Protein Engineering vol.9 no.3 pp.253–263, 1996

Recognition of transmembrane α-helical segments with environmental profiles

Roman G.Efremov[1,2] and Gérard Vergoten[3]

[1]Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Ul. Miklukho-Maklaya, 16/10, Moscow V-437, 117871 GSP, Russia and [3]Université des Sciences et Technologies de Lille, Centre de Recherches et d'Etudes en Simulations et Modélisation Moléculaires (CRESIMM), Bâtiment C8, 59655 Villeneuve d'Ascq Cedex, France

[2]Present address: Université des Sciences et Technologies de Lille, Centre de Recherches et d'Etudes en Simulations et Modélisation Moléculaires (CRESIMM), UFR de Chimie, Bâtiment C8, 59655 Villeneuve d'Ascq Cedex, France

[1]To whom correspondence should be addressed

A method for assessing the environmental properties of membrane-spanning α-helical peptides in proteins has been proposed. The algorithm employs a set of environmental preference parameters derived for amino acid residues based on the analysis of the 3-D structures of membrane domains in bacteriorhodopsin and photoreaction centers *Rhodopseudomonas viridis* and *Rhodobacter sphaeroides*. The resulting 3-D–1-D scores for transmembrane segments are significantly different from those derived for α-helices in globular proteins. The parameters obtained have been used to construct environmental profiles for membrane α-helices in bacteriorhodopsin and photoreaction centers. The profiles successfully recognize their own sequences in several specially designed large databases. The method has been applied to several membrane proteins with unknown spatial structures. Most of their membrane-spanning peptides were efficiently recognized by the profiles. The predicted environment of the residues in the membrane segments fits the experimental data well. The approach is independent of any homology data and can be employed to delineate the membrane segments of a protein with environmental characteristics close to those of bacteriorhodopsin and photoreaction centers. The alignment of these segments with the reference profiles provides a considerable amount of data about their lipid and protein exposure.

Key words: hydrophobic organization/integral membrane proteins/molecular modeling/structure prediction/3-D profile method

Introduction

The prediction of the spatial organization of membrane domains in a protein sequence remains one of the most challenging problems in the field of structural biology. Despite the large number of membrane protein sequences gathered to date, only a few 3-D structures are known to an atomic resolution (for a review see, for example, von Heijne, 1994). This provokes interest in the computer-aided molecular modeling of membrane domains. Of the computational techniques developed for the analysis of the amino acid sequences of membrane proteins, those used most commonly are: (i) methods designed for the identification of putative transmembrane segments (TMS); (ii) algorithms for the assessment of hydrophobic properties and/or probable lipid and protein exposure; and (iii) procedures for the prediction of the spatial arrangement of TMS in membrane bundles.

The first group of techniques includes numerous hydropathy plotting methods (reviewed in Degli Esposito *et al.*, 1990). Usually, the simultaneous application of several such techniques permits a reasonably accurate determination of the TMS boundaries. However, these algorithms give almost no way of deducing the secondary structure of membrane-spanning peptides and their orientation with respect to the bilayer. Often, those hydrophobic or amphiphilic stretches long enough to span the membrane (20–25 residues) are assigned to TM α-helices, while the 3-D structures of membrane moieties in porins (Schulz, 1993) and acetylcholine receptor (Unwin, 1993) reveal a β-barrel architecture.

Additional structural information can be obtained from the treatment of the periodicity of amino acid hydrophobicity (Eisenberg *et al.*, 1982) or variability (Rees *et al.*, 1989; Donnelly *et al.*, 1993; Du and Alkorta, 1994). This approach has been used widely to assess the tendency for a given TMS to form an α-helix or a β-strand and to delineate the most hydrophobic (hydrophilic) and/or variable (conservative) motifs in the TMS sequence. The results obtained can be used as constraints in subsequent molecular modeling studies of the membrane bundles. Despite its predictive power, several shortcomings are inherent with the method. Thus, in the absence of a prominent periodicity in the distribution of the polarity and/or the variability properties of the TMS, the method fails to distinguish between helical and strand conformations, giving similar indices of periodicity for each. In some cases, the calculated dispositions of the most hydrophobic and hydrophilic faces, as well as the orientation of the hydrophobic moment vector, do not reflect all the polarity properties of the segment (Cronet *et al.*, 1993; Du and Alkorta, 1994; Efremov and Vergoten, 1995, 1996). The variability profile method is a powerful tool in the recognition of lipid-accessible and buried residues but it only works for multiple sequence alignments and cannot be applied if no homologous proteins are known.

In summary, only a simultaneous analysis of the membrane protein sequence carried out using different techniques is able to provide a reasonable basis for future molecular modeling studies of the 3-D structure of the membrane domain. This calls for the development of additional independent algorithms for processing the sequences of integral membrane proteins.

Here we propose an alternative approach to assessing the environmental properties and probable packing of α-helical TMS in membrane bundles. Our method is based on the concept of environmental profiles applied successfully to globular proteins (Bowie *et al.*, 1991; Luthy *et al.*, 1992). The main concept of this 3-D profile method is to characterize the environments of residues in known 3-D structures into different categories (e.g. the residue surface area covered by polar/

© Oxford University Press

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

nonpolar atoms, the buried surface area and the local secondary structure) and to develop a scoring table of residues likely to be in a particular environmental class. The characteristics of the residue environment derived from the analysis of the 3-D model are converted into an environmental profile:1-D string of environmental classes. The next step is to estimate the correspondence between a given environmental profile and amino acid sequences in a database. The best-scoring sequences are assumed to share a common 3-D fold with a reference structure.

The main goals pursued here were: (i) to develop 3-D–1-D scores for TM α-helices; (ii) to construct environmental profiles for known TMS in bacteriorhodopsin (BRh) and photoreaction centers; (iii) to test these profiles by screening the databases of sequence fragments containing those from photoreaction centers and BRh; and (iv) to check the environmental profile method by assessing the hydrophobic properties of TM α-helices in several unrelated proteins for which there are experimental constraints on the mutual arrangement of the TMS.

Materials and methods

Environmental parameters for amino acid residues

The coordinates of the TM α-helices in BRh and photoreaction centers (the training set of proteins) were taken from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977; entries 1BRD, 1PRC and 4RCR). Two environmental characteristics were defined for each residue: *AL* was the total surface area accessible to lipid, and *AP* represented the surface area of the side chain covered by polar protein atoms and internal water. The *AL* values were calculated using the program DSSP (Kabsch and Sander, 1983) for whole membrane bundles with internal cavities filled with water. The *AP* parameters were obtained as follows:

$$AP = (F \times ASA) - AL,$$

where F is the fraction of the side chain covered by polar protein atoms and internal and external solvent, and *ASA* is the total accessible surface area for a given type of residue. The F values were calculated using the program Profiles_3D (Biosym Technologies, 1994). In accordance with the values of *AL* and *AP*, all residues in the TMS of BRh and photoreaction centers were attributed to one of seven environmental classes (Figure 1). The score S_{ij} for residue i to be in environmental class j was calculated according to the formula:

$$S_{ij} = \ln\{[P(i:j)/P(i)] + k\},$$

where $P(i:j)$ is the probability of finding residue i in environmental class j, and $P(i)$ is the probability of finding residue i in any environment in the database. Because of the restricted number of residues available for the analysis (698 residues for 29 TMS), we introduced a parameter k to account for zero values of $P(i:j)$ (no residues of type i in environment j):

$$k = 0.0, \text{ if } P(i:j) \neq 0,$$
$$k = 0.1, \text{ if } P(i:j) = 0.$$

The value $k = 0.1$ was chosen empirically to reduce the scores for a few types of charged residues not found in the membrane parts of reference proteins. The need to introduce k originates from the relatively small total number of residues. Boundaries for the environmental categories (Figure 1) were adjusted to maximize the total score for all residues in the database.

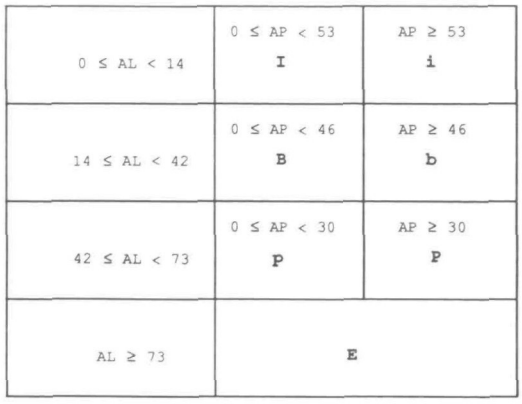

Fig. 1. Definition of environmental classes for residues in TM α-helical segments. Two environmental characteristics were defined for each residue: (i) *AL*, the surface area accessible to lipids; and (ii) *AP*, the surface area covered by polar protein atoms and internal water. Environmental classes: E, preferentially exposed to lipids; P and p, significant lipid exposure, nonpolar and polar protein environment, respectively; B and b, small lipid exposure, nonpolar and polar protein environment, respectively; I and i, buried inside the bundle, nonpolar and polar protein environment, respectively.

Table I. Additional position-dependent scores for residues in the termini of transmembrane segments

Residue	Position in the transmembrane segment[a]		
	±9	±10	±11
Ala	70	50	80
Leu	150	120	100
Ile	–	50	50
Pro	–	–	100
Trp	–	60	50

[a]Residue position is counted from a central residue towards the N- and/or C-termini.

The scores S_{ij} (3-D–1-D scores) were used to generate environmental profiles for all α-helical TMS in BRh and photoreaction centers. The format of the profiles was similar to that used in the program Profiles_3D. In each profile we also accounted for the preferential distribution of several types of residue in the terminal parts of the TMS (Persson and Argos, 1994). First, we calculated the frequencies of occurrence of residues in positions 9–13 when counting from the central residue in the TMS. Then, depending on the frequencies obtained, we added empirically chosen values to the 3-D–1-D scores for some flanking residues which were most often found on the termini (Table I). The environmental profiles based on the 3-D–1-D scores derived from the analysis of globular proteins (Bowie *et al.*, 1991) were computed using the program Profiles_3D with a supplied table of parameters.

Sequence databases

Resulting environmental profiles were used in a compatibility search against several databases of sequence fragments as follows. Each protein sequence was subdivided into fragments 40 residues in length, overlapping by intervals of 15 residues. The length of the last segment was always ⩾40 residues. The resulting sets of fragments were combined in the final database. Such a procedure permits the presence of any 25-residue segment in the database; therefore, all putative TM regions are included. The search for those sequences which correspond

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

well to each of the environmental profiles was performed using a dynamic programming algorithm as it is implemented in the Profiles_3D program running under the Insight II/Homology molecular modeling package (Biosym Technologies, 1994). The alignments of environmental profiles and TMS sequences were considered as significant if their Z scores were >4.0, the fitted length was more than nine residues and the TMS were among the first five best-scoring sequences. The following databases were set up:

(i) GLOB, GLOB-PRC, GLOB-RCR and GLOB-BRD: all included fragments extracted from sequences of a non-redundant set of 185 globular proteins (Abagyan and Totrov, 1994). In addition, the last three databases contained the fragments extracted from the sequences of one of the membrane proteins: photoreaction center of *Rhodopseudomonas viridis* (PRC), photoreaction center of *Rhodobacter sphaeroides* (RCR) and BRh, respectively. The total numbers of fragments in the databases were 2086, 2136, 2135 and 2101, respectively.

(ii) Six sequence databases constructed for the following membrane proteins not included into the training set: lactose permease (Lac) of *Escherichia coli*, coat protein in bacteriophage M13 (cpM13), leader peptidase (Lep) of *E.coli*, aspartate chemoreceptor protein (Tar) of *E.coli*, bovine rhodopsin (Rh) and human glycophorin A (GpA). The 40-residue fragments of each were added to the database GLOB.

The environmental profiles for TM α-helices PRC M3 and M4 have been tested on two corresponding 'minus-one' systems created by removing these segments, respectively, from the original whole database of TMS. The two sets of 3-D–1-D scores were developed as described above, and the environmental profiles for the two helices were constructed based on these scores. Resulting environmental profiles were employed to screen the sequence database GLOB-PRC.

Those interested in obtaining profiles for the TM α-helices as well as the databases of sequence fragments should contact R.G.Efremov by e-mail (roma@pop.univ-lille1.fr).

Results and discussion

Development of 3-D–1-D parameters for residues

Residue environmental classes in TM α-helices. Seven environmental classes defined for residues in TM α-helices taken from known protein structures (Figure 1) can be characterized as follows: (i) class E, preferentially exposed to lipids; (ii) class P, significantly exposed to lipids, nonpolar environment in protein-exposed part of the surface; (iii) class p, significantly exposed to lipids, polar protein environment; (iv) and (v) classes B and b, small lipid exposure, nonpolar and polar protein surroundings, respectively; and (vi) and (vii) classes I and i, buried inside the bundle, nonpolar and polar protein environment, respectively.

As mentioned before, the boundaries of the environmental categories were adjusted to maximize the total score over the database of TM helices. Such calculations were performed for five, six, seven and eight classes. The largest total score per residue was achieved if seven environmental categories were defined. No improvement of the score was obtained when we used corresponding fractions of *ASA* instead of *AL* and *AP*. In addition, because there are grounds to believe that the hydrophobic organization of photoreaction centers and BRh is different (Efremov and Vergoten, 1996), an attempt was made to calculate the 3-D–1-D scores based only on TMS from photoreaction centers (without BRh). However, screening the sequence databases with such environmental profiles did not lead to a better global correspondence between the environmental strings of TM helices and their own amino acid sequences (data not shown).

In several cases, no residues were found in a particular class: e.g. Gln in classes p, P, B and i, Arg in classes P, B and i, Asp was found only in classes i and I, whereas Lys was found only in classes p and i. The main reasons for this were the low preference for charged and strongly polar residues to be in the TMS and the restricted size of the database of known structures. To extend the database, an attempt was made to include TMS sequences from proteins homologous to BRh and photoreaction centers. It was assumed that the environmental parameters of residues in aligned positions are the same as in reference structures. In such a case, the number of residues in the database was 1203 but the maximal total score per residue was lower than that obtained for the original set of 29 TM α-helices (data not shown).

How adequately do the sets of 3-D–1-D scores for TM helices reflect their environmental properties? What is the feasibility of the method? In the subsequent discussion, we will try to answer these important questions. To assess the validity of the parameter set, we used two criteria: (i) the parameter set should demonstrate significant differences in its description of the hydrophobic properties of residues in TM α-helices in comparison with those in globular proteins (it is reasonable to expect the most prominent differences for surface-exposed and deeply buried residues in both classes of protein); and (ii) the environmental profiles constructed on the basis of the 3-D–1-D scores for the TMS should efficiently match their amino acid sequences.

Comparison of 3-D–1-D scoring tables for globular and membrane α-helices. To check the first criterion, we compared the environmental parameters derived for TM α-helices with those obtained for globular proteins (Bowie *et al.*, 1991). The corresponding 3-D–1-D scores are given in Table II. An analysis of the pairwise correlation coefficients (r) between data in columns of this table reveals a most prominent anticorrelation for pairs E_m–E_g (m refers to TM helices, g refers to helices in globular proteins) and I_m–$B3_g$, with r values of –0.80 and –0.72, respectively. The highest values of r were obtained for pairs b_m–$B2_g$ (0.79), E_m–$B2_g$ (0.72), E_m–$B1_g$ (0.71), b_m–$B3_g$ (0.69) and I_m–$P1_g$ (0.67). Therefore, those residues which tend to be exposed on the surface in globular proteins are preferentially buried from lipids inside the membrane bundles, and vice versa. In addition, those residues which are often found in contact with a bilayer in TM helices (class E_m) can be also be found inside the globule in hydrophobic (class $B1_g$) or moderately polar (class $B2_g$) environments. Significant correlations between the 3-D–1-D parameters for several classes of buried (or partially buried) residues in globular and membrane proteins permit us to conclude that the polarity characteristics in their interiors may be similar.

The results described above confirm the observation that the membrane proteins reveal an 'inside out' hydrophobic organization in comparison with globular proteins (Rees *et al.*, 1989), whereas their internal polarity and packing properties are similar. Therefore, comparison of the 3-D–1-D scores for residues in the TM and globular α-helices shows that the parameter set matches criterion (i) described above.

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

Table II. Environmental parameters (3-D–1-D scores) of amino acid residues for different environmental classes in transmembrane and globular α-helices

Residue	Transmembrane α-helices[a]							Globular α-helices[b]					
	E	p	P	b	B	i	I	E	P2	P1	B2	B1	B3
W	83	86	−120	−49	−481	32	−522	−129	86	101	111	−109	−126
F	92	50	−174	46	−126	−21	−138	−85	−22	87	128	−135	−181
Y	127	−386	−63	−61	−424	−7	−465	−88	50	86	27	−55	−170
L	37	37	15	0	9	−23	−78	−30	16	71	130	−46	−137
I	20	46	−8	39	−1	−30	−52	−6	−2	55	111	−59	−236
V	−27	21	30	60	8	−24	−32	30	−29	41	74	−62	−125
M	−39	61	−6	86	0	30	−178	−42	87	102	126	−27	−90
A	−208	−159	25	−203	37	−80	81	76	−44	−65	−77	−2	44
G	−542	−193	−469	−467	24	−114	105	−46	−109	−204	−222	−58	63
P	−410	−360	−37	−335	61	87	−70	−41	−111	−97	−156	−25	5
C	−375	−26	−2	−300	−64	−356	74	95	−138	15	−43	−70	−17
T	−107	−427	57	−101	42	−19	42	39	−69	−67	−172	−13	−20
S	−203	−153	89	−428	56	−23	31	47	−101	−133	−243	−38	16
Q	−6	−325	−301	109	−363	−356	4	−32	16	16	−138	62	29
N	−366	−317	6	−291	13	90	−396	−58	−7	−48	−176	−2	32
E	−366	−17	6	7	−355	90	−27	−43	9	−58	−215	62	60
D	−346	−297	−273	−271	−335	81	−76	−28	−43	−80	−248	29	44
H	−403	−55	38	39	−93	108	−64	−91	61	82	−34	17	−6
K	−346	2	−273	−271	−335	81	−375	−50	56	−94	−137	66	7
R	−131	−12	−357	81	−419	126	−460	−51	110	−11	−180	56	−20

[a]This work.
[b]Parameters are taken from the 3-D–1-D table obtained for globular proteins (Bowie *et al.*, 1991).

Table III. Screening the sequence databases[a] with environmental profiles[b]

N	Profile	Scores for TM helices[c]					Scores for globular helices[d]				
		I	Best-scoring segments[e]	L^f	Z score	g	I	Best-scoring segments[e]	L^f	Z score	g
1	PRC L-1	1	PRC L-1	21	17.93	+	1	1TAB 4–23	20	4.61	−
		2	_h			−	2	MADH 2–19	18	4.15	−
		3	_h			−	3	4GR1 34–47	14	4.07	−
2	PRC L-2	1	3CHY 95–115	21	6.81	−	1	PRC L-2	26	11.08	+
		2	2ER6 176–195	20	5.01	−	2	PRC L-2	17	7.70	+
		3	PRC L-2	24	4.78	+	3	2SCP 22–37	16	5.55	−
3	PRC L-3	1	PRC L-3	21	9.34	+	1	8CPP 245–264	20	6.73	−
		2	2ER6 250–263	14	5.39	−	2	1GOX 60–76	17	4.80	−
		3	PRC L-3	18	5.30	+	3	1LFG 560–576	17	4.77	−
4	PRC L-4	1	PRC L-4	19	8.84	+	1	8ADH 268–281	14	5.54	−
		2	1LH2 88–105	18	4.57	−	2	4PFK 301–318	18	4.90	−
		3	3RP2 116–132	17	4.28	−	3	PRCC 104–119	16	4.71	−
5	PRC L-5	1	3ENL 344–366	23	7.14	−	1	3CRO 7–27	21	5.59	−
		2	2SOD 100–117	18	7.10	−	2	1OVA 287–307	21	5.08	−
		3	PRC L-5	22	6.01	+	3	2PAL 10–24	15	4.99	−
6	PRC M-1	1	PRC M-1	22	7.53	+	1	4P2P 78–98	21	6.35	−
		2	PRC L-5	18	5.18	−	2	5TNC 4–21	18	5.55	−
		3	MADL 42–59	18	4.99	−	3	7ICD 272–288	17	4.40	−
7	PRC M-2	1	PRC M-2	26	8.72	+	1	PRC M-2	27	10.17	+
		2	5RUB 284–307	24	4.81	−	2	PRC M-2	18	5.94	+
		3	PRC M-2	20	4.31	+	3	1ECD 114–134	21	4.85	−
8	PRC M-3	1	PRC M-3	23	13.87	+	1	6TIM 112–134	23	6.65	−
		2	1MVP 74–96	23	7.51	−	2	PRC M-3	17	5.85	+
		3	PRC M-2	15	6.16	−	3	3BCL 51–70	20	5.60	−
9	PRC M-4	1	PRC M-4	19	9.03	+	1	PRC M-4	18	5.70	+
		2	HSAA 13–31	19	5.78	−	2	8CPP 256–272	17	4.86	−
		3	5ACN 78–92	15	4.30	−	3	2FCR 97–112	16	3.96	−
10	PRC M-5	1	1PHH 162–177	16	7.00	−	1	PRC M-5	23	12.97	+
		2	2YHX 47–61	15	6.55	−	2	1BBQ 50–70	21	5.32	−
		3	1NRD 19–35	17	5.90	−	3	2YHX 360–383	23	4.79	−
11	PRC H	1	PRC H	20	12.43	+	1	2TS1	19	4.83	−
		2	PRC H	17	7.49	+	2	1PII 237–252	16	4.15	−
		3	PRC M-2	17	2.72	−	3	1LPE 119–132	14	3.91	−
12	RCR L-1	1	RCR L-1	24	6.65	+	1	1MBO 71–90	20	6.28	−
		2	RCR M-2	17	4.01	−	2	1GOX 153–171	19	5.36	−
		3	RCR M-2	15	2.66	−	3	256B 35–55	21	4.66	−

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

Table III *Continued*

N	Profile	Scores for TM helices[c]					Scores for globular helices[d]				
		I	Best-scoring segments[e]	L[f]	Z score	g	I	Best-scoring segments[e]	L[f]	Z score	g
13	RCR L-2	1	RCR L-2	25	10.32	+	1	RCR L-2	27	8.35	+
		2	RCR L-2	17	6.40	+	2	RCR L-2	18	7.26	+
		3	1COX 72–86	15	5.31	–	3	1SGT 12–32	21	4.68	–
14	RCR L-3	1	RCR L-3	22	9.34	+	1	1CSC 136–151	16	5.71	–
		2	2GBP 224–242	19	6.72	–	2	RCR L-1	17	5.62	–
		3	2ER6 123–139	17	5.45	–	3	1COL 133–150	18	4.81	–
15	RCR L-4	1	RCR L-4	19	9.81	+	1	2RN2 41–55	15	4.69	–
		2	RCR M-1	15	5.56	–	2	2LHB 108–124	17	4.61	–
		3	3BLM 209–224	16	4.52	–	3	1LTT 17–32	16	4.61	–
16	RCR L-5	1	RCR L-5	23	7.80	+	1	RCR L-5	24	7.56	+
		2	1HNE 58–77	20	6.01	–	2	3CRO 42–56	15	6.24	–
		3	1IPD 254–275	22	5.65	–	3	RCR L-5	22	6.03	+
17	RCR M-1	1	RCR M-1	24	16.50	+	1	3ADK 137–156	20	5.24	–
		2	RCR M-1	18	10.41	+	2	1CTF 49–66	18	4.98	–
		3	RCR M-1	15	8.49	+	3	3BLM 141–164	24	4.80	–
18	RCR M-2	1	RCR M-2	29	19.67	+	1	RCR M-2	28	8.86	+
		2	RCR M-2	17	11.51	+	2	6CPA 278–304	27	7.18	–
		3	RCR M-2	22	9.95	+	3	6TIM 37–53	17	6.47	–
19	RCR M-3	1	RCR M-3	22	6.53	+	1	RCR M-3	16	7.88	+
		2	1FNR 194–211	18	4.56	–	2	RCRM 86–99	14	5.66	–
		3	1GMF 73–89	17	4.15	–	3	RCR M-3	14	5.20	+
20	RCR M-4	1	1HDS 109–225	17	5.44	–	1	RCR M-4	18	9.58	+
		2	1CRN 19–37	19	4.79	–	2	RCR L-5	17	7.52	–
		3	1BIA 143–161	19	4.38	–	3	MADH 202–217	16	4.91	–
		4	RCR M-4	16	3.57	+					
21	RCR M-5	1	RCR M-5	19	7.44	+	1	RCR M-5	19	10.86	+
		2	2YHX 88–101	14	5.20	–	2	5ACN 68–82	15	6.37	–
		3	1SGC 24–37	14	4.88	–	3	1GOX 211–229	19	5.56	–
22	RCR H	1	RCR H	26	13.29	+	1	1CLA 197–212	16	5.27	–
		2	RCR H	22	8.54	+	2	5ACN 571–591	21	4.78	–
		3	RCR M-2	21	4.04	–	3	1LH2 11–34	24	4.72	–
23	BRD A	1	1NPX 5–20	16	5.24	–	1	256B 1–18	19	5.41	–
		2	BRD A	19	4.76	+	2	6XIA 312–328	17	4.58	–
		3	1RNB 1–14	14	4.34	–	3	2SCP 34–51	18	4.55	–
24	BRD B	1	BRD B	20	5.97	+	1	1IPD 162–182	21	6.90	–
		2	1LPE 4–17	14	4.96	–	2	HSAA 241–260	20	5.31	–
		3	1LFG 123–138	16	5.04	–	3	8CPP 256–276	21	5.27	–
25	BRD C	1	8CPP 88–104	17	4.99	–	1	BRD 6	21	5.31	–
		2	1-DTX 9–22	14	4.85	–	2	1GOX 238–256	19	4.39	–
		3	1PII 34–51	18	4.18	–	3	1GLY 308–322	15	4.25	–
26	BRD D	1	BRD D	20	5.50	+	1	2LHB 61–75	15	5.93	–
		2	1LMB 62–78	17	5.40	–	2	MADL 15–29	15	5.86	–
		3	3ENL 179–193	15	4.31	–	3	3CNA 92–105	14	5.62	–
27	BRD E	1	1SGC 116–126	11	3.92	–	1	1GD1 150–168	19	5.32	–
		2	–[h]			–	2	2SDH 132–146	15	5.12	–
		3	–[h]			–	3	2FCR 96–109	14	5.10	–
28	BRD F	1	1HDS 125–138	14	3.97	–	1	1GOX 268–283	16	5.77	–
		2	1LH2 1–15	15	3.52	–	2	3ADK 177–194	18	5.30	–
		3	1BIA 240–253	14	3.26	–	3	1LFG 263–279	17	5.27	–
29	BRD G	1	1IPD 193–209	17	7.38	–	1	7ICD 233–249	17	5.72	–
		2	BRD G	22	7.38	+	2	1BBQ 158–175	18	5.16	–
		3	1CLA 180–194	15	7.33	–	3	1SGC 163–177	15	4.85	–

[a]The databases GLOB-BRD, GLOB-PRC and GLOB-RCR of sequence fragments were screened with the environmental profiles of bacteriorhodopsin (BRD) and photoreaction centers *R.viridis* (PRC) and *R.sphaeroides* (RCR), respectively.
[b]The environmental profiles correspond to TM α-helices in PRC, RCR and BRD.
[c]The profiles were generated using a 3-D–1-D table developed here.
[d]The profiles were generated using the 3-D–1-D table of Bowie *et al.* (1991).
[e]Only three best-scoring fragments >14 residues in length were considered (except the results for BRD E, for which only short segments were found).
[f]Length of a sequence fragment.
[g]The sign '+' (or '–') means that a given sequence fragment corresponds (or does not correspond) to the sequence of the profile.
[h]No segments with $L > 14$ were fitted.

Screening the sequence databases containing fragments from photoreaction centers and BRh with environmental profiles. To test how the set of 3-D–1-D scores satisfies criterion (ii) (see above), we have constructed environmental profiles for α-helical TMS in BRh and two photoreaction centers. These profiles were then used for screening the databases of sequence fragments. If the environmental profiles efficiently recognize corresponding TMS sequences, the set of 3-D–1-D parameters

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

can be used in other applications to search for similar environmental templates in membrane proteins with unknown structures.

The procedure was carried out for two sets of environmental profiles: one obtained in our work and the other developed for globular proteins (Bowie *et al.*, 1991). Corresponding 'membrane' and 'globular' profiles were screened against similar databases (GLOB-PRC, GLOB-RCR and GLOB-BRD). The results are presented in Table III. It can be seen that the 'membrane' environmental profiles recognize their amino acid sequences rather better than the 'globular' profiles. Only four (PRC M5, BRD C, BRD E and BRD F) out of 29 environmental strings did not have a good *Z* score (i.e. from among three to four best-scoring segments); in the other 25 cases, a corresponding sequence was found from among three best-scoring peptides. In contrast, the profiles for globular α-helices were able to fit their own sequences with a high *Z* score in only 11 cases. Note that three out of seven membrane-spanning helices in BRh were not recognized in the profiles, whereas only one TMS was not fitted amongst 22 helices in the photoreaction centers. This is consistent with a proposal that the hydrophobic organization of the membrane helix bundle in BRh is significantly different (at least for some TMS) from that in the photoreaction centers (Efremov and Vergoten, 1996).

Interestingly, for TMS PRC L2, M5 and RCR M4, the 'globular' profiles complement the 'membrane' profiles with higher *Z* scores. Therefore, the polarity properties of residues in these segments are closer to those in globular proteins. It is reasonable to expect that this should be the case for TMS screened from lipids by the neighboring helices. We propose that both types of environmental profile should be used to evaluate the hydrophobic characteristics of the putative TMS in new membrane proteins: the 'membrane' profiles can recognize the helices in contact with lipids, whereas the 'globular' profiles can be used to identify those helices buried deep inside the bundle. It should be noted that this approach permits the recognition of TM α-helical peptides whose sequences correspond to that of one of the environmental strings in the proteins of the training set (photoreaction centers and BRh). This is why the main aim of the method is to distinguish those segments in a protein sequence that have environmental properties close to those observed in the known membrane helix bundles.

In this respect it is important to check the redundancy of the set of environmental profiles for TM helices because of the similar spatial organization of the membrane domains in the photoreaction centers. A pairwise comparison of the profiles reveals a high degree of similarity in the following pairs: PRC H/RCR H, PRC L2/RCR L2, PRC M5/RCR M4, PRC M1/RCR L1, PRC L3/RCR M3 and BRD A/BRD E (Figure 2). Hence, a preliminary search in the sequence database can be carried out using only one profile from these pairs. For a more precise fitting, both profiles should be employed because the homologous regions do not span the whole segment lengths (Figure 2). All the homologous profiles reveal a higher percentage identity for the environmental strings than for the sequences (Figure 2). Therefore, the polarity properties of the evolutionarily close bacterial photoreaction centers are conserved rather better than their sequences. It seems reasonable to propose that such stability of the hydrophobic organization of the membrane bundles makes it possible

```
          198  PWHGPSIGFAYGCGLLFAA
PRC M-4     1  BiPBbIpBbIEIIIEBiiI          S-ID = 0.63
               | ||| | | ||||||||           P-ID = 0.74
RCR M-4     1  BIPBbBpPbBEIIIEBiiB
            200  PFHGLSIAFLYGSALLFAM

           12  IAQLVWYAQWLVIWTVVLLY
PRC H       1  BIbpbpEPIEEEIEPBBEEE         S-ID = 0.20
               |    | ||||     ||           P-ID = 0.40
RCR H       1  EIBEBbEBIEEEBPBbpEEp
           12  LASLAIYSFWIFLAGLIYYL

           84  GFWQAITVCALGAFISWMLREVEISR
PRC L-2     1  IIibBIIEIBBIIEPIpBPiIIiiIb  S-ID = 0.73
               ||||  |||||||||  ||| |||     P-ID = 0.77
RCR L-2     2  IIibpBIEIBBIIEPIEIPiIBiiiE
           84  GLWQIITICATGAFVSWALREVEICR

          116  HVPLAFCVPIFMFCVLQVFRPLL
PRC L-3     1  bIiEIpPbPIEbbBIBIBEiiEP     S-ID = 0.35
               |||||  ||| | ||| |||         P-ID = 0.65
RCR L-3     1  iIiEIpBpPIEBbbIBIPEiiPE
          116  HIPFAFAFAILAYLTLVLFRPVM

           10  WIWLALGTALMGLGTLYFLV
BRD-A       1  EEEibEIIPpiIEIBEPiEE         S-ID = 0.05
               ||   ||    |  | | |          P-ID = 0.40
BRD-E       2  pEEIPPIIEEIBEiIEEibE
          138  WAISTAAMLYILYVLFFGFT

           52  GASGIAAFAFGSTAILIILFNM
PRC M-1     1  IEBIEBIEppIPPIEEBBPEBi      S-ID = 0.32
               |||| | | |   ||||  ||        P-ID = 0.59
RCR L-1     1  IEBIPBBEEpPBEIEEBbpEBI
           32  GFFGVATFFFAALGIILIAWSA

          118  PLAFCVPIFMFCVLQVFRPLL
PRC L-3     3  iEIpPbPIEbbBIBIBEiiEP       S-ID = 0.38
               || | ||| | |||  |||          P-ID = 0.62
RCR M-3     1  IEIbPIPIEpbPIBIIPiiEE
          145  AWAFLSAIWLWMVLGFIRPIL
```

Fig. 2. Correspondence between environmental profiles of TM helices. A definition of the environmental classes is given in the legend to Figure 1. The 'I' indicates an identical environment. Numbering corresponds to the residue position in a sequence and in a profile. S-ID and P-ID are the similarity indices for TMS sequences and profiles, respectively.

for the photoreaction centers to keep a common fold in the membrane, which is of prime importance for their function.

An additional test to confirm the validity of the 3-D–1-D parameters derived for the TMS was performed on two 'minus-one' systems, i.e. with TMS PRC M3 and M4 removed from the training set. These TMS were chosen because their profiles do not reveal a prominent correlation with the other environmental profiles. The 3-D–1-D scoring tables were calculated for both 'minus-one' systems, and the new environmental profiles for PRC M3 and M4 were constructed. The profiles recognized their own sequences in the database GLOB-PRC with the highest *Z* scores (6.54 and 5.17 for PRC M3 and M4, respectively). In addition, the exclusion of each of the profiles from the training set only resulted in minor changes of the optimal boundaries between environmental classes. This confirms the validity of the set of 3-D–1-D scores developed here.

To see how the environmental profiles locate their sequences in the presence of numerous other membrane fragments, we used the profiles to screen the database containing (in addition to database GLOB) the sequence segments extracted from membrane proteins of different classes: photoreaction centers, G protein-coupled receptors (GPCRs), P-type ATPases, ligand-gated ion channels, etc. In total, the database contained >2000 sequence segments from 75 membrane proteins (plus ~2000 globular sequence fragments from GLOB). The screening results show that only four TMS in BRh and two in the photoreaction centers were not recognized efficiently by the profiles. Therefore, the addition of a large number of TMS to the sequence database does not significantly affect the

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

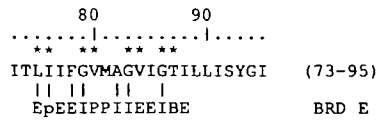

Fig. 3. Alignment of the amino acid sequence of the TMS in GpA with the environmental profile of TM α-helix E in BRh (BRD E). The residues marked with a '*' form the helix–helix interface in the GpA dimer (Lemmon and Engelman, 1992). The symbols 'I', ':' and '.' indicate a strong, significant and moderate correspondence between the residue and the aligned environmental class. The numbering is that corresponding to the residue position in a sequence.

search results compared with those when only one membrane protein sequence was included.

Application of the environmental profiles

An obvious question arises: how efficiently can the profiles be used to find membrane proteins of unknown structure? To provide an answer to this problem, we have applied the environmental profile approach to a number of proteins not included in the training set. The proteins were chosen because the structures of their membrane domains had been studied intensively using different techniques, and molecular models of their membrane moieties had been proposed. This provided a considerable amount of evidence concerning the mutual arrangement of at least some of the TMS and made such systems appropriate for testing using the environmental profile method. In the absence of numerous examples of 3-D models of membrane bundles, this is believed to be a reasonable approach.

The sequence databases containing 40-residue fragments extracted from these proteins were screened with the environmental profiles. The resulting sequence-profile alignments were analyzed and compared with the available data on the arrangement of the TMS. Such an analysis is the subject of the subsequent discussion.

Choice of the sequence databases. We used the following criteria to construct the databases: (i) there should be a large enough number of sequence fragments to provide statistical significance of the search results; and (ii) the database should include sequences from both globular and membrane proteins. Such criteria were proposed in the belief that the typical practical application of the environmental profiles would be to estimate, for a given sequence of a membrane protein, the putative lipid and protein exposures in the TMS. To do this, the whole sequence (or only the sequences of the TMS, if they are well established) of this protein is included in the database, which also contains a large number of sequence segments from globular proteins.

Human glycophorin A (GpA). The mutual orientation of the TM helices in the human GpA dimer was investigated based on the results of site-directed mutagenesis and simulated annealing techniques. A molecular model of the helix–helix interaction inside the membrane has been proposed previously (Lemmon and Engelman, 1992). Screening of the sequence database with the membrane 3-D profiles reveals a high degree of compatibility between the amino acid sequence in the TMS of GpA and the string of environmental parameters in TM helix E of BRh (Figure 3). As seen in Figure 3, TMS GpA is aligned with the 3-D profile BRD E on a 14-residue fragment. Note that this is the only segment in GpA which was found to be among the best-scoring sequences in the database.

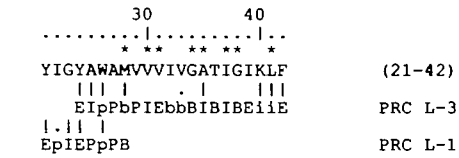

Fig. 4. Sequence profile alignment for the TMS in the M13 coat protein. PRC L-3 and L-1, environmental profiles of TM α-helices L-3 and L-1 in reaction center *R.viridis*. The details are as in the legend to Figure 3.

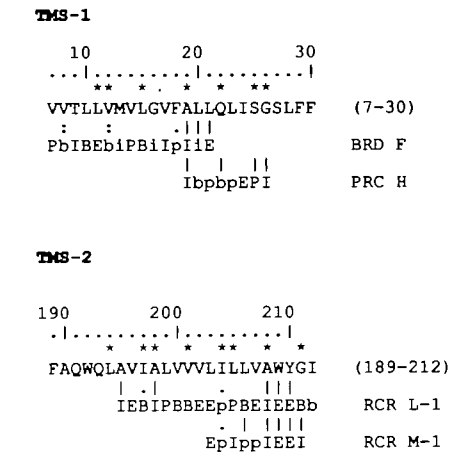

Fig. 5. Sequence-profile alignment for the TMS in the Tar receptor. The profiles are: BRD F, TM helix F in BRh; PRC H, TM helix H in reaction center *R.viridis*; and RCR L-1 and RCR M-1, TM helices L-1 and M-1 in reaction center *R.sphaeroides*. The details are as in the legend to Figure 3.

According to the results of Lemmon and Engelman (1992 and references therein), the helix–helix interface is formed by residues Leu75, Ile76, Gly79, Val80, Gly83, Val84, Gly86 and Thr87. These residues are aligned with environmental classes E, p, I, P, I, E, I and B, respectively (Figure 3). Ile77, Ile85 and Ile88, assigned to the lipid-exposed face of the TM helix in the model of Lemmon and Engelman (1992), are compatible with class E in the 3-D profile.

Comparison of the results obtained here with the experimentally derived topology of the membrane domain of the GpA dimer demonstrates a reasonable degree of agreement between the two techniques. In general, the spatial hydrophobic template of TM helix E in BRh correctly fits the residues in the membrane segment of GpA. The discrepancies are observed for residues Leu75 and Val84: both are assessed as buried in the experiment but as lipid-exposed in the environmental profile approach. At the same time, as follows from the experimentally derived model, these two residues are positioned on the edges of the interface region and are therefore partially accessible to lipids.

M13 coat protein. The helix association in the TM domain of phage M13 coat protein was assessed by Deber *et al.* (1993) using Val → Ala mutations in the hydrophobic segment. As presented in Figure 4, a single membrane-crossing α-helix in the M13 coat protein was recognized efficiently by the environmental profiles PRC L-1 and PRC L-3. As in the case of GpA, no other peptides of the coat protein could be fitted to the 3-D profiles. This means that the extramembrane sequences in these proteins do not reveal compatibility in their polarity properties with the TM α-helices from the reference set. Therefore, the proposed approach can also be used for the

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

identification of putative membrane-embedded helices in a protein sequence.

The molecular model of TM helical dimer proposed by Deber *et al.* (1993) reveals the following residues lying on the protein-interactive face of the TMS: Met28, Val30, Val31, Gly34, Ala35, Ile37, Gly38 and Leu41. According to the sequence-profile alignment (Figure 4), all these residues, except Val31, are totally or significantly shielded from lipids: they are highly compatible with the environmental classes b, I, B, I, I, B and i, respectively. Such categories of residue polarity correspond to amino acids which cannot access the bilayer (classes I and i) or reveal weak contacts with the lipids (classes B and b).

Hence, there is reasonable agreement between the spatial distribution of those residues assessed using the 3-D profiles and those estimated from the mutational data. At the same time, we should note that for putative lipid-exposed residues, the correspondence is somewhat worse. Thus, if the lipid-interactive residues Val29 and Ile39 in the model of Deber *et al.* (1993) were assigned correctly in our method to classes P and E, respectively, some other residues (Ile32, Val33 and Thr36) were not.

E.coli aspartate chemoreceptor (Tar). A molecular model of the Tar intramembrane domain was proposed by Pakula and Simon (1992) based on a mutational analysis. Its membrane part is supposed to be a four-helix bundle formed following the dimerization of two monomers, each containing two TM helices. The TM-1 segments of the protomers interact quite extensively with each other. The TM-2 helices seem to be rather less involved in associating with each other, but interact strongly with the TM-1 helix of the same subunit.

TMS-1 in the Tar receptor was recognized by the environmental profiles BRD F and PRC H, whereas TMS-2 was recognized by the profiles RCR L-1 and RCR M-1 (Figure 5). Most of the residues involved in helix–helix interactions in the model of Pakula and Simon (1992; Val12, Leu15, Ala19, Gln22 and Gly26) are aligned with environmental classes b, B, I, b and I, respectively. The only exception is Leu11, which was predicted as exposed to lipids but was assigned to be on the interface of helices TM-1, TM-1′ and TM-2′ in the model. In addition, a partial lipid exposure was predicted for Ser25, but it was placed between TMS-1 and TMS-2 in the model. There was a good correlation between the lipid-oriented disposition of Leu21, Leu23 and Leu24 in both approaches. At the same time, the intrinsic location of Met13, Gly16 and Leu20, aligned with environmental class i or b, is inconsistent with their arrangement in the model.

A significantly better fit was obtained for the TM-2 helix. All the residues on its lipid-facing side in the model were aligned mainly with environmental class E (Leu199 and Leu206 were aligned with classes P and p/B, respectively). In contrast, all the internal residues in the helix bundle (Ala195, Ile197, Ala198, Val201, Leu205, Ala208 and Gly211) were fitted to polarity class I or B. Note that such a good alignment was observed on the long 18-residue segment (for the RCR L-1 profile). Moreover, the two environmental profiles matching TMS-2 are aligned perfectly relative to each other. In this case, there are strong grounds to believe that the environmental properties and 3-D disposition of helix TM-2 are quite similar to those in TM helices L-1 and M-1 in the reaction center of *R.sphaeroides*.

E.coli leader peptidase (Lep). An interface between two TM

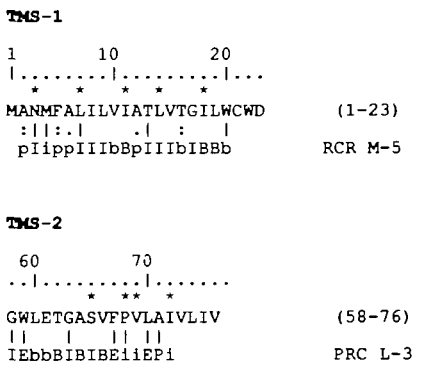

Fig. 6. Sequence-profile alignments for the TMS in the leader peptidase. The profiles are: RCR M-5, TM helix M-5 in reaction center *R.sphaeroides*; and PRC L-3, TM helix L-3 in reaction center *R.viridis*. The details are as in the legend to Figure 3.

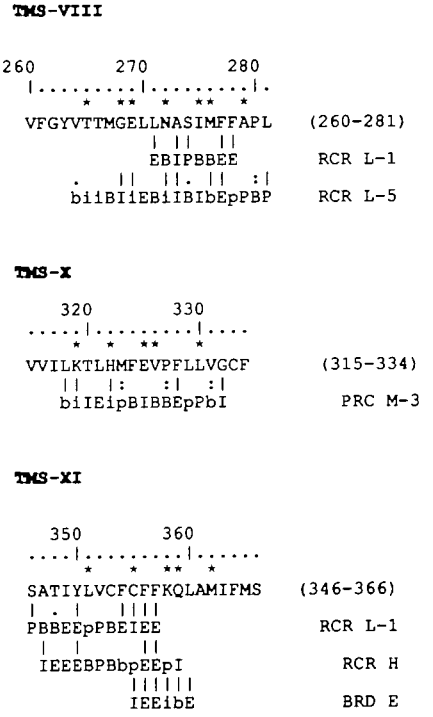

Fig. 7. Sequence-profile alignments for TMS VIII, X and XI in lactose permease. The profiles are: BRD E, TM helix E in BRh; PRC M-3, TM helix M-3 in reaction center *R.viridis*; RCR L-1, L-5, H, TM helices L-1, L-5 and H in reaction center *R.sphaeroides*. The details are as in the legend to Figure 3.

α-helices in the membrane domain of Lep was established by disulfide mapping, and a model of the TM helix–helix hairpin was built by Whitley *et al.* (1993). Curiously, the interface was found to be formed primarily by nonpolar residues which could not be predicted on the basis of a helical amphiphilicity analysis. This is why it was interesting to determine whether it is possible to distinguish between the lipid- and protein-facing sides of the TMS in Lep using our set of environmental profiles.

Figure 6 shows the alignment of the TM sequences in Lep with the well-scoring 3-D profiles RCR M-5 and PRC L-3. It can be seen that residues Asn3, Leu7, Leu14 and Ile18 in TMS-1, which were proposed to be on the helix–helix interface in the model of Whitley *et al.* (1993), are fitted to environmental class i, I or B, and are therefore correctly recognized by the

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

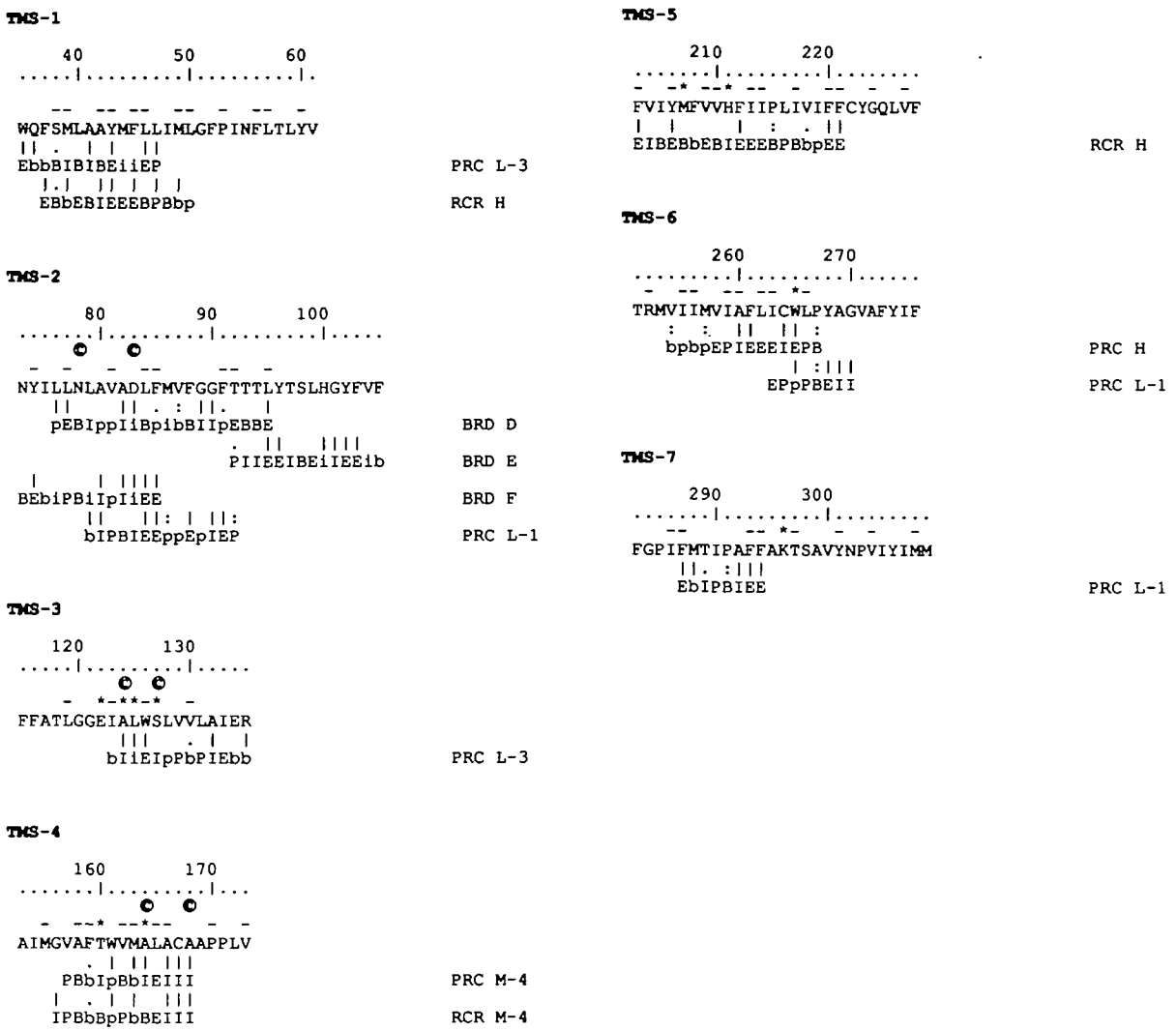

Fig. 8. Sequence-profile alignments for TMS in rhodopsin. Residues in contact with the retinal chromophore are marked with a '*', whereas those exposed to lipids in the 3-D models of GPCR are labeled with a '–'. The symbols indicate polarity-conserved positions in the GPCR sequences (Zhang and Weinstein, 1994). Other details are as in the legend to Figure 3.

profile RCR M-5. At the same time, 'buried' environmental parameters were also attributed to several residues (Ala2, Ala6, Ala12, Thr13, Thr16 and Gly17) placed on the lipid-exposed face in the model. Such a discrepancy can be explained by a rather substantial accessibility of the TM helix RCR M-5 because of its contacts with three neighboring helices in the membrane bundle (Deisenhofer *et al.*, 1985).

The sequence-profile alignment for the second TMS (Figure 6) reveals a good agreement with the Cys-scanning mutagenesis data (Whitley *et al.*, 1993). All the residues lying on the helix–helix interface in the model, namely Ser65, Pro68, Val69 and Ile72, are fitted to environmental category I or i. In contrast, the lipid-exposed residues in the model (Phe67, Leu70 and Ala71) correspond to class E or P. Hence, both the protein- and lipid-interacting faces of this TMS were recognized perfectly in our method.

E.coli lactose permease (Lac). The intramembrane domain of Lac is considered to contain 12 α-helices. The mutual orientation of four TMS, namely VII, VIII, X and XI, has been proposed in the result of Cys-scanning mutagenesis (Frillingos *et al.*, 1994). The residues in helices VIII, X and XI, which were found to be on the interhelical interfaces, are shown in Figure 7. Figure 7 also displays environmental profiles which reveal a strong fit to the amino acid sequences of these TMS. We did not succeed in aligning helix VII with any of the profiles from the reference set. This means that its spatial hydrophobic properties are quite different from those of TM helices in BRh and reaction centers. Thus, the TMS can be shielded completely from the bilayer by the neighboring peptides.

As viewed in Figure 7, almost all the residues positioned in the protein interior in the model of Frillingos *et al.* (1994) were aligned with environmental categories i, I, b and B. The exceptions were Ala279 (helix VIII) and Leu330 (helix X) fitted to the P class, and therefore partially accessible to lipids. Thus, the protein-interacting sides of helices VIII, X and XI were assessed correctly by the 3-D profiles. Unfortunately, at the moment we cannot verify the reliability of our method to recognize the lipid exposure of this four-helix bundle because the orientation of the other eight TMS in Lac has not been established experimentally as yet. For the following residues we predict it is highly probable that they are in contact with a bilayer: Leu270, Leu271, Phe277, Leu281 (helix VIII),

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

Leu321, Phe328 (helix X), Ile349, Tyr350, Phe356, Phe357 and Leu360 (helix XI). All of these residues were assigned to environmental class E. The validity of such a prediction can be tested when the mutational data for the remaining TM helices in Lac becomes available.

Visual rhodopsin (Rh). Rh belongs to a superfamily of GPCRs, integral membrane proteins characterized by seven TMS (for a review see, for example, Donnelly and Findlay, 1994). Numerous experimental and homology modeling studies of GPCRs have allowed the identification of functionally important residues inside the membrane helix bundle. The results of an environmental profile search for Rh are presented in Figure 8. In all, 15 sequence fragments of Rh were found among the best-scoring segments. Except for peptide 92–105, which includes a terminal part of TM helix B and a short extra-membrane loop, all the other fragments correspond to membrane-spanning segments. The residues known to be buried in the vicinity of the retinal chromophore and those proposed to be in contact with a bilayer are indicated. These data were derived from several spatial models of GPCRs (Baldwin, 1993; Donnelly and Findlay, 1994, and references therein; Du and Alkorta, 1994). The polarity-conserved positions identified by Zhang and Weinstein (1994), based on the multiple alignment of GPCR sequences, are also marked.

As seen in Figure 8, of eight residues known to be in the retinal microenvironment, only Trp265 was fitted to the incorrect polarity class E or p, whereas all others were aligned to class I, i or B. All of the polarity-conserved positions correlate with environmental category I or B which is peculiar to the polar protein environment inside the helix bundle. Residues positioned on the lipid-facing sides of TM helices in 3-D models of Rh are attributed mainly to class E or P, but exceptions occur in TMS-1 (Ser38 and Met39), TMS-3 (Ile123 and Val130), TMS-4 (Phe159 and Ala166) and TMS-6 (Ile256). The best sequence-profile fit results were obtained for TMS-2, TMS-5, TMS-6 and TMS-7, although for TMS-7 only a short eight-residue fragment was aligned. At the same time, the lipid exposure in helices 1 and 3 was assessed to be somewhat worse.

Conclusions

The set of environmental parameters for residues in TM α-helical segments developed here was shown to be a powerful tool in assessing the spatial hydrophobic characteristics of membrane-spanning helices. As distinct from the majority of other techniques dedicated to solving this problem, our approach does not use any information related to residue variability and can be used in cases when no homologous sequences are found.

Almost all (25 out of 29) of the environmental profiles from the reference set efficiently recognize their own amino acid sequences in the large databases. The 3-D–1-D scores obtained for residues in TM α-helices are quite distinct from those in globular proteins. The largest differences were observed for deeply buried and solvent-exposed residues in both protein types, whereas for 'intermediate' environmental classes the hydrophobic and packing properties were rather similar. The spatial hydrophobic properties of the TMS can be estimated using both 'membrane' and 'globular' profiles. Membrane profiles characterize the segments and reveal their exposure to lipids, whereas globular profiles recognize those helices buried deep inside the membrane bundles. In the absence of feasible ways of predicting the 3-D structure of the membrane protein domain from the amino acid sequence alone, the environmental profile approach provides an additional insight into the problem.

The efficiency of the environmental profile method was tested on several membrane helix bundles. The results show that the environmental profiles of reference TM α-helices have the ability to distinguish helix–helix interfaces and lipid-exposed stretches in most cases. In addition, the approach made it possible to predict the environmental characteristics of some TM helices which had not been studied experimentally. We hope that these data may be useful in the design of further experiments and the modeling of such membrane moieties. The 'membrane' 3-D profiles efficiently recognize TMS in proteins containing four or more helices in the membrane bundle, whereas the lipid exposure of residues in small, e.g. two-helix, TM domains (e.g. GpA, cpM13 and Lep) is somewhat worse. This is caused by the multi-helical organization of the membrane moieties in the BRh and reaction centers which were used to derive the set of 3-D–1-D parameters.

Several shortcomings are inherent in the method. The most important is the limited number of different spatial/hydrophobic templates that can be derived from the 3-D structures of membrane protein domains. This is why in some instances a number of TMS in a new protein sequence will not be fitted by the current environmental profiles. On the other hand, even if only some TM helices can be characterized using such a technique, this will apply strong constraints on their possible mutual arrangement. Another problem is that in some cases the TMS sequences in the proteins under study were aligned with the profiles only for short intervals (~10 residues) and thus revealed a local polarity pattern, whereas the complete hydrophobic template is not susceptible to this sort of estimation.

As any prediction algorithm has only a limited degree of accuracy, it is important to use the environmental profile approach together with complementary techniques, such as variability profiles, etc. Further improvement of the method can be envisaged. Thus, the appearance of new 3-D structures of membrane proteins will permit us to update the current table of 3-D–1-D scores. Moreover, this will extend the set of environmental motifs for TM helices which is vitally important in the assessment of new membrane helix bundles. Our current studies are also aimed at obtaining environmental templates for membrane-spanning β-strands and applying the environmental profile method to the detection of the degree of similarity between environmental patterns in TMS from different proteins with unknown structure.

Acknowledgements

R.G.E. was the recipient of an invited professor fellowship from the University of Sciences and Technologies of Lille, France.

References

Abagyan,R.A. and Totrov,M.V. (1994) *J. Mol. Biol.*, **235**, 983–1002.
Baldwin,J.M. (1993) *EMBO J.*, **12**, 1693–1703.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Biosym Technologies (1994) *Insight II User Guide*. Biosym Technologies, San Diego, CA, Version 2.3.5.
Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Cronet,P., Sander,C. and Vriend,G. (1993) *Protein Engng*, **6**, 59–64.
Deber,C.M., Khan,A.R., Li,Z., Joensson,C. and Glibowicka,M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 11648–11652.
Degli Esposti,M., Crimi,M. and Venturoli,G. (1990) *Eur. J. Biochem.*, **190**, 207–219.

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024

Deisenhofer,J., Epp,O., Miki,K., Huber,R. and Michel,H. (1985) *Nature*, **318**, 618–624.
Donnelly,D. and Findlay,B.C. (1994) *Curr. Opin. Struct. Biol.*, **4**, 582–589.
Donnelly,D., Overington,J.P., Ruffle,S.V., Nugent,J.H.A. and Blundell,T.L. (1993) *Protein Sci.*, **2**, 55–70.
Du,P. and Alkorta,I. (1994) *Protein Engng*, **7**, 1221–1229.
Efremov,R.G. and Vergoten,G. (1995) *J. Phys. Chem.*, **99**, 10658–10666.
Efremov,R.G. and Vergoten,G. (1996) *J. Protein Chem.*, **15**, 63–76.
Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1982) *Nature*, **299**, 371–374.
Frillingos,S., Sahin-Toth,M., Persson,B. and Kaback,R. (1994) *Biochemistry*, **33**, 8074–8081.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Lemmon,M.A. and Engelman,D. (1992) *Quart. Rev. Biophys.*, **27**, 157–218.
Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) *Nature*, **356**, 83–85.
Pakula,A.A. and Simon,M.I. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 4144–4148.
Persson,B. and Argos,P. (1994) *J. Mol. Biol.*, **237**, 182–192.
Rees,D.C., DeAntonio,L. and Eisenberg,D. (1989) *Science*, **245**, 510–513.
Schulz,G.E. (1993) *Curr. Opin. Struct. Biol.*, **5**, 701–707.
Unwin,N. (1993) *J. Mol. Biol.*, **229**, 1101–1124.
von Heijne,G. (1994) *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167–192.
Whitley,P., Nilsson,L. and von Heijne,G. (1993) *Biochemistry*, **32**, 8534–8539.
Zhang,D. and Weinstein,H. (1994) *FEBS Lett.*, **337**, 207–212.

Received May 19, 1995; revised November 15, 1995; accepted November 22, 1995

Downloaded from https://academic.oup.com/peds/article/9/3/253/1502753 by guest on 23 April 2024