CROATICA CHEMICA ACTA CCACAA **65** (4) 921–932 (1992)

ISSN 0011-1643
UDC 577.1
CCA—2042

Original Scientific Paper

Secondary Structure of Membrane Proteins: Prediction with Conformational Preference Functions of Soluble Proteins

Davor Juretić

Department of Physics, University of Split, Nikole Tesle 12, 58000 Split, Croatia

Received April 7, 1992

Conformational preference functions are derived from the statistical analysis of the data base of soluble protein structures. These functions use local sequence information to modify the Chou-Fasman's preference of a given residue in a protein for secondary conformation. The secondary structure prediction algorithm that compares preferences is constructed. For the testing set of 14 membrane polypeptides the prediction accuracy is 78% in the three state model and 90% for the α–helix residues alone. Correlation coefficients are 0.58 and 0.57 for the α–helix and turn structures, respectively.

INTRODUCTION

The predicted secondary structure of membrane proteins is of limited reliability when available schemes for predicting the secondary structure of soluble proteins are used.[1,2] Are the folding motifs of membrane proteins so different from such motifs in soluble proteins that predictive schemes trained on soluble proteins are inappropriate for membrane proteins? We shall present in this work the predictive scheme for membrane proteins which is based on a novel statistical analysis of the data base of soluble proteins.

The preference functions method has been described in recent publications.[3,6] It is based on the idea that secondary structure preference of an amino acid in a sequence depends on coded properties of its sequence neighbors. The idea that each amino acid in a sequence influences not only its own conformation, but also the conformation of its sequence neighbors, is not a new one. The best secondary structure prediction programs also take into account local sequence patterns,[7,8] but do not allow a choice of folding parameters that can uncover such patterns in local sequence interactions. We have shown previously that prediction of secondary structure segments can be improved when constant preferences are replaced with preferences that are functions of carefully chosen folding parameters.[6]

In this work, the chosen folding parameters are 20 α–helix conformational parameters of Chou and Fasman.[9,10] These parameters are employed to extract conformational preference functions from the data base of known soluble protein structures. Extracted functions can be used in a secondary structure prediction program[5] for any class of proteins, but our hope was that such functions would prove particularly useful for predicting the secondary structure of membrane proteins. It turned out that success in prediction depended also on the method of evaluation of preference functions in the membrane protein primary structure. The empirical and well known observation that β–forming residues (according to Chou-Fasman's scheme) form α–helices instead of β–sheets in most transmembrane segments is used in the present paper to improve the prediction. Preference functions are evaluated for integral membrane proteins with known transmembrane helices by using 20 β–sheet conformational parameters of Chou and Fasman.[9] The results are more accurate than those obtained after application of the original Chou-Fasman scheme,[11] Garnier-Robson scheme[7] or after application of an improved neural network procedure trained on α-class proteins.[12]

METHODS

The Protein Data Base

The data base consisted of 90 different proteins, known at 3 Å or better resolution, from the Brookhaven Protein Data Bank (PDB).[13] Secondary structures (α-helix, β-sheet, turn and undefined) were assigned to all residues using the Kabsch-Sander program DSSP.[14] In the three state model, turn and undefined conformations were considered jointly as coil conformation.

Preference Functions

The sequence environment of a residue, found in a particular secondary conformation, was defined as an arithmetic average of Chou-Fasman's α-helix preferences[9,10] for its four left and four right neighbors. The frequency distributions of all environments from the data base were collected for each amino acid type and for each secondary conformation considered (α-helix, β-sheet, turn and undefined). These distributions were then approximated by normal curves and preference functions constructed as described elsewhere.[5] The data set of parameters needed for each Gaussian curve is given in Table I. For a chosen amino acid i, the preference for a particular conformation j is found as a ratio of a normal curve for that conformation to a sum of all four normal curves. The preference P_{ij} can have a strong dependence on sequence environment x of amino acid i:[5]

$$P_{ij}(x) = \frac{(N/N_j)\,(N_{ij}/\sigma_{ij})\exp\left(-0.5((x-\mu_{ij})/\sigma_{ij})^2\right)}{\sum_{j=1}^{4}(N_{ij}/\sigma_{ij})\exp\left(-0.5((x-\mu_{ij})/\sigma_{ij})^2\right)} \quad (1)$$

N_{ij} is the total number of environments x associated with amino acid i in conformation j, N_j/N is the fraction of conformation j in the protein set, while μ and σ are the average and sample standard deviation of parameters x, respectively. As an example, the procedure of calculating how preference for the alanine in the α-helix conforma-

TABLE I

Parameters for the construction of Gaussian curves derived from the list⁺ of 90 soluble globular proteins*

AA	N_{ij}	μ_{ij}	σ_{ij}	AA	N_{ij}	μ_{ij}	σ_{ij}
α-helix ($N/N_j = 3.4589$)				β-sheet ($N/N_j = 4.5352$)			
ALA	509	1.0539	0.0948	ALA	199	0.9822	0.1056
CYS	82	1.0123	0.1086	CYS	98	0.9626	0.1016
LEU	433	1.0481	0.0969	LEU	314	0.9857	0.1083
MET	98	1.0639	0.0863	MET	71	0.9939	0.1074
GLU	350	1.0571	0.0921	GLU	117	0.9926	0.1077
GLN	186	1.0482	0.0955	GLN	107	0.9580	0.0952
HIS	103	1.0274	0.0936	HIS	78	0.9877	0.1117
LYS	365	1.0624	0.1003	LYS	160	0.9989	0.1047
VAL	291	1.0493	0.0977	VAL	421	0.9814	0.1041
ILE	223	1.0475	0.0950	ILE	295	0.9719	0.1088
PHE	187	1.0507	0.0875	PHE	165	0.9656	0.1051
TYR	134	1.0390	0.0991	TYR	171	0.9636	0.0960
TRP	69	1.0188	0.0883	TRP	75	0.9619	0.0960
THR	219	1.0394	0.0968	THR	258	0.9741	0.1013
GLY	186	1.0481	0.0999	GLY	183	0.9698	0.1024
SER	225	1.0388	0.0958	SER	241	0.9607	0.0992
ASP	243	1.0491	0.0997	ASP	88	0.9641	0.0932
ASN	149	1.0443	0.1052	ASN	101	0.9658	0.0858
PRO	105	1.0482	0.0895	PRO	55	0.9551	0.1064
ARG	176	1.0580	0.0948	ARG	110	0.9893	0.0968
turn ($N/N_j = 3.9458$)				undefined ($N/N_j = 4.2070$)			
ALA	255	0.9783	0.1077	ALA	266	0.9826	0.1049
CYS	72	0.9515	0.1120	CYS	90	0.9381	0.0882
LEU	189	0.9650	0.0957	LEU	221	0.9786	0.1039
MET	37	0.9722	0.0981	MET	50	0.9814	0.1017
GLU	202	0.9932	0.1065	GLY	134	0.9794	0.0996
GLN	141	0.9750	0.1035	GLN	117	0.9681	0.1070
HIS	77	0.9751	0.1187	HOS	96	0.9917	0.0978
LYS	280	0.9977	0.1035	LYS	206	0.9825	0.1049
VAL	135	0.9629	0.0977	VAL	237	0.9698	0.0986
ILE	100	0.9660	0.1060	ILE	157	0.9714	0.1002
PHE	97	0.9890	0.1109	PHE	107	0.9808	0.1060
TYR	113	0.9545	0.1062	TYR	114	0.9513	0.1076
TRP	42	0.9269	0.1067	TRP	43	0.0795	0.0953
THR	221	0.9463	0.1099	THR	242	0.9747	0.1018
GLY	603	0.9847	0.1088	GLY	316	0.9731	0.1066
SER	319	0.9613	0.1079	SER	313	0.9672	0.1024
ASP	290	0.9828	0.1109	ASP	260	0.9884	0.1012
ASN	246	0.9835	0.1100	ASN	204	0.9807	0.1036
PRO	242	0.9828	0.0998	PRO	279	0.9878	0.0977
ARG	134	0.9569	0.0965	ARG	111	0.9748	0.1056

* N_{ij} is the total number of environments x (see Methods) associated with amino acid i in conformation j, N_j/N is the fraction of conformation j in the protein set, while μ and σ are the average and sample standard deviations of parameters x, respectively.

⁺ The list of 90 soluble proteins (PDB code) is: 1abp, 1acx, 1bp2, 1cac, 1cc5, 1ccr, 1ctf, 1ctx, 1eco, 1fbj, 1fc2, 1fx1, 1gcn, 1gcr, 1gp1, 1gpd, 1hho, 1hip, 1hmg, 1hmz, 1ig2, 1ins, 1ldx, 1lz1, 1mbd, 1mlt, 1pp2, 1ppt, 1pyp, 1rhd, 1rn3, 1sn3, 1ubq, 156b, 155c, 2abx, 2act, 2adk, 2alp, 2apr, 2atc, 2aza, 2b5c, 2cdv, 2cga, 2cpp, 2cyp, 2ebx, 2est, 2fd1, 2gn5, 2grs, 2lh7, 2lhb, 2mdh, 2mt2, 2pab, 2pka, 2rhe, 2rhv, 2sbt, 2sga, 2sns, 2sod, 2stv, 2tbv, 2tgt, 3c2c, 3cna, 3cpv, 3cyt, 3fxc, 3gap, 3icb, 3ldh, 3pgk, 3pgm, 3rp2, 4adh, 4dfr, 4fxn, 4sbv, 451c, 5cpa, 5pti, 5rxn, 6pad, 6pcy, 7tln, 8cat.

tion depends on sequence environment x is illustrated in equation (2). Inserting the values from Table I, the preference function for the alanine in helix becomes:

$$P^{\alpha}_{ALA}(x) = 3.4589 \cdot D_1/(D_1 + D_2 + D_3 + D_4) \quad (2)$$

where

$$D_1 = (509/0.0948)\exp(-0.5((x-1.0539)/0.0948)^2)$$
$$D_2 = (199/0.1056)\exp(-0.5((x-0.9822)/0.1056)^2)$$
$$D_3 = (255/0.1077)\exp(-0.5((x-0.9783)/0.1077)^2)$$
$$D_4 = (266/0.1049)\exp(-0.5((x-0.9826)/0.1049)^2)$$

The Procedure for Testing Membrane Proteins

When membrane proteins were tested, the secondary structure prediction program used β-sheet conformational preferences[9] to calculate the sequence environment x of each residue. That value was used as an argument for conformational preference functions (1). The obtained values for α-helix, β-sheet, turn and undefined preference functions were smoothed by computer as 7 point moving average (for α-helix), 5 point moving average (for β-sheet), or 3 point moving average (for the turn and undefined conformations). The resulting numbers were compared for each residue and the conformation with the highest value assigned to a residue. The decision constants were not used except when so stated. When used, it was to add the constant factor to all preferences for a particular conformation. The decision constants for β-sheet, α-helix and coil were, respectively, labelled DCB, DCH and DCC.

Other Secondary Structure Prediction Programs

The Garnier-Robson program with added features for calculating the prediction accuracy and correlation coefficients was written in FORTRAN as described,[7,15] while Chou-Fasman's program was Prevelige algorithm[16] written in C. Kyte-Doolittle program[17] and an improved version of a neural network program written in the C language was also used.[12] The source codes of all programs for finding and using preference functions are available from the author. Programs are written in FORTRAN 77 and can be run on any personal computer.

Test Sets of Membrane Proteins

The test set of proteins contained two lists of membrane proteins: a shorter one with 5 proteins and a longer one with 14 proteins. In the short list, bacteriorhodopsin,[18] rhodopsin,[19] lactose permease,[20] and subunits L and M of the photosynthetic reaction centre[21,22] were tested just as in the previous work.[6] However, more recent and presumably better known structures of bacteriorhodopsin and lactose permease were used in this work. For the longer test set of membrane proteins, primary and secondary structures were extracted from the SWISSPROT data bank. For all of these proteins, the secondary structure of transmembrane segments is either known or assumed to be the α-helical structure. The location of transmembrane segments is based on hydrophobicity analysis and experimental data collected by the cited authors. All residues found in extramembrane segments are assumed to be in the undefined conformation except for the case when their structure is known.

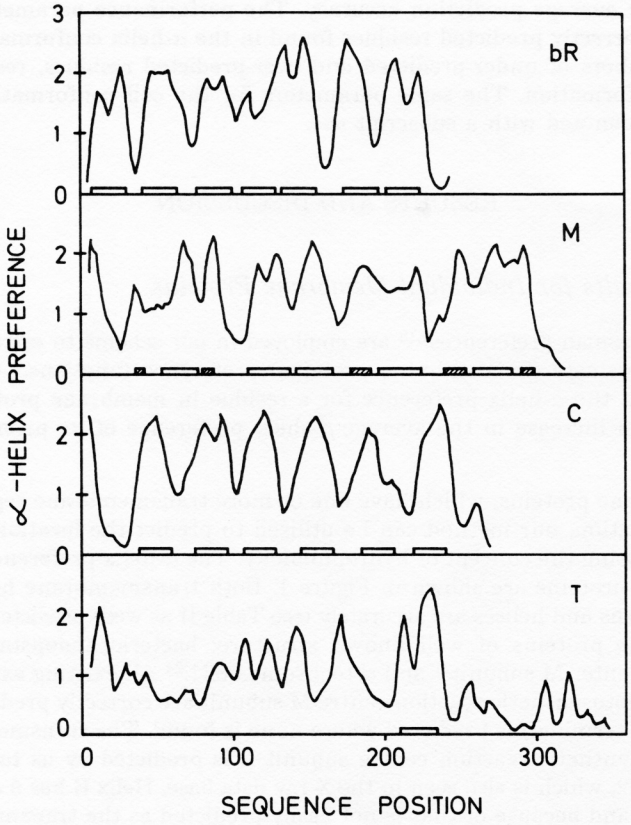

Figure 1. Preference profiles for the α-helix conformation predicted for four membrane proteins: bacteriorhodopsin (bR), photosynthetic reaction centre M subunit (M), cytochrome b561 (C), and T-cell surface antigen (T). Preference functions are calculated (see Methods) by using P_α,conformational parameters of Chou and Fasman[9] and plotted against P_β environments in membrane proteins. P_β environments are calculated from P_β conformational parameters of Chou and Fasman.[9] The preferences are smoothed by computer (as a seven point average) and the resulting points are connected by hand. Experimental data for the α-helical transmembrane segments (bR and M) or the best estimates for the location of such segments from the hydrophobic analysis (C and T) are shown on the x axis in the form of empty boxes while shaded boxes denote α-helical segments found outside the membrane.

Performance Parameters

The correlation coefficient C_α of Matthews[23] is used to estimate how well the predicted secondary structure conformation is correlated with the observed one for each secondary structure of type α. The success rate (in the three state model) is found as percentage Q_3 of correctly predicted residues. An overall Q_3 index (average prediction accuracy) for a list of proteins is calculated as the weighted average of all Q_3 indexes for individual proteins so that longer proteins give a correspondingly larger con-

tribution to the average prediction accuracy. The performance parameter $Q_{3\alpha}$ is the percentage of correctly predicted residues found in the α-helix conformation, while y_α and z_α are numbers of under-predicted and over-predicted residues, respectively, for the α-helix conformation. The same parameters for the coil conformation (turn and undefined) are denoted with a subscript »c«.

RESULTS AND DISCUSSION

Prediction Results for Individual Membrane Proteins

The Chou-Fasman preferences[9,10] are employed in our scheme to extract (in soluble proteins) and evaluate (in membrane proteins) preference functions.[5] Our basic assumption is that the α-helix preference for a residue in membrane protein increases together with an increase in the average β-sheet preference of its primary structure neighbors.

For membrane proteins, which have one or more transmembrane segments in the α-helix configuration, our method can be utilized to predict the location of such segments without using the concept of hydrophobicity. The α-helix preference profiles for four membrane proteins are shown in Figure 1. Both transmembrane helices and extramembrane turns and helices are accurately (see Table II as well) predicted for three integral membrane proteins of well known structure: bacteriorhodopsin,[18] photosynthetic reaction center M subunit[21] and cytochrome b561.[24] All existing extramembrane helices of the photosynthetic reaction centre M subunit are correctly predicted, but one such helix (at N-terminal) is predicted where none is found. The transmembrane helix B of the photosynthetic reaction centre subunit L is predicted by us to extend from residues 85 to 112, which is also seen in the X-ray data base. Helix B has 5 charges in the C-terminal half and because of that is not easily predicted as the transmembrane segment by Kyte-Doolittle[17] and similar schemes.[25]

A set of rules for locating candidate transmembrane segments in the α-helix conformation can be automated or applied manually after preference profiles have been created. For instance, an α-helix preference greater than 1.25 in 14 or more consecutive residues would identify all transmembrane segments in the proteins from Figure 1. The detected α-helix transmembrane segment for the T-cell surface antigen CD2 precursor[26] illustrates that membrane anchors can be predicted as well. The transbilayer segments of the glycophorin A precursor,[27] which is known to span the red cell membrane,[28] and the membrane anchor of the photosynthetic reaction centre H subunit (*Rhodobacter sphaeroides*)[29] are also correctly predicted. The signal peptide of the glycophorin A precursor is predicted as an α-helix extending over the first 20 amino acids on the N-terminal, while transmembrane segment is predicted as an α-helix from residues 91 to 118. The transmembrane segment 12–31 of subunit H is slightly over-predicted as an α-helix extending from residues 11 to 34.

Long helices outside the membrane are not confused with transmembrane helices. One such example is TolA: an *E. coli* membrane protein that contains an extended helical region.[30] Its one membrane spanning region in the N-terminal region is correctly predicted by our program, but in the very long α-helical region, extending from amino acids 48 to 310, the α-helical conformation is not found at all.

TABLE II

Performance parameters for predicting the secondary structure of integral membrane proteins*

A) Prediction results when preference functions are compared with all decision constants set to zero.

Protein+ (# A.A.)	Q_3	C_α	$Q_{3\alpha}$	y_α	z_α	C_c	Q_{3c}	y_c	z_c
ARAB(472)	69	0.56	97	8	107	0.47	40	139	4
BROD(248)	85	0.68	92	14	19	0.66	68	24	10
C561(273)	73	0.69	96	5	40	0.53	49	69	4
CIKA(360)	68	0.59	95	6	78	0.49	53	111	4
CO44(400)	69	0.57	96	8	90	0.49	46	117	5
GALA(494)	70	0.61	93	16	88	0.51	50	132	8
LAC2(416)	72	0.52	96	9	85	0.41	33	107	8
MDR1(372)	62	0.54	98	3	102	0.43	44	138	3
PRCL(273)	74	0.54	83	30	28	0.53	63	34	22
PRCM(323)	76	0.63	90	20	37	0.53	63	41	25
RHOD(348)	68	0.49	89	21	68	0.40	42	91	15
VIRU(97)	74	0.55	89	2	17	0.47	71	23	2
OPS1(373)	74	0.64	90	18	52	0.55	59	79	13
HMDH(240)	75	0.53	88	19	30	0.48	50	40	13
OVERALL	71	0.58	93			0.49	50		

B) Prediction results with an improved neural network program trained on soluble α-class proteins[12]

Protein+ (# A.A.)	Q_3	C_α	Cc
BROD(248)	79	0.51	0.51
PRCL(273	72	0.49	0.45
PRCM(323)	75	0.57	0.53

* Percentage Q_3 of correctly predicted residues. Correlation coefficient C_α of Matthews.[23] Percentage $Q_{3\alpha}$ of correctly predicted residues found in the α-helix conformation. Numbers y_α of under-predicted and z_α of over-predicted residues, respectively, for the α-helix conformation. The same parameters for the coil conformation (turn and undefined) are denoted with subscript »c«. Overall Q_3 and C indexes for a list of proteins are calculated as the weighted average of indexes for individual proteins so that longer proteins give a correspondingly larger contribution to the average prediction accuracy.

+

ARAB - Arabinose–H^+ transporter (*E. coli*)[43]
BROD - Bacteriorhodopsin (*H. halobium*)[18]
C561 - Cytochrome b561 (bovine)[24]
CIKA - S1–S6 segments (181–541 fragment) from potassium channel protein (fruit fly)[44]
CO44 - C-terminal fragment (amino acids 1201–1600) from sodium channel protein (electric eel)[45]
GALA - galactose transporter (without first 80 amino acids at N-terminal)(yeast)[46]
LAC2 - Lactose transporter (*E. coli*)[20]
MDR1 - N-terminal fragment (1–372) of the P-glycoprotein)[47]
PRCL - Photosynthetic reaction centre L subunit (*R. viridis*)[21]
PRCM - Photosynthetic reaction centre M subunit(*R. viridis*)[21]
RHOD - Rhodopsin (human)[19]
VIRU - Matrix M2 protein of influence virus (strain A/Bangkok/1/79)[48]
OPS1 - Opsin RH1 from photoreceptor cells (fruit fly)[49]
HMDH - 3–Hydroxy–3–Methylglutaryl–Coenzyme A reductase (human). N-terminal fragment (first 240 amino acids)[50]

Testing Sets of Integral Membrane Proteins

The performance parameters for 14 integral membrane proteins are listed in Table II. Comparison of parameters y and z reveals that α-helical segments are over-predicted, while coil (turn and undefined conformations) segments are under-predicted. Addition of constant values (»decision constants«) to smoothed preferences can improve performance. For instance, the overall success rate in the three state model (α-helix, β-sheet and coil structure) is increased from 71% to 78% when the following decision constants are used: DCB = –0.2, DCH = 0.0, DCC = 0.2. The correlation coefficient and the accuracy of coil prediction is then also increased to 0.57 and 68%, respectively, but this improvement is due to our initial assumption that all extramembrane residues (except in the photosynthetic reaction centre) are in the undefined conformation. The correlation coefficient for the α-helix structure remained 0.58, while the accuracy of helix prediction decreased to 90%.

In our recent paper,[6] 55 different hydrophobicity scales were tested with respect to their ability to predict the secondary structure of membrane proteins. In each case, the same scale was used to extract preference functions from the data base of soluble proteins and to evaluate and compare these functions for membrane protein sequences. Our scale of conformational parameters, which are formed for each amino acid as an average of the corresponding Chou-Fasman's parameters[9,10] for α-helix and β-sheet, was the best predictor of secondary structure segments in five membrane polypeptides. The best accuracy was 67% for all conformations, while the prediction accuracy and correlation coefficient were 76% and 0.46, respectively, for the α-helix conformation alone. These performance parameters are clearly inferior to the results obtained in the present paper: the overall prediction accuracy of 74%, and α-helix prediction accuracy and correlation coefficient of 91% and 0.56, respectively, for the same testing set of five membrane polypeptides. This result was obtained with all decision constants set to zero.

Comparison with Other Methods

The presented empirical method can be compared with other proposed methods for locating transmembrane helices,[17,25] for deciphering the topography of membrane proteins[31–35] and for secondary structure prediction.[7,8,11] The need to »train« preference functions on a sufficiently large data set of solved protein structures is common with more sophisticated empirical prediction algorithms of statistical nature[7,8] Our algorithm, based on eq. (1), is simpler and also provides for greater flexibility both in the choice of 20 initial folding parameters and in the choice of 20 parameters for the evaluation of preference functions. The training set of crystallographically solved soluble proteins defines the initial set of constant preferences. An optimal set of preferences for the evaluation of preference functions depends on the nature of tested proteins. The training set of proteins does not have to be the set used 15 years ago by Chou and Fasman or the set used in this paper. As long as all protein classes are included in a training set, the results (not shown) are not very sensitive to the choice of the training set. A quite different situation arises during the testing procedure. An optimal set of preferences for the evaluation of preference functions depends strongly on the nature of tested proteins. These preferences can be found either automatically[36] or by using empirical knowledge about the tested proteins as in this paper.

Most of the other secondary structure prediction programs, trained on soluble proteins, have a much weaker performance when applied to membrane proteins. For instance, the Garnier-Osguthorpe-Robson program[7,15] applied to the same set of 14 membrane polypeptides results in a prediction accuracy of 51% (in the three state model), while correlation coefficients for the helix and turn (or coil) structure are only 0.11 and 0.20, respectively. For the testing set of five membrane proteins, the Garnier-Osguthorpe-Robson program predicted 58% of residues in the correct conformation (in the three state model). Correlation coefficients for the α-helix and turn (or coil) structures were 0.23 and 0.25, respectively. The choice of decision constants was: DCH = 100, DCC = 0, DCB = 50, which is appropriate for proteins having more than 50% of helical residues.

The Chou-Fasman-Prevelidge algorithm[16] applied to the same testing set of five membrane proteins predicts only 34% of residues to be in correct conformation. Such prediction results are not better than random association of secondary conformations to the primary structure (in the three state model). In addition, most transmembrane segments are wrongly predicted to be in the β-sheet conformation. However, the Chou-Fasman preferences[9,10] can still be valuable tool for predicting the structure of membrane proteins, as seen in this paper. The most likely reason for this is that statistical preferences contain important determinants for the protein folding process besides the solution hydrophobicity values.[37] The importance of steric factors, the accessible and buried surface area for detecting helical structure in membrane proteins has been pointed out earlier.[5,38]

The performance of an improved neural network program, trained on soluble α-class proteins[12] is quite good (Table IIB). The results tend to confirm that there is similarity of folding motifs for helices in soluble and membrane bound proteins. Our results (Table IIA) for predicting helices in membrane proteins are even better than neural network results. In part, our results also confirm the similarity of folding motifs in soluble and membrane proteins because we have been using α-helix conformational parameters and a set of preference functions, all extracted from a data base of soluble proteins, to predict helices in membrane proteins. In addition, it was simple in the preference functions method to take into account one important difference in folding motifs: that transmembrane helices often contain stretches of β-forming residues.

The fact that for a specific protein class - integral membrane proteins with transmembrane helices, our simple scheme works better than the best available secondary structure prediction algorithms[7,12] provides a clue why our method may be expected to work better in certain situations. Both the conformational preferences and coded folding properties of amino acids in a sequence of tested proteins are used by us. The methods that use hydrophobicity alone to locate secondary structure elements, such as transmembrane helices, do not make use of pattern recognition techniques, while methods that use such techniques do not make use of coded folding properties in tested sequences. The prediction accuracy for secondary structure segments of integral membrane proteins may well depend on the ability of a prediction scheme to find folding parameters that are most important for the folding process of such proteins.

Figure 1 results for the photosynthetic reaction centre illustrate that transmembrane helices, just as extramembrane helices, can be located without using the polarity profile analysis, and often with better accuracy. As another example, the nice separation of transmembrane peaks for the case of cytochrome b561 (Figure 1. C), is not so easily seen in the polarity profile analysis. Furthermore, the high accuracy of turn prediction

in our scheme can be combined with the pattern of minima and maxima in the α-helix preferences for more precise determination of transmembrane segments.

A better positive prediction can be achieved when transmembrane segments contain an unusually high number of charges like in the case of helix B of the photosynthetic reaction center L subunit. Amphiphilic helices are often difficult to predict as transmembrane helices by polarity analysis. Since such helices play a major role in the structure and function of membrane transport proteins and membrane active polypeptides, it is of interest that the membrane spanning ability of amphiphilic helices can be evaluated by our method. The amphiphilic helix of mellitin, which can assume transmembrane orientation[39] is an example of the transmembrane helix being predicted by our criteria. Of the two regions in the nicotinic acetylcholine receptor, MA and M2, which are proposed as ion channel forming regions,[40] only the M2 segment is found by our program with a capability to span bilayer.

An example of better negative prediction is trypsinogen, a soluble protein with an apparent »transmembrane segment« according to the polarity analysis.[25] No such candidate transmembrane segment is found in trypsinogen when preference profiles for trypsinogen are created by using our scheme.

The weakness of the presented empirical method is its poor performance in predicting the location of extramembrane and transmembrane β-sheet structures. For instance, for porin[41] and Omp A fragment[42] from *E. coli*, the overall accuracy of β-strand prediction is $Q_{3\beta}$ = 38%, while the correlation coefficient C_β = 0.26. For 16 soluble proteins of the β-class,[6] the same parameters are 50% and 0.27, respectively.

Another problem of the presented prediction scheme is that extramembrane helices can be predicted (Figure 1) but prediction accuracy does not exceed the performance parameters achieved with the neural network procedure (not shown). Long extramembrane helices are missed (the TolA example mentioned before) just because the prediction scheme has been fine tuned to predict transmembrane helices. The best strategy for membrane proteins may be the application of our method to locate transmembrane helices and then the neural network procedure to locate extramembrane secondary structure elements.

When a structure of a larger number of membrane polypeptides becomes known, it will become possible to extract preference functions from such a data base and to choose separate scales of 20 folding parameters that are most appropriate for predicting either α-helix or β-strand structures. The scheme described in this paper will easily select the optimal hydrophobicity scale or an optimal scale of statistical preferences for each secondary structure considered.

Acknowledgements. – Help from Dr. V. Turk's laboratory (Dr. Eva Žerovnik and Dr. Roman Jerala) at the Josef Stefan Institute in Ljubljana, Slovenia, in finding structures of membrane proteins for analysis is gratefully acknowledged. An improved version of neural network program, kindly made available by Dr. D. G. Kneller from the Department of Pharmaceutical Chemistry, University of San Francisco, CA 94122, U.S.A., was tested on membrane proteins by Bono Lučić at the Rudjer Bošković Institute, Zagreb, Croatia. This work was supported by the Croatian Ministry of Science Grant 1–03–171.

REFERENCES

1. B. A. Wallace, M. Cascio, and L. Mielke, *Proc. Natl. Acad. Sci. USA* **83** (1986) 9423.
2. F. Jahnig, *Prediction of protein structure and the principles of protein conformation,* G. D. Fasman (ed.), Plenum Press, New York, 1989, p. 707.
3. D. Juretić and B. Lee, *Biophys. J.* **55** (2/2) (1989) 354a.
4. D. Juretić., *Period. Biol.* **93** (1991) 279.
5. D. Juretić, B. Lee, N. Trinajstić, and R. W. Williams, *Biopolymers* (1993) in press.
6. D. Juretić, N. Trinajstić, and B. Lučić, *J. Math. Chem.* (1992) in press.
7. J. Garnier, J. Osguthorpe, and B. Robson, *J. Mol. Biol.* **120** (1978) 97.
8. N. Qian and T. J. Sejnowski, *J. Mol. Biol.* **202** (1988) 865.
9. P. Y. Chou and G. D. Fasman, *Annu. Rev. Biochem.* **47** (1978) 251.
10. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47** (1978) 45.
11. P. Y. Chou and G. D. Fasman, *Biochemistry* **13** (1974) 211.
12. D. G. Kneller, F. E. Cohen, and R. Langridge, *J. Mol. Biol.* **214** (1990) 171.
13. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. J. Meyer, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112** (1977) 535.
14. W. Kabsch and C. Sander, *Biopolymers* **22** (1982) 2577.
15. R. W. Williams, A. Chang, D. Juretić., and S. Loughran, *Biochim. Biophys. Acta* **916** (1987) 200.
16. P. Prevelige and G. D. Fasman, *Prediction of protein structure and the principles of protein conformation* G,. D. Fasman (ed.), Plenum Press, New York 1989, p. 391.
17. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157** (1982) 105.
18. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing, *J. Mol. Biol.* **213** (1990) 899.
19. R. P. Birge, *Biochim. Biophys. Acta* **1016** (1990) 293.
20. H. R. Kaback, *Biochim. Biophys. Acta* **1018** (1990) 160.
21. J. Deisenhofer, O. Epp. K. Mikki, R. Huber, and H. Michel, *Nature* **318** (1985) 618.
22. H. Michel, K. A. Weyer, I. Gruenberg, I. Dunger, D. Oesterhelt, and F. Lottspeich, *EMBO J.* **5** (1986) 1149.
23. B. W. Matthews, *Biochim. Biophys. Acta* **405** (1975) 442.
24. M. S. Perin, V. A. Fried, C. A. Slaughter, and T. C. Suedhof, *EMBO J.* **7** (1988) 2697.
25. D. M. Engelman, T. A. Steitz, and A. Goldman, *Ann. Rev. Biophys. Biophys. Chem.* **15** (1986) 321.
26. D. J. Diamond, L. K. Clayton, P. H. Sayre, and E. L. Reinherz, *Proc. Natl. Acad. Sci. U.S.A.* **85** (1988) 1615.
27. S. Kudo and M. Fukuda, *Proc. Natl. Acad. Sci. U.S.A.* **86** (1989) 4619.
28. V. T. Marchesi, H. Furthmayr, and M. Tomita, *Annu. Rev. Biochem.* **45** (1976) 667.
29. J. C. Williams, L. A. Steiner, and G. Feher, *Proteins Struct. Funct. Genet.* **1** (1986) 312.
30. S. K. Levengood, W. F. Beyer, and R. E. Webster, *Proc. Natl. Acad. Sci. U.S.A.* **88** (1991) 5939.
31. G. von Heijne, *Eur. J. Biochem.* **174** (1988) 671.
32. D. C. Rees, L. DeAntonio, and D. Eisenberg, *Science* **245** (1989) 510.
33. M. L. Jennings, *Annu. Rev. Biochem.* **58** (1989) 999.
34. D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, *J. Mol. Biol.* **179** (1984) 125.
35. J. K. M. Rao and P. Argos, *Biochim. Biophys. Acta* **869** (1986) 197.
36. D. Juretić, B. Lučić, and N. Trinajstić, *Croat. Chem. Acta* (1993) in press.
37. M. Charton and B. I. Charton, *J. Theor. Biol.* **102** (1983) 121.
38. D. Juretić and R. W. Williams, *J. Math. Chem.* **8** (1991) 229.
39. C. Kempf, R. D. Klausner, J. N. Weinstein, J. V. Renswoude, M. Picus, and R. Blumenthal, *J. Biol. Chem.* **257** (1982) 2469.
40. P. Ghosh and R. M. Stroud, *Biochemistry* **30** (1991) 3551.
41. E. Schiltz, A. Kreusch, U. Nestel, and G. E. Schulz, *Eur. J. Biochem.* **199** (1991) 587.
42. H. Vogel and F. Jähnig, *J. Mol. Biol.* **190** (1986) 191.
43. M. C. J. Maiden, E. O. Davis, S. A. Baldwin, D. C. M. Moore, and P. J. F. Henderson, *Nature* **325** (1987) 641.

44. O. Pongs, N. Kecskemethy, R. Mueller, I. Krah-Jentgens, A. Baumann, H. H. Kiltz, I. Canal, S. Llamazares, and A. Ferrus, *EMBO J.* **7** (1988) 1087.
45. M. Noda, S. Shimizu, T. Tanabe, T. Takai, T. Kayano, T. Ikeda, H. Takahashi, H. Nakayama, Y. Kanaoka, N. Minamino, K. Kangawa, H. Matsuo, M. A. Raftery, T. Hirose, S. Inayama, H. Hayashida, T. Miyata, and S. Numa, *Nature* **312** (1984) 121.
46. K. Szkutnicka, J. F. Tschopp, L. Andrews, and V. P. Cirillo, *J. Bacteriol.* **171** (1989) 4486.
47. C. Chen, E. J. Chin, K. Ueda, D. P. Clark, I. Pastan, M. M. Gottesman, and I. B. Roninson, *Cell* **47** (1986) 381.
48. R. A. Lamb, S. L. Zebedee, and C. D. Richardson, Cell **40** (1985) 627.
49. J. E. O'Tousa, W. Baehr, R. L. Martin, J. Hirsh, W. L. Pak, and M. L. Applebury, *Cell* **40** (1985) 839.
50. K. L. Luskey and B. Stevens, *J. Biol. Chem.* **260** (1985) 10271.

SAŽETAK

Sekundarna struktura membranskih proteina: Predviđanje s pomoću konformacijskih funkcija sklonosti za topljive proteine

Davor Juretić

Konformacijske funkcije sklonosti dobivene su statističkom analizom baze podataka struktura topljivih proteina. Te se funkcije koriste podacima o primarnoj strukturi ispitivanog proteina da bi se modificirala Chou-Fasmanova sklonost uočenog aminokiselinskog ostatka za sekundarnu strukturu. Izgrađen je algoritam koji predviđa sekundarnu strukturu uspoređivanjem sklonosti. Za proučavani skup od 14 membranskih polipeptida točnost predviđanja iznosi 78% u modelu sa tri stanja i 90% za aminokiselinske ostatke koji imaju konformaciju α-zavojnice, a koeficijenti korelacije su 0.58 za strukturu α-zavojnice i 0.57 za strukturu zavoja.