CROATICA CHEMICA ACTA CCACAA **72** (4) 975–997 (1999)

ISSN-0011-1643
CCA-2631

Original Scientific Paper

Sequence Analysis of Membrane Proteins with the Web Server SPLIT

Davor Juretić,[a,*] *Ana Jerončić,*[a] *and Damir Zucić* [b]

[a] *Physics Dept., Faculty of Natural Sciences, Mathematics and Education, University of Split, N. Tesle 12, HR–21000, Split, Croatia.*

[b] *Faculty of Electrical Engineering, University of Osijek, Istarska 3, HR–31000 Osijek, Croatia*

Received December 23, 1998; revised April 14, 1999; accepted April 19, 1999

In this work, recently solved crystal structures of membrane proteins are examined with respect to the performance of the Web server SPLIT in predicting sequence location, conformation and orientation of membrane associated polypeptide segments. The SPLIT predictor is based on the preference functions method. Preference functions serve to transform the input choice of amino acid attributes into sequence dependent conformational preferences. Transmembrane helical segments are accurately predicted with a good selection of preference functions extracted from the compiled database of non-homologous integral membrane proteins. Unlike other algorithms with similar high accuracy, the SPLIT predictor requires no homology information. With preference functions extracted from soluble proteins, the sequence location of shorter non-transmembrane helices can be also found in membrane proteins. In particular, Richardson's preference functions are even better than hydrophobic moments in finding interface helices at the water/lipid phase boundary. The Internet access for the SPLIT system is at the address: http://pref.etfos.hr/split

Key words: sequence analysis, membrane proteins, prediction, secondary structure, preference functions, transmembrane helix, interface helix, hydrophobic moments, antibacterial peptides

* Author to whom correspondence should be addressed. (E-mail: juretic@mapmf.pmfst.hr)

INTRODUCTION

Different genome projects result in daily addition of new genes and translated protein sequences with an ever increasing flow of genomic information and already significant impact on the world's economy.[1] Approximately 20 to 30% of protein sequences are expected to code for integral membrane proteins.[2] Sequence homology with solved crystal structure helps to model the 3D structure of the tested protein.[3] However, crystal structures of integral membrane proteins, known with high resolution, are still limited in number,[2] so that the degree of sequence homology is often too low to allow 3D modelling of a novel membrane protein sequence.

A more modest goal of sequence analysis is to determine the membrane-associated segments in integral membrane protein. One must answer the question where in the sequence are a) transmembrane segments, b) membrane buried but not membrane spanning segments, and c) surface attached interface segments. In the case of the first question, the answer is provided by algorithms that predict the sequence location of transmembrane segments expected to be in the α-helix conformation.[4–8] Additional information in the form of multiple sequence alignments is usually required for optimal performance.[5–8] Modern algorithms provide topology information also for certain classes of membrane proteins by predicting not only the sequence location of potential transmembrane helical segments, but also their orientation with respect to outer and inner membrane surfaces.[4,5,8]

No explicit prediction of the nature and secondary structure for different classes of membrane-associated segments is attempted by these algorithms. An improved predictor should be able to provide objective and accurate answers to these questions as well. This goal has not been reached yet, but in this work we discuss the capabilities of our Web server, which is versatile in dealing with the above mentioned questions and easy to use. For an operator using such a server it is important to understand its limitations as well as its advantages. We shall illustrate both aspects in the performance of the Web server SPLIT.[9–11]

The Web server SPLIT is very fast because a) it uses very simple preference functions[9,12] and hydrophobic moment functions[11] in its digital predictor, b) it uses the graphics library created by us to enable a fast graphical presentation of results, and c) it does not require multiple sequence alignments as additional information. Since homologous sequences to a novel sequence are often absent in databases of protein sequences, improvements in the speed and accuracy of single-sequence prediction are important. We have recently reported the SPLIT performance in predicting transmembrane helices (TMH) in the photosynthetic reaction center, light-harvesting

protein, cytochrome c oxidase and bc_1 mitochondrial complex, and in predicting membrane-buried but not transmembrane helices in some voltage gated channels.[9–11] In this work, four additional membrane proteins of a recently known structure are tested to learn the predictor's accuracy in predicting the sequence location of observed TMH. In addition, the performance in predicting the sequence location of interface helices, and of other membrane-bound regular structures is examined, and the practical mode of the server's operation is outlined. It is shown that the predictor based on preference functions can complement traditional methods in finding the sequence location of transmembrane and interface helices in integral membrane proteins.

MATERIALS AND METHODS

The Dataset of 31 Integral Membrane Polypeptides with Known Crystal Structure

Membrane polypeptides of known crystal structure are still few in number. Here we use the known structures of subunits H, L and M of the photosynthetic reaction center from *Rhodobacter viridis*[13,14] and from *Rhodobacter sphaeroides,*[15] the light-harvesting protein from *Rhodopseudomonas acidophila*[16,17] and plant light-harvesting protein from *Pisum sativum,*[18] subunits I, II and III of the cytochrome c oxidase from *Paracoccus denitrificans*[19] and subunits I, II, III, IV, VIa, VIc, VIIa, VIIb, VIIc and VIII of the cytochrome c oxidase from bovine heart,[20] bacteriorhodopsin from *Halobacterium salinarium,*[21–23] subunits from beef heart mitochondrial bc_1 complex: 7, 10, 11, cytochrome b, cytochrome c_1, and Rieske protein,[24–27] glycophorin A from human erythrocytes,[28] potassium channel from *Streptomyces lividans,*[29] and ATP synthase subunit c from *Escherichia coli.*[30] Except for the bacteriorhodopsin and glycophorin, the listed polypeptides were **not** seen before by the PREF algorithm[9] during the training procedure. These 31 sequences contained a total of 100 transmembrane helices with 2761 residues in the TMH conformation. Published TMH assignments were used.

Selected 22 Interface Helices

The membrane surface positioned helices were considered to be interface helices. Such helices were selected among non-transmembrane helices from the database of integral membrane polypeptides with known crystal structure (see above). Program RASMOL[31] was used for molecular visualization. It is possible to color amino acids visualized by RASMOL according

to the temperature factor. A small utility program was written to replace experimental temperature factors by hydrophobicity values, based on the Kyte-Doolittle hydropathy scale.[32] A constant value was added to each hydrophobicity, to bring them into a positive range. All values were then multiplied by the same constant factor, so that the final range was from 0 to 90, which is suitable for RASMOL. After coloring the proteins according to the hydrophobicity of side chains, it was possible to determine the approximate position of both membrane interfaces separating the solvent from the lipid phase. Potential interface helices were also visualized with RASMOL and identified with the STRIDE program[33] for secondary structure assignment of known structures. The candidate interface helices were hand-picked according to the following criteria: 1) the center of mass distance from the membrane should not exceed 0.5 nm, 2) there should be no other polypeptide chain between an interface helix and a membrane (but transmembrane helices are regarded as the integral part of a membrane), and 3) the angle between the helix axis and membrane surface should not exceed 50 degrees.

Secondary structure conformation and the segment length of selected segments were in accord with the published assignment in papers where the corresponding high-resolution crystal structures first appeared. We found 50% of selected interface helices in two related photosynthetic reaction center complexes from bacteria. These interface helices are helices cd (149–165) and e (258–268) from subunit L of *Rhodobacter sphaeroides*, helices cd (152–162) and ect (259–267) from subunit L of *Rhodobacter viridis*, helices ab (81–89), cd (178–194) and e (293–302) from subunit M of *Rhodobacter sphaeroides*, and helices ab (81–87), cd (179–190), de' (232–237) and ect (292–298) from subunit M of *Rhodobacter viridis*. Remaining interface helices are helix D (201–210) of the plant light-harvesting complex, helix 39–46 of light-harvesting protein from *Rhodopseudomonas acidophila*, helices 1–7 and 361–367 from subunit I, helix 112–125 from subunit IV, and helix 5–13 from subunit VIIa of the mitochondrial cytochrome c oxidase, helices a (11–20), ab (64–71), cd_1 (138–147), and cd_2 (156–166) from cytochrome b, and helix 4–15 of subunit 10, also from the bovine mitochondrial bc_1 complex.

The SPLIT 3.5 Algorithm

The definition of preference functions and the training part of the procedure leading to extraction of preference functions has been described before.[9,10] It will be only briefly outlined here. The training dataset of 100 non-homologous membrane and soluble proteins contained incompletely known membrane proteins, non-homologous to the testing dataset of membrane proteins.[9] For each amino acid residue, in each sequence, its type, secondary structure and sequence environment were collected. Sequence environment

of a residue was calculated as an average of five left and five right attributes (such as hydrophobicity) of its neighbors. Histograms of sequence environments for all residues were approximated with Gaussian functions.

Conformational preference function for conformation '*j*' of the amino acid type '*i*' found within sequence environments *X* was then defined as:

$$P_{ij}(X) = \frac{(N / N_j)(N_{ij} / \sigma_{ij}) \exp\left[-\left(X - \mu_{ij}\right)^2 / 2\sigma_{ij}^2\right]}{\sum_k \left(N_{ik}/\sigma_{ik}\right) \exp\left[-\left(X - \mu_{ik}\right)^2 / 2\sigma_{ik}^2\right]} \quad (1)$$

where N_j / N is the fraction of conformation '*j*' in the protein dataset, N_{ij} is the number of amino acids found in each conformation, μ_{ij} is the average and σ_{ij} is the sample standard deviation of parameters *X*.

The SPLIT 3.5 algorithm[11] consists of transforming, predicting, filtering and refining modules. By means of preference functions, it first transforms the input choice of amino acid parameters into sequence dependent conformational preferences. A total of 88 scales of amino acid attributes is available on the server's home page with relevant references. Some of these scales are for 20 constant conformational preferences, but in the following text, whenever preferences are mentioned, it is assumed that these values are already transformed sequence dependent preferences.

The predictor part of the algorithm compares preferences for α-helix, β-sheet, turn and undefined conformation at each sequence position and assigns the appropriate secondary structure to the highest preference. Predicted TMH segments are result of the filtering procedure, which rejects too short and splits too long predicted helical segments.

Other conformational profiles are also used to refine the prediction. Ends of the observed TMH are often associated with a raising β-sheet and turn preferences. SPLIT extends the predicted TMH span when the sum of alpha and beta preferences is high (2.0), and stops the extension when a high turn preference (>1.3) is encountered.

High hydrophobic moments[34] are often encountered at TMH termini as well. Hydrophobic moments are calculated at each sequence position i and for each twist angle in the range from 80 to 180 degrees. Hydrophobic moment index, defined as a five times hydrophobic moment, is reported for two standard conformations: α-helix with a 100 degrees twist angle , and β-sheet with a 180 degree twist angle. The hydrophobic moment function $I(k,i)$ is defined as in our recent publication:[11]

$$I(k,i) = 6\mu(k,i) \exp(-(\mu(i)_{\max} - \mu(k,i))^2) \exp(-(\delta(i)_{\text{opt}} - \delta(k,i))^2) \quad (2)$$

where $\mu(i)_{max}$ and $\delta(i)_{opt}$ are the maximal hydrophobic moment and the corresponding optimal twist angle, respectively, while $\mu(k,i)$ and $\delta(k,i)$ are the hydrophobic moment for standard 'k' conformation and the corresponding twist angle, respectively. In the profiles of $I(k)$ values, produced by the server in the numerical output, the average of three values is associated with the central residue in the triplet and denoted as the hydrophobic moment threshold index $I_3(k)$. For $I_3(k) > 2.0$ at TMH termini, the predicted TMH span is also extended. When $I_3(k)$ is very high (>3.5) in the middle of the predicted span, the potential TMH segment is reexamined for the maximal height of α-helix preferences, and rejected if such maximum is less than 2.6.

An extra scale input option enables the predictor to use Richardson's middle helix preferences[35] and the corresponding preference functions, extracted from the database of soluble proteins,[11] for the prediction of interface and extramembrane helices. Sequence dependent Richardson's preferences are denoted as free helix preferences, and are utilized to extend the TMH span when high enough (>1.3).

The prediction accuracy parameter A_{TM} for residues in the TMH structure takes into account the overpredicted o_{TM}, underpredicted u_{TM} and the observed number N_{TM} of residues found in the TMH structure:

$$A_{TM} = (N_{TM} - o_{TM} - u_{TM}) / N_{TM} . \quad (3)$$

Per-segment prediction accuracy is also estimated by using equation (3) when the number of overpredicted and underpredicted TMH segments is known.

Interface helices (see above) were considered predicted when the hydrophobic moment index or the hydrophobic moment threshold index had their maximum equal or higher than 2.0 anywhere along the span of the observed interface helical segment. Positive correct prediction of interface helices with Richardson's preferences occurred when the maximum equal or higher than 0.9 was found inside such observed segments. Correct prediction of β-strand segment was scored when the corresponding preference maximum equal or greater than the threshold value of 1.4 was found along the span of the observed β-strand. The product of transmembrane helix preferences and turn preferences had to be higher than 2.0 to indicate the sequence position of helical ends for helices entering or exiting from the membrane.

The SPLIT Web Server

The original prediction programs,[9–11] written in FORTRAN 77, were wrapped into a modular web server, written in HTML, ANSI C and unix

script language. An independent and portable graphics library was created to enable graphical presentation of the results. The only required input is the protein sequence. Server's speed (predicted conformational profiles are received in seconds) and versatility (many different hydrophobicity scales[36] can be used to calculate the hydrophobic moment[34] and preference profiles) allows easy computer experiments in predicting the secondary structure. The server is accessible at: http://pref.etfos.hr/split

Recommended Amino Acid Attribute Scales and Conformational Profiles

The default choice of scales for operating the server are the Kyte-Doolittle hydropathy scale[32] for calculating conformational preference profiles and the Eisenberg consensus hydrophobicity scale[37] for calculating hydrophobic moments. The same two lists of 88 scales are available for the calculation of preferences and for the calculation of hydrophobic moments, but the rank orders of the scales differ. The default choice of scale is at the top position for each of the two lists. If not specified otherwise, all results presented in this paper have been obtained with the SPLIT 3.5 algorithm version and the above mentioned default choice of amino acid attributes. Notice, however, that the default choice of scales is the most common choice, but not the best choice. For instance, Edelman's scale[38] for calculating conformational preferences[11] and Cornette's PRIFT scale[36] for calculating hydrophobic moments may be used to improve the predictor's performance. All scales except default scales are listed from the top position according to their performance in predicting membrane-spanning segments (first list) and in predicting the sequence location of amphipathic interface helices (second list). An extra scale option (the Richardson scale)[35] can be chosen as the third choice of scales when one wishes to predict the sequence location of interface and extramembrane helices as well as to improve the prediction accuracy for the termini of membrane-spanning helices. Correlation between any two scales can be quickly determined by using the SCACOR routine of the server.

A total of 13 different conformational profiles is available in the Numeric Data Output of the server. Their meaning is described in the SPLIT35 – Output Description. In addition to the three plotted profiles, relevant profiles for the present work can be found as columns 10 (membrane-buried helix times turn preference), 11 to 14 (hydrophobic moment and hydrophobic moment index), and 17 as the last column (Richardson preferences for "free" α-helix when the extra scale option is used).

RESULTS

Performance Tests on Membrane Spanning Helices in Integral Membrane Polypeptides of Known Structure.

All of the 100 observed sequence locations for transmembrane helices (Methods) are associated with α-helix preference maximums. Maximums in the TMH preferences range from 4.75 to 2.40, while maximums in the free helix preferences (Richardson preferences) range from 2.69 to 1.01. Most of TMH preference profiles have only one clear maximum, while free helix preference profiles often exhibit more than one maximum in the sequence region where TMH is observed. Overall per-residue prediction accuracy is clearly improved when both kinds of preference profiles are used in the SPLIT predictor. As measured by our accuracy parameter A_{TM} (Methods), the performance increases from 0.69 (when only the Kyte-Doolittle scale is used) to 0.73 (when Richardson's scale is used as an extra scale too), and to 0.77 (when Edelman's scale is used in combination with the Richardson's scale). The corresponding percentage of correctly predicted TMH residues raises from 76 to 83 and to 85%. Increased per-residue prediction accuracy is gained due to better balance between the underpredicted and overpredicted residues. Per-segment prediction accuracy (Eq. (3)) is high (0.96) for the default choice of amino acid attributes including Richardson preferences, because only one out of 100 TMH is underpredicted and three TMH are overpredicted. For instance, one TMH is ovepredicted, while another is underpredicted in the Rieske protein.[11] Overpredicted TMH in the Rieske protein is the example when a corresponding free helix maximum could not be found in the predicted TMH region. Overpredicted TMH in the cytochrome b corresponds to the sequence location of two surface attached amphipathic helices cd_1 and cd_2. It is rejected as a TMH by the SPLIT algorithm when the PRIFT scale[35] is used (instead of Eisenberg's consensus hydrophobicity scale)[37] to calculate the profile of hydrophobic moments. Another overpredicted TMH in the bovine cytochrome oxidase subunit 1 is not associated with the maximum for the membrane-buried helix within the middle region of the preference profile.

It is also of interest to test separately bacteriorhodopsin, glycophorin A, bacterial potassium channel and ATP synthase subunit c, because no detailed structural knowledge for these polypeptides was available to us when the SPLIT 3.5 predictor was constructed.[9–11] All 12 observed TMH from these four polypeptides are correctly predicted with no overpredictions. Out of 305 amino acid residues observed in the TMH conformation only 19 are underpredicted and 32 are overpredicted with the default scale choice (including Richardson's scale), so that the accuracy parameter is very high $A_{TM} = 0.833$.

Interface Helices

Interface residues in the α-helix configuration are often found at the *N*- or *C*-terminus of membrane spanning helices. Since such segments are often amphipathic, the calculation of hydrophobic moments may be used to achieve a modest increase in the accuracy of TMH prediction.[11] As expected, the TMH prediction accuracy, reported in the 4-th column of part B in Table I, does not vary much when different amino acid attributes are used for the calculation of hydrophobic moments. The best result is achieved with scale 59 that we introduced in an earlier work.[39]

When interface helices are not fused with membrane-spanning helices, it is still of interest to predict their sequence location. A standard set of 22 interface helices is collected from known structures of 31 integral membrane polypeptides (see Methods). This database of helices, oriented approximately parallel and positioned very close to the inner or outer membrane surface, is used to test the performance of different conformational indexes. Since amphipathicity is commonly used for such a purpose, we first created the predictor for amphipathic segments and compared the performance of all 88 amino acid attribute scales available on the server. Our index $I_3(\alpha)$, which locates sequence segments with optimal hydrophobic moments,[11] has a better performance than the hydrophobic moment index itself in finding interface helices for 56 different cases (scales). It gives the same result for 24 scales, and is worse for 8 scales. In all but one of 43 cases (scales) with best performance, our index $I_3(\alpha)$ performes as well or better than the hydrophobic moment (Table I). As the predictor for sequence location of interface helices by means of $I_3(\alpha)$ and/or hydrophobic moment, the Eisenberg consensus hydrophobicity scale[37] comes only 25-th in the rank order of performance. All interface helices are predicted when all 88 scales are considered, but no scale predicts more than 14 out of 22 helices. Two interface helices from the bovine cytochrome oxidase subunit I are predicted only by the Kuhn & Leigh membrane propensity scale (# 43 scale).

All of 22 the interface helices are associated with the maximum in Richardson's α-helix preferences. However, the predictor based on the Richardson preference functions (see Methods) does not see the short interface helix ab in the cytochrome b of the bovine bc_1 complex. It also does not predict the short interface helix 361–367 in subunit I of the cytochrome c oxidase from bovine heart. Reasons for these underpredictions differ. In the case of cytochrome b, the maximum in Richardson's preferences along α-helix strech 64–71 is slightly smaller than the chosen threshold value of 0.90. In the case of subunit I, the TMH predictor used Richardson's preferences to extend the *N*-terminal region of predicted TMH so that the interface helix 361–367 is fused with TMH. The existence of a maximum in Richardson's

TABLE I

Prediction of the sequence position for 22 interface helices. Each row in Table I represents one computer experiment with our SPLIT predictor applied to 31 integral membrane polypeptides. Interface helices are predicted with Richardson's preference functions in section A. Values higher than the threshold value of 2.0 for the hydrophobic moment index (H.M.) and for the hydrophobic moment threshold index $I_3(\alpha)$ are used to predict the sequence position of interface helices in section B, and the best 43 amino acid scales are selected among 88 available scales. The Kyte-Doolittle preference functions and Richardson preference functions are applied in each case to predict the sequence position of transmembrane helices as well. The prediction accuracy for the TMH residues is given in the fourth column as the A_{TM} parameter (Eq. 3).

Scales rank order	# helices detected		performance in TMH prediction	Amino Acid Scale Code / Name
A)	20			(60) RICH, Richardson preferences
B)	$I_3(\alpha)$	H.M.	A_{TM}	
1	14	12	0.737	(17) PONG1, Ponnuswamy hydrophobicity
2	14	9	0.724	(69) MATPO, mean rms fluctational disp. F1
3	13	13	0.737	(27) PRIFT, optimal amphipathic helices
4	13	13	0.725	(79) MARTI, single TMH preferences
5	13	13	0.718	(43) KUHLE, Kuhn membrane propensity
6	13	11	0.737	(66) CHOU6, helix preferences α/β prot.
7	13	7	0.736	(15) CIDA+ , hydrophobicity scale α+β prot.
8	12	12	0.735	(44) DEBER, M/A ratio in membrane prot.
9	12	12	0.725	(52) EDE25, Edelman optimal predictors
10	12	12	0.725	(41) ZAMYA, increase in volume of water
11	12	11	0.729	(3) PONNU, Ponnuswamy hydrophobicity
12	12	11	0.725	(51) EDE31, Edelman optimal predictors
13	12	10	0.732	(32) SWEET, optimal matching hydrop. scale
14	12	10	0.725	(07) GUY-M, average of 4 hydroph. scales
15	12	10	0.723	(22) WOLFE, Wolfeden hydrophobicity scale
16	12	9	0.735	(42) MIJER, average contact energy
17	12	9	0.734	(6) JONES, Jones hydrophobicity scale
18	12	8	0.727	(11) LEVIT, Levitt hydrophobicity scale
19	12	8	0.724	(31) GUYFE, Guy transfer free energies
20	12	7	0.735	(16) CIDAB, Cid hydrophobicity α/β prot.
21	11	12	0.735	(39) MEIRO, C_a distance to protein center
22	11	11	0.738	(85) OSMP1, optimal scale for 1 TMH prot.
23	11	11	0.725	(53) EDE21, Edelman optimal predictors
24	11	10	0.727	(35) NNEIG, Cornette eigenvalues
25	11	9	0.731	(26) EISEN, Eisenberg consensus hydroph.
26	11	9	0.729	(56) FASMB, Chou&Fasman β preferences
27	11	9	0.726	(21) ROSEM, Roseman hydrophobicity scale
28	11	9	0.724	(71) GRANT, Grantham polarity values
29	11	9	0.723	(20) KIDER, hydrophobicity related scale

TABLE I - cont.

Scales rank order	# helices detected		performance in TMH prediction	Amino Acid Scale Code / Name
A)	20			(60) RICH, Richardson preferences
B)	$I_3(\alpha)$	H.M.	A_{TM}	
30	11	8	0.732	(12) GIBRA, hydrophobicity of aa in proteins
31	11	8	0.729	(9) VHEBL, coil to helix in membrane scale
32	11	8	0.726	(45) WERSC, Scheraga ratio of in/out
33	11	7	0.723	(70) WOESE, Woese polarity scale
34	11	7	0.725	(2) FAUPL, Fauchere & Pliska hydrophob.
35	11	7	0.725	(28) HOPPW, antigenic determinant scale
36	10	10	0.725	(54) EDE15, Edelman optimal predictors
37	10	9	**0.743**	(59) JURET, Chou-Fasman values (α+β)/2
38	10	8	0.730	(83) MODKD, modified Kyte-Doolittle scale
39	10	8	0.730	(84) MDK4, modified Kyte-Doolittle scale
40	10	8	0.723	(30) ROSEF, mean fractional area loss
41	9	9	0.726	(86) OSMP2, optimal scale for memb. prot.
42	9	8	0.730	(80) MDK0, Modified Kyte-Doolittle scale
43	9	7	0.718	(87) JACWH2, Jacob & White IFH (0.5) sc.

preferences greater than 0.9 did not help, because TMH prediction by the SPLIT predictor takes precedence. In any case, a positive correct prediction is achieved for 20 interface helices when the predictor based on Richardson's preference functions is used to locate the sequence position of interface helices.

Recognition of Other Structural Motifs in Membrane Proteins

Other types of conformational index profiles produced by the SPLIT algorithm are also useful. For instance, the voltage sensor elements of voltage gated channels[40] are associated with a very high maximum in the conformational index profile for the product of membrane-buried α-helix preference and turn preference (Figure 1). This index is, as a rule, high at sequence regions known to be close to the ends of membrane-spanning helices. For bitopic membrane proteins (with only one TMH), the doublet of maximums in this index is found such that the characteristic membrane-spanning α-helix segment of approximately 20 residues separates these maximums (Figure 2).

Is sequence location of such maximums, always pointing to amino acid residues in the twilight zone of the interface regions (Figures 1 and 2),

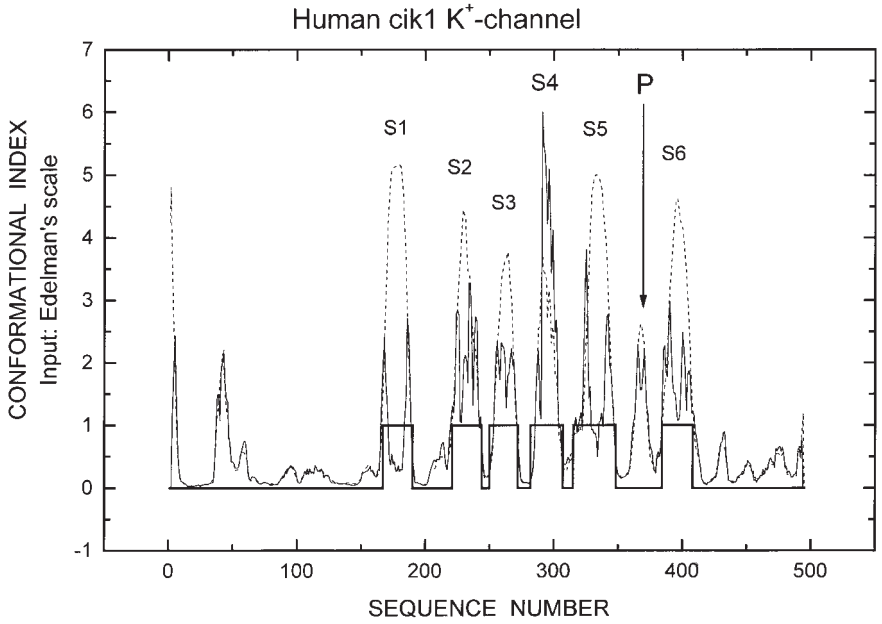

Figure 1. Sequence profile of membrane-buried helix preferences (dashed line) and membrane-buried helix times turn preferences (full line) for human potassium channel cik1. Edelman's scale[38] was used as the input for calculating these preferences, while Richardson's scale[35] and the corresponding preference functions extracted from soluble proteins was used to refine the digital prediction for the sequence location of transmembrane helices (bold line). Functionally most important segments are the membrane-spanning mobile voltage sensor S4 and the pore segment P, thought to contain the pore helix and the selectivity filter.

where the relative dielectric constant must change from the value of 2–3 (nonpolar membrane interior) to 80 (water)? The dataset of interface helices described above is convenient to test the predictor based on this index. Seven out of 22 interface helices can be located in the sequence with this predictor when the Kyte-Doolittle preference functions are used. This is not an impressive result, except for the fact that three of seven correctly predicted interface helices are very difficult to predict with hydrophobic moments (helix 81–89 from the M subunit of the photosynthetic reaction center from *R. sphaeroides*, and helices 1–7 and 361–367 from subunit I of bovine heart mitochondria cytochrome oxidase).

Another class of polypeptides – membrane active peptides, forming the amphipathic α-helix when attached to the membrane surface, have mainly interface seeking residues. Conformational profiles for synthetic antimicrobial peptide PGYa[41] exhibit a symmetric secondary structure with both pep-

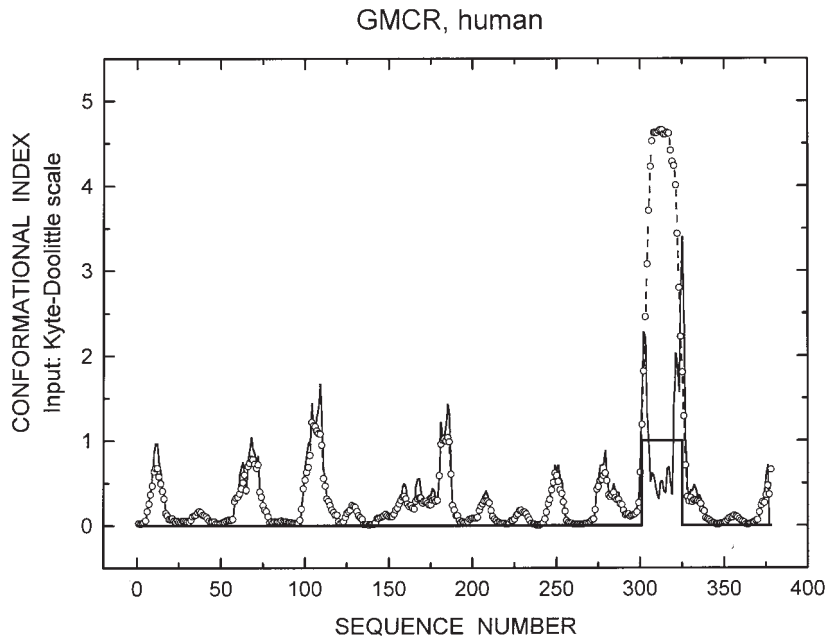

Figure 2. Sequence profile of membrane-buried helix preferences (dashed line with open circles) and membrane-buried helix times turn preferences (full line) for human granulocyte-macrophage colony stimulating factor (receptor). The Kyte-Doolittle preference functions have been used. Predicted transmembrane helix from amino acid 301 to 324 (bold line) agrees with the Swiss-Prot assignment 299–324 for the mature receptor. Sequence location of helix times turn preference maximums alongside the span of the potential membrane-spanning helix corresponds to interface regions where N and C helix termini are breaking through the lipid phase.

tide termini having high preference for the membrane-buried helix, while its middle region is likely to be associated with the interface seeking the amphipathic α-helix (Figure 3). Our threshold index $I_3(\alpha)$ and the hydrophobic moment index provide in this case similar information about the possible sequence location of the amphipathic α-helix. The choice of the amino acid attribute scale for the calculation of hydrophobic moments dictates how high maximums will be found. Maximal hydrophobic moment index of 4.8 at sequence position 8 (Figure 3) decreases to 4.0 at sequence position 15, when Cornette's PRIFT scale[36] is used to calculate hydrophobic moments. Just the opposite happens with hemolytic peptide melittin where the maximal hydrophobic moment of 2.9 at sequence position 17 increases to 3.8 at position 11 when the PRIFT scale is used. Preferences for the membrane-buried helix are sufficiently high in the *N*-terminal part of melittin sequence (Figure 4) for the region to be predicted as the TMH. On the other hand,

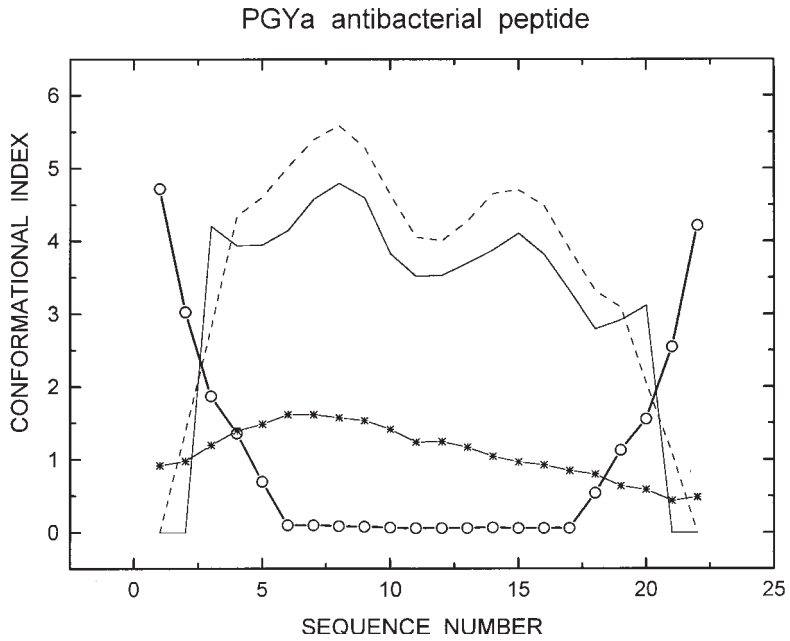

Figure 3. Conformational index profiles for the designed peptide PGYa[41] are for the membrane-buried helix propensity (bold full line with open circles), the hydrophobic moment index for amhipathic α-helix (full thin line), our threshold index $I_3(\alpha)$ for amphipathic α-helix (dashed line), and for Richardson's preferences for the free helix conformation (full thin line with stars).

preferences for the membrane-buried helix are quite low for the middle region of many antimicrobial peptides (only the example of PGYa is shown in Figure 3). This is not so for the free α-helix preferences as calculated with the help of Richardson's preference functions. For instance, these preferences have high values ranging from 1.59 to 3.03 and from 2.23 to 2.88 in the case of PGLa[42] and KLA7[43] antimicrobial peptides, respectively.

To answer the question how accurate is the present version of the SPLIT predictor in predicting β-strands in membrane proteins, we tested the photosynthetic reaction center and porin by using the Kyte-Doolittle preference functions. The percentage of correctly predicted β-strands is similar: 78% in the photosynthetic reaction center polypeptides and 75% in the porin.[44] However, the number of overpredicted β-strands (a total of 36) was considerably higher than the number of the observed (18) and of correctly predicted (14) strands in the photosynthetic reaction center. Hence, our accuracy parameter (Eq. (3)) was considerably lower for the reaction center (–1.0) than for the porin (0.625), where 12 out of 16 membrane-spanning strands

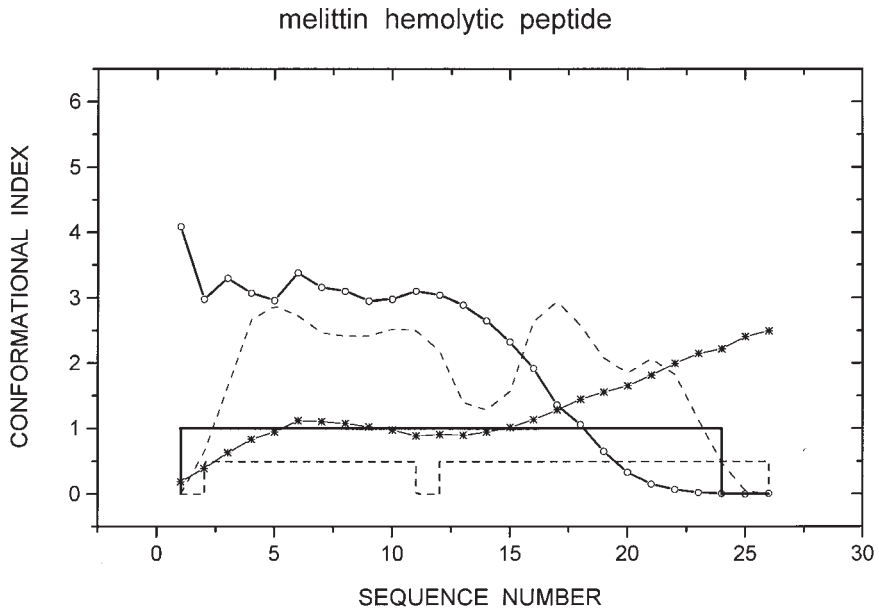

Figure 4. Conformational index profiles for the hemolytic protein melittin. The meaning of profile lines is the same as in Figure 3. Predicted span of transmembrane helix is denoted with the bold line at the 1.0 level. Observed helices are labeled with the dashed line at the 0.5 level.

are correctly predicted, 4 strands are underpredicted and 2 strands are overpredicted. For the recently solved structure of the outer membrane protein A transmembrane domain,[45] six out of eight membrane-spanning strands are predicted at their correct sequence locations (Figure 5).

In the case of the Rieske protein (Figure 6), very high maximums in the turn preference, in the hydrophobic moment for assumed β-sheet conformation and in the threshold index for optimal amphipathic β-strand conformation are all achieved at Ile 74, which is considered to be the pivot point for the movement of the soluble part of the Rieske sequence.[26] The second highest peak in the preferences for the membrane-buried helix (at Thr 43) is flanked on both ends (at Val 39 and at Phe 58) with high values for our threshold index for the α-helix hydrophobic moment (not shown). Richardson's preferences have maximums at Ala 51 and Val 68, respectively, inside and close to the C-terminus of the observed TMH segment, but not anywhere in the sequence region 131–148 with false positive TMH prediction. The maximum at Val 68 points at the short helix from Ala 66 to Ala 70. The other sequence region, 103–115, with high values in Richardson's preferences (1.4) points to helices Lys 103 to Ala 111, and Val 114 to Gln 116.

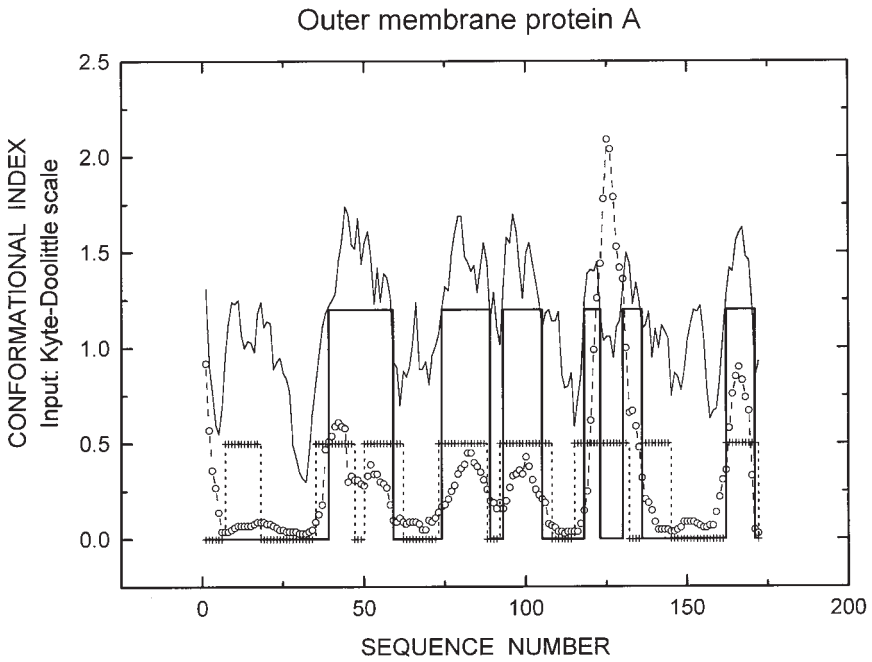

Figure 5. The preference profiles for the outer membrane protein A transmembrane domain. Generally higher β-strand preferences (full bold line) than preferences for the membrane-buried α-helix conformation (dashed line with open circles) predict dominant β-sheet structure for this domain. Horizontal lines at levels 0.5 and 1.2 denote the position of observed transmembrane β-strands (dashed line) and predicted β-strands (full bold line), respectively.

Topology Prediction

With the default choice of scales we can correctly predict *N*-terminus orientation for 26 out of 31 polypeptides with a simple version of the positive-inside rule algorithm.[4,11] Cases with the charge bias of zero (five such cases are found) are interpreted to mean the inside orientation of the *N*-terminus. Since we have a biased sample – only 9 out of 31 polypeptides are observed with outside orientation of their *N*-terminus, our error rate in predicting *N*-terminus orientation would increase if the charge bias of zero were interpreted as the outside orientation. Change in the charge bias when a different hydrophobicity scale is used can help to determine correct transmembrane topology. For instance, using the PRIFT scale[36] to calculate hydrophobic moments, a charge bias of +4 is found for cytochrome b (it is the charge bias of zero with the default choice of scales) and the correct topology of eight transmembrane helices instead of nine.

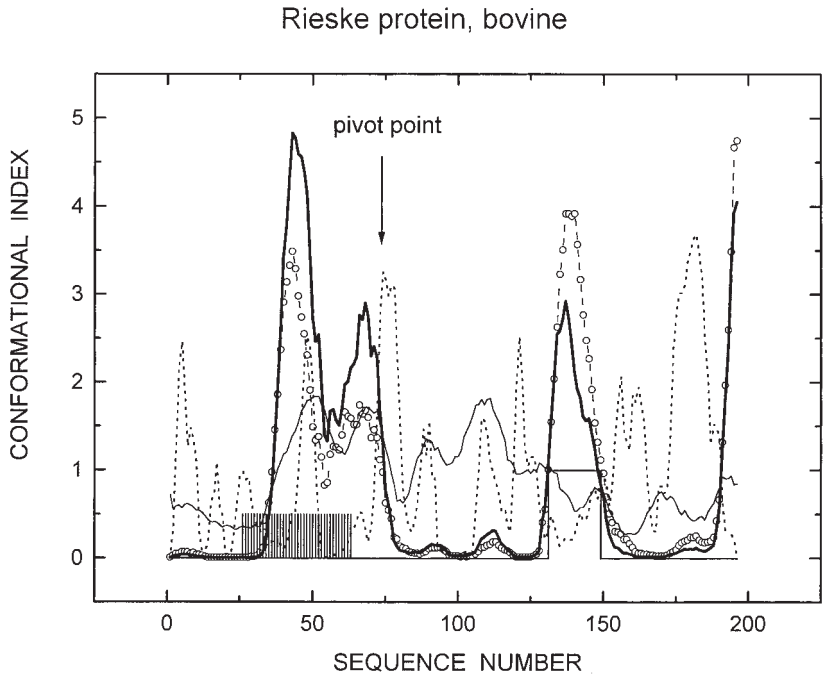

Figure 6. Conformational profiles for mature Rieske protein. Observed TMH (shaded column up to the 0.5 level) is the segment from amino acid 26 to 63, while predicted TMH (bold line at the level 1.0) is the segment from amino acid 131 to 148. Observed TMH is associated with the highest peak (full bold line) in the product of preferences for the membrane-buried helix (the Kyte-Doolittle scale input) and for free helix conformation (Richardson's scale input). The TMH preferences alone (dashed line with open circles) are highest at the sequence positions (131–148) where the hydrophobic β-sheet is known to envelop the iron-sulfur cluster.[24–27] Richardson's preferences (full thin line) do not have a maximum associated with the predicted TMH. The pivot point at Ile 74[26] for the rotation of the functional domain of Rieske protein is seen as the maximum in our threshold index for the β-strand hydrophobic moment (dotted line profile produced with the PRIFT scale[36] input).

DISCUSSION

The presented results indicate what would be the most practical approach to the sequence analysis of membrane proteins by means of preference functions. Success of preference functions in predicting the formation of α-helices must be due to the predominant influence of local interactions. With the default choice of scales, including Richardson's preferences, all of the 100 observed TMH are associated both with an easily selected high TMH preference maximum and with a maximum (often two maximums) for

free α-helix preferences, while overpredicted TMH lack either one or the other of these maximums. To avoid underpredictions of transmembrane segments, it is best to use the Kyte-Doolittle hydropathy scale[32] and the corresponding preference functions. Edelman's optimal predictor scale[38] and the corresponding preference functions increase the per-residue prediction accuracy by avoiding underprediction of residues observed in the transmembrane helix configuration. Richardson's α-helix preferences[35] and the corresponding preference functions for soluble proteins are good additional tools for predicting all α-helices (transmembrane and extramembrane) longer than 5 residues. It is obvious that even in the case of easily predicted membrane-spanning helices, the single amino acid attribute scale is not sufficient. Tests with several different hydrophobicity scales are recommended for each tested sequence. All of the potential transmembrane helical segments can be easily classified as stable (appearing in almost all runs) and unstable (appearing only with the certain choice of hydrophobicity scale). When different results for segment prediction are obtained with several of the best scales, it is advisable to use evolutionary information if available (related homologous sequences), positive inside rule scoring for different topologies[4] and complete information available in the output data file of the SPLIT predictor. Such a procedure would reduce the subjectivity in the choice of different decision (threshold) parameters.

A similar conclusion holds for predicting the sequence location of interface residues. Several different conformational index profiles in the SPLIT numerical output can be used to create the predictor for the sequence location of such residues in the α-helix conformation. Prediction of interface residues protruding through the membrane surface as the *N*- or *C*-terminus of longer membrane-spanning helices is possible by appropriate use of preference functions (see Figures 1 and 2). Prediction of the sequence location of interface helices lying parallel to the membrane surface can be tested when such helices are collected from known crystal structures of membrane proteins. For the dataset of such helices, it is of interest to compare older methods using hydrophobic moments[34] with our own hydrophobic moment threshold functions,[11] and with preference functions method (Table I).

The results in Table I show that:

a) Several of the best scales for calculating hydrophobic moments should be used and the results compared because even the best scales are missing about one third of observed interface helices. Widely used Eisenberg's scale[37] is able to detect the sequence location of only one half of observed interface helices (Table I).

b) Our hydrophobic moment threshold index $I_3(\alpha)$ can be used as an equal or better tool than the hydrophobic moment for the detection of interface helices.

c) Richardson's preference functions extracted from soluble proteins are a better tool for detecting the sequence location of interface helices than hydrophobic moments.

The need to go beyond calculations of the mean hydrophobicity with the Kyte-Doolittle hydropathy values[32] and of the hydrophobic moment with the Eisenberg hydrophobicity values[37] has been also pointed out by other authors.[46–48] These traditional tools for the classification and prediction of membrane-buried, interface and membrane active helices are extended and supplemented in this work with calculations of conformational preference profiles and hydrophobic moment functions based on several different amino acid attributes.

We have judged the performance in predicting the sequence location of interface helices in terms of the percentage of correct predictions. However, it is easy to increase the percentage of correct prediction, for instance by lowering the threshold value. Then, overpredictions are increased and more meaningful performance parameters, such as the A_{TM} (Eq. (3)), can actually decrease. Overpredictions in the case of interface helices can be due to predictions of extramembrane helices that are not included into our dataset of interface helices, or they can be due to completely wrong predictions of helices where none are found. In the case of the best known crystal structures of photosynthetic reaction center from *R. viridis* and *R. sphaeroides*, Richardson's preference functions are predicting a larger number of extramembrane helices, where none exist in the sequence (14 overpredictions) than the best scales used for the hydrophobic moments calculations in Table I (the predictor with the PRIFT scale has 8 overpredictions in the photosynthetic reaction center). Richardson's preference functions can detect almost all α-helix segments in membrane proteins, but these functions are not specific detectors of interface helices, and in proteins with predominant β-sheet structure, can often cause overpredictions of α-helices.

Another class of interface helices appears in antibacterial peptides. Antimicrobial peptides are promising therapeutic agents with very low potential to induce antibiotic resistance,[49] but the problem of their low selectivity in interaction with membranes[50] is still restricting their use. Here, we illustrate the usefulness of combining several conformational index profiles offered by our algorithm to attack the specificity problem by designing novel peptides. Conformational profiles, such as presented in Figure 3, indicate that common motifs in such profiles may exist that are sufficiently different from motifs found in hemolytic peptides (Figure 4) to guide the design and synthesis of peptide antibiotics. Comparison of Figure 3 profiles with conformational profiles associated with transmembrane helices reveals that the buried helix profile and hydrophobic moment profiles are inverted in Figure 3.

Maximal values for hydrophobic moment profiles are in the middle of the antibacterial sequence and maximal TMH preferences are at its *N*- and *C*-terminus. Very low preference for the buried-helix conformation associated with the middle sequence region probably ensures low hemolytic activity of these cationic peptides unless a high membrane potential and high concentration of negative surface charges are encountered. Such conditions are characteristic of the bacterial plasma-membrane and presumably enable a selective entrance and perpendicular orientation of amphipathic monomers with respect to membrane surface. For some critical concentration, spontaneous aggregation of peptide monomers is expected to cause formation of a water filled pore encircled with peptide polar faces.

Concerning topology prediction, we did not take into account that some classes of membrane proteins do not follow the 'positive inside rule'[51] and that this rule should be applied to 2^n topologies arising when n questionable TMH segments are identified.[4] Nevertheless, with the present SPLIT version, the topology prediction of known membrane polypeptides is comparable in performance[11] to the Rost PHDhtm algorithm[52] or to the Jones MEMSAT algorithm.[5]

Our default choice of amino acid scales and preference functions does not wrongly predict transmembrane helices in beta-class membrane proteins such as porins (Figure 5). However, porins are not predicted as membrane proteins and no high accuracy prediction of sequence location for transmembrane beta strands was achieved. Better prediction of the porins secondary structure remains our goal for future improvement of the server SPLIT services.

Although the use of α-helix preferences extracted from soluble proteins may seem out of place in the case of membrane proteins, the example of Rieske protein (Figure 6) and our present and earlier results[11,12] illustrate how TMH prediction can be improved when such preference functions are used. The conformational index, calculated as the product of TMH preferences and Richardson's preferences, exhibits higher and lower values with respect to TMH preferences exactly at the Rieske sequence regions associated, respectively, with TMH underprediction and TMH overprediction. Free helices predicted in soluble and membrane proteins with Richardson's preferences are of interest as possible initiation sites of protein folding, because α-helices may function as independent "seeds for folding".[53]

Acknowledgements. – We are grateful to Bono Lučić of the Ruđer Bošković Institute in Zagreb, who helped us with some references. This work was supported by the Croatian Ministry of Science Grant 177060 to D.J.

REFERENCES

1. J. Enriquez, *Science* **281** (1998) 925–926.
2. D. T. Jones, *FEBS Lett*. **423** (1998) 281–285.
3. A. C. W. May and T. L. Blundell, *Curr. Opin. Biotech*. **5** (1994) 355–360.
4. G. von Heijne, *J. Mol. Biol.* **225** (1992) 487–494.
5. D. T. Jones, W. R. Taylor, and J. M. Thornton, *Biochemistry* **33** (1994) 3038–3049.
6. B. Persson and P. Argos, *J. Mol. Biol.* **237** (1994) 182–192.
7. B. Rost, R. Casadio, P. Fariselli, and C. Sander, *Protein Science* **4** (1995) 521–533.
8. B. Rost, P. Fariselli, and R. Casadio, *Protein Science* **5** (1996) 1704–1718.
9. D. Juretić, B. Lučić, D. Zucić and N. Trinajstić, *Protein Transmembrane Structure: Recognition and Prediction by Using Hydrophobicity Scales Through Preference Functions*, in: C. Parkanyi (Ed.), *Theoretical and Computational Chemistry,* Vol 5. Elsevier Science, Amsterdam, 1998, pp. 405–445.
10. D. Juretić, D. Zucić, B. Lučić, and N. Trinajstić, *Computers Chem*. **22** (1998) 279–294.
11. D. Juretić and A. Lučin, *Journal of Chemical Information and Computer Sciences* **38** (1998) 575–585.
12. D. Juretić, B. K. Lee, N. Trinajstić, and R. W. Williams, *Biopolymers* **33** (1993) 255–273.
13. J. Deisenhofer, O. Epp, K. Miki, R. Huber ,and H. Michel, *Nature* **318** (1985) 618–624.
14. J. Deisenhofer, O. Epp, I. Sinning, and H. Michel, *J. Mol. Biol.* **246** (1995) 429–457.
15. J. P. Allen, G. Feher, T. O. Yeates, H. Komiya, and D. C. Rees, *Proc. Natl. Acad. Sci. USA* **84** (1987) 6162–6166.
16. G. McDermott, S. M. Prince, A. A. Freer, A. M. Hawthornthwaite-Lawless, M. Z. Papiz, R. J. Cogdell, and N. W. Isaacs, *Nature* **374** (1995) 517–521.
17. S. M. Prince , M. Z. Papiz , A. A. Freer , G. McDermott , A. M. Hawthornthwaite-Lawless, R. J. Cogdell, and N. W. Isaacs, *J. Mol. Biol.* **268** (1997) 412–423.
18. W. Kühlbrandt, D. N. Wang, and Y. Fujiyoshi, *Nature* **367** (1994) 614–621.
19. S. Iwata, C. Ostermeier, B. Ludwig, and H. Michel, *Nature* **376** (1995) 660–668.
20. T. Tsukihara, H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa, *Science* **272** (1996) 1136–1144.
21. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing, *J. Mol. Biol.* **213** (1990) 899–920.
22. E. Pebay-Peyroula, G. Rummel, J. P. Rosenbusch, and E. M. Landau, *Science* **277** (1997) 1676–1681.
23. H. Luecke, H. T. Richter, and J. K. Lanyi, *Science* **280** (1998) 1934–1937.
24. S. Iwata, M. Saznovits, T. A. Link, and H. Michel *Structure* **4** (1996) 567–579.
25. D. Xia, C. A. Yu, H. Kim, J. Z. Xia, A. M. Kachurin, L. Zhang, L. Yu, and J. Deisenhofer, *Science* **277** (1997) 60–66.
26. Z. Zhang, L. Huang, V. M. Shulmeister, Y. I. Chi, K. K. Kim, L. W. Hung, A. R. Crofts, E. A. Berry, and S. H. Kim, *Nature* **392** (1998) 677–684.
27 S. Iwata, J. W. Lee, K. Okada, J. K. Lee, M. Iwata, B. Rasmussen, T. A. Link, S. Ramaswamy, and B. K. Jap, *Science* **281** (1998) 64–71.
28. K. R. MacKenzie, J. H. Prestegar and, D. M. Engelman, *Science* **276** (1997) 131–133.

29. D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon, *Science* **280** (1998) 69–77.
30. M. E. Girvin, V. K. Rastogi, F. Abildgaard, J. L. Markley, and R. H. Fillingame, *Biochemistry* **37** (1998) 8817–8824
31. L. A. Sayle and E. J. Milnerwhite, *Trends in Biochemical Sciences* **20** (1995) 374–376.
32. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157** (1982) 105–132.
33. D. Frishman and P. Argos, *Proteins* **23** (1995) 566–579.
34. D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, *J. Mol. Biol.* **179** (1984) 125–142.
35. J. S. Richardson and D. C. Richardson, *Science* **240** (1988) 1648–1652.
36. J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi, *J. Mol. Biol.* **195** (1987) 659–685.
37. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, and W. Wilcox, *Faraday Symp. Chem. Soc.* **17** (1982) 109–120.
38. J. Edelman, *J. Mol. Biol.* **232** (1993) 165–191.
39. D. Juretić, N. Trinajstić, and B. Lučić, *J. Math. Chem.* **14** (1993) 35–45.
40. W. Catterall, *Annu. Rev. Biochem.* **64** (1995) 493–531.
41. A. Tossi, C. Tarantino, and D. Romeo, *Eur. J. Biochem.* **250** (1997) 540–558.
42. W. L. Maloy and U. P. Kari, *Biopolymers* **37** (1995) 105–122.
43. M. Dathe, T. Wieprecht, H. Nikolenko, L. Handel, W. L. Maloy, D. L. MacDonald, M. Beyermann, and M. Bienert, *FEBS Lett*. **403** (1997) 208–212.
44. M. S. Weiss and G. E. Schulz, *J. Mol. Biol.* **227** (1992) 493–509.
45. A. Pautsch and G. E. Schulz, *Nature Structural Biology* **5** (1998) 1013–1017.
46. R. Brasseur, *J. Biol. Chem.* **266** (1991) 16120–16127.
47. M. G. Roberts, D. A. Phoenix, and A. R. Pewsey, *Comput. Appl. Biosci*. **13** (1997) 99–106.
48. D. A. Phoenix, A. Stanworth, and F. Harris, *Biologicheskie Membrany* **15** (1998) 83–89.
49. G. Saberwal and R. Nagaraj, *Biochim. Biophys. Acta* **1197** (1994) 109–131.
50. T. Wieprecht, M. Dathe, M. Beyermann, E. Krause, W. L. Maloy, D. L. MacDonald, and M. Bienert, *Biochemistry* **36** (1997) 6124–6132.
51. Y. Gavel and G. von Heijne, *Eur. J. Biochem.* **205** (1992) 1207–1215.
52. B. Rost, R. Casadio and P. Fariselli, *Refining Neural Network Predictions for Helical Transmembrane Proteins by Dynamic Programming*, in: D. J. States, P. Agarwal, T. Gaasterland, L. Hunter and R. F. Smith (Eds.), *Proceedings Fourth International Conference on Intelligent Systems for Molecular Bilogy*, AAAI Press, Menlo Park, California, 1996, pp. 192–200.
53. L. G. Presta and G. D. Rose, *Science* **240** (1988) 1632–1641.

SAŽETAK

Analiza sekvenci membranskih proteina pomoću Web poslužitelja SPLIT

Davor Juretić, Ana Jerončić i Damir Zucić

U radu se ispituje kvaliteta predviđanja sekvencijske lokacije, konformacije i orijentacije membranskih polipeptida poznate kristalne strukture pomoću web poslužitelja SPLIT. Poslužitelj SPLIT temelji se na metodi sklonosnih funkcija. Navedene funkcije služe za pretvorbu početnog izbora ljestvice aminokiselinskih parametara u konformacijske sklonosti ovisne o sekvencijskoj okolini. Transmembranske uzvojnice točno se predviđaju kada se izvrši dobar izbor sklonosnih funkcija koje se pak dobivaju iz datoteke integralnih membranskih proteina. Za razliku od drugih algoritama s sličnom kvalitetom predviđanja, prediktor SPLIT ne zahtijeva informacije o homologiji. Sekvencijska lokacija kraćih izvanmembranskih uzvojnica također se može naći s pomoću sklonosnih funkcija određenih na skupu topljivih proteina. Posebno, Richardsonove sklonosne funkcije bolji su prediktori od hidrofobnih momenata, čak i onda kada se radi o pogađanju sekvencijskog položaja uzvojnica koje leže na površini membrane. Internet adresa za poslužitelj SPLIT jest:

http://pref.etfos.hr/split