Pergamon

Computers Chem. Vol. 22, No. 4, pp. 279–294, 1998

Printed in Great Britain
0097-8485/98 $19.00 + 0.00

PII: S0097-8485(97)00070-3

Preference Functions for Prediction of Membrane-buried Helices in Integral Membrane Proteins

Davor Juretić,[1]* Damir Zucić,[2] Bono Lučić[3] and Nenad Trinajstić[3]

[1]Physics Department, Faculty of Science and Education, University of Split, N. Tesle 12, HR-21000 Split, Croatia, [2]Faculty of Electrical Engineering, University of Osijek, Istarska 3, HR-31000 Osijek, Croatia and [3]The Rugjer Bošković Institute, P.O. Box 1016, HR-10001 Zagreb, Croatia

(Received 24 July 1997; Accepted 27 October 1997)

Abstract—The preference functions method is described for prediction of membrane-buried helices in membrane proteins. Preference for the α-helix conformation of amino acid residue in a sequence is a non-linear function of average hydrophobicity of its sequence neighbors. Kyte–Doolittle hydropathy values are used to extract preference functions from a training data set of integral membrane proteins of partially known secondary structure. Preference functions for β-sheet, turn and undefined conformation are also extracted by including β-class soluble proteins of known structure in the training data set. Conformational preferences are compared in tested sequence for each residue and predicted secondary structure is associated with the highest preference. This procedure is incorporated in an algorithm that performs accurate prediction of transmembrane helical segments. Correct sequence location and secondary structure of transmembrane segments is predicted for 20 of 21 reference membrane polypeptides with known crystal structure that were not included in the training data set. Comparison with hydrophobicity plots revealed that our preference profiles are more accurate and exhibit higher resolution and less noise. Shorter unstable or movable membrane-buried α-helices are also predicted to exist in different membrane proteins with transport function. For instance, in the sequence of voltage-gated ion channels and glutamate receptors, *N*-terminal parts of known P-segments can be located as characteristic α-helix preference peaks. Our e-mail server: predict@drava.etfos.hr, returns a preference profile and secondary structure prediction for a suspected or known membrane protein when its sequence is submitted.

Key words: integral membrane proteins, secondary structure prediction, α-helices, hydrophobicity, preference functions, voltage-gated ion channels

1. INTRODUCTION

Regular secondary structure of fully saturated hydrogen bond patterns is expected for polypeptide membrane domains (Jähnig, 1989). Ten residues or even less are sufficient to span the membrane as β-strands (Weiss and Schulz, 1992; Cowan and Rosenbusch, 1994), but 20 or more residues are commonly found in membrane-spanning helices (Deisenhofer *et al.*, 1985; Reithmeier, 1995; Pebay-Peyroula *et al.*, 1997).

Hydrophobicity plots (Kyte and Doolittle, 1982) have often been used to reveal putative transmembrane helices (TMH) in integral membrane proteins. Hydrophobicity analysis is still being regarded as the best tool for sequence analysis (White, 1994). Different improved schemes that use hydropathy values to predict transmembrane domains have been developed during the past 10 years (Klein *et al.*, 1985; von Heijne, 1986; Bangham, 1988; Ponnuswamy and Gromiha, 1993). However, tested prediction accuracy is low (Fasman and Gilbert, 1990; Jähnig, 1990; Ponnuswamy and Gromiha, 1993) with polypeptides of known structure from the photosynthetic reaction center (Deisenhofer *et al.*, 1985). Prediction depends on the subjective choice of sliding window length and threshold height. These parameters are often chosen to support experimental observations and homology analysis for tested sequences. Hydrophobicity analysis advocated erroneous secondary structure models of some membrane proteins (Fasman and Gilbert,

*Author for correspondence; e-mail: juretic@-mapmf.pmfst.hr.

1990; Wo and Oswald, 1995b) and predicted transmembrane segments in soluble proteins (Jennings, 1989). Recently found evidence that the transmembrane segment, expected to have the α-helix conformation, can contain only ten or fewer residues (Goldstein, 1996), has caused new problems in attempts to use the hydrophobicity profile as a guide for uncovering the secondary structure. These problems were anticipated by Lodish (1988), who pointed out that inner transmembrane helices in multi-spanning membrane proteins need not be in contact with lipids at all and can span the membrane in much less than 20 residues.

The main difficulty in predicting sequence location and the secondary structure of membrane-spanning segments with modern pattern recognition statistical methods (Edelman, 1993; Jones *et al.*, 1994; Rost *et al.*, 1995) is the limited data base of membrane proteins of known structure. Few known crystal structures of membrane proteins provide the parameters and also (inappropriately) serve as test cases for the predictive methods (Reithmeier, 1995). Prediction accuracy is still not satisfactory. For instance Rost *et al.*'s (1995) neural network predictor overpredicted membrane-spanning segments in the photosynthetic reaction center subunits M and L, underpredicted the first transmembrane helix in plant light-harvesting protein LHC_II, and did not recognize membrane-spanning segments in integral membrane protein FtsH from *Escherichia coli* and spiralin from *Spliroplasma melliferum*, i.e. it did not recognize these proteins as membrane proteins. Important extension is the topology prediction for transmembrane proteins, such as that of Rost *et al.* (1996a), which predicts the orientation of all protein domains with respect to membrane. The "positive inside rule" is then used, which is the observation that positively charged residues are more abundant at the cytoplasmic membrane side (von Heijne, 1992). Overall protein topology is *not* predicted in this work.

The purpose of this study is to illustrate the advantage of using preference functions for predicting sequence position of membrane-buried helices. The preference function method (Juretić *et al.*, 1993) can predict not only sequence location, but also the secondary structure conformation of membrane-buried polypeptide segments. It associates sequence hydrophobicity with statistical propensities for conformational motifs. The essential step in the prediction process is the comparison of preferences for α-helix, β-sheet, turn and undefined conformation for each residue in a sequence. Soluble β-class proteins are used to train the algorithm to predict β-sheets, while membrane proteins are used to train the algorithm to predict membrane-buried helices. Of all the predicted membrane-buried helices, some are associated with a high enough preference peak and large enough peak width to be selected as potential membrane-spanning helices. Other predicted membrane-buried helices are not membrane-spanning. Some are associated with pore-forming P-segments in voltage-gated ion channels (Catterall, 1995). The prediction accuracy for transmembrane helices is superior to different versions of hydrophobicity analysis. Similar results in predicting sequence location of transmembrane segments are obtained as with the neural network and pattern recognition methods without the need to use homologous sequences. Homologous sequences are very often missing for new sequences that are appearing daily from different genome projects (Grey, 1996).

2. MATERIALS AND METHODS

2.1. Amino Acid Attributes and Sequence Environ-environment

Each amino acid from the protein data set was associated with its type, its known or expected secondary structure conformation and with average hydrophobicity (sequence environment) of its five left and five right sequence neighbors. The hydrophobicity of the central amino acid in the sliding window was *not* included for the calculation of its

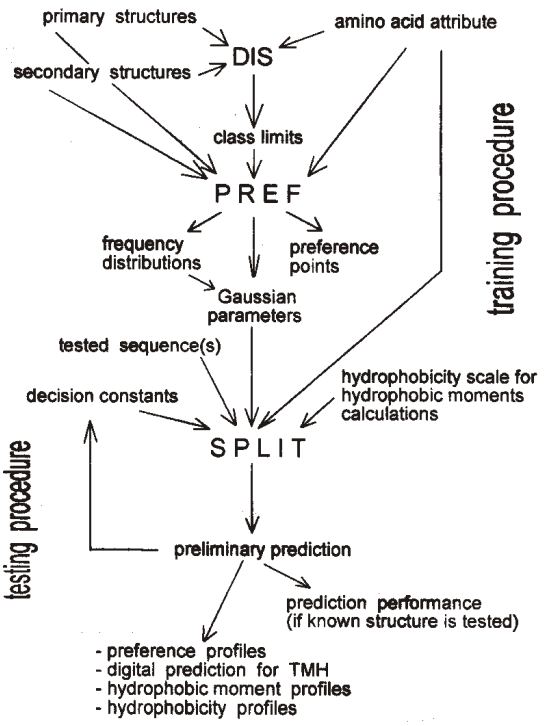

Fig. 1. The flow diagram for the PREF-SPLIT suite of algorithms. The input for the training procedure is a chosen scale of amino acid attributes and a protein data set with assigned primary and secondary structure conformation of each residue. The file with the class limits for collected sequence environments is the output of the algorithm DIS. The PREF algorithm than produces: (a) histograms of environments for each amino acid and each secondary structure (frequency distributions); (b) preference points; and (c) the file with Gaussian parameters such as Table 1. To examine the protein of unknown secondary structure, its sequence is presented to the predictor (SPLIT) taking care that the same scale of amino acid attributes is the input as previously during the training procedure. SPLIT makes its own choice of decision constants if given such freedom. Its filter in the digital predictor splits long predicted helices in two or three transmembrane helical segments (TMH) each having around 20 residues. The output of SPLIT are also numerous performance parameters that are meaningful when tested sequences have a known secondary structure. Preference, hydrophobicity and hydrophobic moment profiles are all included in the output file.

sequence environment. During the training process, the sliding window algorithm, named PREF, collected sequence environments of all amino acids from 172 polypeptides. In the standard procedure, the Kyte–Doolittle hydropathy scale (Kyte and Doolittle, 1982) was used. Other amino acid attributes can be used too among 88 different scales of physical, chemical, biological or mathematical parameters available in the algorithm. To calculate sequence environments, the chosen amino acid scale was normalized with an average of zero and standard deviation of one. All sequence environments were grouped into nine classes. The data file with class limits was created by our fast iterative algorithm (named DIS in the flowchart of Fig. 1) that grouped approximately equal number of environments into each class.

2.2. Protein Databases

The training database consisted of 37 soluble β-class proteins of known crystal structures and 135 integral membrane proteins. Kabsch and Sander's (1983) assignments of secondary structures were used for soluble proteins. We utilized the four-state model with α-helix ("H"), β-sheet ("B"), turn ("T") and undefined ("U") secondary structure. Soluble proteins have been selected from the Protein Data Bank among β-class proteins known with equal or better than 3 Å resolution. When more than one chain was present in the protein, only the first polypeptide chain denoted with the last letter "1" has been selected. Their PDB codes are: 1acx, 1bbp1, 1cd4, 1fdl1, 1hne1, 1mcp1, 1paz, 1pfc, 1rbp, 1rei, 1sgt, 1ton1, 1trm1, 2alp, 2apr, 2aza1, 2fb41, 2fbj1, 2gch1, 2cna, 2gcr, 2i1b, 2ltn, 2pcy, 2pka1, 2ptn, 2rhe, 2rsp1, 2sga, 2sod1, 2tbv1, 3est, 3rp2, 3sgb1, 4ape, 4cms1, 5pep.

The protein data set of 63 integral membrane proteins, with known sequence position of transmembrane helices, common to us, Rost *et al.* (1995) and Jones *et al.* (1994), was used to improve the predictor's performance. Swiss-Prot codes for selected proteins are enclosed here. The letter "s" was added when appropriate to show that signal sequence has been removed: 4f2_human, 5ht3_mouse(s), a1aa_human, a2aa_human, a4_human(s), aa1r_canfa, aa2a_canfa, adt_ricpr, bach_halhm, bacr_halha, cb21_pea, cek2_chick(s), cyoa_ecoli(s), cyob_ecoli, cyoc_ecoli, cyod_ecoli, cyoe_ecoli, edg1_human, fce2_human, glp_pig, glpa_human(s), glpc_human, glra_rat(s), gmcr_human(s), gp1b_human(s), gpt_crilo, hema_cdvo, hema_measa, hema_pi4ma, hg2a_human, iggb_strsp, il2a_human(s), il2b_human(s), ita5_mouse, lacy_ecoli, lech_human, leci_mouse, lep_ecoli, magl_mouse(s), malf_ecoli, motb_ecoli, mprd_human(s), myp0_human(s), nep_human, ngfr_human(s), oppb_salty, oppc_salty, ops1_calvi, ops2_drome, ops3_drome, ops4_drome, opsb_human, opsd_human, opsg_human, opsr_human, pigr_human, ptma_ecoli, sece_ecoli, tcb1_rabit, trbm_human(s), trsr_human, vmt2_iaann, vnb_inbbe.

Another database of 105 integral membrane proteins was selected by us among 4000 such proteins in the Swiss–Prot data base releases 29 and 31 (Bairoch and Boeckmann, 1994). Each of 105 proteins was less than 30% homologous to other proteins from that data set and to all proteins from the 63-protein data set. Swiss–Prot codes for 105 selected proteins are listed below. Appropriate release numbers and the letter "s" (indicating that signal sequence has been removed) are in parentheses: ach1_xenla(29s), acm5_human(29), adt2_yeast(29), ag22_mouse(29), aqp1_human(29), athb_rat(29), athp_neucr(29), atm1_yeast(31), atn1_human(29), atp9_wheat(29), atpl_ecoli(31), b3at_human(29), c561_bovin(29), cadn_mouse(29s), car1_dicdi(29), cb2r_human(29), cd2_human(29s), cd7_human(29s), cd72_human(29), cd8a_human(29s), cgcc_bovin(29), cic1_cypca(29), cikl_drome(29), cox2_parli(29), cox9_yeast(29), cp5a_cantr(29), cxb5_rat(29), cyda_ecoli(29), cydb_ecoli(29), cyf_brara(29), dhg_ecoli(31), dhsc_bacsu(29), divb_bacsu(29), dmsc_ecoli(29), dsbb_ecoli(31), egf_mouse(31), exbb_ecoli(29), fixl_rhime(29), fmlr_rabit(29), frdd_provu(29), ftsl_ecoli(29), ftsh_ecoli(29), furi_human(29s), g2lf_human(29), gaa1_bovin(29), gasr_human(29), gcsr_human(31s), ghr_human(29s), grhr_human(29), ha21_human(29s), hb23_mouse(29), hly4_ecoli(29), hmdh_human(29), imma_citfr(29), isp6_yeast(29), itb1_human(29s), kdgl_ecoli(31), kgtp_ecoli(29), lha1_rhosh(29), lhb4_rhopa(29), ly49_mouse(29), m49_strpy(29s), malg_ecoli(29), mas6_yeast(31), mdr3_human(31), melb_ecoli(29), mepa_mouse(31s), mota_ecoli(29), mpcp_rat(29), mypr_human(29), nals_bovin(29), nk11_mouse(29), nntm_bovin(31), nram_iabda(29), och1_yeast(29), oec6_spiol(29), psaa_pinth(31), psab_pinth(31), psbi_horvu(29), ptgb_ecoli(31), secy_ecoli(29), spc2_canfa(29), spir_spime(29), stub_drome(29), sy65_drome(29), syb1_human(29), synp_rat(29), tal6_human(29), tapa_human(29), tat2_yeast(31), tca_human(29), tcc1_mouse(29), tcrb_bacsu(31), tee6_strpy(29s), tgfa_human(31s), thas_human(29), tnfa_bovin(29), tnr1_human(31), tsa4_giala(31s), ucp_rat(29), va34_vaccc(29), vca1_human(29s), vglg_hrsva(29), vs10_rotbn(29), wapa_strmu(29s).

We used 5-times cross-validation to obtain representative results for the total set of 63 + 105 = 168 integral membrane proteins. Each tested group (33 or 34 proteins) had a similar distribution with respect to the number of transmembrane segments.

The set of 135 membrane proteins, used in the standard training procedure, is obtained when the following 33 proteins are removed from the total set of 168 proteins: cd72, cd7, cd8a, cek2, cp5a, egf, vca1, va34, tsa4, trsr, trbm, ghr, glp, glpa, glpc, gmcr, gp1b, atpl, exbb, cxb5, dsbb, atm1, bach, car1, cb2r, cyda, edg1, fmlr, opsb, athp, gpt, b3at, tat2.

All membrane proteins were of the α-class with approximately known sequence position of membrane-spanning helices. Therefore, only α-helix (for membrane-spanning helix), turn (for four residues next to each helix cap) and undefined (for all remaining residues) conformation were assigned to these proteins.

The reference set of 21 integral membrane polypeptides with known sequence location of trans-

Table 1. Gaussian parameters as the input file* for the SPLIT algorithm

	4.7798	
1324	0.6363	0.2986
278	0.6299	0.3108
1938	0.6271	0.2966
450	0.6058	0.2892
111	0.4364	0.3981
144	0.5298	0.3563
156	0.5660	0.2800
99	0.4406	0.3633
1391	0.6415	0.2954
1377	0.6189	0.3001
1080	0.6054	0.3117
472	0.5546	0.3007
283	0.5767	0.3118
600	0.6719	0.2761
974	0.6679	0.2729
707	0.6297	0.3101
119	0.5074	0.3866
229	0.5691	0.3295
346	0.6289	0.3186
75	0.5072	0.3592
	19.6912	
184	0.1749	0.2724
78	0.0945	0.2995
276	0.0619	0.2791
50	0.1286	0.2349
85	0.1389	0.2802
124	0.1610	0.2639
53	0.1585	0.2439
112	0.1394	0.2724
339	0.0871	0.2582
208	0.0482	0.2781
178	0.1079	0.2978
160	0.0446	0.2764
64	0.1016	0.2827
270	0.1736	0.2759
194	0.1664	0.2828
265	0.1631	0.2749
85	0.2009	0.3334
96	0.2429	0.2808
47	0.2091	0.2277
82	0.2239	0.3199
	7.2830	
550	0.1724	0.3036
124	0.1235	0.3083
517	0.1327	0.3182
142	0.1925	0.3310
360	0.1949	0.3386
317	0.1668	0.3243
176	0.1876	0.3322
512	0.1817	0.3349
389	0.0973	0.3104
283	0.0980	0.3136
301	0.1346	0.3299
277	0.1658	0.2993
142	0.1740	0.3455
511	0.1622	0.3197
758	0.1783	0.3034
701	0.1735	0.3241
451	0.1881	0.3324
458	0.1900	0.3425
471	0.1691	0.3166
533	0.2248	0.3486
	1.6591	
2441	0.0193	0.3086
676	−0.0196	0.3005
2948	−0.0323	0.3005
688	0.0166	0.3167
2393	0.0052	0.3151
1578	0.0025	0.3096
763	0.0353	0.3094
2090	−0.0023	0.3216
2106	−0.0038	0.3020
1701	−0.0089	0.2889
1241	−0.0141	0.2959
1048	0.0090	0.3007
544	−0.0398	0.2924
2281	0.0243	0.3126
2411	0.0450	0.3139
2639	0.0210	0.3088
1907	0.0305	0.3206
1677	0.0591	0.3180
2065	0.0097	0.3235
1816	0.0226	0.3128

*The training with the PREF algorithm produces different numbers in the input file for each choice of hydrophobicity scale and protein database. Presented input file, used in this paper, has been created with the Kyte–Doolittle hydropathy values and protein database consisting of 135 integral membrane proteins and 37 β-class soluble proteins. The number of amino acid residues (the first column), average sequence environment (the second column) and sample standard deviation of sequence environments (the third column) are listed in the format ready to be used by the SPLIT algorithm. Four blocks of 20 values are respectively for the α-helix, β-sheet, turn and undefined conformation. Numbers before each block represent the inverse value of the fraction of corresponding secondary structure conformation in the protein data set. Twenty amino acid types in each block are listed in the order: Ala, Cys, Leu, Met, Glu, Gln, His, Lys, Val, Ile, Phe, Tyr, Trp, Thr, Gly, Ser, Asp, Asn, Pro, and Arg.

membrane helical segments was collected from high-resolution membrane–protein structures determined by X-ray diffraction. It consisted of photosynthetic reaction center subunits H, L and M from *Rhodobacter viridis* (Deisenhofer *et al.*, 1985, 1995) and *Rhodobacter sphaeroides* (Allen *et al.*, 1987), plant light-harvesting complex LHC-II (Kühlbrandt *et al.*, 1994), light-harvesting protein LHA2 from *Rhodopseudomonas acidophila* (McDermott *et al.*, 1995), subunits I, II and III of cytochrome c oxidase from *Paracoccus denitrificans* (Iwata *et al.*, 1995), and subunits I, II, III, IV, VIa, VIc, VIIa, VIIb, VIIc and VIII of cytochrome c oxidase from bovine heart (Tsukihara *et al.*, 1996). These proteins were *not* included in the training data set of proteins.

2.3. Gaussian Parameters, Preferences and Prefer-preference Functions

The PREF algorithm (see the flowchart in Fig. 1) collected histograms of sequence environments for each amino acid type in each of four considered secondary conformations. For each histogram, three Gaussian parameters were extracted: (a) number of sequence environments, (b) average value for sequence environments, and (c) standard deviation for sequence environments. All such parameters were collected from the training data set of 172 proteins in the output file (Table 1). Different Gaussian parameters were collected in the case when 63 membrane proteins (already used to improve the algorithm) were tested for prediction accuracy. These new parameters were extracted from the same set of 37 soluble proteins plus an independent set of 105 membrane proteins.

Gaussian parameters were used by the SPLIT algorithm to replace frequency distributions of environments X with corresponding Gaussian functions. The probability p_{ij} of finding amino acid type "i" in a particular conformation "j" within environment X, was defined as the ratio of the Gaussian function $G_{ij}(X)$ to the sum of Gaussian functions for amino acid type "i" in α-helix, β-strand, turn and undefined conformation:

$$p_{ij}(X) = \frac{(N_{ij}/\sigma_{ij}) \exp[-(X-\mu_{ij})^2/2\sigma_{ij}^2]}{\Sigma(N_{ik}/\sigma_{ik}) \exp[-(X-\mu_{ik})^2/2\sigma_{ik}^2]} \sum_{k=1}^{4} \quad (1)$$

The number of amino acids found in each conformation (N_{ij}), average (μ_{ij}) and sample standard deviation (σ_{ij}) of parameters X are listed in Table 1. Preference functions are obtained as

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

where N_j/N is the fraction of conformation "j" in the protein data set. Preference functions were then utilized by the SPLIT algorithm to evaluate conformational preferences for amino acid residues in a tested protein.

For comparison, nine preference points, $k = 1, \ldots, 9$ were calculated as

$$P_{ijk} = (N_{ijk}/N_{ik}) \cdot (N/N_j) \quad (3)$$

where N_{ik} is the total number of environments "k" belonging to amino acid "i", while N_{ijk} are only those of these environments that are associated with the conformation "j".

2.4. DC Constants, Smoothing and Filtering Procedure

The preliminary prediction results for tested protein were used in the next prediction loop for the determination of decision constants for helix (dch), sheet (dce) and coil (turn and undefined)(dcc) conformation. There were three "if-loops" in the algorithm that could produce decision constants different from zero: dch = 0.3, dce = −0.6 and dcc = 0 were chosen when the predicted helical conformation was greater than 30% and the percentage of charged amino acids less than 20%; dch = −0.2, dce = 0.4 and dcc = 0 were chosen when the percentage of predicted sheet conformation was higher than 25% and the percentage of helical conformation less than 15%; dch = 0.4, dce = −0.2 and dcc = 0.0 were chosen when predicted helical conformation was higher than 25%, protein longer than 300 amino acids and predicted number of transmembrane helices higher than six. Except in explicitly stated cases the algorithm was used with all decision constants equal to zero.

Preferences were smoothed. Seven residue preferences were smoothed for the "H" conformation, five for the "B" conformation and three for the "U" or "T" conformation. The smoothed value was assigned to the residue in the middle of the sliding window. Corresponding decision constants were added to strings of smoothed preference values. Numerical values for smoothed preferences for four conformational states were then compared and a secondary structure assigned to the highest preference. Reported preferences are the smoothed values.

In the standard training and testing procedure, the Kyte–Doolittle hydropathy scale was used to calculate sequence environments and to evaluate preference functions. One example of the predictor output (for melittin) is given in Table 2. To test the prediction accuracy for transmembrane helices it was essential to incorporate digital predictor and filter in the SPLIT algorithm. The main part of the filter was designed to distinguish among normal-length transmembrane helices, short transmembrane helices and membrane-buried helices. Only predicted transmembrane helices were used in prediction accuracy estimates. Short segments of less than 17 residues predicted in the α-helix conformation were rejected as potential transmembrane helices (but not as membrane-buried helices) if (a) corresponding preference peaks did not reach the threshold height of 2.7, or (b) they were found to have more than three charged residues or prolines. Other short segments having 13–16 residues predicted in the α-helix conformation with the α-helix preference peak higher than 2.7 were labeled as short membrane-spanning helices. Predicted helices longer than 27 residues were shortened or *split* into two or three shorter segments depending on their length and predicted maximums in α-helix and turn preference profiles.

Three subroutines were used to split and/or to shorten a predicted long helix. The TURN-BREAK subroutine used maximal turn preference higher than 1.0 to shorten TMH longer than 28 residues or to split TMH longer than 35 residues. The CHARGE-BREAK subroutine examined the first four and last four residues of the putative transmembrane segment for the presence of charged residues and shifted the position of helix caps in the direction of helix middle if turn preferences were greater than 1.0. The FILTER subroutine also created new helix caps closer to the middle of TMH so that the shift in new cap positions was greater for lower α-helix preference and for longer initially predicted TMH. Helical preference was multiplied by the number of residues reached from the cap residue position and the resulting value was compared with (TMH length − 21)/2.

2.5. Performance Measures

The performance parameters for judging the prediction accuracy allowed for overpredictions o and underpredictions u. One such parameter utilized for individual residues is $A_i = (N_i - o_i - u_i)/N_i$ (Ponnuswamy and Gromiha, 1993), where i is the index of chosen secondary conformation (i = TM, when transmembrane helix conformation is examined). There are N_i residues from the protein data base found in the conformation "i", while o_i and u_i residues are respectively overpredicted and underpredicted in that conformation. The A parameter can be a large negative number for poor prediction. The same parameter can be used to measure the prediction accuracy for transmembrane segments: $A_s = (N_s - o_s - u_s)/N_s$, where s denotes the transmembrane segment. There are N_s observed transmembrane segments, u_s underpredicted and o_s overpredicted segments. The simpler performance measure is the fraction of correctly predicted transmembrane helices: $Q_s = N_{cs}/N_s$, where N_{cs} is the number of correctly predicted membrane-spanning helices. An overlap of at least nine common residues in the transmembrane helix conformation was required between predicted and observed helices for the case of correctly predicted transmembrane helix

Table 2. The output file for melittin*

	Melittin, 2 mlt1·sec, 26 amino acids										
	Preferences								Moments		KD profile	
No.	"AA"	"OS"	"PS"	"TMH"	"BET"	"TUR"	"UND"	"H-T"	"MOMA"	"MOMB"	"SW7"	"SW19"
1	G	U	O	4.09	0.19	0.42	0.13	3.67	0.00	0.00	0.00	0.00
2	I	H	O	2.98	0.69	0.48	0.46	2.51	0.00	0.00	0.00	0.00
3	G	H	O	3.30	0.57	0.57	0.52	2.73	1.61	0.85	0.00	0.00
4	A	H	O	3.07	0.61	0.38	0.16	2.69	1.97	0.88	1.37	0.00
5	V	H	O	2.96	0.34	0.47	0.22	2.49	2.06	0.83	2.03	0.00
6	L	H	O	3.38	0.48	1.14	0.30	2.24	1.89	0.63	1.93	0.00
7	K	H	O	3.16	0.47	1.23	0.37	1.93	1.89	0.63	1.89	0.00
8	V	H	O	3.10	0.63	1.27	0.40	1.83	1.86	0.77	1.53	0.00
9	L	H	O	2.95	0.76	0.65	0.34	2.29	2.03	0.98	0.87	0.00
10	T	H	O	2.98	0.76	0.79	0.36	2.19	2.03	0.98	0.87	1.39
11	T	T	O	3.10	0.66	0.87	0.35	2.23	2.18	0.96	1.20	1.65
12	G	H	O	3.04	0.60	0.72	0.29	2.32	2.00	0.80	0.86	1.21
13	L	H	O	2.89	0.55	0.88	0.35	2.01	1.78	0.68	0.86	0.99
14	P	H	O	2.65	0.53	0.87	0.38	1.78	1.88	0.72	1.60	0.69
15	A	H	O	2.33	0.75	1.01	0.53	1.32	1.80	0.77	1.59	0.23
16	L	H	M	1.92	1.04	0.82	0.64	1.09	1.89	0.89	1.51	−0.15
17	I	H	M	1.36	1.32	1.04	0.83	0.32	1.61	1.18	1.61	−0.13
18	S	H	M	1.06	1.50	1.10	1.02	−0.04	1.43	1.45	1.29	0.00
19	W	H	E	0.65	1.44	1.07	1.16	−0.42	1.07	1.06	0.39	0.00
20	I	H	E	0.33	1.17	0.90	1.29	−0.57	1.16	1.05	−0.71	0.00
21	K	H	U	0.15	0.85	0.81	1.39	−0.66	1.28	0.97	−2.00	0.00
22	R	H	U	0.07	0.54	0.75	1.45	−0.69	0.82	1.37	−2.39	0.00
23	K	H	E	0.02	0.22	0.69	1.49	−0.67	1.02	1.89	−2.76	0.00
24	R	H	U	0.01	0.08	0.55	1.52	−0.55	0.82	1.28	0.00	0.00
25	Q	H	U	0.00	0.03	0.50	1.54	−0.49	0.00	0.00	0.00	0.00
26	Q	U	U	0.01	0.01	0.35	1.58	−0.34	0.00	0.00	0.00	0.00

*A one letter amino acid code is used in the second column (AA). Observed structure (OS) is in the third column. Predicted structure (PS) in the fourth column is transmembrane helix configuration labeled with the letter "M" except for highly probable transmembrane helix (TMH) conformation when the letter "O" is used. Conformations "T" and "H" (membrane-buried α-helix not predicted as the part of membrane-spanning helix) are not predicted here. Predicted motif "U" refers to undefined conformation. Residues with the predicted sum of β-preferences (BET) and β-moments (MOMB) higher than 2.0 are labeled with the letter "E". Not-normalized PRIFT scale (Cornette *et al.*, 1987) (input code 27) was used here to calculate hydrophobic moments (Eisenberg *et al.*, 1984), while Kyte–Doolittle hydrophobicity scale (1982) was used to calculate preferences. Columns 5–8 contain smoothed preferences for α-helix (TMH), β-sheet (BET), turn (TUR) and undefined (UND) conformation. Column 9 contains TMH-TUR difference of preferences (H-T) that helps in visual identification of predicted transmembrane helices. Columns 10 and 11 contain numerical values for hydrophobic moments for assumed α-helix (MOMA) and for assumed β-sheet conformation (MOMB). The last two columns contain sliding window averages of Kyte–Doolittle hydropathy values with short window of 7 residues (SW7) and with wide window of 19 residues (SW19).

segment. The performance measure for whole proteins: $Q_p = n_c/n$ gives the percentage of correctly predicted proteins out of the total number of n tested proteins. In the n_c proteins all transmembrane helices are correctly predicted in their sequence positions. The algorithm also reports the number of: (a) residues correctly predicted, overpredicted and underpredicted in the transmembrane helix conformation; (b) transmembrane helical segments predicted, correctly predicted, overpredicted and underpredicted; (c) proteins recognized as membrane proteins; and (d) proteins recognized with 100% correct transmembrane conformation. Any number of membrane proteins can be tested at the same time, because performance parameters are calculated for each individual protein and for all residues from the protein data set.

We have set up an automatic e-mail server: predict@drava.etfos.hr, which will return preference profiles, hydrophobic moment profiles, and hydrophobicity profiles of membrane protein when its primary structure is submitted. FORTRAN source codes for the PREF-SPLIT suite of algorithms version 3.1 and protein data sets used in the training and testing procedure are also available from the first author.

3. RESULTS

3.1. Training Procedure Results

A flow diagram for the PREF 3.1 suite of algorithms is presented in Fig. 1. The input for the training procedure consists of chosen scale of amino acid attributes and protein data set with assigned primary and secondary structure conformation of each residue. The output is the frequency distribution of sequence environments and nine preference points calculated from equation (3) for each amino acid in each secondary conformation. Another output is the file with Gaussian parameters such as the Table 1. Normal distributions (Gaussian functions) are expected to be a good fit for the histograms of sequence environments due to the averaging procedure employed. One example is the histogram and corresponding Gaussian function for glycine in the α-helix conformation (Fig. 2). Preference function for glycine in the α-helix conformation (equation (2)) and observed preference points (calculated according to equation (3)) are compared in Fig. 3. Linear approximation for the dependence of preferences on sequence hydrophobic environment would be clearly inferior. Preference functions are a good fit for the observed preference points for other amino acids too (not shown).

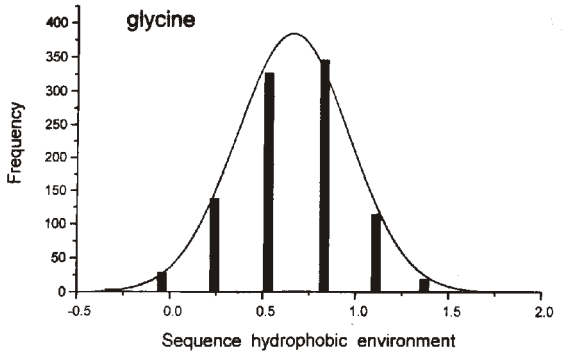

Fig. 2. The histogram and corresponding Gaussian function (full line) for glycine in the α-helix conformation. An average sequence hydrophobic environment on the x-axis is obtained from the Kyte–Doolittle hydropathy scale (1982). Sequence environments are extracted from the data base of 37 soluble and 135 membrane proteins (see Methods). Frequency points are obtained by grouping the environments in classes and counting the number of occurrences of glycine in the α-helix conformation in each class. Normal frequency distribution for 974 glycine environments is determined by their mean: 0.6679, and standard deviation: 0.2729, as found in the Table 1.

3.2. Prediction Tests

3.2.1. Photosynthetic reaction center

For the photosynthetic reaction center subunits L and M, all of 10 observed membrane-spanning segments are predicted by the SPLIT algorithm in their correct sequence location with no overpredicted segments (Figs 4 and 5, upper part). Out of 252 residues in the transmembrane helix conformation, 191 are correctly predicted in such a conformation, 18 are overpredicted and 61 are underpredicted.

Comparison of our preference profiles with the Kyte–Doolittle hydrophobicity profiles for the same reference polypeptides (Fig. 4 and Fig. 5, lower part) illustrates the advantage of using preferences (as the difference between helix and turn preferences) in the upper parts of these figures (Fig. 4 and Fig. 5). Resolution of preference peaks is much better and there is much less noise. These advantages become quantitative when digital prediction of helical membrane-spanning segments (thick line in Fig. 4 and Fig. 5, upper part) is used. The prediction of helical ends is far from perfect. The SPLIT algorithm errs on the conservative side: parts of membrane associated helices are not always recognized as such, but are rarely overpredicted.

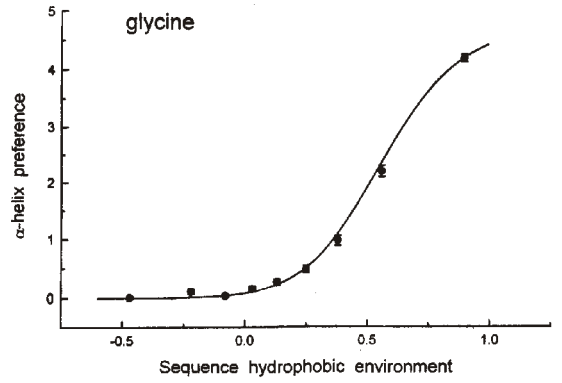

Fig. 3. Preference function and observed preference points for membrane-buried α-helix conformation of glycine. Equation (1) and Table 1 are used to calculate Gaussian functions as explained in the legend of Fig. 1. The preference function (full line) is calculated according to equation (2). Constant preferences are calculated according to equation (3). Error bars are one standard error above and below preference points. The x-axis description can be found in the legend of Fig. 2.

Fine details in the preference profile of each peak are also significant. Minimum associated with the preference peak of the D transmembrane helix corresponds to the Fe ligands His 190 in the subunit L and His 217 in the subunit M. Minimum at predicted *N*-cap of helix E corresponds to the Fe ligands His 230 in the subunit L and His 264 at the subunit M (Deisenhofer *et al.*, 1995).

3.2.2. Melittin

The output file for melittin (Table 2) illustrates what can be obtained (by e-mail) from our predictor. The digital predictor output in column four indicates with letters "O" (stronger preferences) and "M" (weaker preferences) where a transmembrane helix is expected to form in the sequence. Hydrophobicity profiles are also included in the last two columns. Hydrophobicity moment profiles are calculated with the less commonly used PRIFT scale (Cornette *et al.*, 1987). The combination (sum) of hydrophobic moments for assumed β-conformation and β-sheet preference can serve as an empirical parameter for predicting sequence location of β-strands buried in the membrane (Juretić *et al.*, 1998). Prediction of membrane-buried β-strand requires six or more consecutive residues predicted in the "E" conformation (see the legend of Table 2), hence melittin is not predicted to form β-conformation in the membrane. The capabilities of our predictor to predict membrane-bound β-conformation will be discussed in another paper.

3.2.3. Reference polypeptides of known X-ray crystal structure

When 21 integral membrane polypeptides of known crystallographic structure (75 membrane-spanning helices) are examined with the SPLIT predictor, only one transmembrane helix was underpredicted and none overpredicted (Table 3). All polypeptides were predicted as integral membrane proteins—only one was predicted with a lesser number of transmembrane helices than observed. This is a prediction accuracy of 99% ($Q_s = 0.99$, $A_s = 0.99$) for transmembrane helices and 95% ($Q_p = 0.95$) for polypeptides predicted with correct transmembrane structure.

Such high prediction accuracy was not retained for individual residues in observed membrane-spanning helix conformation (Table 3). Of 2081 such residues 1429 were correctly predicted, 652 underpredicted and 79 overpredicted ($A_{TM} = 0.649$).

With free choice of decision constants (see Methods) per-residue performance increased to $A_{TM} = 0.668$. The single underprediction of observed transmembrane segments (of the transmembrane helix XI in the subunit I of *Paracoccus denitrificans* cytochrome c oxidase) was then cor-

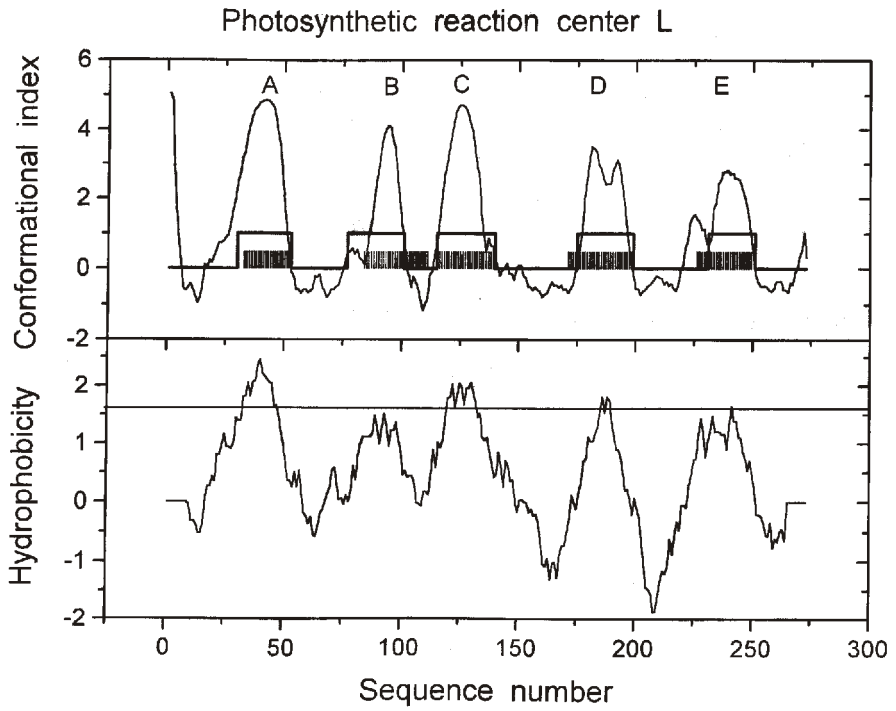

Fig. 4. The conformational index profile for the photosynthetic reaction center L subunit is the preference for the transmembrane helix conformation minus preference for the turn conformation. The preference profile and the digital prediction (thick line at the 1.0 level) for the membrane-buried α-helix conformation are obtained with the PREF-SPLIT algorithm. It uses the Kyte–Doolittle scale (1982) of hydropathy values to calculate preference functions for secondary structure determination. The hatched boxes correspond to observed membrane-spanning helices A–E of the crystal structure. The hydrophobicity profile (lower part of the figure) is obtained by using the Kyte–Doolittle scheme of sliding window average over 19 neighboring residues. The membrane-spanning helix is predicted if sliding window average $H_{19} > 1.58$ (the straight line threshold value) for at least one residue (Jähnig, 1990).

rected. Even without use of decision constants it was obvious from preference profile that high α-helix preference peak (with maximum reaching 3.0) at sequence location 447 to 462 is a good candidate for transmembrane α-helix.

3.3. Comparison With Other Prediction Methods

We have also compared two recent prediction methods, that of Jones *et al.* (1994) and Rost *et al.* (1995) with our own. Features that our predictor lacks (topology prediction) could not be compared. For 83 predicted structures by Jones *et al.* (1994), we calculated the A_s and Q_p performance measures: $A_s = 0.928$ and $Q_p = 79.5\%$. For 69 proteins tested by Rost *et al.* (1995), performance parameters were: $A_s = 0.896$ and $Q_p = 79.7\%$. The A_{TM} parameter was 0.733, when calculated from predictions returned by Rost *et al.*'s (1995) automated service for the subset of 63 proteins (among 69 proteins) used by us also. These results can be compared with our test of 63 proteins common to us and to Jones *et al.* (1994) and Rost *et al.* (1995). Input Gaussian parameters were in this case extracted from a different set of 105 membrane proteins and automatic choice of decision constants was allowed. Performance parameters were: $A_{TM} = 0.740$, $Q_s = 97.9\%$, $A_s = 0.934$ and $Q_p = 84.1\%$. A similar test on the set of 105 proteins, never before seen in the training process for the neural network algorithm, gave a value of $A_{TM} = 0.610$ for Rost *et al.*'s (1995) method. For these proteins we obtained $A_{TM} = 0.682$, $Q_s = 94.7\%$, $A_s = 0.885$ and $Q_p = 75.2\%$ with free choice of decision constants. When all of 168 integral membrane proteins were tested by using the 5-times cross validation procedure, we obtained (also with automatic choice of decision constants) $A_{TM} = 0.712$, $Q_s = 95.3\%$, $A_s = 0.898$ and $Q_p = 77.4\%$.

3.4. Choice Of Filter Parameters, Hydrophobicity Scales And Decision Constants

How do different subroutines and the choice of filter and input parameters in the SPLIT algorithm influence the prediction performance? Careful selection of performance parameters and tests with as large a number as possible of non-homologous integral membrane proteins is important in answering this question. A commonly reported Q_S parameter (the percentage of correctly predicted transmembrane helices) is not sensitive to overprediction of individual residues in the TMH conformation. Table 4 reports only A_{TM} and Q_P parameters (see Methods). Tests were carried out on the data set of 168 proteins (63 + 105). Their Swiss–Prot codes are also given in the Methods section. The smoothing procedure and filter significantly improve the performance, while choice of decision constants different from zero is less important. Listed

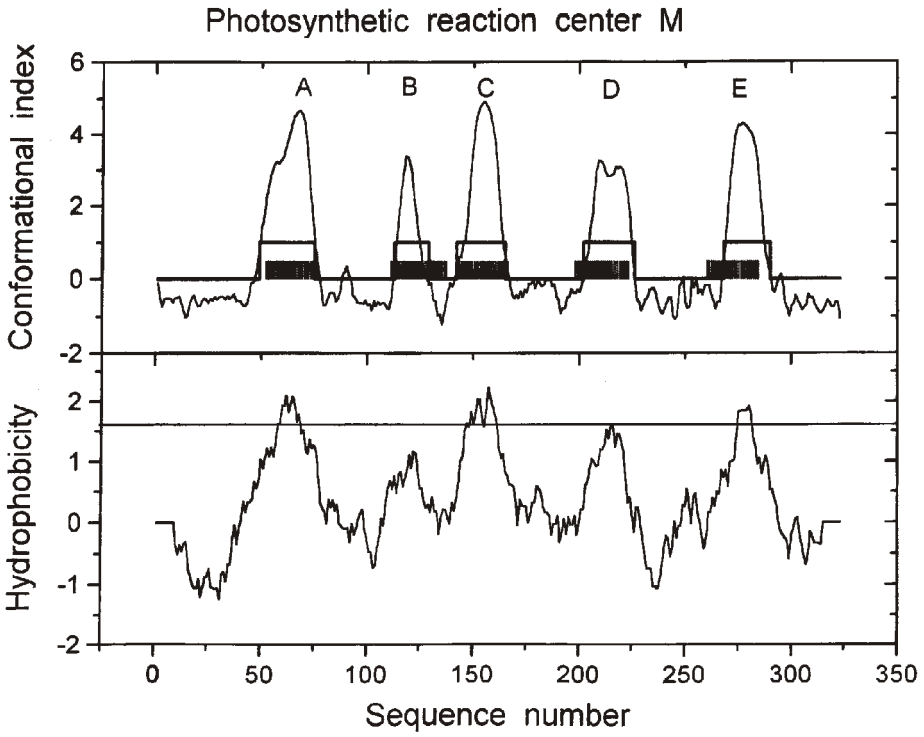

Fig. 5. Predicted and observed membrane-spanning segments for the photosynthetic reaction center M subunit. All labels and procedures used are the same as in Fig. 4.

hydrophobicity scales are well known scales that performed among the top 10 as judged by the A_{TM} parameter. More detailed evaluation of hydrophobicity scales and optimal sliding window length can be found in our earlier work (Juretić *et al.*, 1998).

3.5. Voltage-gated Channels And Receptors

3.5.1. The Shaker potassium channel

Our prediction (Fig. 6) for the *Shaker* potassium channel is that it has five transmembrane helices and three possible membrane-buried helices (at the polypeptide *N*-terminal "ball" region and at the *N*-terminal of S4 and P segments). Turn preferences (the difference between dotted and full line) are very close to zero at the maximums for five predicted transmembrane helices. Three potential membrane buried-helices are too short to span 3 nm of membrane lipid interior. According to prediction rules described in "Methods" these segments are *not* predicted (the digital predictor output is seen as the thick line in Fig. 6) as potential short transmembrane helices. For sequence positions associated with S4 and P-segments, residues with maximal preference for helical conformation have significant potential for the turn conformation as well.

The P-segment residues from *Drosophila Shaker* are: PDAFWWAVVTM(440)TTVGYGDMTP. The middle of the P-segment is thought to be close to methionine 440 (Kürz *et al.*, 1995). Maximal preference for membrane-buried α-helix is found at the alanine 436. For the S4 segment of *Shaker* (residues:
MSLAILRVIRLVR(368)VFRIFKLSRHSKG)
maximal helix preference is also at the *N*-terminal of that segment at arginine 362. The middle position of that segment is closer to arginine 368 (Larsson *et al.*, 1996).

3.5.2. Ionotropic glutamate receptors

We predict three membrane-spanning helices for such receptors. Of several potential membrane-buried short helices seen in preference profiles of these proteins one is at the *N*-terminal part of the proposed P-segment. Only recently it was realized that three transmembrane helices and a P-segment is the most likely transmembrane structure of GluR proteins (Hollman *et al.*, 1994; Bennett and Dingledine, 1995). Figure 7 shows the prediction for transmembrane helices as well as the preference and hydrophobicity profile of the GluR-5. The middle position of the P-segment is thought to be at the phenylalanine 598 of rat GluR-1 (residues: FNSLWFSLGAF(598)MQQGCDISPR) and at the leucine 634 of rat GluR-5 (residues: LNSFWFGVGAL(634)MQQGSELMPK) (Wo and Oswald, 1995a; Sutcliffe *et al.*, 1996). Maximal preference for membrane-buried helical conformation inside the P-segment is found at the Phe-629 of GluR-5, and at the Phe-593 of GluR-1.

3.6. Different Hydrophobicity Scale Predicts Same Conformational Motifs

The correlation coefficient between some hydrophobicity scales is not very high. For instance, the mean fractional area loss of Rose *et al.* (1985) and the Kyte–Doolittle hydrophobicity scale (1982) has a correlation coefficient of 0.84. To answer the question if the choice of different amino acid attributes predicts different conformational motifs, we used Rose's hydrophobicity scale in the PREF-

Table 3. Predicted and observed transmembrane helices (TMH)

Protein*	TMH** predicted		TMH observed	Max TMH preference	Max TMH sequence no.
PRCH_VIR	16–31	Short	12–35	4.04	26
PRCH_SPH	11–31		12–37	4.29	23
PRCL_VIR	26–50		33–53	4.67	41
	85–101	Short	84–111	4.11	94
	114–134		116–139	4.56	125
	176–197		171–198	3.83	182
	232–250		226–249	3.43	239
PRCL_SPH	30–53		32–55	4.65	46
	85–101	Short	83–111	3.67	94
	116–135		116–138	4.57	126
	173–196		171–198	3.85	191
	232–255		225–250	4.48	224
PRCM_VIR	50–74		52–76	4.54	68
	113–127	Short	111–137	3.54	118
	143–165		143–166	4.67	155
	204–225		198–223	3.54	217
	269–288		260–284	4.32	277
PRCM_SPH	50–73		54–78	3.96	58
	113–129	Short	109–139	4.42	120
	148–163	Short	147–168	4.24	156
	203–226		200–226	3.93	211
	268–290		262–286	4.22	276
LHA2	12–36		11–36	4.47	27
LHC-II	46–60	Short	30–64	3.70	53
	95–113		98–118	3.75	106
	162–176	Short	145–174	3.92	169
CX1_PDE	32–52		27–59	4.52	44
	91–112		84–121	4.59	103
	136–151	Short	130–151	4.34	144
	181–204		178–206	3.70	198
	219–244		218–251	4.53	232
	279–297		263–298	4.19	286
	303–326		304–322	4.61	316
	340–363		334–362	3.50	345
	373–394		370–395	4.44	380
	411–432		404–430	4.31	422
		Underp.	441–468	2.98	455
	491–512		483–513	4.62	505
CX2_PDE	36–58		27–59	4.75	52
	78–101		74–105	4.71	89
CX3_PDE	14–33		15–35	4.34	25
	39–65		48–76	4.58	56
	85–108		79–114	4.12	96
	140–161		139–165	4.47	148
	173–187	Short	168–196	4.66	180
	205–230		203–236	4.66	228
	253–270		244–273	4.68	264
COX1_BOV	17–37		12–40	3.66	22
	57–79		51–86	4.60	68
	104–119	Short	95–117	4.08	111
	145–168		141–170	3.71	161
	184–208		183–212	4.66	196
	242–257	Short	228–261	3.87	249
	271–293		270–286	4.18	281
	307–326		299–327	3.67	313
	339–361		336–357	4.51	345
	376–398		371–400	4.28	387
	410–426	Short	407–433	4.05	418
	456–473		447–478	4.61	467
COX2_BOV	27–47		15–45	4.64	35
	64–81		60–87	4.73	73
COX3_BOV	19–33	Short	16–34	3.49	24
	36–52	Short	41–66	3.40	45
	81–102		73–105	4.22	89
	130–145	Short	129–152	3.47	139
	161–176	Short	156–183	4.24	167
	195–219		191–223	4.34	213
	242–258	Short	233–256	4.17	250
COX4_BOV	82–98	Short	77–103	4.47	90
COX6a_BOV	19–33	Short	13–37	3.92	26
COX6c_BOV	20–35	Short	12–52	4.06	27
COX7a_BOV	34–49	Short	26–54	2.88	42
COX7b_BOV	18–35		9–35	4.07	26
COX7c_BOV	22–40		18–44	3.75	27
COX8_BOV	18–34	Short	12–35	4.10	24

*PRC H, L and M are respective subunits of the photosynthetic reaction center from *Rhodobacter viridis* (VIR) and from *Rhodobacter sphaeroides* (SPH). LHA2 and LHC-II are the light-harvesting protein from *Rhodopseudomonas acidophila* and plant light-harvesting protein respectively. CX1, 2 and 3 are subunits I, II and III of the cytochrome c oxidase from *Paracoccus denitrificans* (PDE). COX polypeptides are subunits I, II, III, IV, VIa, VIc, VIIa, VIIc and VIII of the cytochrome c oxidase from bovine heart.

**Decision constants were not used (all were equal to zero). When automatic choice of decision constants is allowed prediction accuracy increases and observed TMH 441–468 from cx1_pde is no longer underpredicted.

Table 4. The dependence of the algorithm's performance on the choice of subroutines and input parameters

Choice*	1	2	3	4	5	6	7	8	9	10
A_{TM}	0.712	0.655	0.693	0.646	0.689	0.694	0.680	0.675	0.666	0.659
Q_p (%)	77	64	68	58	57	59	69	68	63	65

*The effect of each change was separately tested. Tests were always performed on 168 nonhomologous membrane proteins. The choice of parameters and procedures is as follows. 1: The best parameters and complete algorithm with the Kyte–Doolittle hydropathy values (1982) as the input (input code 1). 2: The filter omitted (all filter subroutines omitted). 3: Decision constants all set to zero. 4: Smoothing procedure omitted. 5: Sliding window shorter (7 residues). 6: Sliding window longer (15 residues). 7: Surrounding hydrophobicity scale of Ponnuswamy and Gromiha (1993) (input code 17). 8: Consensus hydrophobicity scale of Eisenberg *et al.* (1984) (input code 26). 9: Mean fractional area loss of Rose *et al.* (1985) (input code 30). 10: Hydropathy values of Engelman *et al.* (1986) (input code 4).

SPLIT procedure for calculating preference functions. It is seen from Fig. 8 that the all important conformational motifs in the *Shaker* sequence are predicted with Rose's scale as well.

4. DISCUSSION

4.1. High Accuracy In Predicting Transmembrane Helices With Preference Functions

The ratio of Gaussians, as probability to find a conformational motif, can be very successful in detecting such motifs (Lupas *et al.*, 1991). The ratio of probabilities as preference for the formation of an α-helix conformation is strongly dependent on sequence hydrophobic environment (Fig. 3). The preference functions method (Juretić *et al.*, 1993) calculates conformational probabilities and preferences as functions of local sequence hydrophobicity. Secondary structure (α-helix, β-strand, turn and undefined) is predicted after comparing preferences for each residue. For many integral membrane proteins the visual comparison of helix and helix minus turn preference profiles is enough to show correct sequence location of membrane-spanning and membrane-buried helices. High α-helix preference peaks (Figs 4–8) are easily resolved by a simple digital predictor with few filtering rules, which calculates performance parameters when known structures are analyzed.

Tests with known crystal structures identified correct sequence position and conformation of all but

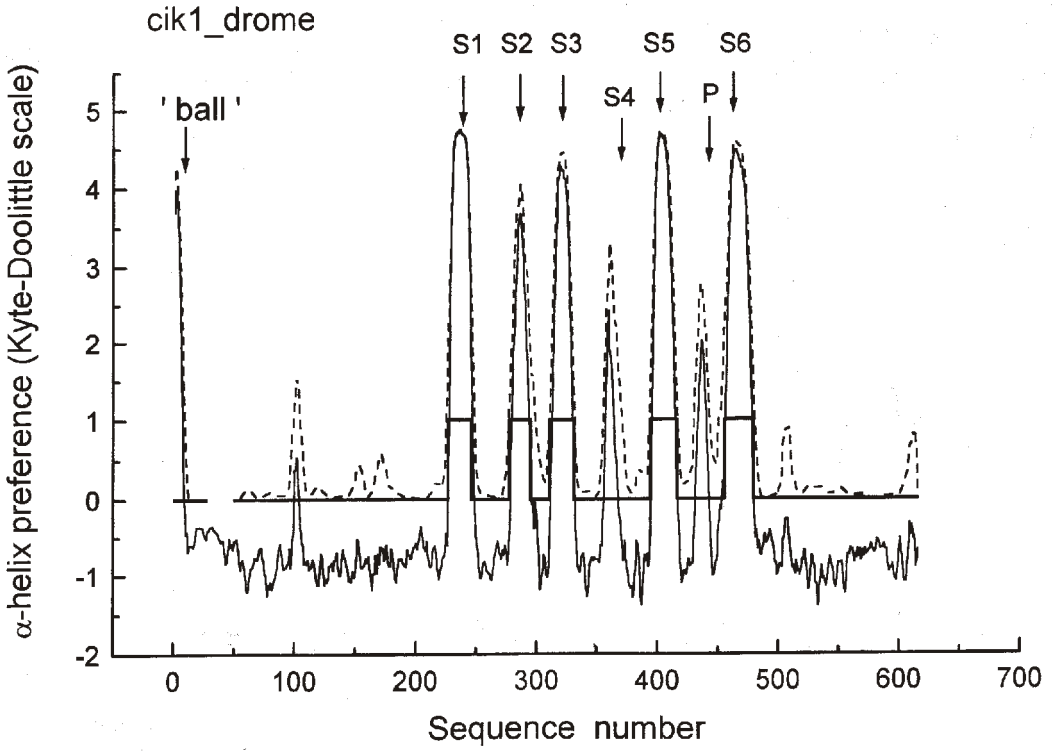

Fig. 6. Predicted α-helix membrane conformation of the *Drosophila Shaker* voltage-gated potassium channel. Dotted line: helix preferences; full thin line: helix–turn preferences; full thick line at the level 1.0: digital prediction for transmembrane α-helix conformation. Stable transmembrane helices are S1, S2, S3, S5 and S6. Maximums in the preference for membrane-buried α-helix conformation next to labels S4 and P correspond to *N*-terminal parts of unstable (movable) voltage sensor S4 segment and pore wall P-segment, respectively. The primary structure of the P-segment extends from D431 to P450, while that of the S4 segment extends from M356 to G381. The "ball" part of the chain and ball inactivation mechanism at the polypeptide *N*-terminal (Amstrong, 1992) is also associated with high preference for membrane-buried α-helix conformation.

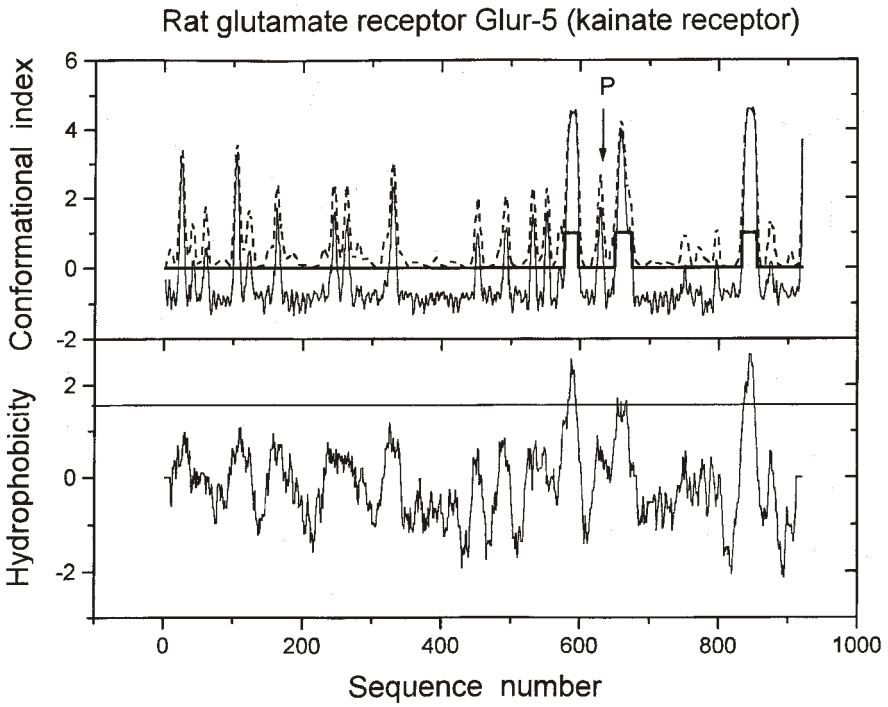

Fig. 7. Hydrophobicity profile and predicted membrane-associated secondary conformation of the rat ionotropic glutamate receptor GluR-5 (kainate receptor). All lines have the same labels and meaning as described in the legend of Figs 4 and 6.

one of 75 transmembrane segments. Importantly, this was done without overpredicting transmembrane segments. Known structures of tested proteins were not included among proteins in the training process, because tested proteins should be "never before seen" by the predictor.

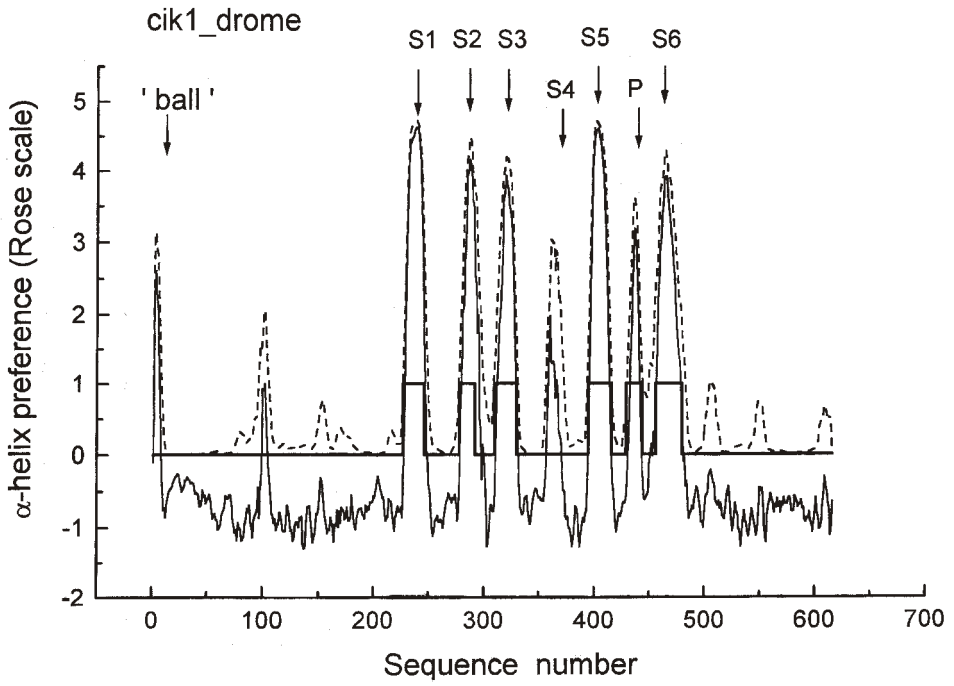

Fig. 8. Predicted membrane-spanning and membrane-buried helices in *Droshophila Shaker* channel when preference functions are derived with the Rose hydrophobicity scale (Rose *et al.*, 1985) of fractional average area loss upon folding in 23 proteins of known crystal structure. All labels are the same as in Fig. 6.

Per-residue prediction accuracy is such that underpredictions dominate for known crystal structures. The crystal structures of cytochrome c oxidase and photosynthetic reaction center are biased samples of integral membrane proteins. These proteins have many cofactors that influence the propensity (probably by increasing it) for the α-helix transmembrane structure. This feature was not taken into account by our predictor, which predicted a total of 23 short membrane-spanning helices at sequence locations with observed transmembrane helices longer than 20 and sometimes even longer than 30 residues (Table 3). Another reason for underpredictions is the optimization of the method for predicting membrane-buried parts of long membrane-spanning helices. In fact, predicted "short transmembrane helix" is usually just the membrane anchoring part of normal-length membrane-spanning helix that reserves a part of its length for the interaction with less hydrophobic protein structures. Per-residue prediction accuracy increased and predicted number of short membrane-spanning helices decreased to 15, when automatic choice of decision constants (see Methods) was allowed.

Reported test results were achieved by using only single sequence information. The knowledge about expected structure of homologous proteins to the tested one (evolutionary information) was not used. Positive-inside rule (von Heijne, 1992, 1995) and comparison of possible topological models according to this rule (Sipos and von Heijne, 1993) were not used. Per-residue prediction accuracy increased to $A_{TM} = 0.712$ when expected sequence locations of transmembrane segments from the Swiss–Prot data base are compared with our predictions. Out of 168 integral membrane proteins, 130 are recognized with correct sequence location and conformation of transmembrane segments. Corresponding per-segment performance parameters are then $Q_s \sim 0.95$ and $A_s \sim 0.90$. Competing methods have reached similar prediction accuracy only after performing the multiple sequence alignment (Persson and Argos, 1994; Rost *et al.*, 1995) and/or after comparing topological models for single sequence and using additional folding determinants such as the positive-inside rule (von Heijne, 1992; Jones *et al.*, 1994; Rost *et al.*, 1996a). The last version of Rost *et al.*'s neural network predictor (Rost *et al.*, 1996b) predicts correct topology for the photosynthetic reaction center (see Introduction), but still fails to recognize spir_spime (spiralin) and ftsh_ecoli protein as membrane proteins. The FtsH protein is predicted with maximal confidence by neural network as soluble protein. The SPLIT algorithm predicts that both of the expected transmembrane segments in the Swiss–Prot entry information are membrane-spanning helices (not shown). Hydrophobicity analysis and available experimental data (Tomoyasu *et al.*, 1995) agree that FtsH is an integral membrane protein. Therefore, high general accuracy of pattern recognition methods does not prevent most serious prediction failures because of wrongly learned rules that are not transparent.

4.2. Comparison With Hydrophobicity Analysis

Hydrophobicity analysis with the same hydrophobicity scale is clearly inferior. Maximums corresponding to helices B, D and E in the photosynthetic reaction center (Fig. 4 and Fig. 5, lower part) are not always recognized as potential transmembrane domains. The threshold value (thin straight line parallel to the x-axis at y = 1.58 height) used with the sliding window average of 19 residues (Jähnig, 1990) may be too high. Insignificant decrease in the threshold value can significantly improve apparent prediction accuracy. One is left with subjectivity, if the choice of sliding window length and the threshold height is completely free, or even worse, with delusion of objectivity, when these parameters are adjusted during training to produce the best results with tested proteins.

Embarrassing blindness to all of the potential membrane-spanning segments in some integral membrane proteins such as the RACTK1 pH sensitive K^+ channel (Suzuki *et al.*, 1994) is the weakness of the wide sliding window (19 residues) Kyte–Doolittle scheme (1982). Some mitochondrial transporters, expected to have six membrane-spanning helices (Walker, 1992) are predicted with only one or none. For instance, in the adenine nucleotide translocator 2 from yeast, none of the hydrophobicity maximums reaches the 1.58 level, while in the brown fat uncoupling protein from a rat only one maximum surpasses that level. Smaller sliding windows of 7, 9 or 11 residues are often used, mainly to define ends of membrane-spanning segments (Reithmeier, 1995). Higher hydrophobicity peaks so produced are also associated with an additional increase in the noise level.

Too many predicted transmembrane segments is a serious problem too. Older predictions of six transmembrane segments for voltage-gated potassium channels similar to *Shaker* from *Drosophila melanogaster* (Jan and Jan, 1989) and four transmembrane segments for ionotropic glutamate receptors (Gasic and Hollman, 1992) were based in part on hydrophobicity plots. For the kainate receptor GluR-5 from rat brain, Swiss–Prot entry information lists seven transmembrane segments at sequence positions 100–120, 243–263, 315–335, **577–597**, 616–636, **654–674** and **835–855**. Unrealistic absence of length distribution in predicted segments reveals that very simple sliding window algorithm has been used. Only bold segments are predicted by us as transmembrane helical segments (Fig. 7). GluR-5 is likely to have the same membrane-buried motifs as other ionotropic glutamate receptors: three transmembrane helices and the P-segment. Four additional potential transmembrane segments in the Swiss–Prot suggested topology for the GluR-5 are probably erroneous assignments. Expected sequence locations of transmembrane segments from the Swiss–Prot data base (designated by FT TRANSMEM in the feature table) should not be taken as the "standard of truth" (Persson and Argos, 1994) when testing the performance of some new predictor. The presence of errors in the Swiss-Prot data base (Juretić *et al.*, 1998) may have caused apparent prediction accu-

racy drop for segment prediction when we shifted from testing crystal structures to testing Swiss-Prot assignments for transmembrane domains. Apparent per-residue prediction accuracy increase for incompletely known structures may be due to a more subtle common mistake in the SPLIT algorithm and in the Swiss–Prot FT TRANSMEM designations: hydrophobicity does not correlate so well with observed transmembrane location of a sequence segment as assumed.

4.3. Prediction Of Unstable Membrane-buried Helices

It is possible that short hydrophobic helices form and enter membrane only after the assembly of protein monomers in a membrane, when stable backbone of membrane-spanning helices has already being formed. Preference functions analysis frequently predicts membrane-buried α-helices associated with considerable turn potential and not long enough to span 3 nm of membrane lipid interior. In the *Shaker* potassium channel monomer our preference profiles (Fig. 6 and Fig. 8) identify all important structural and functional elements. The "ball" part of the chain and ball inactivation mechanism (Amstrong, 1992) is seen in Fig. 6 and Fig. 8 to have high propensity for short membrane-attached regular conformation. We predict membrane-buried helix conformation for the *N*-terminal part of the S4 segment. Only several residues (5–10) are needed in the S4 segment to bridge the membrane (Goldstein, 1996). This voltage-sensor element is surprisingly free to move in the direction perpendicular to the membrane surface under the influence of a transmembrane electric field (Larsson *et al.*, 1996). Its instability in the membrane is indicated in corresponding preference peak as high turn preference and narrow width of less than 10 residues likely to be in the membrane at any time (Figs 6–8). Due to many positive charges present in the S4 segment, our digital predictor did not recognize it as the transmembrane segment.

4.4. Predicted Secondary Structure Of The P-domain

The P-domain has emerged as the most common building block for the pore walls in potassium, sodium and calcium voltage and ligand-gated channels (Catterall, 1995; Goldstein, 1996). Its secondary structure was inferred as β-strand or loop hairpin structure that invaginates into the bilayer interior (Miller, 1991; Bogusz and Busath, 1992). The Chou–Fasman method (Chou and Fasman, 1978) predicted a β-strand–turn–β-strand structure for the pore region of potassium channels (Soman *et al.*, 1995). The last few years has produced theoretical and experimental indications that the *N*-terminal part of some P-domains is not in the β-strand conformation (Guy and Durell, 1994; Kürz *et al.*, 1995; Lu and Miller, 1995). Our tests with voltage and ligand-gated ion channels predicted membrane-buried α-helix conformation at the *N*-terminal part of pore-forming domains in all P-domains found so far (only two examples are shown in Fig. 6 and Fig. 7). This has been supported by recent experiments (Gross and MacKinnon, 1996).

The β-sheet conformation is predicted for the whole pore-forming sequence segment 430–450 in *Shaker* polypeptide by another method that claims high resolution and low noise (Sun and Parthasarathy, 1994). Is our (different) prediction due to the choice of the Kyte–Doolittle hydropathy scale? Rose's scale of mean fractional area loss for each amino acid (Rose *et al.*, 1985) was utilized by the authors of the AutoRegressive Moving Average model of spectral analysis (Sun, 1993). When the Rose scale is used with the preference functions method, the sequence position of the P-segment is even better associated (than in Fig. 6) with narrow high peak in helix-turn preferences (Fig. 8). Large area loss in contact with water of residues in the *N*-terminal to the middle part of the P-segment may be important in membrane-anchoring. With the Rose scale as the input, the digital predictor assigns short transmembrane helix as the secondary structure of that segment. Therefore, our prediction of membrane-buried α-helix conformation for the first half of the P-segment is not dependent (but the prediction of non-transmembrane character is dependent) on the choice of Kyte–Doolittle hydrophobicity scale.

4.5. Choice Of Amino Acid Attribute

The best performance with the Kyte–Doolittle scale in predicting membrane-spanning helices (Table 4) may be due to chosen training and testing procedure. A part of the training procedure is the choice of filter parameters. Their values (see Methods) are chosen during the initial training process with Kyte–Doolittle hydropathy values as the input. Furthermore, the Kyte–Doolittle hydrophobicity analysis helped to determine expected transmembrane segments in the Swiss–Prot data base for some of the membrane proteins that we use in the training and testing procedure. However, evaluation of the 12 best scales in predicting TMH in membrane proteins of known crystallographic structure (Juretić *et al.*, 1998) also selected the Kyte–Doolittle scale as the best.

The flexibility in the choice of amino acid attributes for calculating hydrophobic moments is an important additional in-built feature of our algorithm giving it the capability to recognize surface attached regular structures in the sequence besides membrane-buried structures. The PRIFT scale (Cornette *et al.*, 1987) can often recognize such structures even when the more commonly used Eisenberg consensus hydrophobicity scale for calculating hydrophobic moments (Eisenberg *et al.*, 1984) fails to do so (not shown in this paper). Also, the calculation of hydrophobic moments for assumed β-sheet conformation is essential when the preference function method is used with a goal to predict membrane-buried β-strands (Juretić *et al.*, 1998).

Easy and quick sequence analysis with the SPLIT predictor should uncover or correct assignments for other potential membrane-buried segments. The predicted spectrum of preference peaks can serve as a rough guide or initial guess for the application of other determinants of membrane protein topology, which are known to increase prediction accuracy (such as the positive-inside rule), and for the development of detailed models of the three-dimensional structure.

Acknowledgements—Thanks are due to Burkhard Rost from EMBL, Heidelberg, Germany and Sándor Pongor from ICGEB, Trieste, Italy, who kindly provided databases of soluble and membrane proteins, and to Countess Vivian Grisogono from England and the American Biophysical Society, who helped with journals not available at the University of Split. This work was supported by Croatian Ministry of Science Grants 1-03-171 and 177060 to D.J. and D.Z. and 1-07-159 to N.T. and B.L.

REFERENCES

Allen, J. P., Feher, G., Yeates, T. O., Komiya, H. and Rees, D. C. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6162.
Amstrong, C. M. (1992) *Physiol. Rev.* **72,** S5.
Bairoch, A. and Boeckmann, B. (1994) *Nucleic Acids Res.* **22,** 3578.
Bangham, J. A. (1988) *Analyt. Biochem.* **174,** 142.
Bennett, J. A. and Dingledine, R. (1995) *Neuron* **14,** 373.
Bogusz, S. and Busath, D. D. (1992) *Biophys. J.* **62,** 19.
Catterall, W. (1995) *Annu. Rev. Biochem.* **64,** 493.
Chou, P. Y. and Fasman, G. D. (1978) *Adv. Enzymol.* **120,** 97.
Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. and DeLisi, C. (1987) *J. Mol. Biol.* **195,** 659.
Cowan, S. W. and Rosenbusch, J. P. (1994) *Science* **264,** 914.
Deisenhofer, J., Epp, O., Miki, K., Huber, R. and Michel, H. (1985) *Nature* **318,** 618.
Deisenhofer, J., Epp, O., Sinning, I. and Michel, H. (1995) *J. Mol. Biol.* **246,** 429.
Edelman, J. (1993) *J. Mol. Biol.* **232,** 165.
Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) *J. Mol. Biol.* **179,** 125.
Engelman, D. M., Steitz, T. A. and Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem.* **15,** 321.
Fasman, G. D. and Gilbert, W. A. (1990) *Trends Biochem. Sci.* **15,** 89.
Gasic, G. P. and Hollman, M. (1992) *Annu. Rev. Physiol.* **54,** 507.
Goldstein, S. A. N. (1996) *Neuron* **16,** 717.
Grey, M. W. (1996) *Nature* **383,** 299.
Gross, A. and MacKinnon, R. (1996) *Neuron* **16,** 399.
Guy, H. R. and Durell, S. R. (1994) *J. Gen. Physiol.* **17,** P7a (abstract).
Hollman, M., Maron, C. and Heinemann, S. (1994) *Neuron* **13,** 1331.
Iwata, S., Ostermeier, C., Ludwig, B. and Michel, H. (1995) *Nature* **376,** 660.
Jan, Y. N. and Jan, L. Y. (1989) *Cell* **56,** 13.
Jähnig, F. (1989) In *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. G. D. Fasman, p. 707. Plenum Press, New York.
Jähnig, F. (1990) *Trends Biochem. Sci.* **15,** 93.
Jennings, M. L. (1989) *Ann. Rev. Biochem.* **58,** 999.
Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994) *Biochemistry* **33,** 3038.
Juretić, D., Lee, B. K., Trinajstić, N. and Williams, R. W. (1993) *Biopolymers* **33,** 255.
Juretić, D., Lučić, B., Zucić, D. and Trinajstić, N. (1998) *Theoretical and Computational Chemistry, Vol 5. Theoretical Organic Chemistry*, ed. C. Parkanyi, Elsevier Science, Amsterdam, pp. 405–445.
Kabsch, W. and Sander, C. (1983) *Biopolymers* **22,** 2577.
Klein, P., Kanehisa, M. and DeLisi, C. (1985) *Biochim. Biophys. Acta* **815,** 468.
Kühlbrandt, W., Wang, D. N. and Fujiyoshi, Y. (1994) *Nature* **367,** 614.
Kürz, L. L., Zühlke, R. D., Yhang, H.-J. and Joho, R. H. (1995) *Biophys. J.* **68,** 900.
Kyte, J. and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105.
Larsson, H. P., Baker, O. S., Dhillon, D. S. and Isacoff, E. Y. (1996) *Neuron* **16,** 387.
Lodish, H. F. (1988) *Trends Biochem. Sci.* **13,** 332.
Lu, Q. and Miller, C. (1995) *Science* **268,** 304.
Lupas, A., Van Dyke, M. and Stock, J. (1991) *Science* **252,** 1162.
McDermott, G., Prince, S. M., Freer, A. A., Hawthornthwaite-Lawless, A. M., Papiz, M. Z., Cogdell, R. J. and Isaacs, N. W. (1995) *Nature* **374,** 517.
Miller, C. (1991) *Science* **252,** 1092.
Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. and Landau, E. M. (1997) *Science* **277,** 1676.
Persson, B. and Argos, P. (1994) *J. Mol. Biol.* **237,** 182.
Ponnuswamy, P. K. and Gromiha, M. M. (1993) *Int. J. Peptide Protein. Res.* **42,** 326.
Reithmeier, R. A. (1995) *Curr. Opin. Struct. Biol.* **5,** 491.
Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. and Zehfus, M. H. (1985) *Science* **229,** 834.
Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) *Protein Sci.* **4,** 521.
Rost, B., Fariselli, P. and Casadio, R. (1996a) *Protein Sci.* **5,** 1704.
Rost, B., Casadio, R. and Fariselli, P. (1996b) In *ISMB-96 Proceedings Fourth International Conference on Intelligent Systems for Molecular Biology*, ed. D. J. States, P. Agarwal, T. Gaasterland, L. Hunter and R. F. Smith, p. 192. AAAI Press, Menlo Park, CA.
Sipos, L. and von Heijne, G. (1993) *Eur. J. Biochem.* **213,** 1333.
Soman, K. V., McCammon, J. A. and Brown, A. M. (1995) *Protein Eng.* **8,** 397.
Sun, S. (1993) Protein structure prediction: power spectral analysis approach and reduced representation model. Ph.D. thesis. UMI Dissertation Services, Ann Arbor, MI.
Sun, S. and Parthasarathy, R. (1994) *Biophys. J.* **66,** 2092.
Sutcliffe, M. J., Wo, Z. G. and Oswald, R. E. (1996) *Biophys. J.* **70,** 1575.
Suzuki, M., Takahashi, K., Ikeda, M., Hayakawa, H., Ogawa, A., Kawaguchi, Y. and Sakai, O. (1994) *Nature* **367,** 642.
Tomoyasu, T., Gamer, J., Bukau, B., Kanemori, M., Mori, H., Rutman, A. J., Oppenheim, A. B., Yura, T., Yamanaka, K., Niki, H., Hiraga, S. and Ogura, T. (1995) *EMBO J.* **14,** 2551.
Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996) *Science* **272,** 1136.

von Heijne, G. (1986) *EMBO J.* **5,** 3021.
von Heijne, G. (1992) *J. Mol. Biol.* **225,** 487.
von Heijne, G. (1995) *Bio Essays* **17,** 25.
Walker, J. E. (1992) *Curr. Opin. Struct. Biol.* **2,** 519.
Weiss, M. S. and Schulz, G. E. (1992) *J. Mol. Biol.* **227,** 493.
White, S. H. (1994) *Annu. Rev. Biophys. Biomol. Struct.* **23,** 407.
Wo, Z. G. and Oswald, E. (1995a) *Trends Neurosci.* **18,** 161.
Wo, Z. G. and Oswald, E. (1995b) *J. Biol. Chem.* **270,** 2000.