The Preference Functions Method for Predicting Protein Helical Turns with Membrane Propensity

Davor Juretić* and Ana Lučin

Physics Department, Faculty of Science and Education, University of Split, N. Tesle 12, HR-21000 Split, Croatia

Received September 12, 1997

The prediction of secondary structure conformation of membrane-buried segments of integral membrane proteins is done in this work by using the method of preference functions (Juretić et al. In *Theoretical and Computational Chemistry;* Párkányi, C., Ed.; Elsevier Science: Amsterdam, 1998; Vol 5, Chapter 13, p 405). The method is here extended by predicting sequence location, hydrophobic moments, and transmembrane orientation of helical segments. Independent evaluation of the method with ubiquinol-cytochrome *c* reductase subunits is performed before some already known details of the crystal structure of that respiratory complex are released. Tests with potassium channels revealed that their characteristic super-secondary structure in the pore region can be recognized by performing the analysis with preference functions and with hydrophobic moment threshold functions.

INTRODUCTION

The prediction of sequence location, conformation, and transmembrane orientation of membrane-buried segments is the essential first step in an analysis of membrane protein primary structure.[1] Long, membrane-spanning α-helices, having in some cases more than 30 amino acid residues,[2] are commonly found in integral membrane proteins of known crystal structure. Such helices fold in membrane as stable autonomous folding domains.[3] A segment of 20 residues in the α-helix conformation can easily span membrane lipid interior if oriented perpendicular to membrane surface. In porins six to 10 residues are sufficient to span the membrane as β-strands.[4] Shorter membrane-buried helices that do not span the membrane also appear in some proteins.[5] The prediction of helical turns with membrane propensity will be our main objective in this work.

Hydrophobicity analysis is a very good tool for sequence analysis.[6] In integral membrane proteins of known structure transmembrane segments can be recognized as maximums in averaged hydrophobicities. Such maximums often correspond to observed transmembrane α-helices. However, hydrophobicity analysis was not designed to predict the conformation of transmembrane hydrophobic segments,[6a] and it cannot distinguish hydrophobic segments found in soluble and membrane proteins.[7] Different algorithms based on the hydrophobicity analysis achieved only modest accuracy in predicting sequence location of transmembrane segments.[8]

Functionally important membrane-buried segments in voltage-gated channels, such as the pore-forming P region,[9] or the S4 voltage sensors[10] are not very hydrophobic and present special difficulties for hydrophobicity analysis. Modern pattern-recognition predictors of transmembrane segments[11] give a yes or no answer to the question about the transmembrane nature of the tested segment and do not predict its secondary structure. These predictors are bound to give wrong descriptions of all membrane-buried segments such as the P-segment, because such segments are neither transmembrane nor extramembrane.

This study is using the method of preference functions[12] for predicting the sequence position of membrane-buried helices that are (a) transmembrane helices and (b) shorter helices that invaginate in the membrane as a part of a membrane-buried nontransmembrane domain (such as the pore domain in voltage-gated channels). The preference function associates sequence hydrophobicity with statistical propensities for conformational motifs. The threshold function, introduced in this work, helps to find the sequence location of amphipathic regular structures at the membrane surface and to increase the prediction accuracy for membrane-spanning helices. With a judicious choice of input preference functions achieved prediction accuracy is quite high in our tests with known crystal structures. The prediction of transmembrane topology for polypeptides in the ubiquinol-cytochrome *c* reductase complex is presented in this work before all relevant structural details are released for these polypeptides.

METHODS

The preference functions method[12] works by extracting preference functions from a given database of sequences with known secondary conformation of all residues or at least with expected sequence location of transmembrane segments. The training procedure with the PREF program memorizes secondary conformation, residue type, *and its sequence environment* at each sequence position in all proteins. Sequence hydrophobic environment of each residue is calculated as average hydrophobicity of its five left and five right neighbors. Irrespective of amino acid type, the preference for the α-helix structure for a given residue exhibits nonlinear increase with the increased average hydrophobicity

* Corresponding author.

S0095-2338(97)00073-5 CCC: $15.00
Published on Web 06/30/1998

of its sequence neighbors. Secondary structure is predicted by selecting the conformation associated with the highest preference. Good initial choice for a scale of 20 amino acid attributes largely determines the prediction accuracy with preference functions.

A total of 87 scales of attributes such as hydrophobicity, accessibility, polarity, and conformational preference are available in our algorithm. The database of 63 integral membrane proteins of the α-class and 37 soluble proteins of the β-class was used before[12d] as the training data set of proteins to extract preference functions for all scales considered except two. In the case of the Kyte−Doolittle hydropathy scale[6a] we use optimized training procedure developed previously[12c] with 135 integral membrane proteins and the same set of 37 soluble proteins. Another exception is the Richardson's scale of α-helix preferences[13] and corresponding preference functions that were extracted from the database of 147 soluble proteins. The same database of soluble proteins was used to test the predictor for false positive results. Known structures of membrane proteins used to test the algorithm consisted of 21 polypeptides.[14] These are subunits H, L, and M of the photosynthetic reaction center from *Rhodobacter viridis* and from *Rhodobacter sphaeroides*, the light-harvesting protein from *Rhodopseudomonas acidophila* and plant light-harvesting protein from *Pisum sativum*, the subunits I, II, and III of the cytochrome *c* oxidase from *Paracoccus denitrificans*, and the subunits I, II, III, IV, VIa, VIc, VIIa, VIIc, and VIII of the cytochrome *c* oxidase from bovine heart. These sequences contained a total of 2081 residues in the transmembrane helix (TMH) conformation and 75 TMH which were *not* seen before by the PREF algorithm during the training procedure (extraction of preference functions). The TMH assignments from original published papers were used even when it appeared that TM helices are extremely long (41 residues for the subunit VIc helix from bovine heart cytochrome *c* oxidase[14f]).

Histograms of sequence environments collected by the PREF algorithm are well approximated with Gaussian functions.[12a,b] The probability to find amino acid type "i" in secondary conformation "j" within sequence hydrophobic environment X is then defined as the ratio of corresponding Gaussian for the conformation "j" to sum of Gaussians for all considered secondary conformations:

$$p_{ij}(X) = \frac{(N_{ij}/\sigma_{ij}) \exp[-(X-\mu_{ij})^2/2\sigma^2_{ij}]}{\sum_k (N_{ik}/\sigma_{ik}) \exp[-(X-\mu_{ik})^2/2\sigma^2_{ik}]} \qquad (1)$$

Here 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. Preference functions are defined as

$$P_{ij}(X) = p_{ij}(X)\cdot(N/N_j) \qquad (2)$$

where N/N_j is the inverse fraction of conformation "j" in the protein data set. Extracted preference functions are named according to chosen scale of amino acid attributes used to derive and evaluate these functions.

The model with four secondary conformations ($k = 1,..,4$) has been used: α-helix, β-sheet, turn, and undefined. The β-sheet, turn, and undefined conformation is "learned" by collecting sequence environments of residues from 37 β-class soluble proteins of known crystallographic structure and Kabsch−Sander assignment of secondary structure by the DSSP algorithm.[15] In addition, we assigned α-helix, turn, and undefined conformation in incompletely known membrane proteins. Expected transmembrane segments (according to Swiss-Prot database) were assumed to have an α-helix conformation, four residues on both sides of membrane-spanning segments were assumed to be in the turn conformation, and all remaining residues were considered to be in the undefined conformation. Since α-helix or "H" conformation is "learned" by collecting sequence environments of residues from expected transmembrane segments, predicted "H" conformation should be considered as *membrane-buried a-helix conformation*.

The testing part of the procedure is performed by the SPLIT algorithm which has a filter capable of *splitting* predicted helices that are too long into two or three shorter helices with reasonable length. SPLIT reports a battery of performance parameters for tested proteins of known secondary structure. When we are interested only in the prediction accuracy for transmembrane (TM) structure reported parameter A_{TM}[16] takes into account overpredicted o_{TM}, underpredicted u_{TM}, and observed (or expected) number N_{TM} of residues found in the TMH structure:

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

The A_{TM} parameter is not the percentage and can assume negative values for poor prediction. Unlike the percentage of correctly predicted residues it punishes overprediction and gives a better idea of the prediction quality. However, measures for single residue accuracy do not completely reflect the quality of a prediction.[17] Underpredicted and overpredicted TMH segments are also reported by our algorithm, and the parameter analogous to A_{TM} is calculated. Correct prediction for transmembrane helix is scored when predicted and observed TM helix have at least nine residues in common. Another per-segment measure is the fractional overlap of segments[17b] S^{obs}_{ov}, which is the highest (1.0) for the best overlap of predicted and corresponding observed segments. We use the strict version of this measure for segment prediction accuracy with zero accepted deviation[17b] δ.

The preference profiles for tested protein are created by evaluating preference functions for each residue separately for α-helix, β-sheet, turn, and undefined conformation. Obtained preferences are smoothed: seven values are smoothed for α-helix preferences, five for β-sheet preferences, and three for turn or undefined preferences. The digital predictor[18] compares smoothed preferences for different secondary conformations and assignees the predicted conformation to the highest preference. TM helices are predicted initially (before filter action) when maximal preference is high enough (higher than 2.5) and when number of consecutive residues predicted in the "H" conformation (α-helix) is big enough (14 or more). Short (less than 17 residues) and amphipathic (see below) predicted TMH are labeled as such in the output file.

Additional details about adopted procedure for filtering and splitting potential TMH and for finding optimal choice

of empirical parameters can be found in our two recent papers.[12c,d] There are, however, several novel features. First novelty is the extension of predicted transmembrane helix ends whenever predicted helix cap residue (the first residue at helix NH_2-terminus and the last residue at helix COOH-terminus) has α-helix preference higher than 1.8 and turn preference less than 1.2.

Second novelty is the topology prediction based on the observation[19] that more arginines and lysines are found in internal protein loops than in external protein loops (charge bias). This "positive inside rule" is very useful in determination of transmembrane orientation of all membrane spanning segments in many of the integral membrane proteins.[19] Lysines and arginines were counted in all loop domains outside membrane but not more than 30 residues distant from ends of membrane-spanning helices. Since membrane spanning helices can often reach extramembrane space, we counted all positive charges outside TMH regions where α-helix preference was higher than 2.0. Therefore, only residues with the preference for membrane-buried α-helix higher than the threshold value of two were considered to be located in low dielectric environment (such residues were labeled with the letter "O" in the output data file, while remaining residues in predicted TMH were labeled with the letter "M"). With all of the predicted TMH segments fixed, as in our case, only two possible transmembrane orientations had to be considered: one with protein NH_2-terminus positioned outside and another with protein NH_2-terminus positioned inside space delimited with a membrane. The difference in charges of odd loops (starting with the NH_2-terminus loop) and even loops, or the charge bias, determined the topology. The charge bias greater or equal to zero indicated the inside orientation of the N-terminal.

Third novel feature is the prediction of α-helix and β-strand conformation when corresponding hydrophobic moment was very high.[20] Hydrophobic moments are calculated for all twist angles δ in the range from 80 to 180 degrees with one degree step, for all consecutive 11-residue peptides in tested protein, and their values associated with the central amino acid in the 11-residue sliding window. We use the Eisenberg et al. method[20] for these calculations. However, many of provided 87 scales can be used to calculate hydrophobic moments. Cornette's optimal scale for amphipathic helices, named the PRIFT scale[21] (code # 27) is used in our tests instead of Eisenberg's consensus hydrophobicity scale[20] (code # 26) if not stated otherwise. All amino acid attributes are used by the algorithm as normalized values with the average of zero and standard deviation of one.

Hydrophobic moments μ(k,i), associated with sequence position "i" and secondary conformation "k" (α-helix or β-strand), are used in the SPLIT predictor to evaluate the threshold function I(k,i):

$$I(k,i) = 6\mu(k,i)\exp(-(\mu(i)_{max} - \mu(k,i))^2)\exp(-(\delta(k)_{opt} - \delta(k,i))^2) \quad (4)$$

where $\mu(k,i)_{max}$ and $\mu(k,i)$ are maximal hydrophobic moment and hydrophobic moment for perfect "k" conformation respectively, while $\delta(k)_{opt}$ and $\delta(k,i)$ are optimal twist angle corresponding to maximal hydrophobic moment and twist angle for assumed perfect "k" conformation, respectively. The "perfect" conformation is defined with 100 degree twist angle for the case of α-helix and with 180 degree twist angle for the case of β-strand. The threshold index profiles are obtained with averaged I(k,i) values, so that each average of three values is associated with the central residue in the triplet.

The threshold function (4) and hydrophobic moments are used to extend the ends of predicted TMH and to eliminate potential TMH with high amphipathicity and not enough high hydrophobicity. We used the three point average of $I(\alpha)$ and extended TMH at ends having such an average higher that 2.0, but at the same time we eliminated TMH having such an average higher than 3.0 closer to the middle of predicted segment with maximal α-helix minus turn preference lesser than 2.3. When only one TMH was predicted, it was eliminated if maximal α-helix preference was less than 2.8 or when maximal hydrophobic moment (normalized scale) was higher than 0.65. Additional β-strand conformation was predicted before filter operation for all residues in a sequence with an associated three point average of $I(\beta)$ higher than 2.7 or sum of hydrophobic moment and preference for β-strand conformation higher than 2.0.

The last novelty is the possibility to use two different amino acid scales to calculate preference functions and one scale to calculate hydrophobic moments and threshold functions in the same run. Only the predicted "H" conformation survives after the first scale is used, while reported preference profiles are entirely due to the last scale used. Therefore, the first choice of scale modifies the output of the digital predictor but does not influence the height or sequence position of preference maxima and minima, while the second choice of scale is used only for amphipathicity calculations. It is also possible to eliminate entirely the influence of the first scale by entering appropriate input ("y") after the code for the last scale. For each run a triplet of scale codes is used. We suggest for the standard procedure codes # 60, # 27, and # 1 (see main text).

Computer programs were written in FORTRAN 77. The SPLIT 3.5 routine was wrapped into the Web server written in HTML, Unix shell script language, and ANSI C. A graphic library, created for the SPLIT server, enables the graphical presentation of prediction results when an input sequence is submitted at *<http://drava.etfos.hr/~zucic/split.html>*. An automatic e-mail server is also available at *predict@drava.etfos.hr*.

RESULTS

(a) Tests with Membrane Proteins of Known Crystal Structure. Rigorous tests on never-before-seen 21 integral membrane polypeptides of known crystal structure[14] (see Methods) resulted in 75 correct predictions for sequence location and conformation of membrane spanning α-helices when the Kyte−Doolittle hydropathy scale[6a] was used as the only input scale for extracting corresponding preference functions. Two helices were wrongly predicted as short transmembrane helices (short TM helices, having less than 17 residues, are not observed in this data base according to published assignments), but none was missed and none was overpredicted. All integral membrane polypeptides were recognized as such. Overall per-residue prediction accuracy

of $A_{TM} = 0.73$ corresponded to 423 underpredicted and 138 overpredicted residues in the transmembrane helix conformation. Average predicted TMH length of 23.95 residues was smaller than average observed length of 27.75 residues for membrane-spanning helices. Fractional segment overlap (see Methods) for predicted and observed TMH was 75%. Large average error in length 5.43 ± 4.06 residues and in location 2.65 ± 2.09 residues per predicted TMH, and predominance of *underpredicted* residues indicated that this is still an unbalanced prediction.

Our best result was achieved by using two different sets of preference functions in the same run. We used the Edelman's optimal predictor[22a] (for the width of 25) to extract corresponding preference functions. The value for valine was changed from 0.559 to 0.859 in the original Edelman's scale[22a] to better take into account valine hydrophobicity. Richardson's preference functions were extracted from 147 soluble proteins by using the scale of α-helix preferences[13] (code # 60) for the middle helix regions in soluble proteins. This combination of input scales is called "the best choice of scales" in the following text. We enclose amino acid codes and (in parentheses) corresponding Edelman's (the first number) and Richardson's normalized attributes (the second number): Ala (A = 0.65, 2.09), Cys (C = 1.09, −0.98), Leu (1.53, 0.42), Met (M = 1.62, 1.26), Glu (E = −0.87, −0.70), Gln(Q = −1.16, 0.70), His (H = 0.15, −0.14), Lys (L = −1.03, 0.14), Val (V = 0.94, 0.42), Ile (I = 1.48, 0.42), Phe (F = 0.61, 0.70), Tyr (Y = −0.18, −0.70), Trp (W = 0.12, 1.26), Thr (T = 0.25, −0.14), Gly (G = −0.51, −1.53), Ser (S = −0.16, −1.26), Asp (D = −2.01, −0.14), Asn (N = −0.91, −0.42), Pro (P = −1.18, −2.09), Arg (R = −0.40, 0.70)

The results with the same set of 21 membrane polypeptides were as follows. All observed TMH were correctly predicted, none was overpredicted and none was predicted as short TMH. Overall per-residue prediction accuracy of $A_{TM} = 0.81$ corresponded to 252 underpredicted and 145 overpredicted residues. Fractional segment overlap increased to 83%, and average length of predicted TMH was 26.32 residues. Average error in length and sequence location of predicted TMH was 3.93 ± 3.01 and 1.97 ± 1.23 residues, respectively, per observed helix. On average N-caps were shortened for 1.97 residues, while C-caps were shortened for 0.21 residue per observed helix.

Different assignment of TM helices is possible even in known crystal structures of membrane proteins. For instance, the assignment for the cytochrome *c* oxidase (bovine) TM helices in the Protein Data Bank (PDB) differs slightly from published assignments.[14f] With sequence locations of 28 TM helices taken from the PDB, instead from a published paper,[14f] the prediction accuracy with our best choice of input scales decreases from $A_{TM} = 0.82$ to $A_{TM} = 0.74$, but all TM helices are still predicted. Apparent accuracy decrease is due to four TM helices described in the Tsukihara et al. paper[14f] that are split into two or even three shorter helical segments in the PDB with only one of these segments still considered as the TM helix.

The prediction accuracy decreased from $A_{TM} = 0.81$ to (a) $A_{TM} = 0.79$ and (b) $A_{TM} = 0.77$, when some novel features present in the predictor (Methods) were *not* used for additional prediction or elimination of the α-helix conformation. These are respectively (a) profile of hydrophobic moments and of the threshold index (from eq 4) for hydrophobic moments (b) extension of TM helix caps when high α-helix preference is obtained at the cap position.

Testing all combinations of scales would require several hundreds of thousands of tests. Hence, only the tests with several common scales[22] used to create and evaluate preference functions are performed and reported here. The accuracy parameter A_{TM} in parentheses is reported twice for each scale—first for the chosen scale alone, and second in the run when Richardson's scale is used to modify the output of digital predictor. Listed in the order of performance are Edelman's optimal predictor[22a] for width 25 (code # 52) (0.76 → 0.81), Edelman's optimal predictor scale for width 21 (code # 53)(0.75 → 0.80), modified Kyte−Doolitle scale[12c] (code # 83)(0.72 → 0.77), Kyte−Doolittle hydrophobicity scale[6a] (code # 1)(0.73 → 0.75), Engelman's scale[6c] (code # 4)(0.67 → 0.74), Rose's mean fractional area loss[22b] (code # 30)(0.71 → 74), self-consistent hydrophobicity scale NNEIG[21] (code # 35)(0.71 → 0.73), hydropathy scale for membrane proteins VHEBL[21] (code # 9)(0.68 → 0.73), proportion of residues 95% buried in a protein[22c] (code # 29)(0.69 → 0.73), and surrounding hydrophobicity scale[22d] (code # 3)(0.65 → 0.72). With modified Edelman's scale alone one observed TMH was underpredicted. All observed TMH were correctly predicted with Kyte−Doolittle scale and modified Kyte−Doolitle scale alone, but one TMH was overpredicted when Richardson's scale was used to modify output in each case.

With a fixed best choice of scales one can use Eisenberg et al. consensus hydrophobicity scale[20] instead of PRIFT scale[21] for hydrophobic moment calculations, but the prediction accuracy is then almost the same ($A_{TM} = 0.808$ instead of 0.809). The prediction accuracy decreased to the value of $A_{TM} = 0.78$ when Chou-Fasman scale of α-helix preferences[22e] was used to modify the output with Edelman's scale.[22a]

With Kyte−Doolittle preference functions, as the only input, wrong transmembrane orientation was predicted for three out of 21 polypeptides. These are the subunits II and VIc from bovine cytochrome *c* oxidase and the light-harvesting protein from *Rhodopseudomonas acidophila*. The best choice of preference functions produced the same three erroneous predictions of topology.

The comparison with the MEMSAT algorithm of Jones et al.[11a] gave the following results. When the same reference proteins were tested, it achieved 74 correct predictions of transmembrane helices with one overpredicted and one underpredicted helix. A total of 597 residues were underpredicted and 123 overpredicted as transmembrane helix residues, which corresponded to considerably lower prediction accuracy of $A_{TM} = 0.65$. Eight polypeptides were predicted with wrong in/out orientation.

When refined PHD method of Rost et al.[11d] was tested on the same set of membrane proteins, the results were as follows. Three membrane polypeptides (subunits IV, VIa, and VIc of bovine heart cytochrome *c* oxidase) were not predicted as such. A total number of correctly predicted TMH was 70, while five TMH were underpredicted. Per-residue accuracy was $A_{TM} = 0.60$. This corresponded to 58 overpredicted and 777 underpredicted residues. Prediction of topology by the PHD topology device was also less successful than our simple positive inside rule. For seven

membrane proteins the PHD method predicted wrong transmembrane orientation.

(b) Tests with a Large Number of Membrane Proteins and With Soluble Proteins. Tests performed up to now with only 21 membrane polypeptides of known structure can be extended if Swiss-Prot assignments of transmembrane segments are regarded as expected TM segments in the α-helix conformation. For instance the prediction accuracy is $A_{TM} = 0.66$ with Kyte−Doolittle preference functions when 168 nonhomologous integral membrane proteins used in our previous work[12c,d] are tested. All of the 168 integral membrane proteins are recognized as such. Predicted TM helices are in general extended beyond expected sequence location, so that per-residue prediction accuracy errs in the direction of *overprediction*: a total of 2901 residues are overpredicted, while 1917 are underpredicted as TM helix residues. However, many of expected TM segments are eliminated due to high hydrophobic moment and/or to low maximum in the TMH preference, so that out of 662 expected TM segments 59 are underpredicted. The best choice of preference functions for crystal structures of membrane proteins resulted in the same percentage of correctly predicted TM segments (90%) but in even worse overprediction of TM segments length in the same data set of 168 membrane proteins. The prediction accuracy of only $A_{TM} = 0.56$ is then due to 4624 overpredicted and 2017 underpredicted residues out of the total number of 14 374 residues expected to be in the transmembrane segments.

For the subpopulation of 61 membrane proteins we could compare the observed in/out location of the N-terminal reported by Rost at al.[11c] and our prediction. The Kyte−Doolittle attributes[6a] resulted in the smaller number of wrong in/out assignments (eight wrong assignments) than the best choice of input scales (13 wrong assignments). Seven wrong topological assignments are reported by Rost et al.[11c]

False positive results are sometimes found for soluble proteins of known structure. For 147 such proteins (protein data set used by Rost and Sander[23] plus 21 additional proteins) our predictor with Kyte−Doolittle preference functions finds 23 "transmembrane helices" and 21 "membrane proteins". The Kyte−Doolittle hydrophobicity scale was originally derived[6a] to identify hydrophobic segments in soluble proteins with special structural or functional role. For instance such hydrophobic segment in the horse liver alcohol dehydrogenase (Figure 1), predicted by the SPLIT as TMH, overlaps the NAD(P)-binding $\beta\alpha\beta$ motif from Rossmann-fold domain. A total of 33 "transmembrane helices" in 23 "membrane proteins" were found with our best duet of input scales (Richardson's[13] and Edelman's[22a]) and the same false positive TMH prediction occurred for the NAD-binding region of alcohol dehydrogenase.

(c) Preference Profiles for Cytochrome bc_1 subunits with Kyte−Doolittle Preference Functions. The cytochrome bc_1 complex from mitochondria is the convenient test case. The crystal structures of bovine[24] and chicken[25] heart mitochondria complex will be available soon from the Protein Data Bank. For the bovine cytochrome *b* (Figure 2) our algorithm predicts eight transmembrane helices A to H and matrix location of the NH_2- and COOH-termini (due to charge bias of +4). Predicted TMH sequence positions are 31−54 (A), 85−101 (B), 110−132 (C), 176−200 (D), 225−249 (E), 286−308 (F), 321−338 (G), and 350−371

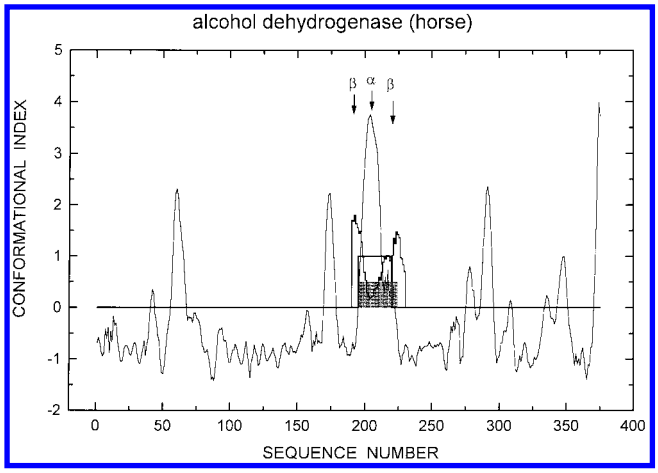

Figure 1. The profile of α-helix minus turn preference (full thin line), regarded as the "conformational index", for soluble protein: alcohol dehydrogenase from horse liver (the PDB entry: 8adh). Conformational preferences for the β-strand structure (jagged line) are shown only in the nicotine adenine dinucleotide (NAD) binding region. Since α-helix preferences are derived (via preference functions) from expected membrane-buried α-helical domains, the digital prediction (full thick line at the level 1.0 from Thr 195 to Arg 219) represents false-positive prediction for membrane-spanning α-helix. Instead of the TMH our predictor has located the hydrophobic segment serving as the binding region for the nicotine adenine dinucleotide (the segment from Thr 195 to Asp 224 with shaded area up to the level of 0.5). This $\beta\alpha\beta$-motif is the core domain of the Rossmann-fold.

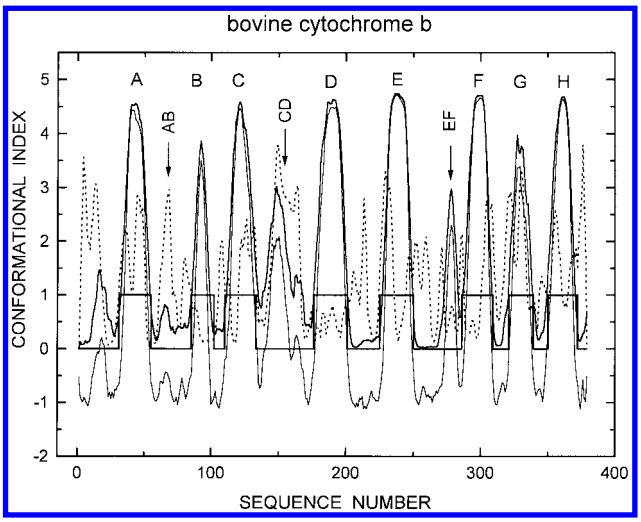

Figure 2. The preference profiles for bovine cytochrome *b* (379 amino acids): digital prediction for transmembrane helices (thick line at the height 1.0), α-helix preference (thick line), α-helix minus turn preference (thin line), and the threshold index (eq 4) for surface-attached amphipathic α-helix structure (dotted line). Preference profiles and digital prediction were obtained with Kyte−Doolittle preference functions extracted from database of soluble and membrane proteins (see Methods). The threshold index was calculated with the PRIFT scale[21] for amphipathic helices as the input.

(H). The threshold index for amphipathic α-helix conformation has one maximum from Ser 69 to Ile 69 in the external loop AB and two such maximums from Ile 146 to Trp 165 in the large external loop CD. Membrane-buried, but not membrane-spanning helix, is predicted in the sequence region from Trp 272 to Leu 281 (with maximum preference close to 3.0 at Tyr 278) in the third external loop EF.

One obvious problem with the digital prediction, using only the Kyte−Doolittle preference functions, is that histidine

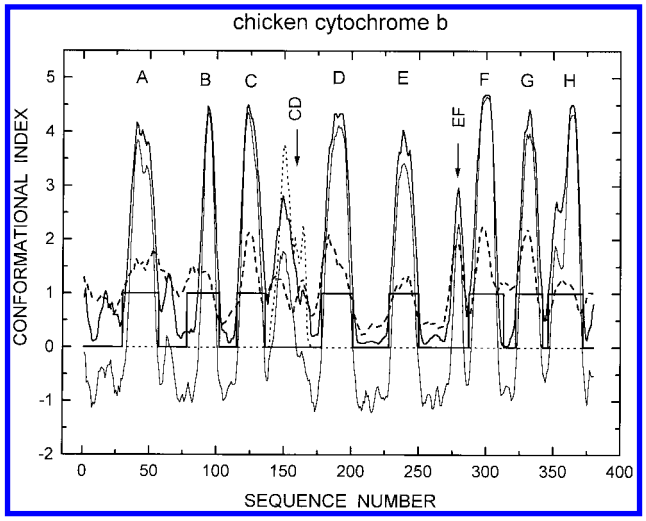

Figure 3. The preference profiles for chicken cytochrome *b* (380 amino acids). The same notation is used as in the Figure 2. The amphipathy at the helical repeat (dotted line) is shown only in the sequence region 140−160 where it is the strongest. Digital prediction with Kyte−Doolittle preference functions is here modified with Richardson's α-helix preferences (dashed line). Richardson's preference functions were extracted from the database of soluble proteins as explained in the text.

ligand 83 of the b_L-type heme should be located inside transmembrane helix B near its N-cap position.[24] When Richardson's preference functions are allowed to modify the prediction, TMH segments 32−54, 75−98, 110−132, 174−200, 225−249, 286−308, 315−338, and 350−371 are predicted. To see how this input modification affects the digital predictor output we tested chicken cytochrome *b* as well (Figure 3). Predicted topology is the same: eight TMH and inside (matrix) orientation of the NH_2-terminus (due to the charge bias of +6). Predicted TMH are 30−56 (A), 78−101 (B), 115−135 (C), 178−200 (D), 228−249 (E), 287−312 (F), 322−341 (G), and 346−371 (H). The couple of amphipathic surface-attached helices is predicted in the large external loop CD (dotted line), and membrane-buried α-helix preference peak in the external loop EF (from Phe 275 to Leu 282) is an important feature too.

Mature iron−sulfur Rieske protein (ucri_bovin) is predicted to have membrane-spanning helix segment 131−149 by the Kyte−Doolittle scale[6a] input. Other input scales gave very similar results. Another α-helix preference maximum of 2.40 at Thr 43 is not predicted as TMH but points to the potential membrane-buried α-helix region from Thr 40 to Val 55. High Richardson's preferences for α-helix conformation (maximum of 1.85 at Ala 51) are associated with this segment too. Homologous membrane-buried helix region close to the NH_2-terminus, that may serve as membrane-anchor, is found in mature Rieske protein sequences from all organisms tested (not shown). Sometimes it is predicted as TMH. For instance predicted TMH close to the NH_2-terminus is the segment 17−35 in *Rhodobacter capsulatus*, with maximal preference of 3.82 at Thr 27 for membrane-buried α-helix, and the 67−89 segment in *Neurospora crassa* with maximal preference of 3.20 at Gly 76.

Bovine cytochrome *c*1 is predicted with one transmembrane helix at the COOH-terminus: 207−221 (or from Pro 196 to Lys 223 when Richardson's preferences are used to modify the Kyte−Doolittle preferences) and outside (intramembrane) orientation of its NH_2-terminus, due to the charge bias of −2. The middle position inside membrane should be close to Gly 213 associated with the buried-helix preference of 4.40. Maximal amphipathicity for the α-helix conformation is found at the Asp 185.

Mature core proteins ucr1_bovin and ucr2_bovin are not predicted with TMH segments. Smaller subunits 6 (ucrq_bovine, 13 kD), 8 (ucrh_bovin, 9.2 kD), and 9 (ucr8_bovin, 8 kD) are also devoid of TMH (with or without modifications caused by Richardson's preferences).

The subunit 10 or ucrx_bovin (7.2 kD, 62 amino acids), subunit 11 or ucry_bovin (6.4 kD, 56 amino acids), and subunit 7 or ucr7_bovin (9.5 kD, 81 amino acids) have potential transmembrane helices. The TMH segment 13−32 or 19−41 is predicted for the subunit 10 by the Kyte−Doolittle and combination of Kyte−Doolittle and Richardson's preferences, respectively. Its predicted orientation of the NH_2-terminus is outside (intramembrane) in accord with charge bias of −2. Maximal α-helix preference for buried α-helix of 4.24 and presumably middle membrane position is at the Val 25. The doublet of arginine residues at positions 15 and 16 may still be in the α-helix conformation, because Richardson's preferences are higher than 1.3 in the whole region from Thr 6 to Ile 42 with maximum higher than 2.0 at Glu 32.

The TMH prediction for the subunit 11 is from residue 22 to 38 with maximum α-helix preference at Ala 29. Modification with Richardson's preferences shift the predicted TMH toward polypeptide's NH_2-terminus: 6−34. The predicted matrix orientation of the NH_2-terminus is in question, since it follows from unresolved charge bias of zero.

The subunit 7 has maximal preference for buried-helix conformation just at the proline doublet position (Pro 50 and 51) of predicted TMH between Ala 43 and Gln 64. The NH_2-terminus position is predicted to be inside (matrix) due to the charge bias of +1. Richardson's preferences for the α-helix conformation are higher than 1.0 from Leu 38 to Thr 60. However, correction due to Richardson's preferences is small: predicted TMH is then 41−64.

(d) Preference Profiles of Pore Domains with Edelman's Preference Functions. *Shorter membrane-buried nontransmemnbrane helices* are also often predicted by our algorithm in integral membrane proteins. The example from previous section is the membrane-buried helix in the external loop EF of cytochrome *b*. Another example is our prediction of membrane-buried helix in the H5 pore segment (the P-segment) of different potassium channels (Figures 4−6). Expected folding motif strand-turn-strand for the P-segment[26] was not confirmed by the X-ray analysis of the potassium channel KcsA from *Streptomyces lividans*.[5b] Instead, the pore helix is found at the sequence location in KcsA (Figure 4) which is homologous to sequence domain in *Shaker* predicted by us as the membrane-buried helix.[12d]

Folding motifs in the pore region can be recognized in very different organisms by using our preference and conformational index profiles[12d] (Figures 4−6). Predicted transmembrane orientation for the NH_2-terminus is cytoplasmic for all potassium channels tested here when Edelman's preference functions are used. Corresponding charge bias is +9 for the cik1_human, +10 for the *Shaker* channel, and +8 for the KcsA channel.[5b]

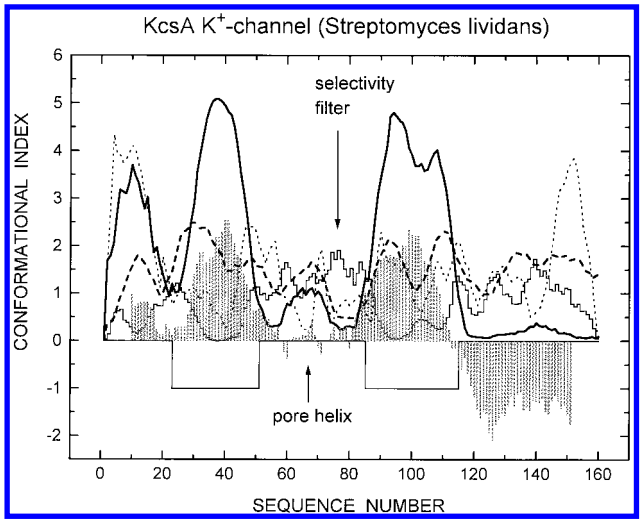

Figure 4. Preference and conformational index profiles for the KcsA K^+-channel of *Streptomyces lividans*. Thick full line and thin line at the level −1.0 are respectively Edelman's preference for membrane-buried α-helix conformation and TMH digital prediction. Jagged thick line is the preference for the β-strand conformation, while thick dashed line is Richardson's preference for α-helix conformation. Dotted line is our threshold index for amphipathic α-helix structure. Shaded area represents direct Kyte−Doolittle average of 19 hydropathy attributes.

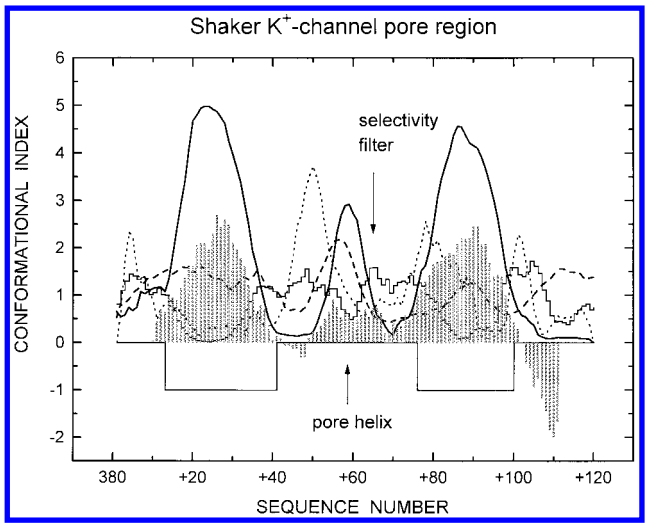

Figure 5. Preference and conformational index profiles (same as in the Figure 4) for the pore region of *Drosophila Shaker* K^+-channel.

Going from NH_2- to COOH-terminus in the pore domain the surface-attached turret region is first encountered in the solved structure of KcsA[5b] followed with short membrane-buried helix and even shorter β-strand serving as the selectivity filter. The whole pore domain is sandwiched in the sequence between two membrane-spanning helices. Strong maximums exist in the KcsA sequence for membrane-spanning helices predicted by SPLIT to extend from Ser 23 to Ala 50 and from Thr 85 to Phe 114 (Figure 4). The pore helix cannot be located at all by using direct Kyte−Doolittle average of hydropathy values,[6a] but the propensity for buried α-helix can be seen as the smaller maximum in the pore region when either Edelman's or Richardson's preference functions are used.

Better identification of candidates for the turret region (as an amphipathic span) and for pore helix (as membrane-buried helix) is achieved in the *Drosophila Shaker* channel (Figure 5) and in its human cousin cik1_human (Figure 6). Associated maximums in amphipathicity are located at the Ser 428 of the *Shaker* channel, and at the Ile 359 of the cik1_human channel. The pore helix of the *Shaker* channel is predicted from Ala 432 to Thr 442 with maximal preference for membrane-buried helix at Val 438. Membrane-buried helix is predicted from Pro 360 to Thr 371 for human cik1 channel with a maximum at Val 368.

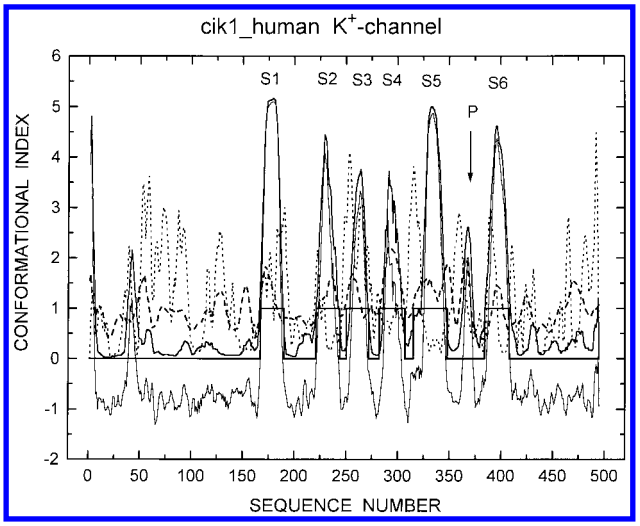

Figure 6. Conformational profiles for the potassium channel cik1_human are obtained with Edelman's preference functions, while digital prediction has been modified with Richardson's preferences. Full thick line and full thin line are respectively preferences for buried α-helix and for buried α-helix minus turn preference. The thick line at the level 1.0 is our prediction for the THM segments S1 to S6. Notice that digital prediction is based on additional filtering rules and on modifications and/or extensions due to Richardson's preferences (thicker dashed line). Therefore, it does not have to have exact correspondence to visual identification of preference peaks. The pore forming helix, where arrow with a label P is pointing, is predicted to extend from Pro 360 to Thr 371. Dotted line has the same meaning as in the Figure 4.

Predicted pore helix has considerable turn potential (one example is shown in Figure 6, where turn preference is the difference between thick and thin full line) even at the maximum of corresponding helix peak. Another aid in the recognition of nontransmembrane character for pore helix is reduced width of associated preference peak, which is less than half of the corresponding width associated with each of two nearby TMH peaks (when measured at the height of 1.0.

Maximal preference for the β-strand conformation (jagged line in Figures 4 and 5) is found next to the sequence location for pore helix. The maximum is at the Val 76 for the KcsA channel, just before the GYG motif of the selectivity filter and at the Gly 444, which is the first glycine of the GYG motif for *Shaker*. Short β-strand is predicted in the cik1_human pore domain too (not shown), as the TVGYG motif which extends from Thr 372 to Gly 376.

Drosophila Shaker and human potassium channel cik1 are expected to have six membrane-spanning helices S1 to S6.[10,27] With Edelman's scale,[22a] as the input, membrane-spanning helices in the *Shaker* channel are predicted as 223−248 (S1), 279−301 (S2), 311−337 (S3), 353−376 (S4), 392−419 (S5) and 455−478 (S6). Predicted TMH locations in the cik1 channel are 167−189 (S1), 221−243 (S2), 250−271 (S3), 282−306 (S4), 315−347 (S5), and 384−407 (S6).

With the Kyte−Doolittle scale,[6a] as the input, highly charged voltage sensor segment S4 is not predicted as the TMH segment. The consequence of such underprediction is the predicted outside location for the NH_2-terminus (due to the charge bias of −7).

(e) Correcting Errors in Protein Databases with Preference Functions. The *precursor* of $GABA_A$ α-5 receptor subunit from rat with 464 amino acids (PIR locus B34130) illustrates how easily errors can be spotted in protein databases when preference functions analysis is applied. Expected TMH are listed in the Protein Identification Resource Protein Sequence Database (PIR) as 229−250 (M1), 255−276 (M2), 288−310 (M3), and 399−420 (M4). The analysis with Edelman's preference functions predicts extracellular location of the NH_2-terminus (with the charge bias of +12) and four TMH at sequence positions (227−253, 259−275, 282−314, 400−421) in *mature* receptor with 433 amino acids without signal sequence of 31 residues. When precursor polypeptide is analyzed with our predictor, all potential transmembrane domains are shifted toward COOH-terminus for 31 residues, and the prediction is radically different from expected TMH positions listed in PIR. It appears that the PIR entry for $GABA_A$ α-5 is using the double standard for counting sequence positions. For the majority of sequence positions it is using the NH_2-terminus of a precursor subunit as the first sequence position, but for transmembrane domains the NH_2-terminus of mature subunit (without signal peptide) is used to start counting residues. This error causes the carbohydrate binding site at Asn 236 to be erroneously located near the middle of the first expected transmembrane domain 229−250. The Swiss-Prot database gives correct sequence locations for expected transmembrane segments of the $GABA_A$ α-5 receptor subunit precursor.

DISCUSSION

Sequence analysis tools for membrane proteins can be tested with already known or soon to be known X-ray structures that are the best "standard of truth". Analysis with preference functions is superior to hydrophobicity analysis available for instance as the SOAP algorithm.[6a] It is competitive with modern pattern recognition methods for detection of membrane-spanning domains in membrane proteins.[11] In this paper we show that Jones et al. MEMSAT algorithm[11a] and Rost et al. PHDhtm refined algorithm[11d] are less accurate than the SPLIT 3.5 algorithm when tested on 21 integral membrane polypeptides of known structure. SPLIT predicts all of the 75 observed membrane-spanning segments as membrane-buried α-helices at their correct sequence positions. Our requirement of having at least nine residues in common between predicted and observed (expected) transmembrane domain is more strict than Rost at al. criteria to have at least three[11b] or five[11d] such residues before correct segment prediction is scored. Per-residue prediction accuracy in crystal structures shows underprediction, which is common to other prediction methods.[11a] Our best result for prediction accuracy on crystal structures of $A_{TM} = 0.809$ represents significant improvement with respect to the $A_{TM} = 0.668$ result achieved with the earlier version (3.1) of the SPLIT algorithm.[12c] However, for the Swiss-Prot "standard of truth" tests on 168 membrane proteins indicated higher accuracy ($A_{TM} = 0.712$) with the SPLIT 3.1. It appears that any improvement leading to the decrease in the number of underpredicted residues for membrane proteins of known crystal structure must also lead to increased unbalance between overpredicted and underpredicted TM helix residues in the Swiss-Prot database. Errors found by us in the Swiss-Prot database[12b−d] and in the PIR database (this paper) and recent improvements in the X-ray analysis of membrane proteins argue for the choice of solved crystal structures as the "standard of truth" for testing the algorithm.

Can solved structures be used not only to test but also to train the algorithm as well? One advantage of the preference functions method is that it does not require more than 30−40 polypeptides for extracting preference functions in the four-state model of helix, strand, turn, and undefined conformation.[12c] By using the jack-knife statistical method, as Rost and Sander did in protein secondary structure prediction,[23] it will be possible in the near future to extract preference functions from known structures of membrane proteins and to test the predictor's accuracy with each of these structures.

The present version of our algorithm uses "positive-inside rule"[19] to predict transmembrane topology with 86% accuracy. For many integral membrane proteins there is no need to predict the NH_2-terminus orientation, because cytoplasmic or extracellular location of both terminals is known from experiments. Such information can be used to improve the prediction accuracy for transmembrane segments, because one can vary the choice of input attributes and reject predicted topological models that are not in accord with observations. In any case a single choice of input attributes (amino acid scale) is not enough for accurate prediction of membrane-spanning segments in different classes of membrane proteins.

One of our objectives in this work, to find sequence location of membrane-spanning helices, should not be confused with the aim of locating transmembrane domains of membrane-spanning helices. Already solved crystal structures of membrane proteins[14] testify that many membrane-spanning helices are so long that their N-caps or C-caps or even both caps are located outside 3.5 nm of membrane lipid interior. Also some membrane proteins span the membrane in such a manner that membrane dielectric interior with low dielectric constant is thinner than the usual 3.5 nm just at the position where some internal helices span the membrane.[28] Our objective was to improve the prediction for the whole span of TM helix even if it is partially extramembrane and/or partially not in direct contact with phospholipid fatty acids. However, to predict the whole length of the membrane-spanning helix it is not enough to rally only on the correspondence between hydrophobicity and propensity for the α-helix formation in the low-dielectric environment. The failure to appreciate this point would contribute errors of underprediction in attempts to associate TMH spans with hydrophobicity maximums in the sequence. Our adopted scheme in this work was to predict first the hydrophobic nucleus of potential TMH and then to extend or shorten its span in both directions by using different devices including β-strand preferences, turn preferences, and even preferences for the α-helix conformation extracted as Richardson's preference functions from soluble proteins. Different as-

signments for TM helices (from published papers, deposited in the PDB or from Swiss-Prot database, only membrane span or whole span) will of course change apparent prediction accuracy for all predictors.

Our algorithm does not use multiple sequence alignments to improve the prediction accuracy, as Rost et al.[11b] and Persson and Argos[29] methods do. The weak point of algorithms that do not use evolutionary information may become a strong point during sequence analysis when homologous sequences are few or entirely lacking. Namely, the lack of knowledge of homology to the tested sequence would be expected to decrease declared prediction accuracy of pattern recognition methods using such knowledge. This is indeed the case.[12c]

The high rate of false positive results of 14% soluble proteins predicted as membrane proteins appears because one long hydrophobic segment is found with similar rate in soluble proteins of known structure. With the Kyte−Doolittle scale[6a] input only 1% of 147 soluble proteins are wrongly predicted as membrane proteins *with two or more TMH* and none with *three or more TMH*. A better choice of amino acid scale is also possible[12c] when the goal is to distinguish polytopic membrane proteins (with more than one TMH) from soluble proteins in putative protein-coding sequences. On the other hand, a predicted "membrane-buried" regular structure known to be located in the soluble protein or in the extramembrane part of membrane protein may be an important functional hydrophobic pocket such as the binding place of dinucleotides or metal centers (see Figure 1 and results for iron−sulfur protein).

We did not perform accuracy analysis for prediction of an amphipathic regular structure by means of the threshold function. Introduction of the threshold function (eq 4) is based upon the assumption that regular amphipathic structure in the membrane tends to acquire maximal possible hydrophobic moment. Indirect evidence in support of using threshold functions is increased accuracy in predicting transmembrane secondary structure obtained in this work.

The best objective test for any predictor is to submit predicted motifs for the membrane protein whose secondary structure is due to be released soon but is still on hold at the time of analysis. Such an example is that of the ubiquinol-cytochrome *c* reductase complex from bovine heart mitochondria.[24] Some structural details are already known. The SPLIT predictor with Kyte−Doolittle input scale[6a] can find 12 out of 13 membrane-spanning helices of the bc_1 complex, most of them in their expected sequence location and transmembrane orientation. One false-positive result is the prediction of TMH in the soluble part of Rieske iron−sulfur protein, just at the sequence position of the loop β4-β5 which contributes Cys 139 and His 141 to the [2Fe-2S] cluster.[30] This segment with very hydrophobic β-strand from tryptophan 132 to valine 138 forms part of the hydrophobic core of central β-sheet in Rieske protein. The TMH serving as the membrane-anchor is not seen by the digital predictor when mature Rieske protein from beef is examined. This is likely to be one false-negative result or TMH underprediction for the bc_1-complex.[24] Homology analysis and comparisons with Rieske proteins from other species, where TMH closer to NH_2-terminus is indeed predicted, convinced us that the sequence region from Thr 40 to Val 55 may originate TMH.

Prediction of membrane-spanning helices (of normal length and short) is essential for testing protein topology in the membrane and for testing prediction accuracy in proteins of known topology. However, the SPLIT predictor output file contains sequence profiles of preferences (helix, strand, turn, undefined, helix-turn) and amphipathic parameters (hydrophobic moments, thresholds for amphipathic structure, optimal twist angles) in addition to secondary structure prediction. Several of these conformational index profiles plotted together for a tested polypeptide are more informative and closer to the raw sequence analysis data than the prediction of membrane-spanning helices, which is the higher-level output of the filtering procedure (see Methods). Predicted preference profiles for membrane-buried helices, that are not membrane spanning, are also very instructive about possible sequence position of functionally important segments. For cation channels from a brain involved in neural communication by means of action potentials, voltage sensor elements, inactivation segments, and P-segments are the most important functional motifs.[9b] These are the motifs needed to open the gate for cations after change in membrane voltage and to close the gate again. Hydrophobicity analysis and modern programs for predicting topology of membrane proteins, such as the above-mentioned MEMSAT[11a] and PHD algorithms,[11b] often fail to recognize the importance of S4 and P regions and consequently are unable to predict correct topology for these proteins. Buried-helix, free-helix, amphipathic helix, β-strand, and turn preference profiles can be used jointly to recognize supersecondary structure of the pore domain as in Figures 4−6. Our prediction that all of the P-segments found so far have a membrane-buried α-helix associated with the pore[12c] has been verified recently in the case of the KcsA potassium channel from *Streptomyces lividans*.[5b]

Shown examples of conformational index profiles for pore domains (and many more that are not shown) reveal several additional points. The first observation is that KcsA K^+ channel is the special case with short pore helix positioning the selectivity filter close to the extracellular membrane surface. In other cases of potassium channels we predict a longer pore helix, which should position the selectivity filter closer to the middle of the membrane phospholipid bilayer. The selectivity filter is always associated with a maximum in the β-strand preference and often with a maximum in our threshold index $I(\beta)$ for amphipathic β-strand conformation. The turret region from the KcsA pore domain can be better recognized in the sequence of other potassium channels, where it is associated with maximum in the threshold index $I(\alpha)$. The last observation about high turn preference for the whole pore domain span and for the whole S4 helix span may be connected with known remarkable movements of the S4 helix across membrane[10] and ensuing conformational dynamics of the pore region, which opens the gate for potassium ions.

CONCLUSION

Three new numerical methods are described in this paper: a method for calculating preference functions to predict transmembrane helices and topology of integral membrane proteins, a method for evaluation of the hydrophobic moment threshold function with a goal to find

potential membrane-bound amphipathic structures, and a procedure for recognition of potential nontransmembrane buried α-helices and pore-forming segments. These methods are included in the algorithm with a single output data file that now runs as an automatic Web and e-mail server. Any triplet of amino acid scales for hydrophobicity or preference values can be used as an input to run the algorithm with a sequence of choice.

ACKNOWLEDGMENT

The authors acknowledge the help of Damir Zucić, from University of Osijek, Croatia, in setting up Web and e-mail server and of Bono Lučié, from The Institute Rugjer Bošković, Zagreb, Croatia, who found some articles mentioned in this work. The kind help of Edward A. Berry with bc_1-complex sequences is also acknowledged. This work was supported by Croatian Ministry of Science Grant 1-03-171.

REFERENCES AND NOTES

(1) Jähnig, F. In *Prediction of protein structure and the principles of protein conformation*; Fasman, G. D., Ed.; Plenum Press: New York, 1989; Chapter 18, p 707.

(2) (a) Deisenhofer, J.; Epp, O.; Miki, K.; Huber, R.; Michel, H. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature* **1985**, *318*, 618−624. (b) Reithmeier, R. A. Characterization and modeling of membrane proteins using sequence analysis. *Current Opinion Struct. Biol.* **1995**, *5*, 491−500.

(3) Popot, J.-L. Integral membrane protein structure: transmembrane α-helices as autonomous folding domains. *Current Opinion Struct. Biol.* **1993**, *3*, 532−540.

(4) (a) Weiss, M. S.; Schulz, G. E. Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* **1992**, *227*, 493−509. (b) Cowan, S. W.; Rosenbusch, J. P. Folding pattern diversity of integral membrane proteins. *Science* **1994**, *264*, 914−916.

(5) (a) Gross, A.; MacKinnon, R. Agitoxin footprinting the *Shaker* potassium channel pore. *Neuron* **1996**, *16*, 399−406. (b) Doyle, D. A.; Cabral, J. M.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* **1998**, *280*, 69−77.

(6) (a) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydrophatic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132. (b) Klein, P.; Kanehisa, M.; DeLisi, C. The detection and classification of membrane-spanning proteins. *Biochim Biophys Acta* **1985**, *815*, 468−476. (c) Engelman, D. M.; Steitz, T. A.; Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **1986**, *15*, 321−353. (d) White, S. H. Global statistics of protein sequences: Implications for the origin, evolution, and the prediction of structure. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 407−439.

(7) Jennings, M. L. Topography of membrane proteins. *Annu. Rev. Biochem.* **1989**, *58*, 999−1027.

(8) (a) Fasman, G. D.; Gilbert, W. A. The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem.* **1990**, *15*, 89−92. (b) Jähnig, F. Structure predictions of membrane proteins are not that bad. *Trends Biochem. Sci.* **1990**, *15*, 93−95.

(9) (a) Miller, C. 1990: *annus mirabilis* of potassium channels. *Science* **1991**, *252*, 1092−1096. (b) Catterall, W. Structure and function of voltage-gated ion channels. *Annu. Rev. Biochem.* **1995**, *64*, 493−531.

(10) Larsson, H. P.; Baker, O. S.; Dhillon, D. S.; Isacoff, E. Y. Transmembrane movement of the *Shaker* K^+ channel S4. *Neuron* **1996**, *16*, 387−397.

(11) (a) Jones, D. T.; Taylor, W. R.; Thornton, J. M. A model approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **1994**, *33*, 3038−3049. (b) Rost, B.; Casadio, R.; Fariselli, P.; Sander, C. Transmembrane helices predicted at 95% accuracy, *Protein Science* **1995**, *4*, 521−533. (c) Rost, B.; Fariselli, P.; Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science* **1996**, *5*, 1704−1718. (d) Rost, B.; Casadio, R.; Fariselli, P. In *Proceedings Fourth International Conference on Intelligent Systems for Molecular Bilogy*; States, D. J., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R. F., Eds.; AAAI Press: Menlo Park, CA, 1996; pp 192−200.

(12) (a) Juretić, D.; Lee, B. K.; Trinajstić, N.; Williams, R. W. Conformational preference functions for predicting helices in membrane proteins. *Biopolymers* **1993**, *33*, 255−273. (b) Lučić. B.; Juretić, D.; Trinajstić, N. Recognition of membrane protein structure from amino acid sequence. In *From Chemical topology to Three-Dimensional Geometry;* Balaban, A. T., Ed.; Plenum Press: New York, 1997; Chapter 5, pp 117−158. (c) Juretić, D.; Lučić, B.; Zucić, D.; Trinajstić, N. In *Theoretical and Computational Chemistry;* Párkányi, C., Ed.; Elsevier Science: Amsterdam, 1998; Vol. 5, Chapter 13, pp 405−445. (d) Juretić, D.; Zucič, D.; Lučić, B.; Trinajstić, N. Preference functions for prediction of membrane-buried helices in integral membrane proteins. *Comput. Chem.* **1998**, In press.

(13) Richardson, J. S.; Richardson, D. C. Amino acid preferences for specific locations at the ends of alfa Helices. *Science* **1988**, *240*, 1648−1652.

(14) (a) Deisenhofer, J.; Epp, O.; Sinning, I.; Michel, H. Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis. J. Mol. Biol.* **1995**, *246*, 429−457. (b) Allen. J. P.; Feher, G.; Yeates, T. O.; Komiya, H.; Rees, D. C. Structure of the reaction center from *Rhodobacter sphaeroides R-26*: The protein subunits. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6162−6166. (c) McDermott, G.; Prince, S. M.; Freer, A. A., Hawthornthwaite-Lawless, A. M.; Papiz, M. Z.; Cogdell, R. J.; Isaacs, N. W. Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **1995**, *374*, 517−521. (d) Kühlbrandt, W.; Wang, D. N.; Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **1994**, *367*, 614−621. (e) Iwata, S.; Ostermeier, C.; Ludwig, B.; Michel, H. Structure at 2.8 Å resolution of cytochrome *c* oxidase from *Paracoccus denitrificans*. *Nature* **1995**, *376*, 660−668. (f) Tsukihara, T.; Aoyama, H.; Yamashita, E.; Tomizaki, T.; Yamaguchi, H.; Shinzawa-Itoh, K.; Nakashima, R.; Yaono, R.; Yoshikawa, S. The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 Å. *Science* **1996**, *272*, 1136−1144.

(15) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(16) Ponnuswamy, P. K.; Gromiha, M. M. Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int. J. Peptide Protein Res.* **1993**, *42*, 326−341.

(17) (a) Thornton, J. M.; Flores, T. P.; Jones, D. T.; Swindells, M. B. Prediction of progress at least. *Nature* **1992**, *354*, 105−106. (b) Rost, B.; Sander, C.; Schneider, R. Redefining the goals of protein structure prediction. *J. Mol. Biol.* **1994**, *235*, 13−26.

(18) The term "digital prediction" is used for automatic assignment by the algorithm of secondary conformation for amino acid residue or sequence segment in a protein.

(19) (a) von Heijne, G. Membrane protein structure prediction-hydrophobicity analysis and the positive inside rule. *J. Mol. Biol.* **1992**, *225*, 487−494. (b) von Heijne, G. Membrane proteins: From sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 167−192.

(20) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **1984**, *179*, 125−142.

(21) Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659−685.

(22) (a) Edelman, J. Quadratic minimization of predictors for protein secondary structure: Application to transmembrane α-helices. *J. Mol. Biol.* **1993**, *232*, 165−191. (b) Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H. Hydrophobicity of amino acid residues in globular proteins. *Science* **1985**, *229*, 834−838. (c) Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1976**, *105*, 1−14. (d) Ponnuswamy, P. K.; Prabhakaran, M.; Manavalan, P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* **1980**, *623*, 301−316. (e) Chou, P. Y.; Fasman, G. D. Prediction of protein secondary structure. *Adv. Enzymol.* **1978**, *47*, 45−148.

(23) Rost, B.; Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.

(24) Xia, D.; Yu, C.-A.; Kim, H.; Xia, J.-Z.; Kachurin, A. M.; Zhang, L.; Yu, L.; Deisenhofer, J. Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. *Science* **1996**, *277*, 60−66.

(25) Berry, E. personal communication.

(26) Soman, K. V.; McCammon, J. A.; Brown, A. M. Secondary structure prediction of the H5 pore of potassium channels. *Protein Eng.* **1995**, *8*, 397−401.

(27) Browne, D. L.; Gancher, S. T.; Nutt, J. G.; Brunt, E. R.; Smith, E. A.; Kramer, P.; Litt, M. Episodic ataxia/myokymia syndrome is associated with point mutations in the human potassium channel gene, KCNA1. *Nature Genet.* **1994**, *8*, 136−140.

(28) Goldstein, S. A. N. A structural vignette common to voltage sensors and conduction pores:*canaliculi*. *Neuron* **1996**, *16*, 717−722.
(29) (a) Persson, B.; Argos, P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* **1994**, *237*, 182−192. (b) Persson, B.; Argos, P. Topology prediction of membrane proteins. *Protein Science* **1996**, *5*, 363−371.
(30) Iwata, S.; Saznovits, M.; Link, T. A.; Michel, H. Structure of a water soluble fragment of the Rieske iron−sulfur protein of the bovine heart mitochondrial cytochrome bc_1 complex determined by MAD phasing at 1.5 Å resolution. *Structure* **1996**, *4*, 567−579.

CI970073A