine and tryptophan). However, we categorized those residues into two groups, strongly polar amino acids (arginine, lysine, histidine, glutamic acid and glutamine) and weakly polar amino acids (tyrosine and tryptophan), because the polar groups of tyrosine and tryptophan have only a single electric dipole, whereas arginine, lysine, histidine and glutamic acid have a full elementary charge and glutamine has two electric dipoles. Amphiphilicity index was shown by two parameters, A and A′ that represent strongly and weakly polar amino acids, respectively.

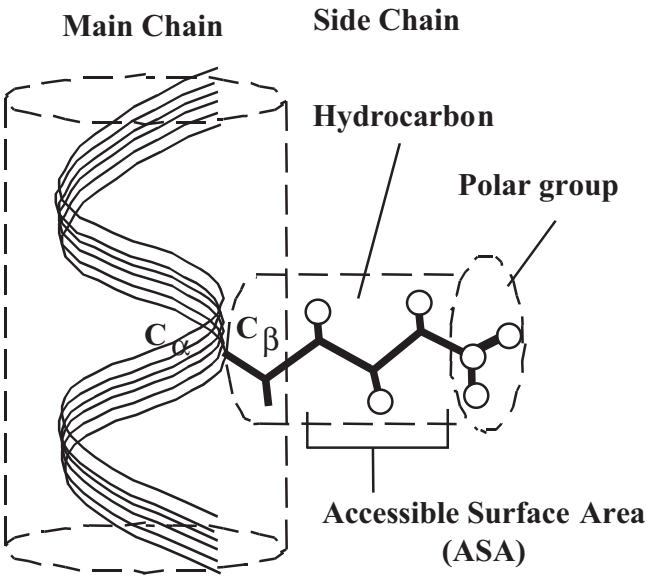

Fig. 1. Schematic diagram of a side chain of a polar residue (lysine), together with the ribbon model of the main chain. The accessible surface area of the non-polar stem beyond the γ-carbon was calculated in order to determine the amphiphilicity index values.

Amphiphilicity plot of amino acid sequences

In order to visualize the local preference of amino acids for transmembrane regions, we calculated the local averages of the hydropathy index, $\overline{H}$, and amphiphilicity index, $\overline{A}$ and $\overline{A'}$ (for strongly and weakly polar residues, respectively), and separately plotted as a function of sequence number:

$$\overline{H}(i) = \left(\sum_{j=i-3}^{i+3} H(j) \right) \Big/ 7, \qquad (2)$$

$$\overline{A}(i) = \left(\sum_{j=i-3}^{i+3} A(j) \right) \Big/ 7, \qquad (3)$$

Table 1. Dataset of membrane proteins whose 3D-structure is known

Protein	ID	Number of chains	Number of helices
F1F0 ATPsynthase (*Escherichia coli*)	1A91	1	2
Cytochrome bc1 complex (bovine)	1BGY(chain C, D, G, J, K)	5	8, 1, 1, 1, 1
Bacteriorhodopsin (*H. salinarium*)	1AT9	1	7
Ca ATPase, SR (rabbit)	1EUL	1	10
Cytochrome *c* oxidase (bovine)	1OCC (I, II, III, IV, VIa, VIc, VIIa, VIIb, VIIc, VIII)	10	12, 2, 7, 1, 1, 1, 1, 1, 1, 1
Cytochrome *c* oxidase (*Thermus thermophilus*)	1EHK (chain I, II)	2	13, 1
Fumarate reductase (*W. succinogenes* 1)	1QLA (chain C)	1	5
Fumarate reductase (*E. coli*)	1FUM (15 kD anchor, 13 kD anchor)	2	3, 3
Glycophorin A (human)	1MSR	1	1
Halorhodopsin (*H. salinarium*)	1E12	1	7
KcsA potassium channel (*S. lividans*)	1BL8	1	2
Light harvesting complex (*R. acidophila*)	1KZU (chain A, B)	2	1, 1
Light harvesting complex (*R. molischianum*)	1LGH (chain A)	1	1
MscL ion channel (*Mycobacterium tuberculosis*)	1MSL	1	2
Reaction center (*R. viridis*)	1PRC (chain M, L, H)	3	5, 5, 1
Rhodopsin (bovine)	1F88	1	7
Total		34	117

$$\overline{A'}(i) = \left(\sum_{j=i-3}^{i+3} A'(j) \right) \Big/ 7, \qquad (4)$$

where i is the sequence number of the center of the seven-residues window.

Dispersion diagram of amphiphilicity index versus hydropathy index

Two physicochemical features characterize a transmembrane helix. One feature is a hydrophobic segment that corresponds to the non-polar environment of the membrane, and the other feature is clusters of amphiphilic residues around the ends of the peak of hydrophobicity. In order to examine the applicability of the amphiphilicity index to transmembrane helix prediction, we calculated the weighted average, $\langle A \rangle$, of the amphiphilicity index using the following equation:

$$\langle A \rangle = \left[\sum_{i=k_N}^{k_N+4} \overline{A}(i) + \sum_{i=k_{C-4}}^{k_C} \overline{A}(i) \right] \Big/ 10 + \left[\sum_{i=k_N}^{k_C} \overline{A'}(i) \right] \Big/ (k_C - k_N + 1) \qquad (5)$$

in which k_N and k_C are the residue numbers of the N- and C-termini of a helical region, respectively. The weight used to average amphiphilicity index A is different from that used for A', because strongly polar residues are very rare in the non-polar stretch of the amino acid sequence, whereas weakly polar residues are found at almost equal frequency in helical and non-helical regions.

The other parameter of the dispersion diagram is the average hydropathy index, $\langle H \rangle$, which is calculated using the following equation:

$$\langle H \rangle = \left[\sum_{i=k_N}^{k_C} \overline{H'}(i) \right] \Big/ (k_C - k_N + 1). \qquad (6)$$

All helical regions longer than 19 residues were plotted in a dispersion diagram of $\langle A \rangle$ versus $\langle H \rangle$, and separation between helices in soluble and membrane proteins was examined.

Data set for analyses

We used three sets of amino acid sequence data. The first set included the most reliable data: that of membrane proteins of known 3D structure, whose atomic coordinates are recorded in the Protein Data Bank (PDB; Girvin *et al.*, 1998; Iwata *et al.*, 1998; Kimura *et al.*, 1997; Toyoshima *et al.*, 2000; Tsukihara *et al.*, 1996; Soulimane *et al.*, 2000; Lancaster *et al.*, 1999; Iverson *et al.*, 1999; Kolbe *et al.*, 2000; Doyle *et al.*, 1998; McDermott *et al.*, 1995; Koepke *et al.*, 1996; Chang *et al.*, 1998; Deisenhofer *et al.*, 1985; Palczewski *et al.*, 2000). Table 1 shows the names and PDB codes of membrane proteins, and the numbers of chains and helices they contain. Redundancy of data was removed with the cutoff of 30% homology. The total numbers of membrane proteins and transmembrane helices were 34 and 117, respectively. The local propensities were calculated for this data set. The second data set which included data of 148 membrane proteins reported by Möller *et al.* (2000) were also used, in order to increase the number of membrane protein data (706

transmembrane helices) used in the dispersion diagram analysis. There is overlap of 22 data (63 helices) between these two data sets. The ftp site of the Möller data set is ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane. The third data set included data of 397 soluble proteins, altogether containing 489 helices longer than 19 residues, obtained from PDBselect and used as reference data in the dispersion diagram (Hobohm *et al.*, 1992). The data sets can be found at the following URL: http://sosui.proteome.bio.tuat.ac.jp/~sosui/proteome/dataset/index.html.

RESULTS

We defined our amphiphilicity index of amino acids according to (1). The values of amphiphilicity index and hydropathy index are shown in Table 2. Positive amphiphilicity index values were obtained for seven polar residues (lysine, arginine, histidine, glutamic acid, glutamine, tryptophan and tyrosine), whereas small polar residues (aspartic acid, aspargine, serine and threonine) had an amphiphilicity value of zero from its definition. Table 2 shows the amphiphilicity values of strongly polar residues, A, and weakly polar residues, A', in different columns, because the amphiphilicity of a molecule or a group generally depends on not only the size of the hydrocarbon chain but also the polarity of the hydrophilic region. The last column represents the hydropathy index (Kyte and Doolittle, 1982). The positively charged residues arginine, lysine and histidine have high amphiphilicity values, since they each have a net elementary charge linked by a long hydrocarbon chain. In contrast, the polar groups of aspartic acid and aspargine are directly connected to the main chain through a β-carbon, resulting in an amphiphilicity value of zero. The hydrophobic stems of glutamic acid and glutamine are of intermediate size, and their amphiphilicity values are thus of intermediate value.

Amphiphilic molecules or groups usually prefer at the interface between polar–non-polar environments. Therefore, we examined the correlation between the amphiphilicity index of amino acids in Table 2 and the local propensities of amphiphilic residues at the membrane–water interface. As shown in Figure 2, transmembrane helices were divided into three regions: two end regions (five residues long) and a central region (all other residues in the middle). We calculated the propensities of amino acids for these three regions of helices and a loop region ten residues long. Propensities for the intrahelical regions were then normalized with the corresponding propensities for loop segments, in order to compare the position preference of each amino acid. Relative propensity was calculated according to the following equation:

$$p_\alpha = P_\alpha / P_l. \tag{7}$$

Table 2. Amphiphilicity and hydropathy indices of amino acids

Amino acid	A	A'	H
Lysine (K)	3.67	0	−3.9
Arginine (R)	2.45	0	−4.5
Histidine (H)	1.45	0	−3.2
Glutamic acid (E)	1.27	0	−3.5
Glutamine (Q)	1.25	0	−3.5
Aspartic acid (D)	0	0	−3.5
Asparagine (N)	0	0	−3.5
Trptophan (W)	0	6.93	−0.9
Tyrosine (Y)	0	5.06	−1.3
Serine (S)	0	0	−0.8
Threonine (T)	0	0	−0.7
Proline (P)	0	0	−1.6
Glycine (G)	0	0	−0.4
Alanine (A)	0	0	1.8
Methionine (M)	0	0	1.9
Cysteine (C)	0	0	2.5
Phenylalanine (A)	0	0	2.8
Leucine (L)	0	0	3.8
Valine (V)	0	0	4.2
Isoleucine (I)	0	0	4.5

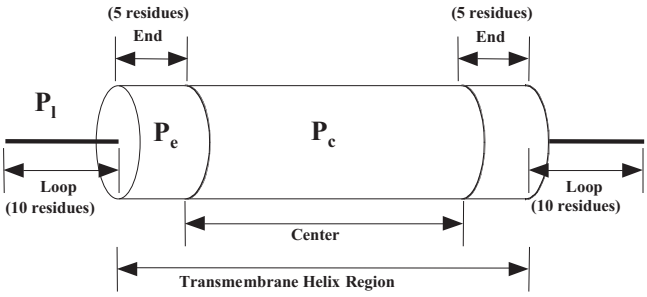

Fig. 2. Three regions of a transmembrane helix, central and end regions, together with loop segments 10 residues long were examined in order to estimate the local amino acid preference.

Parameters P and p indicate propensity in a particular region and relative propensity normalized by the corresponding value in loop segments, respectively. The subscript α represents a central or end region of a transmembrane helix and l denotes a loop segment. The loop segments of membrane proteins are similar to soluble protein in the meaning that they are exposed to water. The correlation coefficient between the Dayhoff's index of amino acids (Dayhoff *et al.*, 1978) for soluble proteins and the propensity of the residues in loop segments of membrane proteins was as large as 0.99 with only a small systematic decrease for loop segments. Since the conclusion did not change by the normalization factors, we used the relative propensity in (7) for the analysis.

Figure 3 shows the relative propensities of occurrence

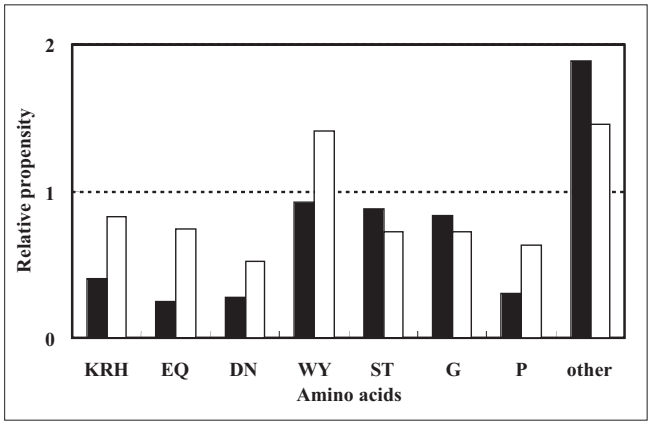

Fig. 3. Relative propensities of eight kinds of amino acids in the central (solid bars) and end (open bars) regions of transmembrane helices: positively charged residues (arginine, lysine, histidine); glutamic acid and glutamine; aspartic acid and aspargine; large neutral residues (tyrosine, tryptophan); small neutral residues (serine, threonine); glycine; proline; and other hydrophobic residues.

of eight kinds of amino acids: (1) positively charged residues (arginine, lysine, histidine); (2) glutamic residues (glutamic acid, glutamine); (3) aspartic residues (aspartic acid, aspargine); (4) aromatic polar residues (tyrosine, tryptophan); (5) small polar residues (serine, threonine); (6) glycine; (7) proline and (8) other hydrophobic residues (alanine, valine, leucine, isoleucine, phenylalanine, methionine, cysteine). We classified amino acids by the size of their side chain as well as their polarity. Based on the polarity, we classified polar amino acids into the same two categories as in the evaluation of the amphiphilicity index: strongly polar residues and weakly polar residues. Strongly polar residues were classified into three groups by the size of their hydrocarbon stems: positively charged residues; glutamic acid and glutamine; and aspartic acid and aspargine. Weakly polar residues were categorized into two groups: tyrosine and tryptophan; and serine and threonine.

As shown in Figure 3, the relative propensities of polar residues in the central region are dependent on polarity, whereas their relative propensities at the end regions show good correlation not only with polarity but also the size of their non-polar group. Strongly polar residues (arginine, lysine, histidine, glutamic acid, glutamine, aspartic acid, aspargine) were rare in the central regions of transmembrane helices, as indicated by the fact that their relative propensity values are much smaller than 1.0. The propensities of weakly polar residues (tyrosine, tryptophan, serine, threonine) were almost the same for central regions as for loop segments. This difference in relative propensity for central regions between strongly and weakly polar residues is to be expected, because the partitioning of polar groups depends on the strength of their polarity. However, the propensity of polar residues for the end regions of transmembrane helices cannot be explained by polarity alone. Large polar residues showed much greater propensities than small ones, regardless of the relative strength of polarity.

The order of increasing relative propensity at the membrane–water interface region was the same as the order of increasing amphiphilicity value as follows: (aspartic residues) < (glutamic residues) < (arginine, lysine, histidine) for strongly polar residues and (serine, threonine) < (tyrosine, tryptophan) for weakly polar residues. Although in the case of weakly polar residues the propensity for the central region of transmembrane helices was almost the same as that for loop segments, the trend of amino acid preference at the end regions was also consistent with amphiphilicity index values: the propensities of tyrosine and tryptophan were larger at the end regions than at the center, while those of serine and threonine changed in the opposite direction.

Statistical preference of large polar residues for the ends of transmembrane helices does not necessarily mean this tendency is applicable to the prediction of transmembrane regions. Unless this preference applies to most transmembrane helical regions, it will be difficult to use it to improve the accuracy of membrane protein prediction. Figure 4 shows a plot of moving average of amphiphilicity index values for cytochrome *c* oxidase subunit III (Tsukihara *et al.*, 1996) and rhodopsin (Palczewski *et al.*, 2000), along with their so-called hydropathy plots. The boxes in the plots represent transmembrane regions, as defined in the PDB. The peaks of the average amphiphilicity index which coincide with the end points of transmembrane helices are shaded in Figure 4 to aid examination of whether amphiphilic residues are actually preferable at helix ends (which are at the membrane–water interface). The shapes of the amphiphilicity plots and hydropathy plots of transmembrane helices were not the same for each helix. However, there is a general tendency for there to be a significant peak in hydrophobicity in the central region of a transmembrane helix, while peaks in amphiphilicity values tend to occur very near the end points of transmembrane helices.

Averaged profiles of those parameters were calculated by averaging all profiles of transmembrane helices. Figure 5 shows the plots of averaged hydropathy index and averaged amphiphilicity index, with the lengths of transmembrane helices normalized to 20. The positions −10 and 10 correspond to the cytoplasmic and cell surface ends of helices, respectively. The averaged profiles in Figure 5 are consistent with predicted preference; there is a significant peak of hydrophobicity index in the central region, and the peaks in amphiphilicity values are near

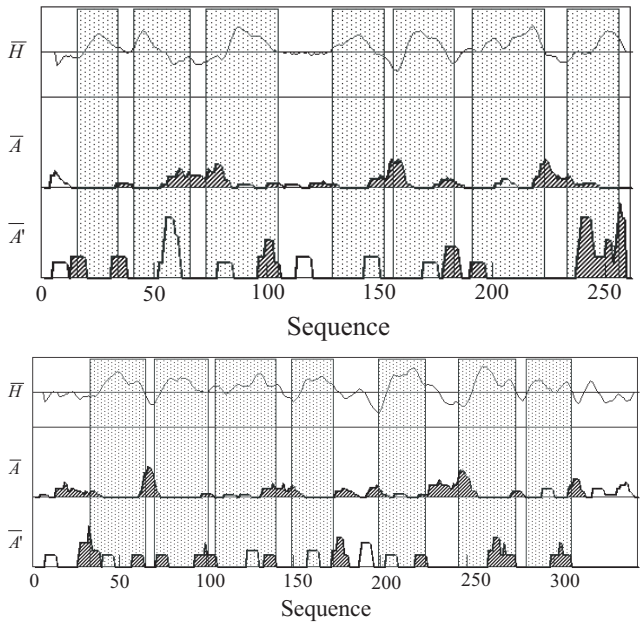

Fig. 4. Plots of local averages of hydropathy and amphiphilicity values, which were calculated for a 7-residue window, as a function of amino acid sequences. Transmembrane helical regions are represented by gray boxes in the plots for cytochrome *c* oxidase subunit III (1occ) (a) and rhodopsin (1f88) (b). Hatching indicates peaks of amphiphilicity plots for sequences containing helix ends.

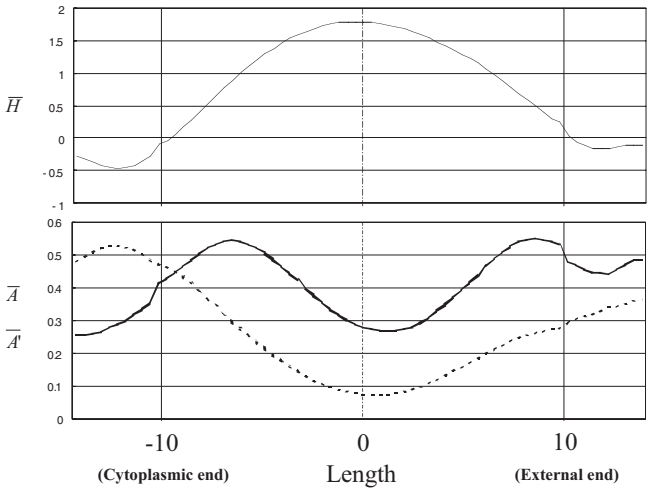

Fig. 5. Averaged profiles of hydropathy and amphiphilicity plots, in which the lengths of all transmembrane helices of membrane proteins in Table 2 have been normalized to 20. The origin represents the midpoints of helices, and the cytoplasmic and external ends of helices correspond to the positions −10 and 10, respectively. The solid line represents average hydropathy values, while the amphiphilicity values A and A' are represented by dotted and gray lines, respectively.

the membrane surface. Peaks in amphiphilicity values and minimum values of hydrophobicity both occur very near the end points of transmembrane helices (Wallin *et al.*, 1997). However, it should be noted that a profile constructed using amphiphilicity values of strongly polar residues is somewhat different from one constructed using values of weakly polar residues. Although both profiles show high values of amphiphilicity at the end points of transmembrane helices, the peak in values of strongly polar residues is located just outside the helical region, whereas the peak in values of weakly polar residues is apparently located inside the helical region. There is also significant asymmetry in profiles made using values of strongly polar residues, with higher average amphiphilicity values at the cytoplasmic side than at the cell surface side.

The applicability of the amphiphilicity index to the problem of transmembrane helix prediction was demonstrated by plotting a dispersion diagram of average amphiphilicity index $\langle A \rangle$ versus average hydrophobicity index $\langle H \rangle$, which are defined by (5) and (6), respectively, using values for helical segments from both membrane proteins and soluble proteins. Figure 6 shows an example of such a dispersion diagram, including data for 793 transmembrane helices of 167 membrane proteins and 489 helices (all longer than 19 residues) of 397 soluble proteins which are larger than 100 residues (Mitaku and Hirokawa, 1999). In general, it was possible to distinguish transmembrane helices from helices of soluble proteins by average hydrophobicity $\langle H \rangle$ alone. However, the region of overlap was rather large, and the introduction of the second parameter, average amphiphilicity, $\langle A \rangle$, significantly improved discrimination between these two types of helices. In Figure 6, a boundary line was drawn that clearly separated the two types of helices:

$$\langle A \rangle = -1.25\langle H \rangle + 1.625. \tag{8}$$

The number of errors of separation was only 20 out of a total of 1282 helices. Figure 6 apparently shows that the introduction of the amphiphilicity index will be useful for improving the prediction systems of transmembrane helices. However, other definitions of the amphiphilicity index is possible in general. For example, the average number of amphiphilic residues may be used instead of the average amphiphilicity index $\langle A \rangle$ defined by (5). The dispersion diagram of Figure 6 using the two definitions of amphiphilicity index for comparison showed that the amphiphilicity index defined in this work was better in discriminating transmembrane helices (data not shown).

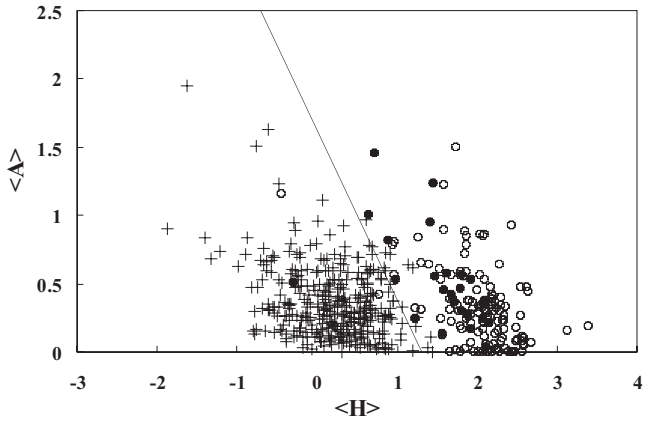

Fig. 6. Data of helical segments from three kinds of databases of amino acid sequences were plotted in a dispersion diagram of average amphiphilicity index $\langle A \rangle$ versus average hydropathy index $\langle H \rangle$. Data are represented by three different symbols: (●), membrane proteins of known 3D-structure; (○), membrane proteins reported by Möller *et al.* (2000); (+), soluble proteins from PDBselect. A solid line represents the boundary between helices in membrane proteins and those in soluble proteins.

DISCUSSION

An amphiphilic molecule generally consists of two parts: a polar group and a non-polar group (Tanford, 1980). The chemical potential of an amphiphilic molecule greatly increases, when a polar group of the molecule is translocated into a hydrophobic environment and a non-polar group is exposed to water. Therefore, an amphiphilic molecule segregates from both pure non-polar solvents and water, preferring the interface between polar and non-polar environments.

Preference for this interface is also expected for amino acid side chains which have a polar group and a hydrocarbon stem. If we assume that the dielectric constant of the hydrophobic environment of the membrane is approximately 4, the self-energy of an electric elementary charge is calculated to be approximately 30 kcal mol^{-1} (Israelachvili, 1991). This self-energy is much greater than the thermal energy, which is approximately 0.6 kcal mol^{-1}. Therefore, the Boltzmann factor of the distribution of strongly polar residues to the central region of the membrane has to be very small, and the small relative propensity of occurrence of strongly polar residues in the central region of the membrane is generally considered to be the result of a genetic optimization process influenced by the Boltzmann factor. In contrast, the transfer energy of the hydrocarbon stem of lysine, for example, is 3.7 kcal mol^{-1}, as estimated from the solvent-accessible surface area (Table 2). Because the hydrocarbon stems of side chains are smaller than the non-polar group of common detergents, their transfer energy value is only a few times greater than their thermal energy. However, a cluster of amphiphilic side chains in an amino acid sequence produces a high transfer energy value. It is logical to conclude that a cluster of amphiphilic residues would stabilize the end region of a transmembrane helix at the membrane–water interface.

In order to test this concept of stabilization of transmembrane helices by amphiphilic side chains at the helix end regions, we performed three analyses of amino acid sequences of membrane proteins: (1) analysis of local propensity of occurrence of amino acids at central and end regions of helices; (2) calculation of moving average of amphiphilicity and hydrophobicity; and (3) discrimination of transmembrane helices using a dispersion diagram of $\langle A \rangle$ versus $\langle H \rangle$.

In the first analysis, comparing the distribution of polar residues in three distinct regions of helices, we concluded that large polar residues, which are amphiphilic, are more preferable than small polar residues in helix end regions. It was previously reported that positively charged residues are commonly observed at the end regions of transmembrane helices, particularly on the cytoplasmic side (von Heijne, 1992). The present analysis of 150 transmembrane helices confirmed this characteristic distribution of positively charged residues. The present finding that propensities of occurrence of tryptophan and tyrosine were higher for end regions than central regions are also consistent with findings reported previously by Schiffer *et al.* (1992); Braun and von Heijne (1999) and Ridder *et al.* (2000). The most interesting result was that glutamic residues (glutamic acid and glutamine) are more common than aspartic residues (aspartic acid and aspargine) in end regions, while propensities of glutamic residues and aspartic residues for the central regions of transmembrane helices are almost the same. This fact, which has not been pointed out previously, is a good indication of the importance of amphiphilic residues for stabilization of transmembrane helices at the membrane–water interface.

We defined an index of amphiphilicity, which is a physically well-defined parameter, in order to construct a system for predicting membrane proteins on the basis of the concept of stabilization of transmembrane helices by amphiphilic residues. Amphiphilicity plots of membrane proteins of known 3D structure indicated that this kind of plot is useful in finding the end region of transmembrane helices. Rhodopsin and cytochrome *c* oxidase subunit III, whose amphiphilicity plots are shown in Figure 4, contain seven transmembrane helices, but there is no sequence homology between them. The two proteins contain 14 transmembrane regions, 16 loop segments (including amino- and carboxyl-terminal segments) and 28 helix end regions. Two-thirds of the end regions of these proteins coincide with peak values of average amphiphilicity index

for strongly polar residues; only a third of the total number of central and loop regions contain such peaks. Similarly, more than half of the end regions are associated with peak values of average amphiphilicity for weakly polar residues. Thus, peak values of amino acid amphiphilicity are frequently observed at polar–non-polar interfaces of membrane proteins, indicating that the amphiphilicity index is useful for the prediction of transmembrane helices.

Finally, we devised a dispersion diagram of average amphiphilicity index $\langle A \rangle$ versus average hydropathy index $\langle H \rangle$, by which 1282 helical segments were analysed. Transmembrane helices were well discriminated from helices of soluble proteins by this diagram, as shown in Figure 6. This dispersion diagram is used as a basic algorithm in the second step of the SOSUI system (Hirokawa *et al.*, 1998). The full algorithm of the SOSUI system, including the selection of candidate segments, is described elsewhere.

ACKNOWLEDGEMENTS

We are grateful to Dr Sonoyama of Tokyo University of Agriculture and Technology for his contributions to our critical discussions. This study was partly supported by a grant-in-aid for special projects in genome science from the Ministry of Education, Sports, Science and Technology (Mombukagakusho).

REFERENCES

Braun,P. and von Heijne,G. (1999) The aromatic residues Trp and Phe have different effects on the positioning of a transmembrane helix in the microsomal membrane. *Biochemistry*, **38**, 9778–9782.

Brown,T.A. (1999) *Genomes*. J & L Composition, Filey, North Yorkshire, UK.

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

Deisenhofer,J., Epp,O., Sinning,I. and Michel,H. (1985) Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3A resolution. *Nature*, **318**, 618–624.

Dayhoff,M.O., Hunt,L.T. and Hurst-Calderone,S. (1978) Composition of proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, **5** (Suppl. 3), p. 363.

Doyle,D.A., Morais,C.J., 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.

Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.

Engelman,D.M., Goldman,A. and Steitz,T.A. (1982) The identification of helical segments in the polypeptide chain of bacteriorhodopsin. *Meth. Enzymol.*, **88**, 81–88.

Girvin,M.E., Rastogi,V.K., Abildgaard,F., Markley,J.L. and Fillingame,R.H. (1998) Solution structure of the transmembrane H+-transporting subunit c of the F1F0 ATP synthase. *Biochemistry*, **37**, 8817–8824.

Hirokawa,T., Seah,B.-C. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.

Hobohm,U., Scharf,M. and Sander,C. (1992) Selection of a representative set of structures from the Brookhaven protein data bank. *Protein Sci.*, **1**, 409–417.

Israelachvili,J.N. (1991) *Intermolecular and Surface Forces*, 2nd edn, Academic, London.

Iverson,T.M., Luna-Chavez,C., Cecchini,G. and Rees,D.C. (1999) Structure of the *Escherichia coli* fumarate reductase respiratory complex. *Science*, **284**, 1961–1966.

Iwata,S., Lee,J.W., Okada,K., Lee,J.K., Iwata,M., Rasmussen,B., Link,T.A., Ramaswamy,S. and Jap,B.K. (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science*, **281**, 64–71.

Killian,J.A. and von Heijne,G. (2000) How proteins adapt to a membrane–water interface? *TIBS*, **25**, 429–434.

Kimura,Y., Vassylyev,D.G., Miyazawa,A., Kidera,A., Matsushima,M., Mitsuoka,K., Murata,K., Hirai,T. and Fujiyoshi,Y. (1997) Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature*, **389**, 206–211.

Kolbe,M., Besir,H., Essen,L.-O. and Oesterhelt,D. (2000) Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. *Science*, 1390–1396.

Koepke,J., Hu,X., Muenke,C., Schulten,K. and Michel,H. (1996) The crystal structure of the light-harvesting complex II (B800-850) from *Rhodospirillum molischianum*. *Structure*, **4**, 581–597.

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

Lancaster,C.R., Kroger,A., Auer,M. and Michel,H. (1999) Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 A resolution. *Nature*, **402**, 377–385.

McDermott,G., Prince,S.M., Freer,A.A., Hawthornthwaite-Lawless,A.M., Papiz,M.Z., Cogdell,R.J. and Isaacs,N.W. (1995) Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature*, **374**, 517–521.

Mitaku,S. (1993) The role of hydrophobic interaction in phase transition and structure formation of lipid membrane and proteins. *Phase Transitions*, **45**, 137–155.

Mitaku,S. and Hirokawa,T. (1999) Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and length of proteins. *Protein Eng.*, **12**, 953–957.

Mitaku,S., Hoshi,S. and Kataoka,R. (1985) Spectral analysis of amino acid sequence. II. Characterization of α-helices by local periodicity. *J. Phys. Soc. Jpn.*, **54**, 2047–2054.

Mitaku,S., Ono,M., Hirokawa,T., Seah,B.-C. and Sonoyama,M. (1999) Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by a prediction system SOSUI. *Biophys. Chem.*, **82**, 165–171.

Möller,S., Kriventseva,E.V. and Apweiler,R. (2000) A collection of well characterized integral membrane proteins. *Bioinformatics*, **16**, 1159–1160.

Nakai,K., Kidera,A. and Kanehisa,M. (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, **2**, 93–100.

Palczewski,K., Kumasaka,T., Hori,T., Behnke,C.A., Motoshima,H., Fox,B.A., Le Trong,I., Teller,D.C., Okada,T., Stenkamp,R.E., Yamamoto,M. and Miyano,M. (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**, 739–745.

Reithmeier,R.A. (1995) Characterization and modeling of membrane proteins using sequence analysis. *Curr. Opin. Struct. Biol.*, **5**, 491–500.

Ridder,A.N.J.A., Morein,S., Stam,J.G., Kuhn,A., de Knuijff,B. and Killian,J.A. (2000) Analysis of the role of interfacial tryptophan residues in controlling the topology of membrane proteins. *Biochemistry*, **39**, 6521–6528.

Schiffer,M., Chang,C.H. and Stevens,F.J. (1992) The functions of tryptophan residues in membrane proteins. *Protein Eng.*, **5**, 213–214.

Segrest,J.P., De Loof,H., Bohlman,J.G., Brouillette,C.G. and Anantharamaiah,G.M. (1990) Amphipathic helix motif: classes and properties. *Proteins*, **8**, 103–117.

Soulimane,T., Buse,G., Bourenkov,G.P., Bartunik,H.D., Huber,R. and Than,M.E. (2000) Structure and mechanism of the aberrant ba(3)-cytochrome *c* oxidase from *Thermusthermophilus*. *EMBO J.*, **19**, 1766–1776.

Tanford,C. (1980) *The Hydrophobic Effect; Formation of Micelles and Biological Membrane*. Wiley, New York.

Toyoshima,C., Nakasako,M., Nomura,H. and Ogawa,H. (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature*, **405**, 647–655.

Tsukihara,T., Aoyama,H., Yamashita,E., Tomizaki,T., Yamaguchi,H., Shinzawa-Itoh,K., Nakashima,R., Yaono,R. and Yoshikawa,S. (1996) The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 A. *Science*, **272**, 1136–1144.

Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.

von Heijne,G. (1992) Membrane protein structure prediction hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.*, **225**, 487–494.

Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029–1038.

Wallin,E., Tsukihara,T., Yoshikawa,S., von Heijne,G. and Elofsson (1997) Architecture of helix bundle membrane proteins: an analysis of cytochrome *c* oxidase from bovine mitochondria. *Protein Sci.*, **6**, 808–815.