CROATICA CHEMICA ACTA CCACAA **66** (1) 201–208 (1993)

ISSN 0011–1643
CCA–2132 *Original Scientific Paper*

Predicting Membrane Protein Secondary Structure: Preference Functions Method for Finding Optimal Conformational Parameters

Davor Juretić[+], *Bono Lučić** *and Nenad Trinajstić**

[+]*Physics Department, Faculty of Science, University of Split, N. Tesle 12, 58000 Split, Croatia.*

Received April 9, 1992.

An automated iterative method is developed for predicting secondary conformation in membrane proteins. The initial set of parameters are α-helix preferences and associated conformational preference functions extracted from the data set of known soluble protein structures. The secondary structure segments are assigned to each of 14 tested membrane proteins by using the prediction method, which evaluates and compares preference functions in the tested protein. A new set of parameters are then calculated which is based on the predicted protein structure from the previous iterative cycle. The method takes advantage of the similarities in local sequence patterns found in the tested proteins. Residues in membrane proteins are predicted with 84% accuracy and with the correlation coefficient for the α-helix structure equal to 0.68, which is a considerably better performance than that of neural network programs or Garnier-Robson's algorithm.

INTRODUCTION

A knowledge of protein three-dimensional structure is necessary to propose realistic models for protein function. More than 300 protein structures have been determined from the X-ray diffraction patterns of crystallized proteins, but the number of known sequences with unknown three-dimensional structure is about 100 times larger. This situation is even worse for membrane proteins because three-dimensional structures of sufficient resolution to allow detailed analysis are available for only a few of them.[1–3]

* Rudjer Bošković Institute. P.O.B. 1016, 41001 Zagreb, Croatia

A goal of the theoretical methods of structure determination is to predict the tertiary structure of folded protein. A protein folds spontaneously into the three-dimensional structure under the direction of its amino acid sequence,[4] but folding rules are still largely unknown. A more modest goal of locating regular secondary structure elements along the chain has been vigorously pursued during the last 20 years because an accurate secondary structure prediction can be used as the first step in tertiary structure prediction by energy minimization or other predictive techniques.[5] In addition, recent experimental work has suggested that secondary structure formation precedes tertiary organization.[6]

Methods for secondary structure prediction often use statistical analysis of the data base of crystallographically solved soluble proteins to predict the secondary structure of unsolved protein.[7–9] Amino acids have distinct conformational preferences.[10] A statistical method that can use conformational preferences is the recently described preference functions method.[11–13] It has been applied to prediction of the membrane proteins secondary structure[14] and shown to be much more accurate than Chou-Fasman's method.[7] Indeed, the membrane proteins structure is not well predicted by other statistical methods.[15] Instead, hydrophobicity analysis is widely used to predict the location of transmembrane segments in the primary structure of such proteins.[16–18] Recently, computational neural networks have proved superior to any other method in predicting the α-helix structure in soluble proteins,[19,20] and have been applied to predicting the secondary structure in membrane proteins as well.[21,22]

The aim of this work is to examine whether preference functions can be used to exploit similarities in folding patterns of integral membrane proteins that have well known locations of helical segments. We have shown before that the preference function method can be used to select an optimal amino acid scale for predicting the secondary structure of a chosen set of test proteins.[23] In this work, a chosen set of integral membrane proteins is used for an automated training procedure which results in optimal amino acid scales for all proteins of the set. The results obtained for the prediction of α-helices are considerably better than those obtained with neural network programs, trained on α-class proteins,[20] or Garnier-Robson's[8,24] algorithm.

METHOD

Two training sets of proteins are used: a set of 90 soluble proteins with secondary structures assigned by the DSSP program of Kabsch and Sander,[25] and a set of 14 integral membrane proteins (Table I). Since only membrane proteins with residues predominantly in two conformational states, helix and coil, are tested, all secondary structures of soluble proteins are modified so that only helix and coil conformations are retained. The method of preference functions as well as additional details concerning selection and preparation of protein data bases have been described previously.[13,14] The present data base of 14 membrane proteins has more recent structures for photosynthetic reaction center subunits (Table I), which is the only difference. Another data base of soluble proteins, all of α-class and all known to equal or better than 2.5 Å resolution, has been also selected from the Brokhaven Protein Data Bank (PDB).[26] These 55 proteins are listed here with their PDB code: 1abp, 1bp2, 1cc5, 1ccr, 1cpv, 1csc, 1cyc, 1ecd, 1fdh, 1hds, 1hmq, 1gox, 1hho, 1hmd, 1lrd, 1lz1, 1mba, 1mbd, 1mbs, 1omd, 1p2p, 1pmb, 1ppt, 1r69, 1sdh, 1ycc, 1ypi, 256b, 2ccy, 2cdv, 2cpp, 2cro, 2cyp, 2dhb, 2lh1, 2lhb, 2lz2, 2lzm, 2mhr, 2mlt, 2ts1, 2utg, 2wrp, 3adk, 3c2c, 3cln, 3hhb, 3icb, 3ins, 3pgm, 451c, 4cts, 5cyt, 5tnc, 7xia.

In this work, the method of preference functions[13,14] uses preferences as preference environments. Preference environments x are calculated by averaging four left and four right preferences of neighboring residues in the sequence for all residues in the data base of soluble protein structures. The frequency distributions of environments for each amino acid type and for each secondary conformation are approximated using normal curves $G(x)$.[13] For a chosen amino acid type i, the preference for a particular conformation j is found as a ratio of a normal curve for that conformation to a sum of normal curves for helix and coil conformations:

$$P_{ij}(x) = \frac{(N/N_j)\, G_{ij}(x)}{G_{i1}(x) + G_{i2}(x)}$$

where

$$G_{ij}(x) = (N_{ij}/\sigma_{ij}) \exp\{-0.5[(x-\mu_{ij})/\sigma_{ij}]^2\}$$

N_{ij} is the total number of environments x in the frequency distribution, 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 correlation between the structure and amino acid sequence environment x is found when preference function $P_{ij}(x)$ depends strongly on x.

New FORTRAN programs SPH, PDIS, SP1 and SP2 were written for this work. Programs SPH and PDIS are used on soluble proteins for extracting an initial set of α-helix preferences (SPH) and for finding class limits for preference environments that would leave a roughly equal number of environments in each class (PDIS). Two previously described programs, PR and FREQ,[13] are used after SPH and PDIS, to extract conformational preference functions from the data set of soluble proteins. Preference functions are utilized by iterative programs SP1 and SP2.

The same training set of membrane proteins is used both to produce an optimized set of conformational parameters (program SP1), and to predict the secondary structure of individual proteins from that set (program SP2). The membrane protein to be tested is first removed from the training set (jack knife test[27]). Program SP1 uses the overall correlation coefficient for helical residues of all membrane proteins, except the tested one, to stop the iteration when correlation starts decreasing. The output of SP1, 14 optimized sets of conformational parameters (with a choice of the above mentioned training set of membrane proteins), is used as input for program SP2, which predicts the structure of the »unknown« protein after two additional iterations on that protein. Prediction results are smoothed along the sequence by averaging seven preferences in the case of helix and five preferences in the case of coil conformation.

In each iterative cycle the old scale of conformational parameters is »mixed« with a new one derived from the prediction results. A »mixing« factor of 0.6 was used for all calculations, which means that the conformational parameter for each amino acid type retained 60% of the initial value from the previous iterative cycle.

New conformational parameters are calculated from predicted secondary structures found in the previous iterative cycle just as preferences would be calculated from known secondary structures.[10] The SP1 program uses the predicted structures of all proteins (except for the one to be tested with the SP2 program) to calculate a new set of parameters, while the SP2 program iterates only one tested protein at a time. Both

programs use the same set of conformational preference functions, but these functions are evaluated by new conformational parameters in each iterative cycle and the results for helix and coil preferences are compared to predict the secondary structure.

The procedure for testing protein of a completely unknown secondary structure is slightly different. Such protein must also have both primary and secondary structures present in the protein file. All residues in its secondary structure are then assumed to be in the unknown (U) conformation. During the second stage of the training procedure (iterative programs), such protein must be present in the list of membrane proteins of known or partially known secondary structures. Of course, all the reported performance parameters would be meaningless in that case.

The performance parameters used in this work are the correlation coefficient C_h of Matthews[28] for the α-helix structure, the percentage Q_2 of correctly predicted residues, and the percentage Q_{2h} of correctly predicted residues in helical conformation. The C_h and Q_2 values reported in our two state model (helix and coil) can be compared with the previously reported C_α and Q_3 values for membrane proteins that have very few, if any, β-sheet residues.[14] The overall performance parameters are calculated as the weighted average of parameters for individual proteins. The source codes of FORTRAN programs mentioned in this work are available from the first author.

RESULTS AND DISCUSSION

The final results are presented in Table I. The larger number of iterative cycles with program SP2 increase prediction accuracy for the photosynthetic reaction center[29] (C_h above 0.6 for both subunits), but a decrease occurs for some other proteins so that, on average, it does not help to continue iteration on individual proteins in the absence of some physical, chemical or structural criteria that can tell us when the iteration procedure with the protein of unknown structure should stop.

To compare these results with the results obtained in other laboratories, we have used the best algorithm that was available to us: the neural network method trained on α-class proteins.[20] These results are also provided in Table I. Naturally, the disadvantage of the neural network method in this comparison was that it was trained only on soluble proteins. This observation serves to underline the fact that the neural network method works best when it can be trained on the largest number of proteins belonging to the same class as the tested proteins and that the set of membrane proteins of solved structure is extremely limited and, therefore, not suitable as a training set for that method.

Our method works even with only two proteins if they are homologous. For instance, the same iterative procedure applied to the training and testing of the photosynthetic reaction center L and M subunits[29] results in correlation coefficients for helical residues of 0.68 and 0.59 for L and M subunits, respectively. As it can be seen from Table I, these performance parameters decrease when the training set of proteins used for the extraction of optimized preferences is enlarged with proteins belonging to the same class but not necessarily homologous to reaction center subunits. Sequence homology is not used directly by our algorithms, but as a secondary structure homology between the predicted and experimental structures for the training set of membrane proteins. There is a feedback mechanism that modifies initial preference as long as the above mentioned homology increases.

The performance parameters obtained with the Garnier-Robson algorithm[8,24] applied on our list of membrane proteins are not given in Table I, but we give here the overall values of 52% for the accuracy, and 0.12 for the correlation coefficient for residues in helical conformation. These results were obtained using decision constants appropriate for α-class proteins.[8]

In this work, optimal conformational parameters for evaluating preference functions are found automatically. Therefore, both the initial and final choice of conformational parameters are not arbitrary, but defined by the chosen training set of proteins and derived by our programs SPH and SP1, respectively. The final set of conforma-

TABLE I

Results for predicting the secondary structure of integral membrane proteins

Protein code*	Iteration results+			Neural network results#		
	Q_2(%)	Q_{2h}(%)	C_h	Q_3(%)	Q_{3h}(%)	C_h
BROD	85	86	0.67	79	87	0.51
PRCM	81	79	0.61	75	79	0.57
PRCL	72	70	0.43	72	69	0.49
LAC2	87	92	0.71	59	74	0.10
RHOD	89	88	0.79	48	64	−0.07
CIKA	89	93	0.78	65	94	0.44
GALA	85	87	0.70	58	75	0.19
VIRU	85	94	0.84	61	100	0.41
ARAB	82	88	0.65	57	85	0.14
C561	90	93	0.80	71	90	0.45
CO44	81	85	0.63	63	85	0.31
MDR1	82	97	0.67	50	85	0.18
OPS1	85	88	0.70	59	76	0.21
HMDH	87	84	0.74	62	79	0.06
Overall	84	87	0.68	63	79	0.28

*The nonstandard code used by us for integral membrane proteins listed below:

BROD – Bacteriorhodopsin (H. halobium).[2]
PRCL – Photosynthetic reaction center L subunit (R. viridis).[29]
PRCM – Photosynthetic reaction center M subunit (R. viridis).[29]
LAC2 – Lactose transporter (E. coli).[30]
RHOD – Rhodopsin (human).[31]
CIKA – S1–S6 segments (181–541 fragment) from the potassium channel protein (fruit fly).[32]
GALA – galactose transporter (without first 80 amino acids at the N-terminal)(yeast).[33]
VIRU – Matrix M2 protein of the influenza virus (strain A/Bangkok/1/79).[34]
ARAB – Arabinose-H^+ transporter (E. coli).[35]
C561 – Cytochrome b561 (bovine).[36]
CO44 – C-terminal fragment (amino acids 1201–1600) from the sodium channel protein (electric eel).[37]
MDR1 – N-terminal fragment (1–372) of the P-glycoprotein[38] (human).
OPS1 – Opsin RH1 from photoreceptor cells (fruit fly).[39]
HMDH – 3-Hydroxy-3-Methylglutaryl-Coenzyme A reductase (human). N-terminal fragment (first 240 amino acids).[40]

+ The performance parameters Q_2, Q_{2h}, and C_h are, respectively, prediction accuracy, prediction accuracy for the α-helix residues and the correlation coefficient[28] for the α-helix. The correlation coefficient for the coil (or undefined) conformation is not listed, because it is equal to C_h in our two state (helix and coil) model.

The three state model (helix, sheet and coil) was used by the neural network program[20], but performance parameters are directly comparable with our two-state model results because in all 14 proteins only several residues from photosynthetic reaction center subunits are in the β-sheet conformation.

tional parameters for each of 14 membrane proteins included preferences close to 1.4 for isoleucine, valine and phenylalanine, which represented a significant increase over initial values (close to 1.0) for the α-helix preferences for these amino acids. The iterative procedure has arrived, in effect, at the values of conformational parameters that are closer to β-sheet preferences than to α-helix preferences. The results in Table I are in accord with a recent observation[14] that α-helix preference functions extracted from the soluble proteins data base can be used for accurate prediction of α-helical segments if these functions are evaluated with β-sheet preferences in tested membrane proteins. The advantage of finding the optimal conformational parameters for each tested protein in this work, with respect to our earlier results[14] is seen as a 10 point increase in overall accuracy (percentage) and an 0.10 increase in the correlation coefficient for α-helix residues.

In order to check how much the results depend on the choice of the training set of soluble proteins, we selected 55 soluble proteins of α-class (listed in Methods) and extracted α-helix preferences and α-helix preference functions from such a data base. The results obtained by our iterative programs applied to the same set of 14 membrane proteins were similar: the overall accuracy 84% and overall helix correlation coefficient 0.67. The selected data base of α-class proteins included 22 α-class proteins that Kneller *et al.*[20] used for training and testing the neural network program. The same set of 22 proteins was used for extracting α-helix preferences and α-helix preference functions. By using these preference functions, the same accuracy of 84% and helix correlation coefficient of 0.67 were obtained in our iterative procedure with 14 membrane proteins. This performance in testing membrane proteins is much better than that of the neural network program trained on the same set of soluble proteins (Table I).

To check prediction on soluble proteins, the remaining 33 proteins of α-class were used to extract α-helix preference functions. These functions were evaluated by our iterative programs on the Knellers data set of 22 α-class proteins. Overall performance parameters of 65% for accuracy and of 0.33 for helix correlation coefficient indicated that our present procedure should not be used without modification if protein classes other than integral membrane proteins with transmembrane α-helices are examined.

Acknowledgements. — We are grateful to C. Sander and R. Schneider of the European Molecular Laboratory, Heidelberg, Germany, for the use of protein data bases DSSP and HSSP. We acknowledge support by the Ministry of Science of the Republic of Croatia *via* grants 1–03–171 (D. J. and B. L.) and 1–07–159 (B. L. and N. T.).

REFERENCES

1. J. Deisenhofer, O. Epp, K. Mikki, R. Huber, and H. Michel, *Nature* **318** (1985) 618.
2. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing, *J. Mol. Biol.* **213** (1990) 899.
3. M. S. Weis, A. Kreusch, E. Schiltz, U. Nestel, W. Welte, J. Weckesser, and G. E. Schulz, *FEBS Lett.* **280** (1991) 379.
4. C. B. Anfinsen, E. Haber, M. Sela, and F. H. Jr White, *Proc. Natl. Acad. Sci., U.S.A.* **47** (1961) 1309.
5. T. A. Webster, R. H. Lathrop, and T. F. Smith, *Biochemistry* **26** (1987) 6950.
6. J. B. Udgaonkar and R. L. Baldwin, *Nature (London)* **335** (1988) 694.
7. P. Y. Chou and G. D. Fasman, *Biochemistry* **13** (1974) 222.
8. J. Garnier, J. Osguthorpe, and B. Robson, *J. Mol. Biol.* **120** (1978) 97.

9. F. R. Maxfield and H. A. Scheraga, *Biochemistry* **18** (1979) 5138.
10. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47** (1978) 45.
11. D. Juretić and B. Lee, *Biophys. J.* **55** (2/2) (1989) 354a.
12. D. Juretić, *Period. Biolog.* **93** (1991) 279.
13. D. Juretić, B. Lee, N. Trinajstić, and R. W. Williams, *Biopolymers* **33** (1993) 255.
14. D. Juretić, *Croat. Chem. Acta* **65** (1992) 921.
15. B. A. Wallace, M. Cascio, and L. Mielke, *Proc. Natl. Acad. Sci. USA* **83** (1986) 9423.
16. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157** (1982) 105.
17. D. M. Engleman, T. A. Steitz, and A. Goldman, *Ann. Rev. Biophys. Biophys. Chem.* **15** (1986) 321.
18. D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, *J. Mol. Biol.* **179** (1984) 125.
19. N. Qian and T. J. Sejnowski, *J. Mol. Biol.* **202** (1988) 865.
20. D. G. Kneller, F. E. Cohen, and R. Langridge, *J. Mol. Biol.* **214** (1990) 171.
21. H. Bohr, J. Bohr, S. Brunak, M. J. Cotterill, B. Lautrup, L. Norskov, O. H. Olsen, and S. B. Petersen, *FEBS Lett.* **241** (1988) 223.
22. R. Casidio, P. Fariselli and M. Compiani, *Biophys. J.* **61** (2/2) (1992) 350a.
23. D. Juretić and B. Lučić, *J. Mat. Chem.* (1993) in press.
24. R. W. Williams, A. Chang, D. Juretić, and S. Loughran, *Biochim. Biophys. Acta* **916** (1987) 200.
25. W. Kabsch and C. Sander, *Biopolymers* **22** (1983) 2577.
26. 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.
27. B.Efron, *The-Jack Knife, the Bootstrap and Other Resampling Plans*, Society of Industrial and Applied Mathematics, Philadelphia 1982.
28. B. W. Matthews, *Biochim. Biophys. Acta* **405** (1975) 442.
29. J. Deisenhofer and H. Michel, *Science* **245** (1989) 1463.
30. H. R. Kaback, *Biochim. Biophys. Acta* **1018** (1990) 160.
31. R. P. Birge, *Biochim. Biophys. Acta* **1016** (1990) 293.
32. 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.
33. K. Szkutnicka, J. F. Tschopp, L. Andrews, and V. P. Cirillo, *J. Bacteriol.* **171** (1989) 4486.
34. R. A. Lamb, S. L. Zebedee, and C. D. Richardson, *Cell* **40** (1985) 627.
35. M. C. J. Maiden, E. O. Davis, S. A. Baldwin, D. C. M. Moore, and P. J. F. Henderson, *Nature* **325** (1987) 641.
36. M. S. Perin, V. A. Fried, C. A. Slaughter, and T. C. Suedhof, *EMBO J.* **7** (1988) 2697.
37. 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.
38. C. Chen, J. E. Chin, K. Ueda, D. P. Clark, I. Pastan, M. M. Gottesman, and I. B. Roninson, *Cell* **47** (1986) 381.
39 J. E. O'Tousa, W. Baehr, R. L. Martin, J. Hirsh, W. L. Pak, and M. L. Applebury, *Cell* **40** (1985) 839.
40. K. L. Luskey and B. Stevens, *J. Biol. Chem.* **260** (1985) 10271.

SAŽETAK

Predviđanje sekundarne strukture membranskih proteina: metoda sklonosnih funkcija za nalaženje najboljih konformacijskih parametara

Davor Juretić, Bono Lučić i Nenad Trinajstić

Razvijena je automatska iterativna metoda za predviđanje sekundarne strukture membranskih proteina. Početni skup parametara čine sklonosti aminokiselina za konformaciju α-uzvojnice i pridružene im funkcije konformacijskih sklonosti koje se dobiju iz baze podataka topljivih proteina poznate strukture. Elementi sekundarne strukture pridruže se svakomu od 14 testiranih membranskih proteina tako da se izračunaju i uspoređuju numeričke vrijednosti za funkcije sklonosti u tim proteinima. Zatim se izračuna novi skup parametara koji se osniva na predviđenoj sekundarnoj strukturi iz prethodnoga iteracijskog ciklusa. Ova metoda koristi se sličnostima koje postoje u primarnim strukturama skupine proteina za testiranje. Točnost predviđanja konformacije aminokiselinskih ostataka u membranskim proteinima jest 84%, a koeficijent korelacije za strukturu α-zavojnice iznosi 0.68, što je bitno bolji rezultat od onoga koji se može dobiti primjenom programa neuronske mreže ili Garnier-Robsonova algoritma.