C. Párkányi (Editor) / *Theoretical Organic Chemistry*
Theoretical and Computational Chemistry, Vol. 5

Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions

Davor Juretić,[a] Bono Lučić,[b] Damir Zucić[c] and Nenad Trinajstić[b]

[a]Department of Physics, Faculty of Science, N. Tesle 12, HR-21001 Split, Croatia

[b]The Rugjer Bošković Institute, P.O.B. 1016, HR-10001 Zagreb, Croatia

[c]Faculty of Electrical Engineering, Kneza Trpimira 2b, University of Osijek, HR-3100 Osijek, Croatia

1. INTRODUCTION

The problem of structure prediction for proteins involves secondary structure prediction based on sequence analysis as the first step [1]. Secondary structure prediction algorithms [2,3] that worked reasonably well with soluble proteins were considered inadequate for membrane proteins [4]. Recently, different artificial neural network algorithms have been used to predict secondary structure in globular soluble proteins [5-7] and sequence location of transmembrane segments (TMS) in integral membrane proteins [8,9]. When a small data base of structural features is used to train such algorithms there is always a danger of overtraining and that is precisely the case with integral membrane proteins. The structure of only several membrane proteins is known with high enough resolution for unambiguous assignment of secondary structure features [10-15]. Enlarging the data base by the use of, for example, the SWISS-PROT sequence data base [16] assignments of potential TMS as the 'standard of truth' is also connected with serious problems: in general, secondary structure information is not provided and erroneous assignments may be present in the data base.

Training is not necessary for simpler algorithms for analysis of hydrophobicity profiles [17-21]. However, recent improvements of sliding window algorithms [22] optimize all variable parameters by using the very restricted number of integral membrane proteins of known structure. Such procedure leads to overtraining and to a significant drop in prediction quality for unrelated proteins. In general, overprediction of putative TMS, and a need for subjective decision about their location and length is a common deficiency of hydrophobicity plots. As often observed [18], hydrophobicity alone is not enough to detect membrane spanning domains and their secondary structure conformation, because folding into the TMS conformation is controlled by the primary structure context. Sequence folding codes may be simpler for globular membrane proteins [23] than for globular soluble proteins, but paucity of known membrane protein structures is still making it very difficult to recognize such codes. Recognition of putative TMS from hydrophobicity plots may seem to be easy, but prediction of such segments must be accompanied with prediction accuracy assessment to be meaningful. In spite of these shortcomings hydrophobicity plots are still considered among most promising approaches to successful structure prediction schemes [24].

For a large number of deduced sequences, coming out daily from different genome projects, *theoretical sequence analysis is the only possible method* for predicting TMS and deciphering transmembrane topology. At present membrane-embedded domains can be predicted with good accuracy but this is not the case with the secondary structure of these domains. This is important deficiency of the sliding window methods based on sequence hydrophobicity, because only the secondary structure information can serve as the starting point for predicting protein assembly into the final three-dimensional structure. In this report we shall describe a theoretical method based on hydropathy analysis that accurately predicts not only the sequence location of transmembrane segments but their secondary structure conformation as well.

The conformation of TMS seems to be α-helical for most membrane proteins [12,13,25-27], but there are some proteins, such as porins, that have TMS in the β-sheet conformation [28]. There may also exist proteins with both helical and β-strand transmembrane segments or with transmembrane helical segments combined with still unknown topology of membrane buried β-strands [29-31]. This work is focused on the prediction of transmembrane helical segments (TMH), but our algorithms do allow the prediction of transmembrane or surface attached β-strands (TMBS) as well.

Our approach is to associate given amino acid type both with its secondary structure conformation and with hydrophobicities of its sequence neighbors in a carefully selected reference set of membrane and soluble proteins. Conformational preferences are then calculated. Since preferences depend not only on amino acid type but also on amino acid attributes and local sequence context, our predictor is using preference functions [32]. Very high preferences for the α-helix conformation (often higher than 4.0) are then associated with residues known to have very hydrophobic sequence environment inside transmembrane segments with known sequence location. Based on chosen scale of 20 amino acid attributes (such as hydrophobicity, polarity, statistical preferences), secondary structure conformation is first predicted as α-helical, β-sheet, turn or undefined conformation in given protein sequence and secondly those segments are selected that have high preference for the membrane-embedded conformation.

In effect, our method predicts 6 different secondary structure conformations: α-helical, β-sheet, turn, undefined, TMH and TMBS. Only primary structure segments with predicted long uninterrupted stretches of α-helical residues with high maximum preference for helical configuration are considered as candidates for the TMH. Longer β-strands are also predicted and, at least in porins, are never confused with TMH. We have no false positive predictions of TMH in porins, but we do have false positive predictions of TMH in some soluble proteins.

By using the cross-validation statistical procedure and Kyte-Doolittle hydropathy scale, the prediction results for TMH in the training data base of 63 membrane proteins common to us and to Rost *et al.* [9] and also to Jones *et al.* [33] were similar in accuracy by all three methods. When training data base is enlarged to 168 proteins, we maintain the 95% accuracy for predicted transmembrane helices and almost 80% (78.6%) of proteins are predicted with 100% correct transmembrane topology. When 168 proteins are divided in the above mentioned training set of 63 proteins and an independent test set of 105 proteins, all performance parameters for TMH prediction associated with a set of 105 proteins exhibited a decrease which was smaller in our case than for Rost *et al.* [9].

2. METHODS

2.1. Selecting protein data bases for training and for testing

Rost *et al.* training list of proteins [9] and SWISS-PROT sequence data base [16] release 29 and 31 were used to select training and testing sets of proteins. We examined more than 4000 proteins with transmembrane domains mostly selected from the SWISS-PROT data base release 29. A total of 168 integral membrane proteins were finally selected. These are in alphabetical order with the SWISS-PROT release code or letter 'r' (for the Rost *et al.* proteins [9]). Letter 's' is added when appropriate to indicate that signal sequence has been removed: 4f2_human(r), 5ht3_mouse(rs), a1aa_human(r), a2aa_human(r), a4_human(rs), aa1r_canfa(r), aa2a_canfa(r), ach1_xenla(29s), acm5_human(29), adt_ricpr(r), 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), bach_halhm(r), bacr_halha(r), c561_bovin(29), cadn_mouse(29s), car1_dicdi(29), cb2r_human(29), cb21_pea(r), cd2_human(29s), cd7_human(29s), cd72_human(29), cd8a_human(29s), cek2_chick(rs), 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), cyoa_ecoli(rs), cyob_ecoli(r), cyoc_ecoli(r), cyod_ecoli(r), cyoe_ecoli(r), dhg_ecoli(31), dhsc_bacsu(29), divb_bacsu(29), dmsc_ecoli(29), dsbb_ecoli(31), edg1_human(r), egf_mouse(31), exbb_ecoli(29), fce2_human(r), 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), glp_pig(r), glpa_human(rs), glpc_human(r), glra_rat(rs), gmcr_human(rs), gp1b_human(rs), gpt_crilo(r), grhr_human(29), ha21_human(29s), hb23_mouse(29), hema_cdvo(r), hema_measa(r), hema_pi4ma(r), hg2a_human(r), hly4_ecoli(29), hmdh_human(29), iggb_strsp(r), il2a_human(rs), il2b_human(rs), imma_citfr(29), isp6_yeast(29), ita5_mouse(r), itb1_human(29s), kdgl_ecoli(31), kgtp_ecoli(29), lacy_ecoli(r), lech_human(r), leci_mouse(r), lep_ecoli(r), lha1_rhosh(29), lhb4_rhopa(29), ly49_mouse(29), m49_strpy(29s), magl_mouse(rs), malf_ecoli(r), malg_ecoli(29), mas6_yeast(31), mdr3_human(31), melb_ecoli(29), mepa_mouse(31s), mota_ecoli(29), motb_ecoli(r), mpcp_rat(29), mprd_human(rs), myp0_human(rs), mypr_human(29), nals_bovin(29), nep_human(r), ngfr_human(rs), nk11_mouse(29), nntm_bovin(31), nram_iabda(29), och1_yeast(29), oec6_spiol(29), oppb_salty(r), oppc_salty(r), ops1_calvi(r), ops2_drome(r), ops3_drome(r), ops4_drome(r), opsb_human(r), opsd_human(r), opsg_human(r), opsr_human(r), pigr_human(r), psaa_pinth(31), psab_pinth(31), psbi_horvu(29), ptgb_ecoli(31), ptma_ecoli(r), sece_ecoli(r), 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), tcb1_rabit(r), tcc1_mouse(29), tcrb_bacsu(31), tee6_strpy(29s), tgfa_human(31s), thas_human(29), tnfa_bovin(29), tnr1_human(31), trbm_human(rs), trsr_human(r), tsa4_giala(31s), ucp_rat(29), va34_vaccc(29), vca1_human(29s), vglg_hrsva(29), vmt2_iaann(r), vnb_inbbe(r), vs10_rotbn(29), wapa_strmu(29s).

Of these proteins 80 had a single transmembrane helix, 6 had 2, 6 had 3, 14 had 4, 4 had 5, 13 had 6, 24 had 7, 5 had 8, 2 had 10, 3 had 11, 9 had 12, and 2 had more than 12 TMH. There were 662 expected transmembrane segments with 14359 residues in 'observed' transmembrane helix configuration among total number of 67155 residues. In the selection process preference was given to proteins with transmembrane domains without associated label (such as 'putative'). Proteins with 'probable' or 'potential' transmembrane domains were also collected in the case when such segments were sandwiched between protein domains of known cytoplasmatic and extracellular identity. Signal sequences, claimed as such in the SWISS-PROT, were omitted.

In order to facilitate comparisons with other statistical methods the training data base of proteins contained a subset of the training data base selected by Rost *et al.* [9] that was already a subset of the training data base selected by Jones *et al.* [33]. The omission of some proteins first by Rost and then by us decreased the size of the original data base from 83 (Jones *et al.*, [33]) to 69 (Rost *et al.*, [9]) and to 63 (this work). While Rost omitted 14 proteins because they were less well determined experimentally than other proteins included for the training procedure by Jones, a few additional polypeptides were omitted by us because of their known X-ray structure that we later used for rigorous testing of our algorithm. Two proteins longer than 1000 amino acids were omitted, too. Some proteins from the 168 protein list have been taken without N-terminal or C-terminal amino acids in undefined conformation so that their final length is also less than 1000 residues. These are: atn1_human with omitted 23 C-terminal amino acids, cic1_cypca with only first 720 amino acids taken out of 1852, egf_mouse with omitted 400 N-terminal amino acids, mdr3_human with omitted 279 C-terminal amino acids and nntm_bovin with omitted 200 N-terminal amino acids.

We took care that all polypeptides selected by us show less than 30% similarity with any other polypeptide used in the training process. Of 10 proteins from the test list of membrane proteins with the best known structure only one (plant light-harvesting complex) had its twin (cb21_pea) in the data base of 168 proteins. The similarity was judged by the HSSP data base of [34]. An exception to that rule are several of 63 proteins selected by Rost *et al.* [9]. These are hema_cdvo and hema_measa with 40% similarity, opsb and opsd with 41% similarity, opsb and opsg with 37% similarity, opsb and opsr with 37% similarity, opsd and opsg with 36% similarity, opsd and opsr with 35% similarity, ops4 and ops3 with 68% similarity, aa1r and aa2a with 44% similarity and opsg and opsr with 97% similarity. All of 105 proteins selected by us from SWISS-PROT releases 29 and 31 were less than 30% similar to each other and less than 30% similar to any other protein of the complete set of 168 protein. The rule of less than 30% similarity among tested proteins was not maintained for some special data bases of such proteins, such as the above mentioned collection of integral membrane proteins of the α-class with known high resolution X-ray structure. On the other hand, with the exception of light-harvesting complex, we always made sure that no tested protein was more than 30% similar to any protein from the training list of proteins.

All potential transmembrane segments in the reference set of 168 integral membrane proteins were considered to be in the α-helix conformation during training process. Five residues next to each observed TMH were considered to be in the turn conformation, while all other residues were regarded as present in the undefined conformation. Soluble proteins and membrane proteins with solved structure were analyzed with the Kabsch and Sander program DSSP [35], which assigned the secondary structure. All helical conformations 'H', 'I', and 'G' were lumped into α-helical 'H', all beta into 'B', all turn into 'T', while all remaining residues were considered to be in the 'U' conformation. Transmembrane helices broken with turn residues in the SWISS-PROT data base or by the DSSP algorithm, but reported as transmembrane helices in the original papers, were considered to be unbroken string of 'H' residues.

Two sets of soluble globular proteins were selected from Protein Data Bank (PDB) for testing for false positive results. In the set SOLU1 of 187 such proteins resolution for each protein was equal or better than 3 Å. Secondary structure conformations were determined by the DSSP algorithm. In the set SOLU2 of 147 proteins (protein data set used in [7] plus 21 additional proteins) only proteins known with equal or better than 2.5 Å resolution were included. There was less than 25% pairwise similarity. Three different secondary structures were determined as described by Rost and Sander [7]. Both data sets are available in the Supplementary Material.

Two sets of β-class soluble proteins were used for training.

a) The first set of 37 such proteins SOLB1 has been selected from Protein Data Bank among soluble 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:

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.

b) The second set of 39 such proteins SOLB2 has been selected from SOLU2 data set:

1azu, 1bbp_A, 1bds, 1bmv_1, 1bmv_2, 1cbh, 1cd4, 1cdt_A, 1fc2_D, 1fdl_H, 1mcp_L, 1rnh, 1sh1, 2alp, 2gcr, 2i1b, 2ltn_A, 2ltn_B, 2mev_1, 2mev_3, 2pab_A, 2pcy, 2pka_A, 2rsp_A, 2sod_B, 2stv, 3ait, 3ebx, 3hla_B, 3hmg_A, 4cms, 4cpa_I, 4rhv_1, 4rhv_3, 4sgb_I, 5er2_E, 5hvp_A, 6hir, 9api_B.

The data base of the best known 10 integral membrane proteins with transmembrane helical segments (BESTP) consisted of photosynthetic reaction center subunits H, L and M from *Rhodobacter viridis* [12,36] and *Rhodobacter sphaeroides* [25], plant light-harvesting complex LHC-II [13], light-harvesting protein LHA2 from *Rhodopseudomonas acidophila* [27], and two human class I histocompatibility antigens 1b14 [37] and 1a02 [38]. The X-ray structure for the single transmembrane segment of each histocompatibility antigen was not determined, but since all the rest of the three-dimensional structures of these proteins was determined by the X-ray crystallography, we considered the combination of experimental and theoretical methods used to describe their structure powerful enough to include these proteins among the best known integral membrane proteins.

The data base PORINS consisted of seven porins and two defensins all with known or proposed transmembrane β-strand structure. The porins with known X-ray structure were porin from *Rhodobacter capsulatus* [10,28] and porins PhoE and OmpF from *Escherichia coli* [39,11]. Porins with proposed transmembrane β-barrel topology were anion-selective porin Omp32 from *Comamonas acidovorans* [40], outer membrane protein OmpA from *Escherichia coli* K12 membrane (membrane-embedded fragment residues 1 to 177, [41,42]), and mitochondrial outer membrane porin from human B-lymphocytes [43] and from *Neurospora crassa* [44]. Two defensins of known structure were HNP-3 [45] and defensin from larvae of the dragonfly *Aeschna cyanea* [46].

2.2. Main performance parameters used to judge the prediction quality

a) Parameters for individual residues are composed of correct positive predictions p, correct negative predictions n, overpredictions o and underpredictions u for all residues found in the protein data base. One such parameter is the fraction of residues predicted in correct secondary conformation:

$$Q_3 = (p_1 + p_2 + p_3)/N$$

where secondary conformations are helix, beta and everything else (turn, undefined or coil) found in the data base having a total of N residues. Another such parameter [47] is

$$A_i = (N_i - o_i - u_i)/N_i$$

where *i* is the index of chosen secondary conformation with N_i residues from protein data base found in that conformation, while o_i and u_i are respectively overpredicted and underpredicted residues in that conformation. While lower bound for the Q parameter is 0, the A parameter can be large negative number for poor prediction. For α-helical, β-strand, turn, undefined and TMH conformation A parameters are respectively A_h, A_b, A_t, A_u and A_{TM}.

b) Parameters for TMH segments as prediction units. Parameter of the A type measures prediction accuracy for transmembrane segments instead of prediction accuracy for individual residues:

$$A_s = (N_s - o_s - u_s)/N_s$$

where s denotes transmembrane segment. There are N_s observed transmembrane segments, u_s underpredicted and o_s overpredicted segments. Even simpler performance measure is the fraction of correctly predicted TMH:

$$Q_s = N_{cs}/N_s$$

where N_{cs} is the number of correctly predicted TMH. There must be an overlap of at least 9 residues in the TMH conformation between predicted and observed TMH for the case of correctly predicted TMH.

c) Protein topology parameters. If there are n_c proteins with 100% correctly predicted transmembrane topology (all TMH correctly predicted in correct sequence positions) out of the total number of n tested proteins, than a very useful parameter is

$$Q_p = n_c / n$$

Our algorithm also reports absolute values of: a) residues correctly predicted, overpredicted and underpredicted in the TMH conformation, b) transmembrane helical segments predicted, correctly predicted, overpredicted and underpredicted as TMH, c) proteins recognized as membrane proteins and d) proteins recognized with 100% correct topology.

2.3. Hydrophobic moment profile

Hydrophobic moment profile is calculated as described by Eisenberg *et al.* [48] to collect information about possible amphipathic helices and strands. We used only the PRIFT scale (# 27 in Table 5) to find hydrophobic moments. Scales used for the calculation of hydrophobic moments were not normalized. The PRIFT scale produces high moments (sometimes higher than 2.0) for sequence segments known to be highly amphipathic. An ideal α-helix twist angle of 100 was used to associate α-helix hydrophobic moments with all sequence positions. Less ideal angle of 162 (more appropriate to the β-barrel structure) was used to produce sequence profile of β-strand moments.

2.3.1. The training procedure for the preference functions method

The prediction is based on the method of preference functions [32]. The PREF suite of algorithms in the present version (PREF 3.0) consists of training and testing algorithms called PREF (PREference Functions) and SPLIT (predicted long helices are SPLIT into two or three TMH), respectively. The first obligatory step in the PREF algorithm is the choice of amino acids scale of 20 values. Secondly, data sets of proteins are selected to train the algorithm. Standard training procedure uses the Kyte and Doolittle hydropathy scale [17], 168 integral membrane proteins listed above and 37 soluble β-class proteins (SOLB1). With a chosen scale, sequence environment is calculated for each amino acid type associated with one of four secondary conformations (helix, sheet, turn and undefined) at each sequence position, as an average over hydrophobicity values of neighboring 10 amino acids. The amino acid attribute of the central amino acid residue in the sliding window is not taken into account to calculate sequence environment. This is being done for the whole data set of proteins. Collected sequence environments are grouped into nine classes so that about equal number of environments is collected into each class. For the best scales, histograms of frequency distributions for environments for the same amino acid type differ significantly for different secondary conformations. This is most easily seen if frequency distributions are replaced with corresponding Gaussian functions. For each amino acid type in each secondary conformation three Gaussian parameters are extracted from observed frequency distributions. These are: a) the number of sequence environments, b) average value for sequence environments and c) standard deviation for sequence environments. All such parameters (3 x 20 x 4 if four different folding motifs are considered) are collected in the file with Gaussian parameters (enclosed in the Supplementary Material, Table III).

2.3.2. The testing procedure

Preference functions are calculated in the SPLIT algorithm as described before (ref. [32]; equations (2) and (3)). For instance, up to the constant factor, the preference function for alanine in helix conformation is found as the ratio of the Gaussian function for alanine in helix conformation to sum of Gaussian functions for alanine in all four conformations. The constant divisor is the frequency of helix conformation in protein data set. For tested protein preference functions are evaluated for all amino acids and for all four conformations. Ratio of Gaussians, as probability to find conformational motif, can be very successful in detecting such motifs, when overlap of corresponding distributions is not too great, or, in other words, when different conformations can be associated with different scores or averages (as proved by Lupas *et al.* [49] in their statistical method for detecting coiled-coil structures).

2.3.3. Decision constants choice

The automatic choice of decision constants (DC) is the standard feature of the testing procedure by the SPLIT algorithm. In the first prediction loop, preliminary prediction results for tested protein are used for the automatic determination of decision constants for helix (dch), sheet (dce) and coil (dcc) conformation (turn or undefined). Each choice of decision constants is made sequentially and independently of previous choices in the following order: Constants dch = 0.3, dce = -0.6 and dcc = 0 are chosen when predicted helical conformation is greater than 30% and percentage of charged amino acids is less than 20%. Constants dch = -0.2, dce = 0.4 and dcc = 0 are chosen when percentage of predicted sheet conformation is higher than 25%, while the percentage of predicted helical conformation is less than 15%. In the case if predicted helical conformation is higher than 25%, protein is longer than 300 amino acids and predicted number of transmembrane helices is higher than 6, then chosen decision constants are dch = 0.4, dce = -0.2 and dcc = 0. For all other possibilities decision constants are all set to zero. The algorithm is used without decision constants by setting initially all decision constants to zero, only when so noted in the text!

2.3.4. Collection of environments and smoothing procedure

Except when testing the window length influence on prediction performance a sliding window length of 11 residues was used throughout this report in such a way that central residue in the window was omitted from averaging procedure. Resulting sequence environments are then used for the evaluation of preference functions. In practice it is advantageous to smooth these preferences before comparing preference profiles. Seven residue preferences are smoothed for the 'H' conformation, five for the 'B' conformation and three for the 'U' or 'T' conformation. The smoothed value is always assigned to the residue in the middle of the sliding window. Corresponding decision constants are added to strings of smoothed preference values. Numerical values for smoothed preferences for four conformational states are then compared and secondary structure is assigned to the highest preference. In the remaining text whenever preferences are mentioned or reported it should be understood that we have in mind the sequence profile of smoothed preferences.

2.3.5. Filtering procedure

Unrealistic assignments of a single isolated residue assuming helical or beta sheet conformation, among two left and two right neighbors in nonregular conformations, are corrected by introducing nonregular ('U') conformation for such residues. Isolated residues in 'B' conformation surrounded by two left and two right residues in 'H' conformations are reassigned as residues in helical conformation. Similarly the BBHBB pattern is transformed into BBBBB. Two arginines neighbors or two proline neighbors are assigned to nonregular ('U') conformation whenever found with helix preference less than 3.0.

The essential part of the algorithm recognizes transmembrane helix conformation as the fifth possible conformation. The first appearance of the 'H' conformation is memorized and subsequent 'H' residues are counted if not interrupted by any other conformational assignment. The value for maximum helical preference is also memorized in each helical segment. String of helical residues is considered as possible transmembrane helix if found to be longer than 12 residues with maximum preference for helical conformation higher than 2.7. Residues predicted in the 'B' conformation as neighbors to candidate TMH segment are used to elongate it at both ends. Even shorter helical segments (from 9 to 14 residues in length) are memorized and fused together if less than six residues apart with at least one maximum helix preference higher than 2.7. Total number of predicted C caps is memorized at this stage and used as information about total number of predicted transmembrane helices in the protein. The percentage of predicted helical and sheet residues with respect to protein sequence is also calculated in order to determine decision constants for the next prediction cycle.

In order to compare predicted and observed transmembrane helices automatic extraction of observed transmembrane segments with probable helical conformation is also performed. It was necessary to consider all uninterrupted transmembrane segments longer than 13 and shorter than 38 residues as potential transmembrane helical segments.

The main part of the filter is designed to reexamine potential transmembrane helical segments and to shorten or split TMH of unrealistic length. All candidates for predicted TMH are divided into five groups: short segments having 13 to 16 residues, normal length segments having 17 to 27 residues, long segments having 28 to 35 residues, very long segments having 36 to 54 residues and obviously wrong predictions of segments longer than 54 residues.

Short TMH are eliminated if their TMH preference peak is less than 2.7, and also in the case three of residues E, P, K, D, R are present in the segment with maximum helix peak less than 4.0. Normal length TMH are shortened from both ends in the case when any of charged amino acids: arginine, lysine, aspartic or glutamic acid are found inside first four and last four positions of the putative transmembrane segment. In addition, turn preference had to be greater than 1.0 for these amino acids for shortening to take effect. New N and C caps are positioned at the first residue inside segment (going from old helix caps in the direction of helix middle) that could remain in the helical conformation. We shall call this subroutine the CHARGE-BREAK subroutine. Also disregarded are TMH that are too short (shorter than 17 amino acids) after CHARGE-BREAK routine, and of not enough high helical preference peak (less than 2.7).

In the case if length of putative TMH remains equal or greater than 24 FILTER subroutine is applied. It shortens TMH on both ends until helix preference becomes too high. Helical preference is multiplied with number of residues reached from the cap residue position and resulting value compared with (TMH length -21)/2. The FILTER creates new helix cap positions closer to the middle of TMH. The shift in the new cap positions is greater for lower helical preference and for longer TMH.

Long TMH's, having 28 to 35 residues, are shortened by using the TURN-BREAK subroutine. In brief, residues inside helix and next to each cap are examined with respect to their turn preferences. If maximum turn preference is greater than 1.0 then corresponding cap position is shifted to the position next to turn preference maximum in the direction of helix middle. In the case if remaining TMH is still longer than 24 residues, than CHARGE-BREAK and FILTER subroutine is applied. Predicted helical segments longer than 35 residues are broken into two or three segments with the TURN-BREAK subroutine and with a help of additional similar routine for finding maximum α-helix preference, while both TURN-BREAK and FILTER routine is used to shorten remaining segments that are still too long.

After ending the main filter routine the 'T' conformation is assigned to four residues next to each predicted helix cap. Also peaks in helix preference higher than 4.0 are examined for the whole sequence. Additional (overlooked) TMH is assigned as 15 residue segment centered around such peak if a) less than three of K, P, D, R, E residues are present in such a segment, if b) such segment is at least 20 residues removed from sequence terminals, and if c) TMH was not previously predicted in that position.

2.3.6. Predicting transmembrane β-strands (TMBS)

The SPLIT algorithm was optimized for predicting transmembrane α-helices by using the Kyte-Doolittle hydropathy scale to create profile of α-helix preferences. The digital version of prediction for transmembrane α-helices is designated as the TMH predictor. Predicted profile of β-strand preferences can be used to find sequence location of potential membrane-embedded or surface-attached β-strands. The score for potential membrane-attached β-strand conformation is found by summing up β-sheet preference and β-sheet hydrophobic moment (calculated using PRIFT scale [50]) for each sequence position. The digital version of the prediction for potential membrane-embedded β-strands (TMBS predictor) is then found as collection of sequence segments at least 6 residues long with each residue-associated score higher than 2.0.

2.3.7. Adopted cross-validation technique

The prediction performance statistics is better for larger number of proteins tested. All proteins included in the training data base can be used for testing as well if the jack-knife or cross-validation technique is adopted (see ref. [7] concerning the necessity of using this statistical technique to estimate the prediction performance). We used 5-times cross-validation to obtain representative results for the reference set of 168 integral membrane proteins after extracting preference functions with the Kyte-Doolittle hydropathy scale. The same set of 37 soluble proteins of the β-class (SOLB1) was always included in the training data base of proteins. It was noticed that prediction results are sensitive to the type of transmembrane topology. Therefore, for the 5-times cross-validation, all proteins were grouped according to expected number of transmembrane segments. We took care that each group of tested membrane proteins (33 or 34 proteins) has similar distribution of proteins with respect to their transmembrane topology as the total set of 168 membrane proteins. For instance smaller reference set of 135 proteins, used in the 'best' training procedure, is left when following 33 proteins are removed from the original reference set: 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. In some explicitly stated cases the 2-times cross validation procedure was used such that training set of 168 proteins was divided into 63 proteins selected by Jones *et al.* [33] and Rost *et al.* [9], and 105 proteins selected by us.

Both training and testing process take only several minutes on the PC equipped with the 486 processor in the case when up to 200 proteins are used. The FORTRAN source code, files with Gaussian parameters and protein data bases used in this report are available via Internet (see Supplementary Material).

3. RESULTS

3.1. Conformational preference for transmembrane α-helix is strongly dependent on sequence hydrophobic environment for most amino acid types

When only transmembrane segments, expected to be in the helical conformation, are used to train the algorithm to predict helical segments, then preference for the α-helix conformation ('H') is at the same time the preference for the TMH conformation. In our case the training part of the algorithm uses so small percentage of the observed 'H' conformations in soluble proteins (because only soluble proteins of the β-class are used) that we can still consider the 'H' conformational preference as the transmembrane helical segment conformational preference. It appears that some amino acid types passively acquire the conformation dictated by their neighbors (Figure 1), while others (mainly charged amino acids) are able to resist to some extent (Table 1). Extremely secure dependence of TMH preference on the hydrophobic sequence environment is found for 12 amino acid types (Table 1). Only

Arg, Lys, Asp and Glu have the F factor (a statistical measure for the dependence of preference on hydrophobicity of sequence neighbors) less than 50.

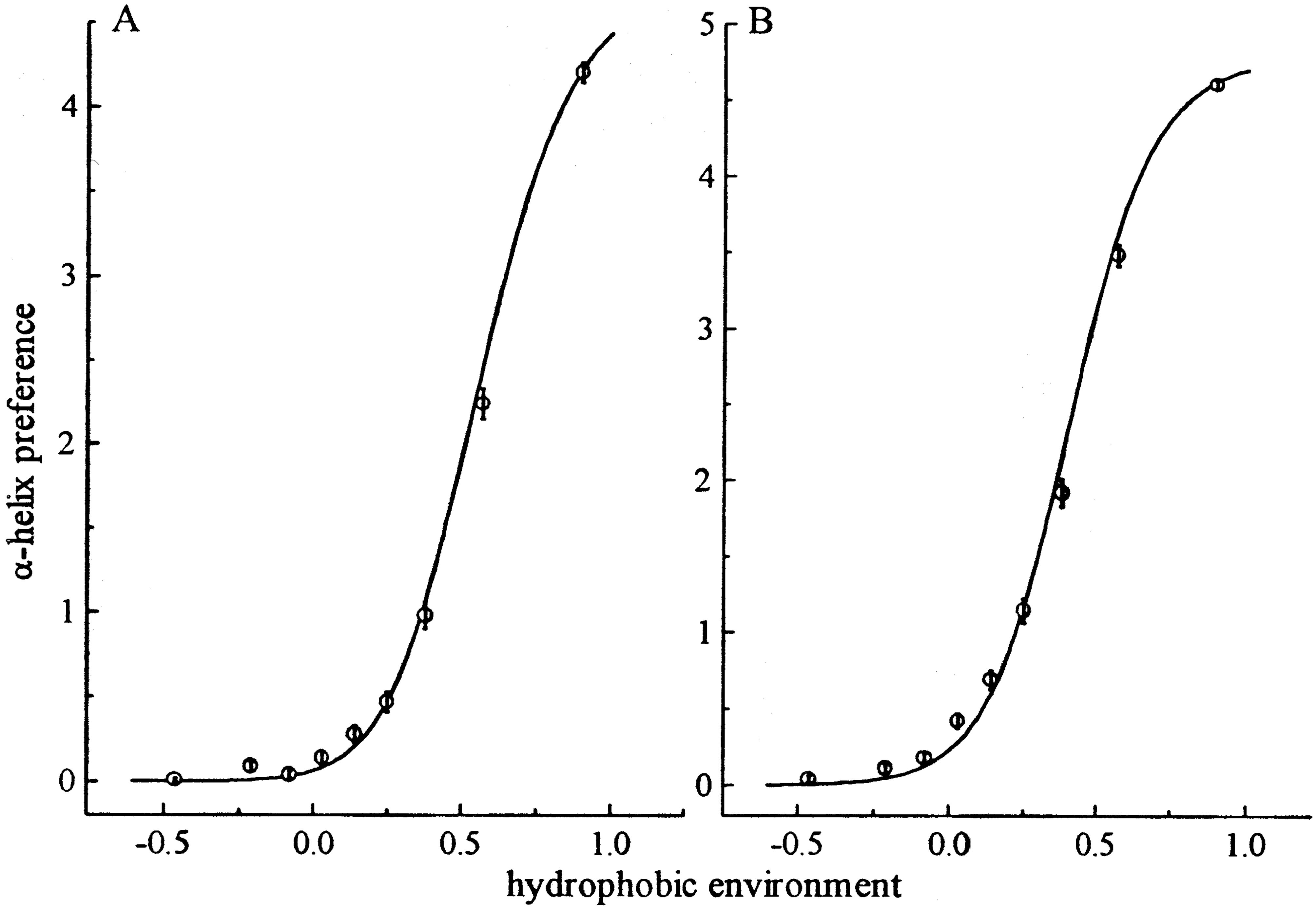

Figure 1: Very strong dependence of the α-helix conformational preferences on average hydrophobic sequence environment. Standard training procedure (Methods) was used. Observed preferences for glycine (Figure 1A) and leucine (Figure 1B) are shown as open points. Confidence limits, shown as bars above and below preference points, were calculated as described by Ptitsyn [51] so that it was 67.5% certain that observed preferences would fell between these values. The preference functions for leucine and glycine are shown as full lines.

From Figure 1 it is quite clear that linear approximation for the dependence of TMH preference on sequence environment is not so good as the preference function approximation. Similar results are obtained for other 18 amino acids (not shown). It is also clear that preference functions can be regarded as good but not the best nonlinear fit to observed preference points. For the four state model of secondary structure the preference function is obtained as the ratio of one Gaussian function to four Gaussian functions (Methods). Normal distribution (Gaussian function) is expected to be good fit for the histogram of sequence environments [32,52] due to averaging procedure used to produce histograms for each amino acid type and each secondary structure. However, cases of nonrandom distribution of amino acid types among sequence environments for particular secondary structure motifs have been observed [52] as well as the cases of too small number of sequence environments for the particular class of sequence environments, chosen amino acid type, secondary structure and

Table 1
Statistical parameters derived for the linear approximation of the dependence of helix preference on hydrophobic environment

Amino acid type	Slope: b	Standard error in slope: s_b	F parameter $(b/s_b)^2$:
Ala	2.726	0.034	6360
Arg	0.088	0.033	7
Asn	0.601	0.052	131
Asp	0.210	0.038	31
Cys	2.441	0.050	2387
Gln	0.337	0.045	57
Glu	0.173	0.037	22
Gly	1.605	0.029	3040
His	0.739	0.070	111
Ile	3.624	0.029	15941
Leu	3.381	0.025	18042
Lys	0.192	0.038	25
Met	2.806	0.055	2568
Phe	3.393	0.035	9292
Pro	0.754	0.042	319
Ser	1.110	0.030	1340
Thr	0.880	0.030	870
Trp	2.815	0.078	1288
Tyr	2.302	0.059	1510
Val	3.317	0.025	16961

protein data set used for the training procedure (not shown). Both possibilities can produce less than ideal fit of preference function to experimental data, and are discussed in a recent paper [52].

3.2. Expected and predicted length distribution for transmembrane helical segments

The transmembrane segments (TMS) and transmembrane helical segments (TMH) are not necessarily identical in lengths. Our predicted TMH could be longer and could be shorter than usual length of TMS of 19-22 residues. Figure 2 illustrates in the form of histogram that expected lengths of TMS could also be different from expected 19-22 residues.

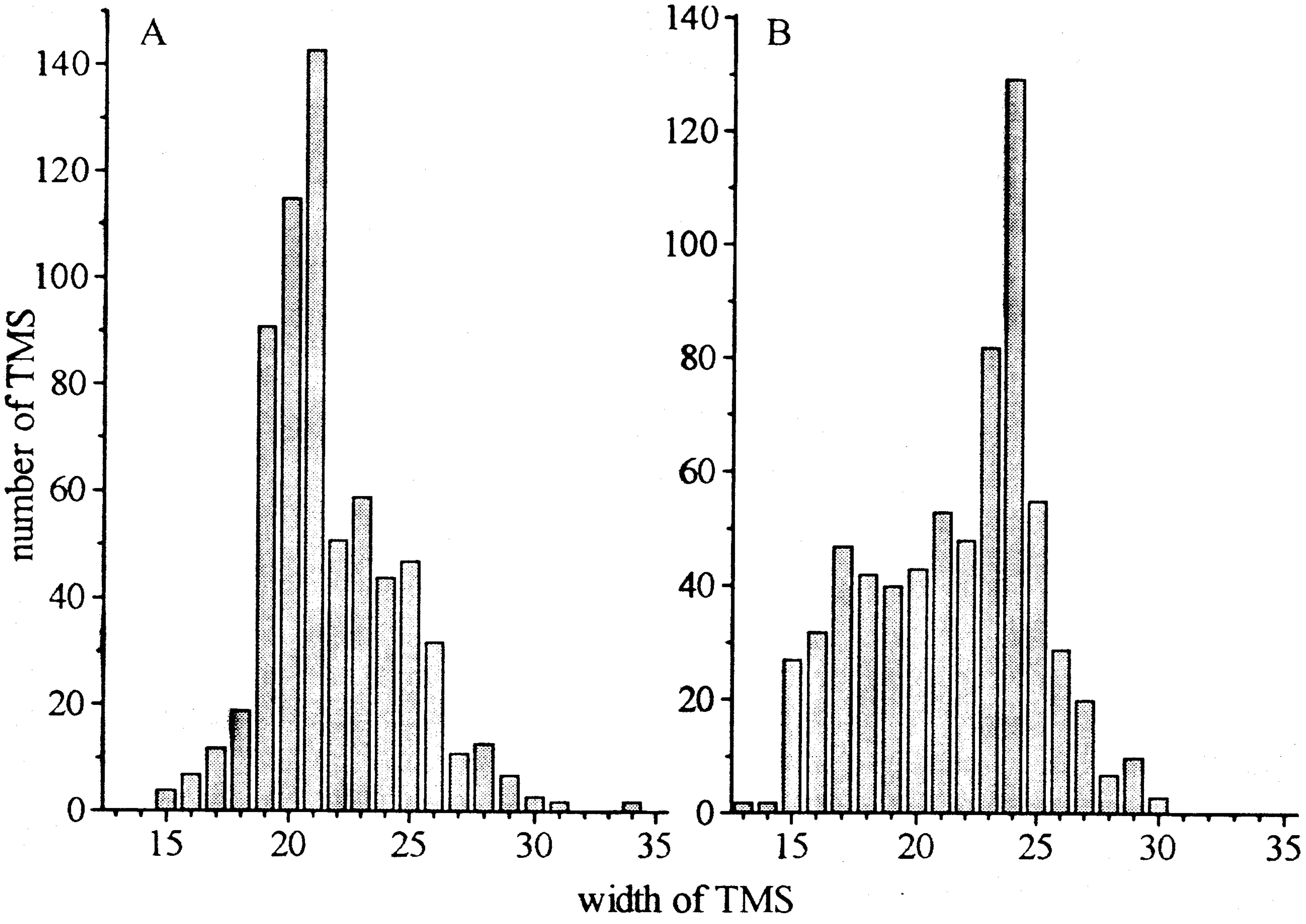

Figure 2: The length distribution of TMS in 168 proteins expected (Figure 2A) and predicted as TMH by us (Figure 2B). Two-times cross validation procedure (Methods) was used.

Both expected TMS and predicted TMH are often too short to span the membrane as α-helices or are so long that extramembrane parts in such segments must exist. Helical configurations other than α-helix should not be excluded for potential transmembrane segments [31]. For instance, it was pointed out [23] that 15 residue segment could span the bilayer as a 3_{10} helix. It is also possible that some transmembrane segments in α-class membrane proteins are in reality helical segments that pass through a part of membrane depth [53] or through whole membrane depth extending outside membrane or are in tilted orientation with respect to orthogonal direction from membrane surface. It appears from TMH length distribution in 10 integral membrane proteins of the best known structure too (not

Table 2
How results depend on the W parameter (sliding window length)^a

W	Q_s	A_{TM}	Q_p
7	94.0	0.689	57
9	95.2	0.700	64
11	95.5	0.704	67
13	94.7	0.697	58
15	93.1	0.694	59
17	93.0	0.693	57
19	92.1	0.675	57

^{a}Preference functions were extracted from the data base of 63 membrane proteins selected by Rost *et al.* [9] and 37 soluble proteins of the β-class (SOLB1, Methods) by using the PREF algorithm versions with sliding window length from 7 to 19 residues.

shown) that some membrane protein structures must be able to use hydrophobic segments of nonstandard lengths.

3.3. What is the optimal choice of the sliding window size?

In order to find the optimal length of the sliding window we varied the W parameter (sliding window length) from 7 to 19 (Table 2). Tests were done with the version of the SPLIT predictor that had corresponding length of the sliding window in each case. Only proteins having two or more transmembrane segments were used to test the predictor. There were 88 such proteins from our list of 168 proteins. All three performance parameters A_{TM}, Q_s, and Q_p (Methods) agree that a window size 11, requiring averaging of 5 left and 5 right sequence neighbor attributes, is optimal. Window size 11 is the half way between optimal size of 7 residues found by Degli Esposti *et al.* [54] and optimal window size of 15 residues found by Persson and Argos [55].

3.4. How do the results depend on different devices used in the SPLIT algorithm?

Table 3 results compare the importance of different devices used in the SPLIT algorithm. Chosen smoothing procedure is very important, while main filter procedure is next in importance. Subroutines 'FILTER', 'CHARGE-BREAK', 'TURN-BREAK' (Methods) and routine for finding maximum preference for the α-helix configuration were all eliminated to examine the importance of the main filter procedure. Automatic choice of decision constants for each tested protein helps to improve the prediction accuracy and the improvement is most obvious when A_{TM} and Q_p parameters are compared in the presence of the decision constants device (first row) and in its absence (fourth row).

Table 3
The dependence of prediction results on different devices used in the SPLIT algorithm[a]

SPLIT algorithm	A_{TM}	Q_s	Q_p	# predicted TMH	# proteins with correct prediction
With no change	0.712	95.0	76.8	665	129
With no smoothing	0.646	87.0	57.8	613	97
Without main part of the filter	0.655	95.3	63.7	596	107
With all DC = 0	0.693	92.3	67.9	649	114
Without 'FILTER' subroutine	0.701	95.0	76.8	665	129
Without additional parts of the filter	0.705	94.1	74.4	659	125

[a]Each device is separately eliminated from the algorithm before testing the prediction on the complete data set of 168 membrane proteins. The best Gaussian parameters file was obtained after 5-times cross validation procedure applied as described in the Methods section (the 'best' training procedure), but cross-validation was not performed. Refer to the Methods section for performance parameters.

'FILTER' subroutine alone seems to be important only in adjusting the positions of transmembrane helical caps. Additional parts of the filter, such as fusing short predicted helices, that may be the part of longer transmembrane helix, and extracting very short predicted helices with very high α-helix preference, are of minor importance. The Q_S parameter, or percentage of TMH that are correctly predicted, can be very misleading as the measure of prediction accuracy in the absence of a good filter, because it then increases together with the increase (overprediction) of residues predicted in the TMH conformation.

Table 4
Several scales of amino acid attributes used in this report

AA code	KYTDO[a]	MODKD[b]	CPREF[c]
Ala	1.8	1.10	0.6942
Arg	-4.5	-5.10	-1.4344
Asn	-3.5	-3.50	-0.7786
Asp	-3.5	-3.60	-1.1296
Cys	2.5	2.50	0.3427
Gln	-3.5	-3.68	-1.0870
Glu	-3.5	-3.20	-1.2480
Gly	-0.4	-0.64	-0.0549
His	-3.2	-3.20	-0.9697
Ile	4.5	4.50	1.7999
Leu	3.8	3.80	1.1403
Lys	-3.9	-4.11	-1.1850
Met	1.9	[illegible]	1.3557
Phe	2.8	2.80	1.3171
Pro	-1.6	-1.90	-0.5091
Ser	-0.8	-0.50	-0.2812
Thr	-0.7	-0.70	-0.2030
Trp	-0.9	-0.46	0.8475
Tyr	-1.3	-1.3	0.3693
Val	4.2	4.2	1.0138

[a,b,c] Scale acronyms are defined in Table 5.

3.5. What are the best scales of amino acid attributes?

As expected many different hydrophobicity scales are good predictors of transmembrane helical segments. The same scale is used during training and testing procedure. Each scale is normalized with average zero and standard deviation of one when called by the algorithm. As an example 20 values for the Kyte-Doolittle scale (KYTDO) are given together with modified Kyte-Doolittle scale (MODKD) and with normalized scale of constant preferences (CPREF) that were extracted from the reference data set of 168 membrane proteins (Table 4).

The list of 30 scales in Table 5 is our selection of the best predictor-scales from almost 100 scales that are available in the algorithm.

Table 5
Evaluation of hydrophobicity scales[a]

Scale #	Acronym: Attribute	Reference	Performance parameters		
			A_{TM}	Q_s	Q_p
83	MODKD: Modified Kyte-Doolittle hydropathy scale	This work (Table 4)	0.711	95.7	76.8
1	KYTDO: Hydropathy values	[17]	0.704	95.9	78.6
100	CPREF: TMH preferences from training data base	This work (Table 4)	0.699	96.4	73.2
17	PONG1: Surrounding hydrophobicity scale	[47]	0.680	94.1	68.5
26	EISEN: Consensus hydrophobicity scale	[18]	0.675	95.0	67.8
9	VHEBL: Hydropathy scale for membrane proteins	[56]	0.672	95.2	68.5
35	NNEIG: Self-consistent hydrophobicity scale	[50]	0.671	93.7	66.7
29	CHOTH: Proportion of residues 95 percent buried	[57]	0.670	93.8	66.7
30	ROSEF: Mean fractional area loss	[58]	0.666	93.7	63.1
52	EDE25: Optimal predictors for width 25	[22]	0.660	94.6	66.7
53	EDE21: Optimal predictors for width 21	[22]	0.660	94.6	65.5
4	ENGEL: Hydropathy values	[59]	0.659	94.7	64.9

Continued on next page.

Table 5 - Continued

49	HEIJN: Hydrophobicity scale for TMS	[60]	0.658	94.4	63.7
71	GRANT: Polarity scale	[61]	0.658	93.2	62.5
44	DEBER: M/A ratio in membrane transport proteins	[62]	0.656	93.2	62.5
7	GUY-M: Average of four hydrophobicity scales	[63]	0.655	92.1	64.3
70	WOESE: Polarity scale	[64]	0.652	94.0	61.9
3	PONNU: Surrounding hydrophobicity scale	[65]	0.652	91.5	59.5
8	KRIGK: Ethanol to H_2O hydrophobicity scale	[66]	0.650	93.1	60.1
28	HOPPW: Antigenic determinant scale	[67]	0.649	94.1	63.1
5	JANIN: Free energy of transfer from protein interior	[68]	0.645	92.9	61.3
16	CIDAB: Hydrophobicity scale for proteins of α/β class	[69]	0.645	90.5	63.7
31	GUYFE: Transfer free energy for 6 layers in proteins	[63]	0.645	91.4	61.9
42	MIJER: Average contact energy	[70]	0.645	91.2	61.9
12	GIBRA: Solvent accessibility in proteins	[71]	0.645	92.1	60.1
2	FAUPL: Solution hydrophobicities	[72]	0.643	92.6	60.1
19	PONG3: Combined membrane hydrophobicity scale	[47]	0.642	94.9	63.7
27	PRIFT: Statistical scale for amphipathic helices	[50]	0.642	93.2	61.3
21	ROSEM: Self-solvation free-energy changes	[73]	0.639	92.1	61.9
78	CASSI: Structure-derived hydrophobicity scale	[74]	0.635	93.4	58.3

[a]For a chosen scale of amino acid attributes each of 168 membrane proteins was tested once without being used in the training procedure as described in the Methods section.

3.6. The prediction results with Kyte-Doolittle preference functions

Full details of prediction results for each of 168 reference membrane proteins are enclosed in the Supplementary Material (Table IV). We used cross-validation (5-fold, Methods) and the KYTDO scale (# 1). All of 168 proteins were correctly predicted as membrane proteins having at least one transmembrane segment. With 100% correct transmembrane topology 130 proteins were predicted. A total of 631 transmembrane helices were correctly predicted out of a total number of 662 expected transmembrane segments. Only 36 TMH were overpredicted and 31 underpredicted. Of individual residues in TMH configuration 12273 out of 14374 were correctly predicted, 2033 overpredicted and 2101 underpredicted. The performance parameters (Methods) are then: $A_{TM} = 0.712$, $Q_s = 95.3\%$, $Q_p = 77.4\%$, $A_s = 0.898$.

As an example of complete information provided by the predictor the predicted preference profiles and hydrophobic moment profiles for the gef_ecoli protein (outside reference list of 168 proteins and without assigned transmembrane domain in the SWISS-PROT data base) are given in Table 6 as unmodified output file. The gef protein can stimulate cell killing [75] after overexpression and oligomerization in the membrane environment. In addition to predicted α-helix transmembrane segment from residues 6 to 24 there is also the 31-46 segment predicted in the β-strand conformation. The 31 to 45 segment may be another potential membrane-embedded segment possibly involved in dimerization or oligomerization process in the membrane environment that can lead to cell killing.

Table 6
Complete prediction results for the gef_ecoli protein by using the Kyte-Doolittle hydropathy scale through preference functions[a]

	AA	PS	PTM	PH	PB	PT	PU	MA	MB	H-T	PB+MB-2
1	M	C	U	0.00	0.00	0.33	1.58	ND	ND	-0.33	ND
2	K	C	U	0.02	0.29	0.62	1.49	ND	ND	-0.60	ND
3	Q	C	U	0.06	0.54	0.81	1.42	1.38	1.08	-0.75	-0.38
4	H	C	U	0.48	0.79	1.06	1.33	1.21	0.33	-0.57	-0.88
5	K	C	T	1.02	0.82	1.30	1.12	1.07	0.21	-0.27	-0.97
6	A	H	M	1.66	0.88	1.31	0.89	1.26	0.39	0.35	-0.73
7	M	H	O	2.32	0.74	0.86	0.60	1.17	0.34	1.47	-0.91
8	I	H	O	3.00	0.53	0.49	0.32	0.61	1.00	2.51	-0.47
9	V	H	O	3.65	0.28	0.19	0.10	0.53	0.52	3.46	-1.19
10	A	H	O	4.17	0.11	0.08	0.02	0.79	0.60	4.10	-1.29
11	L	H	O	4.57	0.02	0.04	0.01	0.73	0.74	4.53	-1.23
12	I	H	O	4.71	0.01	0.02	0.00	0.92	0.92	4.69	-1.07
13	V	H	O	4.75	0.01	0.02	0.00	1.07	0.59	4.73	-1.41
14	I	H	O	4.75	0.01	0.02	0.00	1.11	0.60	4.73	-1.39
15	C	H	O	4.74	0.01	0.02	0.00	0.97	0.56	4.72	-1.43
16	I	H	O	4.74	0.01	0.05	0.01	0.96	0.56	4.68	-1.43
17	T	H	O	4.72	0.01	0.06	0.01	1.11	0.38	4.66	-1.61
18	A	H	O	4.64	0.01	0.06	0.01	1.22	0.55	4.57	-1.44
19	V	H	O	4.33	0.06	0.05	0.01	1.07	0.48	4.28	-1.46

Continued on next page.

Table 6 - Continued

20	V	H	O	3.93	0.21	0.15	0.05	1.67	0.14	3.78	-1.65
21	A	H	O	3.58	0.43	0.44	0.21	1.72	0.16	3.13	-1.42
22	A	H	O	3.15	0.74	0.72	0.42	1.05	0.55	2.43	-0.71
23	L	H	O	2.50	1.07	0.85	0.59	1.17	0.48	1.65	-0.45
24	V	H	M	1.88	1.21	0.94	0.65	0.99	0.48	0.94	-0.30
25	T	C	T	1.31	1.26	1.34	0.76	0.97	0.44	-0.03	-0.30
26	R	C	T	0.98	1.20	1.77	0.90	1.02	0.23	-0.78	-0.57
27	K	C	T	0.71	1.19	1.87	1.07	0.81	0.52	-1.16	-0.29
28	D	C	T	0.41	1.12	1.46	1.14	0.83	0.51	-1.05	-0.37
29	L	C	U	0.18	1.05	1.03	1.26	0.65	0.34	-0.85	-0.61
30	C	C	U	0.18	1.26	0.77	1.34	0.95	0.59	-0.59	-0.15
31	E	B	E	0.17	1.37	0.76	1.35	0.89	0.63	-0.58	0.01
32	V	B	E	0.16	1.43	0.91	1.29	0.77	0.68	-0.75	0.12
33	H	B	E	0.11	1.27	0.92	1.24	0.83	0.77	-0.81	0.04
34	I	B	E	0.14	1.46	1.07	1.25	0.70	0.74	-0.93	0.20
35	R	C	E	0.16	1.28	1.02	1.28	0.45	1.27	-0.86	0.55
36	T	B	E	0.24	1.33	1.30	1.24	0.52	1.34	-1.06	0.67
37	G	B	E	0.23	1.40	1.29	1.19	0.59	1.16	-1.06	0.55
38	Q	B	E	0.31	1.44	1.39	1.08	0.57	1.10	-1.08	0.54
39	T	B	E	0.44	1.63	1.21	1.14	0.79	0.91	-0.77	0.54
40	E	B	E	0.49	1.64	1.10	1.07	0.53	0.91	-0.61	0.55
41	V	B	E	0.64	1.83	1.07	1.06	0.64	0.86	-0.43	0.69
42	A	B	E	0.65	1.77	1.02	1.00	0.95	0.55	-0.36	0.32
43	V	B	E	0.63	2.03	1.06	0.96	0.90	0.44	-0.44	0.47
44	F	B	E	0.63	1.78	1.08	1.00	0.71	0.36	-0.45	0.14
45	T	B	E	0.52	1.97	1.17	1.01	0.65	0.33	-0.65	0.30
46	A	B	B	0.38	1.55	1.17	1.12	1.16	0.40	-0.79	-0.05
47	Y	C	E	0.34	1.20	0.95	1.25	1.09	0.95	-0.61	0.15
48	E	C	U	0.15	0.81	0.74	1.39	0.61	1.15	-0.59	-0.04
49	S	C	U	0.03	0.22	0.52	1.51	ND	ND	-0.49	ND
50	E	C	U	0.04	0.16	0.39	1.54	ND	ND	-0.36	ND

[a]One letter amino acid codes are used in the second column (AA). Predicted structure (PS) in the third column can be α-helix (H), β-sheet (B) or coil (C) structure that includes turn and undefined structure. Residues predicted in the transmembrane helix configuration (PTM) in the fourth column are labeled with letter 'M' except for highly probable TMH conformation when letter 'O' is used. Residues with a potential to form transmembrane β-strands are labeled with letter 'E' in the fourth column. The coil (C) conformation from third column is specified as undefined (U) or turn (T) conformation in the fourth column. Fifth to eighth column contain smoothed preferences for α-helix (PH), β-sheet (PB), turn (PT) and undefined (PU) conformation. The columns 9 and 10 contain numerical values for hydrophobic moments calculated in the case of assumed α-helix configuration (MA) and for moments calculated for assumed β-sheet configuration (MB). Last two columns contain PH-PT difference of preferences (H-T) that helps in visual identification of predicted transmembrane helices and PB+MB-2.0 scores that help in prediction of potential membrane-embedded β-strands.

Since interaction of transmembrane helices is not directly taken into account by us it is possible that our prediction for proteins expected to have large number of transmembrane helices systematically err on the side of underprediction. One such example may be the calcium channel subunit cic1_cypca in which fourth and tenth potential transmembrane segments are not predicted. Another such example is the human erythrocyte anion exchanger b3at_human in which our prediction of 13 transmembrane helices is associated with one underpredicted TMH (residues 460-479) according to the 14 TMH topological model of Wang *et al.* [76]. Underpredicted segment has three Glu residues and not enough high preference peak, so that it is rejected by algorithm's filter procedure, but can be recognized from preference profile (not shown) as potential TMH segment. Earlier models for a monomer of the Band 3 dimmer [77] predicted only 12 membrane-spanning α-helices, but one of the authors [53] later observed that 'inner' helices can be easily overlooked when sufficiently long hydrophobic segments are sought, since such helices can span the membrane only partially and without direct contact with lipid environments.

Binding of ligands or cofactors is not taken into account too, but it can conceivably change the potential for formation of regular secondary structure for sequence segment that interacts with a ligand or cofactor. Underprediction was seen for the tromboxane A synthase (thas_human), a member of the P-450 family, which probably binds heme-thiolate at the position 479. The fifth transmembrane segment, that is underpredicted both by us and Rost *et al.* [9], starts with residue 480 in the tromboxane A synthase topological model reported by the SWISS-PROT data base.

Gross errors in the topological models adopted by the authors and by the SWISS-PROT or some other data base can be easily detected by our algorithm. We have very strong prediction of three transmembrane helical segments (14-36, 139-160 and 166-189) for the TOLQ protein from *Escherichia coli*. Only the first TMS from residues 23 to 43 is correctly predicted according to the SWISS-PROT assignment of the bitopic transmembrane topology for that protein. Interestingly, very similar protein exbb_ecoli has SWISS-PROT release 29 assignment of three transmembrane segments too. Two commonly used methods for predicting transmembrane helices, that of Eisenberg *et al.* [18] and that of Rao and Argos [20] also predict three transmembrane segments for tolq_ecoli, while Rost *et al.* method [9] predicts four transmembrane segments for that protein. Small number of homologues for that protein (only 3) and a need to filter predicted 'transmembrane segment' having 66 residues is the likely cause for the proposed four helix model by the automatic *E-mail* service of Rost *et al.* [9].

3.7. Testing for false positive predictions in membrane and soluble proteins of crystallographically known structure

Ten integral membrane proteins of well known structure (BESTP, Methods) have been tested first. Only the Kyte-Doolittle and our modification of the Kyte-Doolittle scale (MODKD, # 83) were able to predict all od these ten membrane proteins with 100% correct transmembrane topology, i.e. all transmembrane helices were correctly predicted at their observed sequence locations and there were no overpredicted TMH (Table 7). Only the Chothia buried surface scale (CHOTH, # 29) did not recognize one of ten membrane proteins as the membrane protein (the subunit H from the photosynthetic reaction center from *R. viridis*). Nine long extramembrane helices in these 10 proteins were not predicted as TMH by any of 12 tested amino acid scales. That these sensitive tests of our predictor do not depend on the chosen training procedure was checked by using different training procedures. After training the algorithm on 63 proteins selected by Rost *et al.* [9] or on 105 proteins selected by us with the addition of 37 β-class soluble proteins (SOLB1) the results were very similar (not shown). Another sensitive test was made possible when the crystal structure of cytochrome c oxidase from *Paracoccus denitrificans* [14] became known during work on this report.

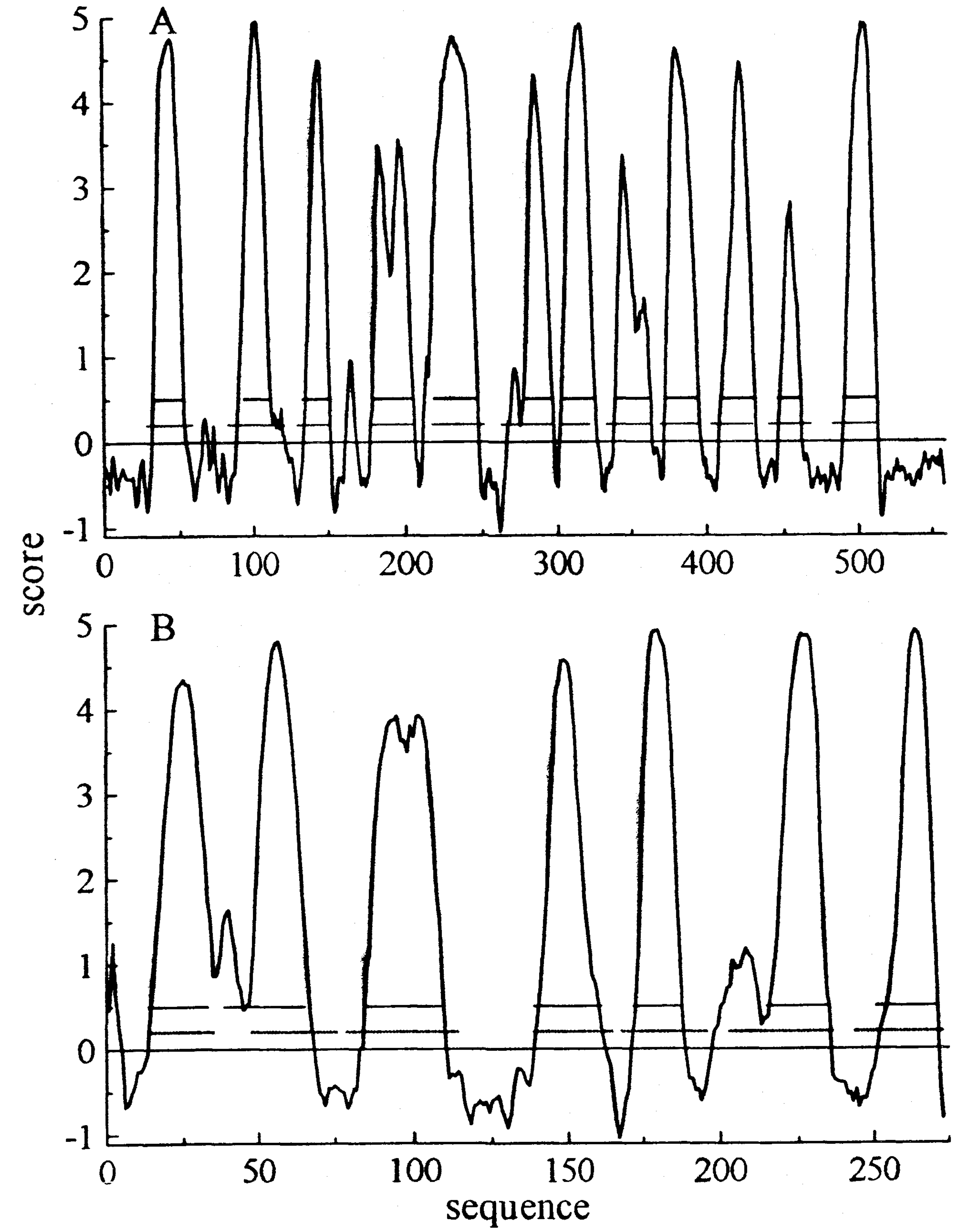

Figure 3: Score profiles for cx1b_parde (Figure 3A) and for cox3_parde (Figure 3B) of cytochrome oxidase from *Paracoccus denitrificans* [14] are obtained by substraction of turn preferences from α-helix preferences (full line). Digital predictions, as outcome of the best training procedure for the SPLIT algorithm with Kyte-Doolittle hydropathy scale (Methods), are shown as bold horizontal bars at the score level 0.5. Observed location of TMH segments are shown as bold horizontal bars at the score level 0.2.

With our best file of Gaussian parameters ('best' training procedure, Methods) we correctly predicted all of 12 TMH in cx1b_parde (Figure 3A) and all of 7 TMH in cox3_parde (Figure 3B) without single overpredicted TMH. Subunit IV was not tested, while in the cox2_parde two TMH were predicted correctly and two overpredicted. Predicted 'TMH' at residues 12 to 30 is the signal sequence. Predicted 'TMH' at residues 192 to 216 has atypical flat profile with maximum height less than half of other peaks. Three observed β-strands: 190-194, 200-204 and 209-216 at that position are not seen by the algorithm when it makes the automatic choice of decision constants (Methods) such that β-structure is depressed. Setting all decision constants to zero eliminated this erroneous TMH prediction. When these 3 polipeptydes are added to 10 considered above, the total score is 49 correctly predicted TMH out of the total number of 49 observed TMH (not counting the signal sequence), with one, easy to detect, overprediction. This result did not change in the case when 105 integral membrane proteins were used to train the algorithm and to extract corresponding file with Gaussian parameters, but one TMH was overpredicted in the cox3 at sequence segment 198-213 when 63 or all of 168 proteins were used in the training process. Setting all decision constants to zero eliminated this overprediction as well in both cases. Standard training procedure with the MODKD scale (Table 4) produced 100% correct topology for subunits cx1b and cox3 and the same two overpredicted TMH in the cox2.

Table 7
Test of best 12 amino acid attributes in predicting TMH in membrane proteins of known structure[a]

Scale #	A_{TM}	# correct TMH pred.	# predicted TMH	# correct M. P. pred.	# predicted M. P.
1	0.695	28	28	10	10
83	0.693	28	28	10	10
52	0.682	27	28	8	10
53	0.679	27	28	8	10
29	0.676	27	27	9	9
17	0.644	27	30	7	10
35	0.626	27	29	7	10
100	0.616	27	29	7	10
30	0.603	27	31	6	10
4	0.547	27	31	5	10
9	0.527	27	32	5	10
26	0.523	27	32	6	10

[a]Tested proteins (BESTP, Methods) had 28 observed TMH with 717 residues in the TMH conformation. Standard training procedure was used with each choice of amino acid attribute. Code numbers for amino acid scales are listed in Table 5.

By using our standard training procedure the tests were performed on membrane proteins of known or partially known structure with transmembrane β-strands and on soluble proteins of known structure. For seven tested porins and two defensins (PORINS, Methods) we tested 12 best scales used in Table 7. Only the scale # 4 [59] predicted one transmembrane segment in the α-helix conformation (residues 119 to 133 in the porin sequence from *Rhodobacter capsulatus*). For two different sets of soluble proteins SOLU1 and SOLU2 (Methods) prediction results are collected in Table 8 as percentage of proteins falsely predicted to be membrane proteins. The best scales for TMH prediction in membrane proteins still falsely predicted 11-12% of soluble proteins as being membrane proteins with at least one transmembrane helix.

Table 8
The prediction performance of 12 best amino acid scales (Table 5) on soluble proteins[a]

Scale #	SOLU1[b]	SOLU2[b]
17	11.2	12.2
53	11.2	12.2
30	11.2	13.6
52	11.2	13.6
83	12.8	13.6
100	13.4	14.3
35	15.0	16.3
1	13.9	17.7
29	19.3	17.7
9	21.4	19.0
26	20.3	19.7
4	25.7	23.8

[a] Only the percentage of proteins predicted with one or more transmembrane helices is reported. Code numbers for amino acid scales are listed in Table 5.

[b] Data base of soluble proteins of known structures (see Methods).

3.8. Cross-validation, overtraining and sensitivity to the choice of protein data base

After standard training procedure tests were performed separately on the subsets of 80 proteins having only one observed TMH and 88 proteins having more than one TMH. All performance parameters registered higher prediction accuracy for 80 proteins having only one transmembrane segment. The best result of A_{TM} = 0.778, Q_s = 97.5% and Q_p = 92.5% was achieved in the 2 times cross-validation procedure when training was done on 88 proteins having more than one TMH and 37 soluble proteins of β-class. Interestingly, training and testing on the same data set of 80 membrane proteins (with 37 soluble β-class proteins included as always in the training procedure) produced huge overprediction of predicted TMH and very poor performance parameters A_{TM} = 0.285 and Q_p = 52.5%. The percentage of accurately predicted transmembrane helices remained the same: Q_s = 97.5% or 78 correctly predicted TMH of 80 observed, but total number of predicted TMH jumped from 84 to 139, while number of overpredicted TMH jumped from 6 to 61! Commonly used Q_s parameter gives obviously wrong picture of the prediction performance in this case. More surprising result is such extreme advantage of cross-validation procedure versus training and testing on the same data set of integral membrane proteins. Slight advantage of the cross-validation procedure is seen too when all of 168 reference proteins are used for training and for testing (compare performance parameters in the first two rows of Table 9). Needless to say, we always expect a decrease in the prediction performance when training is no longer performed on the same data set that is used for testing procedure.

Table 9
Different training procedures[a]

	A_{TM}	Q_s	A_s	Q_p	# prot tested
a) Five-times cross validation (Supplementary Material Table IV).	0.712	95.3	0.898	77.4	168
b) No cross-validation. All of 168 membrane proteins used to train and to test.	0.709	94.7	0.891	76.2	168
c) No cross-validation. Best training procedure (Methods).	0.712	95.0	0.896	76.8	168
d) Two-times cross validation: 63 proteins to train and 105 to test and vice versa.	0.704	95.9	0.903	78.6	168
e) Train on 105 proteins. Test on 63.	0.740	97.9	0.934	84.1	63
f) Train on 63 proteins. Test on 105.	0.682	94.7	0.885	75.2	105
g) Train on 105 proteins. Test on 105.	0.693	94.0	0.878	73.3	105
h)Train on 63 proteins. Test on 63.	0.737	97.5	0.905	76.2	63
i) No cross-validation. Soluble proteins SOLB2 used instead of SOLB1 during training procedure.	0.705	94.4	0.890	74.4	168

[a]The Kyte-Doolittle scale is used in each case. See Methods for performance parameters.

The clue is offered by such training procedure when only 16 residues next to each side of a transmembrane segment are used to extract sequence environments. Then it becomes possible to use 80 proteins having single TMH both for training and for testing and to obtain high performance parameters: A_{TM} = 0.777, Q_s = 97.5% and Q_p = 92.5%. It would seem that dominant contribution of sequence environments from extramembrane parts of membrane proteins with single TMH must be reduced if balanced training is to be achieved. This can be done either directly by omitting residues from the training process that are far removed from expected transmembrane segments or indirectly by choosing the training data base of membrane proteins with balanced contribution of residues in transmembrane and in extramembrane positions.

That need for balanced training is not the whole explanation becomes clear when the PREF algorithm is modified in such a way that is always collects exactly the same number of environments associated with different secondary structure motifs. Again poor prediction results are obtained when bitopic proteins are used for training and for testing (not shown). When all of 168 membrane and 37 soluble proteins are used in a balanced training procedure prediction results for 168 proteins remain similar for the TMH prediction (A_{TM} = 0.702), but overall prediction of all secondary structures is dramatically improved (Q_3 = 0.775) meaning that turn and undefined residues are much better predicted.

The extraction of preference functions, as the training procedure, is not a very powerful training procedure and it is not expected to lead to overtraining. We shall test this assumption by performing still another two-times cross-validation test in which 168 membrane proteins are divided into 63 proteins used by Rost *et al.* [9] and 105 proteins used by us. Table 9 lists performance results for different combinations of training and testing procedures.

Table 9 results indicate that extraction of preference functions as the part of the training procedure does not lead to overtraining, because training on an independent set of unrelated proteins can produce even better results. It is still possible that either automatic or subjective choice of filter parameters leads to overtraining. All our filter parameters were trained on the subset of 63 proteins and with the choice of the sliding window length of W = 9 residues. A drop in prediction accuracy when W = 11 (window length used in all presented results) is used, for the same subset of 63 proteins, was indeed observed (not shown). Since W = 11 seems to be optimal for much larger group of transmembrane segments (Table 2) it is indeed possible to increase apparent prediction accuracy by variation of filter parameters. To avoid such a danger we did not try to optimize filter parameters for a final choice of sliding window length (W = 11) and protein data base (168 proteins).

Having a larger reference set of nonhomologous proteins for extracting preference functions will not increase prediction accuracy. Safe lower limit is difficult to estimate, but is probably no more than 30-40 such proteins. In terms of residues considered for extraction of preference functions only about 4500 residues were enough to achieve very high prediction accuracy (A_{TM} = 0.777) in the case of bitopic (single-span) membrane proteins. A different set of soluble proteins in the training list of proteins may change slightly the prediction performance (last row in Table 9).

3.9. Comparisons with other methods

An automated FTP service was used to obtain the predictions for all of our 168 integral membrane proteins by using the Rost *et al.* method [9]. A total of 11870 residues were correctly predicted in the TMH conformations, 2436 residues were overpredicted, 2512 residues were underpredicted, while 50335 residues were correctly predicted not to be in the TMH conformation. One of many different performance parameter that can be constructed by using these data is the A_{TM} parameter (Methods). Its value is A_{TM} = 0.656, which is inferior to our value of 0.712 (Table 9) for the same parameter. However, when tested on the subset of 63 proteins used by Rost *et al.* [9] the A_{TM} parameter, calculated from predictions returned by automated service, becomes 0.733, which is comparable to our value of A_{TM} = 0.740 for the same subset of proteins (Table 9). Similar test on the subset of 105 proteins, never before seen in the training process for the neural network algorithm, gave quite a low value of A_{TM} = 0.610 for the Rost *et al.* method [9]. That value is lower than our value of A_{TM} = 0.682 for the same subset of 105 proteins (Table 9). All of 63 proteins selected by Rost *et al.* [9] are also predicted as membrane proteins, but their method does not recognize 2 out of 105 membrane proteins selected by us. Underprediction of membrane proteins is due to serious underprediction of transmembrane helices: 50 of observed 419 TMH are underpredicted and 11 overpredicted by Rost *et al.* [9]. For comparison our Table 9 results (row f) for A_s are obtained for the case of 21 underpredicted and 25 overpredicted TMH in the same test set of 105 proteins.

The prediction results for three commonly used prediction methods: that of Rao and Argos [20], that of Eisenberg *et al.* [18], and that of Rost *et al.* [9] can be compared with our results listed in Table 7 for the data set of 10 best known membrane proteins with observed 717 residues in the membrane-spanning helix conformation (Supplementary Material, Table V). Eisenberg's algorithm overpredicts 5 helices in the subunits M, and L from the photosynthetic reaction center, and has correspondigly low performance parameter for all ten proteins: A_{TM} = 0.470 (195 underpredicted and 185 overpredicted residues). Rao and Argos algorithm [20] has better performance of A_{TM} = 0.562, but large number of residues (314) is still underpredicted or overpredicted. Rost *et al.* neural network algorithm [9] used some subunits of the photosynthetic reaction center for training and achieved a much better result: A_{TM} = 0.702 with 108 underpredicted and 106 overpredicted residues. Only one helix was underpredicted (the N-terminal transmembrane helix from the light harvesting center).

Residues in three transmembrane helices of the LHC-II are underpredicted by all four methods, the likely reason being increase in helix preference due to binding of chlorophylls which is not taken into account by these methods. The whole first helix is underpredicted in Rost *et al.* [9] and Rao and Argos method [20]. It may seem strange that Rost *et al.* method [9] can predict 7 of 35 residues in the first transmembrane helix of the cb21_pea protein, but cannot repeat even such partial success when shorter but otherwise identical LHC-II polypeptide is tested. Filter elimination of signal sequences (see Discussion) and/or too short potential transmembrane segments in the Rost *et al.* [9] procedure becomes critical when polypeptides lacking complete N-terminal section in front of a potential TMS are considered.

Our result A_{TM} = 0.695 (Table 7, Kyte-Doolittle scale) for all of 10 membrane proteins becomes A_{TM} = 0.714 (56 underpredicted and 23 overpredicted residues, all 11 TMH correctly predicted) when only H, M an L subunits of the photosynthetic reaction center from *Rhodopseudomonas viridis* are considered. This can be compared with Fasman and Gilbert [78], and Ponnuswamy and Gromiha [47] evaluation of many different methods for predicting transmembrane helices when these same three polypeptides are used as very restricted 'standard of truth'. The Kyte-Doolittle [17], Sieved Kyte-Doolittle [21] and Klein-Kanehisa-DeLisi procedure [19] are associated with prediction accuracy lower than A_{TM} = 0.7, while von Heijne [79], Engelman-Steitz-Goldman [59], Esposti-Crimi-Ventruoli [54], and Ponnuswamy-Gromiha procedure [47] are associated with higher prediction accuracy.

We have also compared two powerful prediction methods, that of Jones *et al.* [33] and Rost *et al.* [9] with our own (JLT) by testing greater number of proteins whose expected transmembrane structure is taken from the SWISS-PROT data base. For 83 proteins used by Jones *et al.* [33] one can extract A_s and Q_p performance parameters as A_s = 0.928 and Q_p = 79.5%. For 69 proteins tested by Rost *et al.* [9] A_s and Q_p parameters are 0.896 and 79.7%, respectively. For 63 proteins tested by us these parameters are A_s = 0.934 and Q_p = 84.1% (Table 9).

Overprediction of transmembrane segments in large eukaryotic proteins having single transmembrane achoring segment is common deficiency of many prediction methods [33]. Our algorithm overpredicts six and underpredicts two transmembrane segments in the data base of 80 membrane proteins expected to have single transmembrane helix. For instance, in the case of epidermal growth factor receptor precursor: egfr_human Jones *et al.* [33] overpredicts two transmembrane segments. Our method adds to correct prediction of the segment 646 to 668 an incorrect prediction for residues 777 to 798. Rost *et al.* prediction 648-666 without overpredicted segments is even better [9]. The price paid for reduced overprediction in single-span proteins is seen much better when Rost *et al.* method [9] is tested on never-before-seen data set of 105 membrane proteins containing 48 single-span proteins. Then two proteins: ftsh_ecoli and spir_spime are not predicted as membrane proteins, because Rost at al. [9] do not find a single transmembrane segment in these proteins. Cell division protein ftsh is strongly predicted with two transmembrane helices at correct sequence location by our method (Supplementary Material, Table IV). Spiralin is predicted by us as membrane protein, but with transmembrane segment at the N-terminal (residues 3 to 21) instead from residues 165 to 184 (SWISS-PROT assignment).

Underpredictions of the last transmembrane segment in G-protein coupled receptors with seven transmembrane segments are also commonly seen by our and other methods [33]. This is the case with a2aa_human, aa2a_canfa, acm5_human, car1_dicdi, ops1_calvi and ops2_drome for our prediction. The seventh helix in the superfamily of seven-helix protein G-coupled receptors contains retinyl-lysine in the case of opsins or may be adjacent to a potential acylation site [80]. As a rule it can be recognized for potential TMH from preference profile as the last of 7 sharp peaks (often with characteristic minimum pointing at sequence position of functionally important lysine residue) even if the digital version of the predictor cannot predict it.

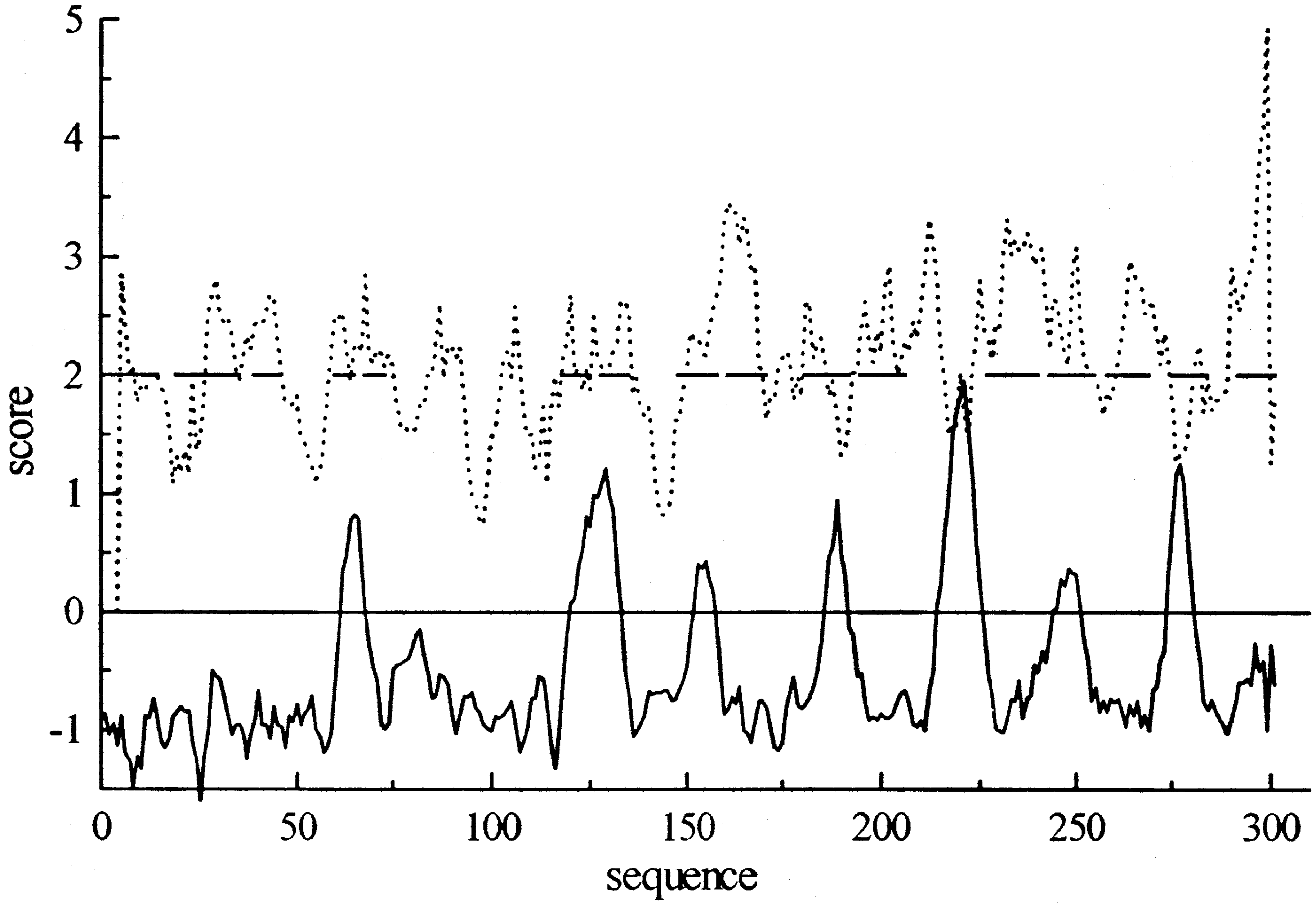

Figure 4: Score profiles for porin from *Rhodobacter capsulatus* are obtained by subtraction of turn preferences from helical preferences (full line) and as sum of β-sheet preferences and hydrophobic moment scores for assumed β-sheet conformation (dotted line). Kyte-Doolittle scale [17] is used to calculate preferences, while PRIFT scale [50] is used to calculate hydrophobic moments. Observed transmembrane strands are shown as bold horizontal bars at the score level 2.0.

3.10. Using prediction profiles with both α and β motifs

Unrealistic initial assumption that only α-helices exist as transmembrane polypeptide structure can be tested by using predictions for membrane or surface attached β-strands (Methods) as well. All tests in this section are done with decision constants fixed at zero. Previously unseen possibilities for β-strand formation in the membrane environment become apparent from profiles of summed β-preferences and β-hydrophobic moments (Figure 4 and 5). Dotted line in Figures 4 and 5 can be regarded as the score profile for potential formation of membrane-buried or membrane attached β-strands ('E' structure in the fourth column of Table 6). As before we used the Kyte-Doolittle scale for preference calculation and the PRIFT scale to find hydrophobic moments for assumed β-structure. Revealed potential for the β-structure formation in the membrane is quite robust with respect to the change in the choice of hydrophobicity scale for preference calculation, notwithstanding the complexity of the scores profile. Above mentioned combination of scales predicted correctly 79% of membrane-buried β-strand residues in 9 membrane β-class proteins (PORINS, Methods), 72% of such residues in three best known porins (porin from *R. capsulatus* OmpF and PhoE) and 76% of such residues in the *R. capsulatus* porin. When algorithm is allowed to make its own choice od decision constants these percentages raise to 87, 82 and 76 respectively. Only one membrane-embedded β-strand (the 15-th) is underpredicted in the *R. capsulatus* porin (Figure 4), but there are two pairs of strands that are fused in our prediction. For three best known porins 7 β-strands are underpredicted, 4 overpredicted and 7 pairs of strands are predicted fused.

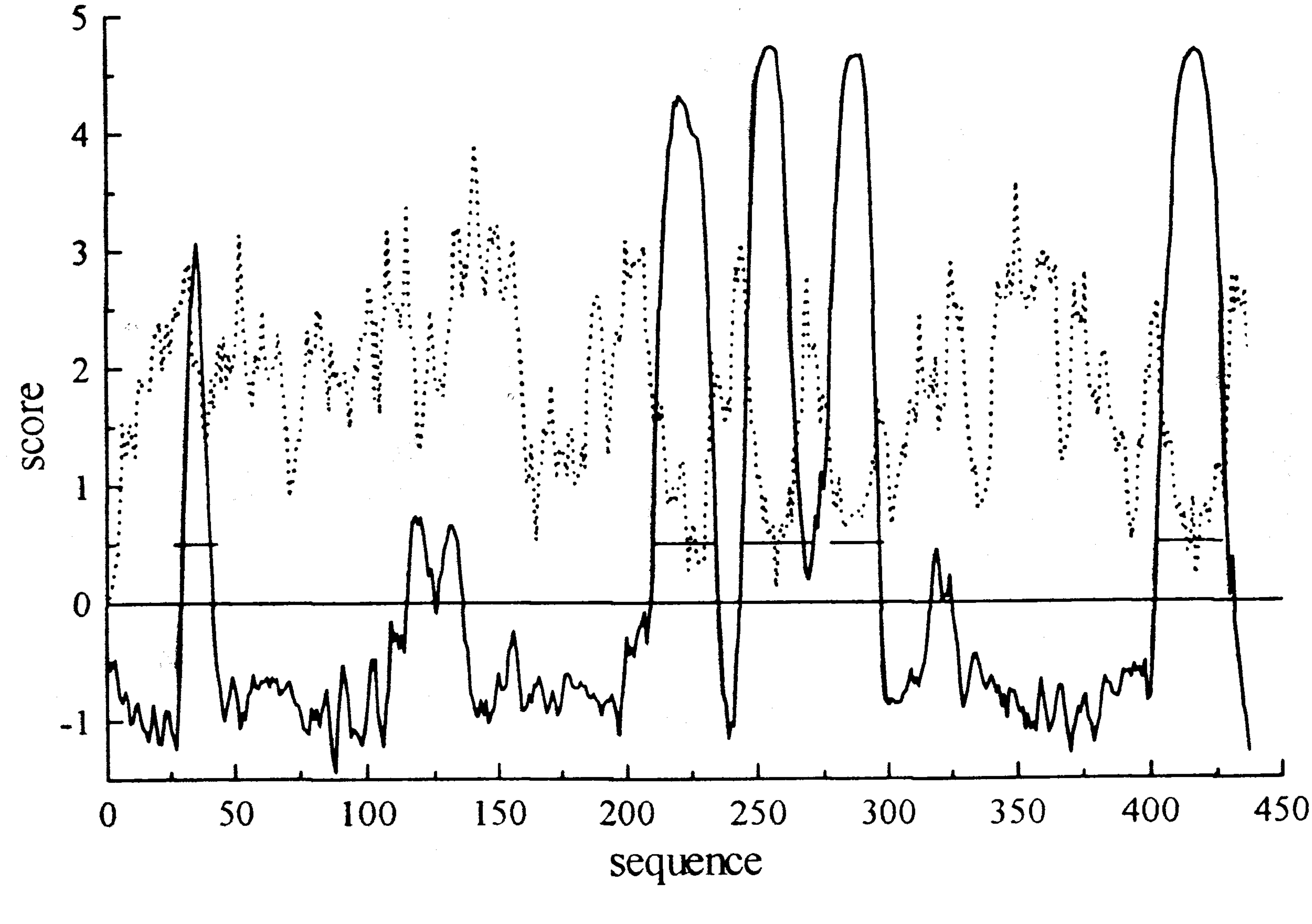

Figure 5: Nicotinic acetylcholine receptor ach1_xenla profiles for finding potential tansmembrane α-helices and β-strands. Same conditions and same notation is used as for the Figure 4. Predicted transmembrane α-helices are shown as bold horizontal lines at the score level 0.5.

Underprediction of transmembrane helices, according to the SWISS-PROT reference standard, was very serious but variable, for proteins belonging to mitochondrial carrier family that are all expected to have six transmembrane helices [81-87]. The digitalization process in the algorithm, that decides whether given segment is in the TMH conformation or not, is the cause of instability in prediction performance for borderline cases. Preference and hydrophobic moment profiles contain considerably more information about the arrangement of potential transmembrane segments. Preference profiles (not shown) for the adt2_yeast, adt1_bovin, adt_neucr, mpcp_rat, ucp_rat, m2om_bovine, and txtp_rat agree that no more than 3 to 5 transmembrane helices can be predicted for each mitochondrial carrier protein and that second expected transmembrane helix can never be predicted by using our method. Accepted topological model for these proteins in the NBRF data base is three hydrophobic transmembrane α-helices for the brown fat uncoupling protein and phosphate carrier [88]. A β-strand that spans the membrane or three β-hairpins have been proposed for the adenine nucleotide translocator [89,90,83]. All mitochondrial carrier proteins have a tripartite structure, with three similar repeats about 100 residues each [82,91]. Our prediction profiles often better exhibit the tripartite symmetry for the profile of potential membrane attached or transmembrane β-strands than for predicted transmembrane α-helices (not shown).

The question of how many TMH segments are in the nicotinic acetylcholine receptor subunits has been going on for a number of years [92]. Earlier reviews [93] supported the four-TMH model. The possible existence of a scaffold of membrane associated β-strands supporting smaller number of transmembrane helices (may be only one) has been raised after low resolution electron microscopy studies [29]. One recent review [30] concludes that of four proposed TMH: M1, M2, M3 and M4 only M2 and M4 are the TMH while M1 and M3 most probably form β-structures. M2, M3 and M4 are α-helical according to Blanton and Cohen [92]. Our TMH predictor strongly predicts all of M1 to M4 segments as TMH segments in the ach1_xenla (Figure 5). High potential (dotted line) for the formation of membrane-buried β-strands is found in sequence domains 101-116, 138-159 and 341-364. Predicted percentage of α-helix transmembrane configuration (24%) is less than 34% [94] or 44.5% [95] suggested by circular dichroism experiments for the whole protein α-helix conformation, but similar to 25% suggested recently by hydrogen/deuterium exchange experiments [96]. Observed percentage of β-sheet residues (29% reported by Moore *et al.* [94], 34% reported by Chang *et al.*, [97]) is higher than predicted 96 residues (22%) in the potential membrane-embedded β-sheet conformation by our TMBS predictor. Observed percentage of β structures in the transmembrane domains alone (40% if β-turns are included according to Görne-Tschelnokow *et al.* [98]) is probably enough for the formation of six membrane buried β-strands in the presence of four transmembrane helices. Potential transmembrane sequence segments of α and β type, that are predicted by our algorithm, must be able to form novel combination of transmembrane regular structure.

Membrane import machinery protein mas6_yeast was predicted with a maximum of only two short transmembrane helices 101-116 and 201-215, instead of four expected, but with many potential amphipathic β-strands. High peak in β-amphipathicity just next to the LDL or IDI motif is found at the mas6_yeast residue # 69, mpcp_rat (from mitochondrial carrier family) residue # 83 and ach1_xenla residue # 350. Observed amphipathic β-structure of a leucine rich repeat peptide LRP32 also contains LDL motif [99]. This motif may be important in protein-lipid interactions because the peptide LRP32 integrates into lipid bilayers, probably as oligomer, forming amphiphatic β-sheet and promoting ion conductances.

The tonb_ecoli protein may be the molecular machine which transduces protonmotive force into mechanical energy [100]. Its proposed transmembrane topology with two potential TMH and three potential transmembrane β-strands [100] is key to the understanding how it connects inner bacterial membrane to outer membrane receptor proteins. We predict only one TMH for residues 13 to 32, probably anchored in the inner bacterial membrane, and several β-strand segments mostly close to the C-terminal, which can interact with outer membrane due to unusually long rigid and highly charged domain which connects these two domains. Our proposed structure for TonB is similar to proposed structure for the TolA protein [101] from *Escherichia coli*, which is also thought to connect inner and outer membrane. Very long connecting domain II of TolA (residues 48 to 310) has been modelled as an α-helical tether. Our prediction of transmembrane segment 14-33 in the TolA agrees with expected span 14-35 [101]. No TMH is predicted by us in the domain II region. This domain is associated with high preference for extramembrane α-helical conformation, but with very low preference for our transmembrane 'H' conformation.

4. DISCUSSION

The observation that conformational preferences are specified by the contexts - local segment primary structure, amino acid attributes, the three-dimensional environment in protein and environmental media, has been discussed before [102-105]. Algorithms that do take into account context-dependence of preferences [106] generally perform better for secondary structure prediction. In this report simple mathematical representation of context dependence is obtained through preference functions that are analytical functions of the surrounding sequence hydrophobicity or of any other amino acid attribute. Furthermore, preference functions are used to predict secondary structure motifs. It has turned out that for integral membrane proteins preference functions are excellent predictors of transmembrane segments in helical conformation. In fact preference functions are much better predictors than the hydrophobicity scale chosen to extract these functions.

A case in point is the application of the Kyte-Doolittle hydrophobicity scale directly and indirectly through preference functions. For the best known membrane proteins direct application of the Kyte-Doolittle algorithm and of its improved versions [107] is inferior to the performance of our algorithm that was also used with Kyte-Doolittle hydrophobicity scale. For instance, helix B is not predicted as hydrophobic helix in subunits M and L of the photosynthetic reaction center but only as an amphiphilic membrane-spanning helix [107]. Helix F from bacteriorhodopsin could not be predicted even after change in the window size, but again only as an amphiphilic helix [107]. We did not use the hydrophobic moment calculations for predicting transmembrane helices, but only as a help in predicting potential membrane buried β-structures. We predict all of 11 transmembrane helices from subunits L, M, and H from both bacterial sources (*Rhodopseudomonas viridis* and *Rhodobacter sphaeroides*) without overpredicting membrane-spanning helices as happens when hydrophobic moment analysis is used in the predictor [18]. In 10 integral membrane proteins of known structure all observed transmembrane helices are predicted by us at their correct sequence location and none of nine long extramembrane helices are confused with transmembrane helices.

Transmembrane helical segments are predicted by us with a high accuracy in 168 integral membrane proteins. All of 168 tested membrane proteins are recognized as such, because at least one transmembrane segment is predicted in each protein. No TMH is predicted in porins.

There are several reasons why preference functions, based on a chosen hydrophobicity scale, are better predictors of transmembrane segments than that hydrophobicity scale. Helix formation in a suitable environment is an cooperative process when nearby residues in a sequence are not independent. In other words the preference for helix conformation of each residue strongly depends on hydrophobicities of its sequence neighbors (Figure 1). The sigmoidal shape of preference function dependence on average hydrophobicity, such as shown in the Figure 1, is found for all amino acid types (not shown). It is suggestive of an cooperative nonlinear process. This cooperative effect is most pronounced for transmembrane segments of integral membrane proteins.

For bitopic membrane proteins, having only one transmembrane segment, local sequence information should be enough to predict the sequence location of such a segment. The prediction accuracy of 97.5% reported for such segments (our result) is impressive only in the case when there is very little overprediction. In the case of bitopic membrane proteins we have found two different traninig procedures for extracting preference functions that result in high prediction accuracy without large overprediction. Obviously, such training procedures cannot be included in algorithms that use the same hydrophobicity scale, but do not use preference functions. One possible answer to the initial question is that preference functions are so much better than simple use of hydrophobicity scale, because preference functions are firmly connected with protein data base used for training and with secondary structure features present or expected in that data base. Therefore, another important advantage of preference functions is the possibility to enhance amino acid attributes or secondary structure preferences through training process that ends with extraction of preference functions. In our recent work [52] we demonstrated that enhancement of the Chou-Fasman type constant preferences for transmembrane configuration, leads also to high prediction accuracy for transmembrane segments, even if prediction model (two state model) and training procedure (without soluble proteins) was completely different. Evidence that transmembrane helices are autonomous folding domains [108] helps to clarify why many different methods of sequence analysis are good predictors of transmembrane segments that are potential TMH segments.

Inability to distinguish an α-helix from β-strand transmembrane structure is even more serious weakness of hydrophobicity analysis. To build any reasonable topological model for membrane protein we must know what is the secondary structure of its transmembrane segments. Such information cannot be the output of any other algorithm that uses hydrophobicity scale, without additional training that attempts to correlate amino acid attributes with conformational motifs in proteins of known structure. Residues known to prefer β-strand conformation in soluble proteins [109], are very frequent residues in known transmembrane segments [62,110]. It is possible that some membrane proteins with transmembrane helices have had predominantly β-structure before being incorporated in the membrane [111]. When algorithms, trained on soluble proteins, attempt to predict secondary structure of membrane proteins, transmembrane segments known to be helical are often broken or predicted as β-strands. Therefore, the training process that includes membrane proteins of known or partially known structure is absolutely essential for the recognition of transmembrane structural motifs.

The need for more extensive training procedure was recognized by neural network programmers, but they trained their algorithms only too well. Overtraining is more subtle, but equally serious problem, that can greatly diminish prediction usefulness. A case in point is Rost *et al.* neural network algorithm [9] whose performance is significantly decreased when tested on never-before-seen set of proteins (Results section). Since we did not use evolutionary information (alignments of similar proteins) our choice of 105 integral membrane proteins was unintentionally such that average number of possible homologues per one protein (as average weighted number of alignments that do take into account sequence lengths) was smaller in that group of proteins (14 per one protein) than in the set of 63 proteins selected by Rost *et al.* (23 per one protein) [9]. This would partly explain the decreased performance when 105 membrane proteins are tested with Rost *et al.* method whose accuracy depends on available evolutionary information [9].

There are several reasons why overtraining may have happened during Rost *et al.* procedure despite careful cross-validation procedure [9]. Firstly, the pairwise homology among chosen proteins was not always less than 30%, as documented before (Methods). In the original set of 69 membrane proteins used by Rost *et al.* there was a subset of 47 proteins that had less than 30% pairwise similarity with all other proteins from that subset and had on average only 13 homologues per each protein [9]. The prediction accuracy, as measured for that subset of proteins with the A_{TM} parameter (Methods), was only $A_{TM} = 0.665$ as compared with $A_{TM} = 0.736$ for all of 69 proteins. The remaining subset of 22 proteins (mainly opsins) with more than 30% pairwise similarity and with an average of 32 homologues per protein was predicted with much higher prediction accuracy of $A_{TM} = 0.814$. For membrane proteins, considerably less than 30% similarity in the sequence may be needed, when we want to exclude very similar folding motifs. Failure to exclude similar proteins will cause an artificial increase in prediction accuracy in the case when similar proteins are predicted with higher than average accuracy, no matter what prediction method is used. Secondly, multiple alignment procedure, as a part of the training and testing process, was specific for the chosen protein data base of 69 proteins [9]. It increased prediction accuracy for that data base, but it does not have to do so for a set of nonhomologous never-before-seen proteins that for instance are not associated with similarly large average number of homologues per each protein from that data base. Thirdly in the data set of only 69 membrane proteins the number of objects determining prediction accuracy is really quite small: not more than 20 to 30 transmembrane helices that are difficult to predict by using any prediction method. The prediction accuracy becomes quite high when such specific patterns are learned, either through direct training procedure or through the choice of filter parameters. Unfortunately, neural network parameters learned in the process become very specific for such patterns that may not repeat easily in proteins outside training data set. A known disadvantage of neural network algorithm is its inability to tell us what it learned, in this case how it become capable of correct prediction of transmembrane helices most difficult to predict.

Signal sequences are, as a rule, not predicted as transmembrane segments by the neural network algorithm [9]. In our data base of 168 integral membrane proteins there are 32 proteins with signal sequences at the N-terminal (labeled with letter 's', Methods). Rost *et al.* wrongly predict only 3 such proteins as having the transmembrane segment at the sequence location of known signal sequence (cyoa_ecoli, myp0_human and wapa_strmu) [9]. We predict all of 32 signal sequences except two as transmembrane helices. Overprediction happens because very high preference for transmembrane helix conformation is often associated with signal sequences. Somewhat shorter length of signal sequences does not help, because many correct predictions of transmembrane helices are initially associated with predicted short segments (12 to 16 residues that have high preference for transmembrane helix). Filter modification with negative weight at protein N-terminal can easily eliminate most of false positive predictions of TMS at the location of known signal sequences [52]. We did not use such modifications in this work because it would lead to difficult to detect underpredictions of real TMH at the N-terminal, while overprediction of TMH is easily detected when it happens at the location of known signal sequence. One advantage of omitting filter modifications with respect to signal sequences is that potential signal sequences are predicted as TMH with the same high accuracy as all other TMH, but then additional information from experiments is needed to decide if potential TMH near N-terminal is indeed true TMH or signal sequence. Another advantage is that primary structures without transit polypeptide, or without N-terminal signal sequence next to first potential transmembrane segment can be tested with assurance that first TMH will not be underpredicted due to omission of the N-terminal segment. Underprediction of the whole first TMH, containing 35 amino acid residues, happens in the LHC-II sequence taken from the Nature article [13] or in the cb22_pea sequence without transit polypeptide, when Rost *et al.* method [9], optimized to eliminate signal sequences from consideration, is presented with such truncated versions of polypeptides.

Errors in the SWISS-PROT assignment of transmembrane segments will reduce the prediction performance for all prediction methods that use this data base as 'standard of truth'. Such errors can indeed happen. We discussed the case of tolq_ecoli protein from *Escherichia coli*, which has only one transmembrane segment according to SWISS-PROT version 29 assignment, but is strongly predicted by us with three transmembrane segments in helical conformation. The same topology of three transmembrane helices is currently accepted in the SWISS-PROT data base for very similar exbb_ecoli protein.

Fortunately, many different theoretical and experimental procedures were used in SWISS-PROT assignments for the proteins finally chosen by us, so that for the purpose of our weak training procedure this set of proteins can be considered as reference set, but probably not as the 'standard of truth'. Observed and predicted length distribution of transmembrane segments in protein data base (Figure 2) may indicate that considerable room is still left for improving the algorithm. However, average length of expected transmembrane segments in our test set of 168 membrane proteins (21.7 residues) is quite close to predicted average length (21.5 residues). In any event, the absence of length distribution for predicted transmembrane segments that is in-built in some of simpler algorithms using hydrophobicity scales is quite unrealistic.

The TMH predictor underpredicts some of expected transmembrane segments in voltage-gated channels [112] (cic1_cypca case was mentioned in the section 3.6). Closer analysis revealed that underpredicted TMS are highly charged S4 segments known to span the membrane with less than 10 residues [113]. Although often missed by the TMH predictor essential parts of channel machinery, such as S4 and P-segments of the *Shaker* potassium channel pore [114, 115], are clearly resolved by our preference profiles (in preparation).

The main goal of this work was accurate prediction of transmembrane helical structures, but we do realize that membrane proteins may exist that have both α-helices and β-strands as transmembrane structure. Preference function method is capable of predicting separately α-helical and β-strand conformation of segments that have potential to become membrane buried. Known structures of β-class soluble proteins are used in order to extract β-sheet preferences and as a help in extracting turn preferences. The reason why we had to enlarge the data base of membrane proteins with soluble proteins of the β-class is very simple. Few porins of known structure were not enough to serve as the training data base for the extraction of β-strand preference functions. Therefore, as the best substitute we used soluble proteins of the β-class. It is not an disadvantage to use much more abundant information available in soluble proteins of known structure. The number of nonhomologous proteins used to train preference functions for one secondary structural motif, can serve as the rough estimate of what is the minimal number of proteins that must be used during training procedure by our method (30 to 40 integral membrane proteins and the same number of soluble proteins of the β-class).

We have used a very simple procedure to predict transmembrane β-strands in porins. As observed before [107,116] it is useful to take into account hydrophobic moment for assumed β structure when the goal is to predict such a structure. The standard training and testing procedure with the Kyte-Doolittle scale gives reasonably good results with porins and defensins in terms of predicting transmembrane β-strands, but overprediction of membrane β-structure happens in the photosynthetic reaction center subunits in the case when decision constants are fixed to zero values (not shown). Preliminary results with a choice of the Cid *et al.* [69] hydrophobicity scale are encouraging both in terms of increased accuracy in predicting TMBS and in terms of a low percentage of wrongly predicted TMH in soluble proteins (only 4 to 5% for our data sets of soluble proteins). At any rate, the prediction of β-strands, turn and undefined conformations as well as the calculation of hydrophobic moment profile for assumed α-helix and β-strand conformation helped to locate transmembrane helices and other potential membrane-embedded regular structures.

One application of our standard training and testing procedure is for the nicotinic acetylcholine receptor, where M1, M2, M3 and M4 segments are all strongly predicted as transmembrane helices, but in addition there are several sequence domains with a potential for membrane-embedded β-strands (Figure 5). Another application has been described in the case of mitochondrial carrier family proteins. In many proteins from this family that have a known tripartite structure we have seen such a structure revealed in great details through profile of summed β-moments and β-sheet preferences (not shown). Contrary to the proposed six-helix model for these proteins thought to be required to take account of the threefold repeat [82,83] tripartite symmetry does not require the presence of two transmembrane helices in each of three domains. A small change in the primary structure or even in polypeptide environment may be enough to transform one regular structure into another in one of three domains without significant change in the tripartite symmetry. Functional asymmetry of three domains is known to exist in these proteins and some experimental evidence already exists that movement of loops in and out of the membrane can regulate transport activity of the mitochondrial ADP/ATP carrier [117].

Our algorithm can give partial answer to the question what attributes are optimal predictors for specific folding motifs. Kyte-Doolittle type hydropathy values and Chou-Fasman type conformational preferences are two obvious answers to the question what amino acid attributes are good predictors for majority of transmembrane helices. Indeed, three such scales MODKD, KYTDO and CPREF (Table 4), are on the very top of the list of the best amino acid scales (Table 5). Performance parameters that punish overprediction (A_{TM} and Q_p) give advantage to hydropathy values. Modifications to the Kyte-Doolittle values in the MODKD scale increase prediction accuracy by increasing Trp and decreasing Ala importance for the formation of TMH. Surrounding hydrophobicity scale for membrane proteins (PONG1) takes into account actual hydrophobic environment in the three-dimensional protein structures. It produces less of false-positive TMH predictions when tested through preference functions on soluble proteins (Table 8). It appears that this scale can be used when an alternative to Kyte-Doolittle scale [17] is sought, because very popular Engelman *et al.* scale [59] is associated with up to 25% of false-positive TMH predictions (Table 8). Optimal scale for identification of amphipathic helices (PRIFT) is obviously not optimal for the recognition of TMH. Solution hydrophobicity scales such as FAUPL are clearly inferior to protein derived scales such as PONG1, CHOTH or ROSEF. A good performance of scales that measure water-accessible surface area loss upon protein folding has been noticed before [32]. More interesting are relatively high A_{TM} scores for polarity scales GRANT and WOESE and for the antigenic determinant hydrophilicity scale HOPPW. It would be quite interesting, but outside the scope of this work, to see if some transmembrane helices, difficult to predict by hydrophobicity analysis, are well predicted by polarity or hydrophilicity attributes. Such job can be easily done by using the PREF suite of algorithms, version 3.0. Even scales with inferior performance, such as the Cid *et al.* scale [69], are potentially very useful when different folding motifs in the membrane are being sought: transmembrane β-strands instead of TMH.

The filter parameters of our optimal predictor for transmembrane helices were optimized by using the Kyte-Doolittle scale and a reference set of 63 integral membrane proteins having one or more of long transmembrane segments, for which experimental and theoretical analysis indicated an α-helix configuration. Optimization of parameters was done by trial and error procedure and certainly was not perfect. Automatic procedures for finding optimal parameters for the TMH predictor were recently developed within the framework of preference functions method [52]. We did not use such procedures due to their inherent shortcomings: the danger of overtraining the predictor is then increased and due to the size of optimization problem different order of parameter optimization can lead to different results. In any case, it is quite possible that some other scale of amino acid attributes could have been chosen initially in the optimization process to produce higher prediction accuracy than the KYTDO scale. The natural choice of scale associated with a chosen reference set of proteins is the scale of statistical preferences, such as the CPREF scale, that can be extracted from that data base of proteins.

To summarize, the practical advantages of using the PREF suite of algorithms are as follows:

- It is much less expensive in computer time than a neural network algorithm.
- It works with equal expected high accuracy in the case when very few or no homologues are known.
- It has the potential to identify those physical, chemical or protein-derived statistical properties that are the most important for segment folding into the TMH configuration.
- Well known Kyte-Doolittle scale [17] can be used throughout, except in the case when specific need exists to test other amino acid attributes.
- All stages of prediction process are associated with transparent rules that are objective, automatic and easily inspected.
- There is an automatic recognition of different folding types of integral membrane proteins and automatic choice of decision constants for each type which improves the prediction accuracy.

- Thirty to forty membrane proteins and same number of soluble proteins of known structure are sufficient to train the algorithm.
- Accurate prediction of transmembrane helical segments is superimposed on the prediction of all other secondary structure elements of interest.
- Peaks in the transmembrane helical preference of lesser height and width can be used for identification of primary structure segments of special interest such as signal sequences and pore-forming segments (in preparation).
- Membrane-embedded or surface-attached β-strands can also be recognized from the sum of prediction profiles for β-strand preferences and of hydrophobic moments for assumed β-strand conformation.

The negative aspects or disadvantages are as follows:

- Balanced training procedure is needed. Including many more extramembrane than transmembrane residues in the training data set is wrong not only because of unbalanced training procedure, but also because we know that undefined conformation is forced upon us for extramembrane residues due to our lack of knowledge.
- A high percentage of soluble proteins are falsely recognized as membrane proteins (from 12 to 17%).
- Only one conformation is predicted with high accuracy: transmembrane helix conformation. Predictions of other regular or irregular conformations are not associated with the same high accuracy.
- The monotopic membrane proteins [118], that cross only one bilayer but not two, such as the prostaglandin H2 synthase [119], and self-inserting membrane proteins or toxins [120], such as colicin A [121], diphtheria toxin [122], beetle δ-endotoxin [123] and annexin [124] are associated with poor prediction (not shown).

Several improvements to the proposed method can be envisaged.

a) Multiple alignment was not used. It should improve prediction accuracy for a single tested protein when thirty to forty homologous proteins exist. As already shown before [125], the PREF method can use training data set of proteins specific for protein to be tested.

b) The prevalence of positively charged residues in the interior loops [60,79] or 'positive inside rule' is shown to improve prediction accuracy of our algorithm [52], but was not used in the present work. The predictor can become informative about the direction of membrane crossing, especially in the case of plasma membrane proteins of bacterial origin, when 'positive inside rule' is taken into account.

c) It is not known if mixed type α/β or α+β structure can exist as transmembrane structure and if so what combinations of α-helix segments and β-strand segments may join to form transmembrane structure. Extracting preference functions from large enough data base of porins and related proteins with β-strand transmembrane structure will soon be possible. Then, appropriate modification of PREF-SPLIT algorithm, along lines suggested in this report, will serve to predict sequence location of both transmembrane α-helices and transmembrane β-strands.

Availability of the prediction with preference functions. We have set up an automatic electronic mail server at the Internet address: predict@drava.etfos.hr. The server will return complete prediction results, such as given in Table 6, when provided with the sequence of your protein. For further information, send the word *help* to the server. Questions, comments and suggestions should be sent to juretic@mapmf.pmfst.hr or zucic@mia.os.carnet.hr.

ACKNOWLEDGEMENTS

We are grateful to Sandor Pongor from ICGEB, Trieste, Italy and to Burkhard Rost from EMBL, Heidelberg, Germany for data bases of membrane and soluble proteins kindly provided for our use. Thanks are due to Vera Gamulin and Boris Lenhard from Rugjer Bošković Institute in Zagreb, Croatia, who helped us with Eisenberg and Rao and Argos analysis [20] that was carried out with the PCGENE software. This work was supported by the Croatian Ministry of Science and Technology grants 1-03-171 to D.J. and 1-07-159 to B.L. and N.T.

SUPPLEMENTARY MATERIAL AVAILABLE via INTERNET

Two data bases of soluble proteins of known structure used to find false positive prediction results (Table I and Table II). Gaussian parameters needed for evaluation of preference functions based on the Kyte-Doolittle hydropathy scale [17] (Table III). Table with detailed prediction results for transmembrane helices in 168 integral membrane proteins (Table IV). Table with a detailed comparison of prediction results for 10 best known membrane proteins for our and three other algorithms (Table V). All these tables together with the FORTRAN 77 source code are available from the anonymous ftp server mia.os.carnet.hr in the /pub/pssp directory. The anonymous login is ftp and the e-mail address is accepted as password. The list of files with short descriptions is contained in the 00index.txt file.

REFERENCES

1. F. Eisenhaber, B. Persson and P. Argos, Crit. Rev. Biochem. Mol. Biol., 30 (1995) 1.
2. P.Y. Chou and G.D. Fasman, Biochemistry, 13 (1974) 211.
3. J. Garnier, D.J. Osguthorpe and B. Robson, J. Mol. Biol., 120 (1978) 97.
4. B.A. Wallace, M. Cascio and D.L. Mielke, Proc. Natl. Acad. Sci. U.S.A., 83 (1986) 9423.
5. N. Qian and T.J. Sejnowski, J. Mol. Biol., 202 (1988) 865.
6. D.G. Kneller, F.E. Cohen and R. Langridge, J. Mol. Biol., 214 (1990) 171.
7. B. Rost and C. Sander, J. Mol. Biol., 232 (1993) 584.
8. R. Lohmann, G. Schneider, D. Behrens and P. Wrede, Protein Sci., 3 (1994) 1597.
9. B. Rost, R. Casadio, P Fariselli and C. Sander, Protein Sci., 4 (1995) 521.
10. M.S. Weiss, A. Kreusch, E. Schiltz, U. Nestel, W. Welte, J. Weckesser and G.E. Schulz, FEBS Lett., 280 (1991) 379.
11. S.W. Cowan, T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R.A. Pauptit, J.N. Jansonius and J.P. Rosenbusch, Nature, 358 (1992) 727.
12. J. Deisenhofer, O. Epp, K. Miki, R. Huber and H. Michel, Nature, 318 (1985) 618.
13. W. Kühlbrandt, D.N. Wang and Y. Fujiyoshi, Nature, 367 (1994) 614.
14. S. Iwata, C. Ostermeier, B. Ludwig and H. Michel, Nature, 376 (1995) 660.
15. T. Tsukihara, H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono and S. Yoshikawa, Science, 272 (1996) 1136.
16. A. Bairoch and B. Boeckmann, Nucl. Acids Res., 22 (1994) 3578.

17. J. Kyte and R.F. Doolittle, J. Mol. Biol., 157 (1982) 105.
18. D. Eisenberg, E. Schwarz, M. Komaromy and R. Wall, J. Mol. Biol., 179 (1984) 125.
19. P. Klein, M. Kanehisa and C. DeLisi, Biochim. Biophys. Acta, 815 (1985) 468.
20. J. Rao and P. Argos, Biochim. Biophys. Acta., 869 (1986) 197.
21. J.A. Bangham, Anal. Biochem., 174 (1988) 142.
22. J. Edelman, J. Mol. Biol., 232 (1993) 165.
23. D.M. Engelman and T.A. Steitz, Cell, 23 (1981) 411.
24. S.H. White, Annu. Rev. Biophys. Biomol. Struct., 23 (1994) 407.
25. J.P. Allen, G. Feher, T.O. Yeates, H. Komiya and D.C. Rees, Proc. Natl. Acad. Sci. U.S.A., 84 (1987) 6162.
26. R. Henderson, J.M. Baldwin, T.A. Ceska, F. Zemlin, E. Beckmann and K.H. Downing, J. Mol. Biol., 213 (1990) 899.
27. G. McDermott, S.M. Prince, A.A. Freer, A.M. Hawthornthwaite-Lawless, M.Z. Papiz, R.J. Cogdell and N.W. Isaacs, Nature, 374 (1995) 517.
28. M.S. Weiss and G.E. Schulz, J. Mol. Biol., 227 (1992) 493.
29. U. Unwin, J. Mol. Biol., 229 (1993) 1101.
30. F. Hucho, U. Görne-Tschelnokow and A. Strecker, Trends Biochem. Sci., 19 (1994) 383.
31. S.W. Cowan and J.P. Rosenbusch, Science, 264 (1994) 914.
32. D. Juretić, B.K. Lee, N. Trinajstić and R.W. Williams, Biopolymers, 33 (1993) 255.
33. D.T Jones, W.R. Taylor and J.M. Thornton, Biochemistry, 33 (1994) 3038.
34. C. Sander and R. Schneider, Nucl. Acids Res., 22 (1994) 3597.
35. W. Kabsch and C. Sander, Biopolymers, 22 (1983) 2577.
36. J. Deisenhofer and H. Michel, Science, 245 (1989) 1463.
37. D.R. Madden, J.C. Gorga, J.L. Strominger and D.C. Wiley, Cell, 70 (1992) 1035.
38. M.A. Saper, P.J. Bjorkman and D.C. Wiley, J. Mol. Biol., 219 (1991) 277.
39. B.K. Jap, J. Mol. Biol., 205 (1989) 407.
40. S. Gerbl-Rieger, H. Engelhardt, J. Peters, M. Kehl, F. Lottspeich and W. Baumeister, J. Struct. Biol., 108 (1992) 14.
41. F. Jähnig, in Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., ed.) pp 707-717, Plenum Press, New York, NY, 1989.
42. G. Ried, R. Koebnik, I. Hindennach, B. Mutschler and U. Henning, Mol. Gen. Genet., 243 (1994) 127.
43. V. De Pinto and F. Palmieri, J. Bioenerg. Biomembr., 24 (1992) 21.
44. C.A. Mannella, M. Forte and M. Colombini, J. Bioenerg. Biomembr., 24 (1992) 7.
45. C.P. Hill, J. Yee, M.E. Selsted and D. Eisenberg, Science, 251 (1991) 1481.
46. P. Bulet, S. Cociancich, M. Reuland, F. Sauber, R. Bischoff, G. Hegy, A. Van Dorsselaer, C. Hetru and J.A. Hoffmann, Eur. J. Biochem., 209 (1992) 977.
47. P.K. Ponnuswamy and M.M. Gromiha, Int. J. Peptide Protein Res., 42 (1993) 326.
48. D. Eisenberg, R.M. Weis and T.C. Terwillinger, Proc. Natl. Acad. Sci. U.S.A., 81 (1984) 140.
49. A. Lupas, M. Van Dyke and J. Stock, Science, 252 (1991) 1162.
50. J.L. Cornette, K.B. Cease, H. Margalit, J.L. Spouge, J.A. Berzofsky and C. DeLisi, J. Mol. Biol., 195 (1987) 659.
51. O.B. Ptitsyn, J. Mol. Biol., 42 (1969) 501.

52. B. Lučić, N. Trinajstić and D. Juretić, in From Chemical Topology to Three-Dimensional Geometry (A.T. Balaban, ed.) pp 117-158, Plenum Press, New York, NY, 1997.
53. H.F. Lodish, Trends Biochem. Sci., 13 (1988) 332.
54. M. Degli Esposti, M. Crimi and G. Venturoli, Eur. J. Biochem., 190 (1990) 207.
55. B. Persson and P. Argos, J. Mol. Biol., 237 (1994) 182.
56. G. von Heijne and C. Blomberg, Eur. J. Biochem., 97 (1979) 175.
57. C. Chothia, J. Mol. Biol., 105 (1976) 1.
58. G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, Science, 229 (1985) 834.
59. D.M. Engelman, T.A. Steitz and A. Goldman, Annu. Rev. Biophys. Biophys. Chem., 15 (1986) 321.
60. G. von Heijne, J. Mol. Biol., 225 (1992) 487.
61. R. Grantham, Science, 185 (1974) 862.
62. C.M. Deber, C.J. Brandl, R.B. Deber, L.C. Hsu and X.K. Young, Arch. Biochem. Biophys., 251 (1986) 68.
63. H.R. Guy, Biophys. J., 47 (1985) 61.
64. C.R. Woese, D.H. Dugre, S.A. Dugre, M. Kondo and W.C. Saxinger, Cold Spring Harbor Symp. Quant. Biol., 31 (1966) 723.
65. P.K. Ponnuswamy, M. Prabhakaran and P. Manavalan, Biochim. Biophys. Acta, 623 (1980) 301.
66. W.R. Krigbaum and A. Komoriya, Biochim. Biophys. Acta, 576 (1979) 204.
67. T:P. Hopp and K.R. Woods, Proc. Natl. Acad. Sci. U.S.A., 78 (1981) 3824.
68. J. Janin, Nature, 277 (1979) 491.
69. H. Cid, M. Bunster, M. Canales and F. Gazitua, Protein Eng., 5 (1992) 373.
70. S. Miyazava and R.J. Jernigan, Macromolecules, 18 (1985) 534.
71. J. Garnier and B. Robson, in Prediction of Protein Structure and the Principles of Protein Conformation (G.D. Fasman, ed.) pp 417-465, Plenum Press, New York, NY, 1989.
72. J.-L. Fauchere and V. Pliška, Eur. J. Med. Chem. - Chim. Ther., 18 (1983) 369.
73. M.A. Roseman, J. Mol. Biol., 200 (1988) 513.
74. G. Casari and M. Sippl, J. Mol. Biol., 224 (1992) 725.
75. L.K. Poulsen, A. Refn, S Molin and P. Andersson, Mol. Microbiol., 5 (1991) 1627.
76. D.N. Wang, V.E. Sarabia, R.A.F. Reithmeier and W. Kühlbrandt, EMBO J., 13 (1994) 3230.
77. R.R. Kopito and H.F. Lodish, Nature, 316 (1985) 234.
78. G.D. Fasman and W.A. Gilbert, Trends Biochem. Sci., 15 (1990) 89.
79. G. von Heijne, EMBO J., 5 (1986) 3021.
80. T.M. Savarese and C.M. Fraser, Biochem. J., 283 (1992) 1.
81. M. Klingenberg, Trends Biochem. Sci., 15 (1990) 108.
82. J.E. Walker, Curr. Opin. Struct. Biol., 2 (1992) 519.
83. M. Klingenberg, J. Bioenerg. Biomembr., 25 (1993) 447.
84. M. Klingenberg, Arch. Biochem. Biophys., 270 (1993) 1.
85. D.R. Nelson, J.E. Lawson, M. Klingenberg and M.G. Douglas, J. Mol. Biol., 230 (1993) 1159.
86. B. Jank, B. Habermann, R.J. Schweyen and T.A. Link, Trends Biochem. Sci., 18 (1993) 427.
87. F. Palmieri, FEBS Lett. 346 (1994) 48.

88. J.-L. Popot and C. de Vitry, Annu. Rev. Biophys. Biophys. Chem., 19 (1990) 369.
89. W. Bogner, H. Aquila and M. Klingenberg, Eur. J. Biochem., 161 (1986) 611.
90. H. Aquila, T.A. Link and M. Klingenberg, FEBS Lett., 212 (1987) 1.
91. G. Brandolin, A. Le Saux, V. Trezeguet, G.J.M. Lauquin and P.V. Vignais, J. Bioenerg. Biomembr., 25 (1993) 459.
92. M.P. Blanton and J.B. Cohen, Biochemistry, 33 (1994) 2859.
93. B. Traxler, D. Boyd and J. Beckwith, J. Membr. Biol., 132 (1993) 1.
94. W.M. Moore, L.A. Holladay, D. Puett and R.N. Brady, FEBS Lett 45 (1974) 145.
95. G.D. Fasman, Biopolymers, 37 (1995) 339.
96. J.E. Baenziger and N. Méthot, J. Biol. Chem., 270 (1995) 29129.
97. E.L. Chang, P. Yager, R.W. Williams and A.W. Dalziel, Biophys. J., 41 (1983) 65a.
98. U. Görne-Tschelnokow, A. Strecker, C. Kaduk, D. Naumann and F. Hucho, EMBO J., 13 (1994) 338.
99. D.D. Krantz, R. Zidovetzki, B.L. Kagan and S.L. Zipursky, J. Biol. Chem., 266 (1991) 16801.
100. P.E Klebba, J.M. Rutz, J. Liu and C.K. Murphy, J. Bioenerg. Biomembr., 25 (1993) 603.
101. S.K. Levengood, W.F. Beyer and R.E. Webster, Proc. Natl. Acad. Sci. U.S.A., 88 (1991) 5939.
102. S.-C. Li and C.M. Deber, Int. J. Peptide Protein Res., 40 (1992) 243.
103. G.E. Arnold, A.K. Dunker, S.J. Johns and R.J. Douthart, Proteins, 12 (1992) 382.
104. L. Zhong and W.C.Jr. Johnson, Proc. Natl. Acad. Sci. U.S.A., 89 (1992) 4462.
105. H. Wako and T.L. Blundell, J. Mol. Biol., 238 (1994) 693.
106. J.-F. Gibrat, J. Garnier and B. Robson, J. Mol. Biol., 198 (1987) 425.
107. F. Jähnig, Trends Biochem. Sci., 15 (1990) 93.
108. J.-L. Popot, Curr. Opin. Struct. Biol., 3 (1993) 532.
109. P.Y. Chou and G.D. Fasman, Advan. Enzymol., 47 (1978) 45.
110. C.M. Deber, A.R. Khan, Z. Li, C. Joensson and M. Glibowicka, Proc. Natl. Acad. Sci. U.S.A., 90 (1993) 11648.
111. L.L. Randall and S.J.S. Hardy, Science, 243 (1989) 1156.
112. W. Catterall, Annu. Rev. Biochem., 64 (1995) 493.
113. S.A.N. Goldstein, Neuron, 16 (1996) 717.
114. H.P. Larsson, O.S. Baker, D.S. Dhillon and E.Y. Isacoff, Neuron, 16 (1996) 387.
115. A. Gross and R. MacKinnon, Neuron, 16 (1996) 399.
116. M.M. Gromiha and P.K. Ponnuswamy, Int. J. Peptide Protein Res., 42 (1993) 420.
117. E. Majima, K. Ikawa, M. Takeda, M. Hashimoto, Y. Shinohara and H. Terada, J. Biol. Chem., 270 (1995) 29548.
118. M.L. Jennings, Annu. Rev. Biochem., 58 (1989) 999.
119. D. Picot, P.J. Loll and M. Garavito, Nature, 367 (1994) 243.
120. J. Li, Curr. Opin. Struct. Biol., 2 (1995) 545.
121. M.W. Parker, J.P.M. Postma, F. Pattus, A.D. Tucker and D. Tsernoglou, J. Mol. Biol., 224 (1992) 639.
122. S. Choe, M.J. Bennett, G. Fujii, P.M.G. Curmi, K.A. Kantardjieff, R.J. Collier and D. Eisenberg, Nature, 357 (1992) 216.
123. J. Li, J. Carroll and D.J. Ellar, Nature, 353 (1991) 815.

124. R. Huber, R. Berendes, A. Burger, M. Schneider, A. Karshikov, H. Luecke, J. Romisch and E. Paques, J. Mol. Biol., 223 (1992) 683.

125. D. Juretić, B. Lučić and N. Trinajstić, Croat. Chem. Acta, 66 (1993) 201.