Basic Charge Clusters and Predictions of Membrane Protein Topology

Davor Juretić,*,† Larisa Zoranić,† and Damir Zucić‡

Physics Department, Faculty of Natural Sciences, Mathematics and Education, University of Split, N. Tesle 12, HR-21000, Split, Croatia, Faculty of Electrical Engineering, University of Osijek, Istarska 3, HR-31000 Osijek, Croatia, and School of Medicine, University of Osijek, Huttlerova 4, HR-31000 Osijek, Croatia

Received November 12, 2001

The topology predictor SPLIT 4.0 (http://pref.etfos.hr) predicts the sequence location of transmembrane helices by performing an automatic selection of optimal amino acid attribute and corresponding preference functions. The best topological model is selected by choosing the highest absolute bias parameter that combines the bias in basic charge motifs and the bias in positive residues (the "positive inside rule") with the charge difference across the first transmembrane segment. Basic charge motifs, such as the BBB, BXXBB, and BBXXB motifs in α-helical integral membrane proteins, are significantly more frequent near cytoplasmic membrane surface than expected from the Arg/Lys (B) frequency. The predictor's accuracy is 99% for predicting 178 transmembrane helices in all membrane proteins or subunits of known 3D structure.

INTRODUCTION

Membrane proteins are encoded by 20−30% of all open reading frames found in different organisms.[1] Integral membrane proteins are usually built from bundles of transmembrane (TM) α-helices. The sequence location and transmembrane orientation of membrane-spanning helices is very important information for research workers concerned with the structure and function of such proteins.[2] Sequence analysis can locate potential transmembrane segments, for instance by using their greater average hydrophobicity, and it can also provide the additional information about their transmembrane orientation.[3] The resulting two-dimensional topological model is very useful as a schematic representation of protein arrangement in the membrane. It can be used as the starting point in attempts to predict the three-dimensional structure. The prediction accuracy reported in topology prediction papers[3−8] depends on the method used to select favorites among topological models. One favorite selection criterion is the magnitude of positive charge bias.[3] Namely, in topological models with the best experimental support arginines and lysines are found more often in cytoplasmic loops than in extracytoplasmic regions.[9,10] This observation, named the "positive inside rule", is used to predict cytoplasmic (IN) or extracytoplasmic (OUT) location of protein amino terminus.

When the inside/outside charge bias is different from zero it is an overall property of a sequence with a known or predicted transmembrane segments. In other words, global heuristics, such as the positive-inside bias, cannot be associated with specific sequence location. Other topological determinants have been described in the literature, such as the charge difference across the first transmembrane segment[11] and asymmetry in the amino acid composition in outside and inside loops.[12] These determinants also do not point to specific sequence locations. On the other hand, sequence locations where a small number of amino acids have obligatory cytoplasmic or extracytoplasmic positions are known to exist in integral membrane proteins. The problem with topological determinants specific for sequence location is that they have a high specificity for similar sequences belonging to a particular class of membrane proteins. In general, such determinants are not common and cannot be extended to majority of membrane proteins.

In this paper we describe the clusters of the basic residues predominantly found in cytoplasmic loops as additional topological determinants that are often present in membrane proteins. These determinants are specific for sequence locations near cytoplasmic membrane surface. The first goal in this paper is to prove that high frequency of basic charge motifs cannot be explained away with a prevalence of positively charged residues in the cytoplasmic loops. The bias in cytoplasmic/extracytoplasmic basic charge motifs, together with other topological determinants, is then used to select the optimal topological model for a given integral membrane protein. Models with different sequence location and number of predicted transmembrane helices are created by using an automatic run of different amino acid attributes in the preference functions method.[13,14] We shall show that our predictor is one of the best performing computational methods for topology predictions.

MATERIALS AND METHODS

Protein Data Set. Experimental evidence was stricter for the testing data set of proteins, while nonredundancy was stricter for the training data set. We removed all annotated signal sequence. Three data sets were used for testing. We used 158 membrane proteins collected earlier[5,7,8,15] and 79 membrane protein sequences collected recently as the MPtopo database.[16] The subset of 41 sequences from the MPtopo database are the X-ray determined structures. An additional

* Corresponding author phone: 385-21385133 local 118; e-mail: juretic@pmfst.hr.
† University of Split.
‡ University of Osijek.

10.1021/ci010263s CCC: $22.00
Published on Web 05/10/2002

data set of 11 sequences from membrane proteins of known structure, not included in the subset of 41 sequences, nor in the test set of 158 proteins, was collected from published papers,[17] from the best characterized proteins in the Möller et al. database[18] and from the most recent additions to the MPtopo database.[16] These sequences are (in the Swiss-Prot codes) as follows: COAB_BPFD, COAB_BPPF1, COAT_BPPF3, COXA_THETH, 1B14_HUMAN, COX4_PARDE, AQP1_HUMAN, GLPF_ECOLI, UCRI_BOVIN, CB21_PEA, and MSBA_ECOLI. The combined data set of 41 and 11 sequences of known X-ray structure was split into the subset of 36 sequences that are not from inner mitochondrial membrane (Table 3) and 16 sequences from inner mitochondrial membrane. In proteins of known X-ray structure the position of helices with respect to putative membrane location was examined with the program GARLIC from our laboratory (free source code available at the Internet address: http://pref.etfos.hr).

For the training data set we used membrane proteins with assigned topological orientation from the version 38 of the Swiss-Prot database.[19] There were 4737 such proteins. In the first selection process we selected 687 proteins that had good annotation and belonged to (a) different protein classes, (b) diverse kingdoms: eukaryotic, bacterial, and archaea, (c) diverse membrane location, and (d) judging by the length of potential transmembrane segments of about 20 residues did not have transmembrane β-strands.

Although Swiss-Prot is an excellent source of protein sequences and annotated information,[20] it inevitably presents some erroneous information, especially in sequence annotation. For instance nontransmembrane pore P-segment is often assigned as a transmembrane segment in the Swiss-Prot version 38. This is the case with glutamate receptors NMZ1_HUMAN, NMZ1_MOUSE, NME1_RAT, GLK2_HUMAN, GLK1_HUMAN, and others. These sequences were not removed from protein data sets. Instead, the transmembrane segment was not assigned to the P-segment, leaving a total of three transmembrane segments for sequences with these Swiss-Prot codes. Such correction is in accord with three TM topology model revealed in experiments.[21,22] It is also in accord with the annotation of TM segments now present in the Swiss-Prot (release 40) for the NMZ1 receptors but not for the NME1, GLK2, and GLK1 that are still annotated as having four TM segments, the second being the pore segment.

Sequence with what we felt was wrong IN/OUT topology assignment was removed only in one case. That was the CMLA_PSEAE drug resistance translocase with assigned OUT topology. All others sequences belonging to drug resistance translocase family with 12 TM segments have IN topology. The second selection step eliminated sequences having greater than 25% identity, The redundancy of the training set was checked by pairwise alignment of each protein pair with the CLUSTALW program.[23] Standard scoring matrix and slow (accurate) alignment options were used.

Sequences already included in the test data set of 158 proteins were left out from the training data set. This procedure left 265 nonhomologous proteins in the training set with assigned IN/OUT topological orientation. Two other test data sets of 79 and 11 proteins shared only 5 and 1 proteins, respectively, with the training set. There were 187 eukaryotic, 76 bacterial, and 2 archaeal proteins, all but 7 from the plasma membrane. The remaining seven proteins originated from the endoplasmic reticulum, the sarcoplasmic reticulum, the rod disk membrane, the lysosomal membrane, and the viral membrane. Sorted by the number of (potential or expected) transmembrane helices (TMH) there were 111 proteins with 1 TMH, 37 with 2, 9 with 3, 8 with 4, 2 with 5, 11 with 6, 48 with 7, 6 with 8, 4 with 10, 1 with 11, 22 with 12, 2 with 13, 3 with 14, and 1 with 24 TMH.

The testing set of 158 proteins was also reduced by eliminating all homologous proteins (with identity higher than 25%) and proteins homologous to any sequence in the training set. This procedure left a total of 96 nonhomologous proteins in the reduced testing set. Together with a training set of 265 proteins we had 361 nonhomologous proteins of known (assigned) IN/OUT topology. The subsets of proteins with 6 and 12 transmembrane domains are all associated with expected IN orientation (only the SECD_ECOLI is assigned OUT topology[8]). Proteins with seven TM segments are almost all associated with expected OUT orientation (only the cyda and cyoe proteins from *E. coli* have IN orientation).

The training procedure was not automatic, and it was not optimized. It consisted of using the prediction performed with the database of 265 nonhomologous proteins, whenever we had to fix some free parameter (such as the choice of amino acid scales). To extract preference functions[13] from soluble proteins we used PDB files to create a data set of 100 such proteins determined by X-ray analysis (1.7 Å resolution or better). The secondary structures were retrieved for these proteins and for membrane proteins of known crystallographic structure by using the program STRIDE.[24] A small utility program was written to convert the STRIDE output into a format suitable for our program. The structural motifs were reduced to helix (H), beta (E), turn (T), and undefined (U). The 3_{10} helices were treated the same way as α-helices.

All of above-mentioned data sets of membrane (158, 79, 41, 52, 36, 16, 265, 96, and 361) and soluble (100) proteins are available as lists of PDB or Swiss-Prot codes and corresponding primary and secondary structures at the following Internet address: http://pref.etfos.hr/split-4.0/extra/. Several corrections to data presented in the Tusnády and Simon[8] and MPtopo database[16] as well as other Supporting Information can be found at the same address.

Counting of Positive Charges Occurring Singly or as Pairs, Triplets, Quadruplets, and Clusters. Arginines and lysines are counted not only in extramembrane loops but also when found in helix caps among the first and last six residues in each known or expected transmembrane helix. In extramembrane loops the basic residues are counted among 30 residues before and after the transmembrane segment. This rule is modified in the case of proteins with only one transmembrane segment. Then, arginines and lysines are counted among 15 extramembrane residues on the each side of the transmembrane segment. Histidine can also have positive charge, but since it is not easy to predict from sequence inspection when it is positive, it is not included in counting procedure.

It is convenient to use the same symbol B for arginine and lysine and X for any other amino acid. Motifs collected in Table 1 are isolated basic charges, charge pairs, triplets, and quadruplets. The BXXXBB triplet is not included in counting triplets for the determination of the triplet-inside

Table 1. Distribution of Basic (B) among Other (X) Amino Acids Close to Putative Membrane Surface in 361 Membrane Proteins

	cytoplasmic loops			extracytoplasmic loops		
		random			random	
motifs	real	observed	expected	real	observed	expected
XXXBXXX singlets	1384	1769 ± 32	1751	1155	1176 ± 26	1208
2B motifs	1846	1515 ± 20	1505	352	342 ± 15	321
BB	848	597 ± 20	577	121	127 ± 8	114
BXB	516	499 ± 22	498	96	112 ± 10	107
BXXB	482	419 ± 16	430	135	104 ± 9	100
3B motifs	912	604 ± 18	589	81	77 ± 9	63
BBB	182	88 ± 12	79	11	10 ± 3	7
BXBB	123	74 ± 7	68	5	8 ± 3	7
BBXB	109	73 ± 8	68	10	9 ± 3	7
BXBXB	55	60 ± 8	59	6	8 ± 3	6
BXXBB	102	60 ± 7	59	5	8 ± 3	6
BBXXB	99	58 ± 7	59	10	9 ± 3	6
BXBXXB	57	49 ± 6	51	6	7 ± 3	6
BXXBXB	57	51 ± 8	51	6	6 ± 2	6
BXXBXXB	49	40 ± 6	44	12	5 ± 2	6
BXXXBB	79	51 ± 7	51	10	6 ± 2	6
4B motifs	191	71 ± 10	62	5	5 ± 3	3
BBBB	36	14 ± 5	11	0	1 ± 1	0.5
BXBBB	25	11 ± 3	9	0	1 ± 1	0.4
BBXBB	26	12 ± 3	9	0	1 ± 1	0.4
BBBXB	24	11 ± 3	9	1	0.5 ± 1	0.4
BXXBBB	30	8 ± 3	8	1	0.5 ± 1	0.4
BBXXBB	25	8 ± 3	8	2	1 ± 1	0.4
BBBXXB	25	9 ± 3	8	1	0.5 ± 1	0.4

bias (motif bias). Remaining nine triplets (Table 1) are counted as overlapping units in which any two successive B are separated with no more than two X. Clusters are counted as separate not overlapping units when three or more B are found so that any two successive B were separated with no more than two X. For instance, according to such a definition, the cluster of basic charges BXBXXBB contains three overlapping charge pairs BXB, BXXB, and BB and two overlapping triplets BXBXXB and BXXBB. For the Table 1 results, motifs and positive charges are collected within the same limits, namely in six helix cup residues and in up to 30 extramembrane residues. For all other results triplets are collected starting from the TM segment middle and also up to 30 residues outside it. The observed B, X frequencies are used to calculate the expected frequency of different motifs according to the binomial distribution.

Randomization of Basic Charges. Random number generator subroutine is used to randomize the sequence position of basic residues Arg/Lys in all cytoplasmic loops and (separately) in all extracytoplasmic loops in each protein. The total number of randomized residues is the same as the number of residues counted in cytoplasmic or extracytoplasmic domains. Also, counting restrictions of six membrane and 15 or 30 extramembrane residues are maintained for random placement of Arg/Lys near membrane surface. Randomization was performed 50 times to find the average number of basic charge clusters created by this procedure as opposed to the number of corresponding clusters counted in unchanged sequences from the same data set of proteins.

Overall Bias. We use three basic topological determinants. Basic determinants are positive charge bias (C.B.), charge difference across the first transmembrane segment (CH.D.), and the bias in triplet motifs of basic residues (M.B.). The charge/triplet bias is calculated as the difference between charges/triplets in odd and even extramembrane loops. Since

Scheme 1

Scale #	Acronym: Attribute
19	PONG3: Combined membrane hydrophobicity scale of Ponnuswamy and Gromiha [25]
27	PRIFT: Cornette statistical scale for amphipathic helices (Cornette) [30]
17	PONG1: Surrounding hydrophobicity scale of Ponnuswamy and Gromiha [29]
12	GIBRA: Gibrat solvent accessibility in proteins [31]
3	PONNU: Surrounding hydrophobicity scale of Ponnuswamy [32]
85	OSMP1: Juretić optimal scale for membrane proteins with one TM segment [14, 33]
43	KUHLE: Kuhn and Leigh membrane propensity scale [34]
4	ENGEL: Engelman hydrophathy values [35]
54	EDE15: Edelman optimal predictors for width 15 [36]
53	EDE21: Edelman optimal predictors for width 21 [36]
51	EDE31: Edelman optimal predictors for width 31 [36]
52	EDE52: Edelman optimal predictors for width 25 [36]
82	MKD2: Juretić modified Kyte-Doolittle hydrophathy scale [14]
81	MKD1: Juretić modified Kyte-Doolittle hydrophathy scale [14]
1	KYTDO: Kyte-Doolitle hydropathy values [25]

the motifs bias is considered to have the greatest weight, it is multiplied with the factor three and an overall bias, named COMB2, is defined as

$$\mathrm{COMB2} = \mathrm{C.B.} + \mathrm{CH.D.} + 3{*}\mathrm{M.B.} \qquad (1)$$

A positive or zero overall bias value is considered to be the "IN" prediction for the location of the N-terminus, while the negative value is the "OUT" prediction.

Creation of Topological Models by Using Different Amino Acid Attributes. Amino acid attributes are listed in the main page for the SPLIT server (http://pref.etfos.hr/split) according to their performance in predicting sequence location of transmembrane segments. The SPLIT 4.0 predictor is using 15 attributes (Scheme 1) at the top or close to the top of that list, but in reverse order so that best scales are last to be associated with a corresponding bias. The numerical values for 20 amino acids in all amino acid scales are available at the same address. The Kyte-Doolittle hydrophobicity scale[25] is used as the first and the last scale. These scales are used to find potential transmembrane helices, through the preference functions method as described previously.[26−28] In this work only old preference functions were used to predict TM helices[14] since we did not utilize new amino acid scales.

The Choice of an Optimal Scale. As a general rule the SPLIT 4.0 finds optimal attribute as associated with the highest absolute overall bias value (1). The absolute bias values are compared, and the last scale with the highest bias is chosen. The training procedure is used to define additional rules for selecting the optimal scale and potential transmembrane helices. The tested sequence is regarded as the soluble protein sequence if at least 14 out of 15 scales do not find any potential TM helix. When a small number of potential transmembrane segments is predicted (less than five), then the bias per one potential TM segment (relative bias) is used as the selection parameter. When the majority of scales predicted the same number of potential TM helices, then one of these scales associated with the maximal bias is chosen as the optimal. However, in the case when the maximal bias value is considerably higher (more than 1.5 times higher) from the next highest bias associated with different number of predicted TM helices, then the scale giving the maximal bias is chosen.

The prediction with the Kyte-Doolittle scale, as the first scale, served to decide whether to use with all other scales the corrections (mostly in length extension, but sometimes in different number of predicted TM helices) obtained with Richardson middle helix preferences[37] and corresponding preference functions.[27] When five or eleven TM segments are initially predicted then such corrections are used. The prediction with the Kyte-Doolitle scale, as the last scale, served to channel the decision toward that scale, as the default scale, if some other attribute did not win due to higher bias.

In addition, some of the potential TM segments are eliminated if the polar residue score inside the segment is too high (K,R scored 5; E,D,Q,N scored 4; P,S,H scored 3; T scored 2; Y,W,C,M,G scored 1; A,V,L,F,I scored 0), the segment length too small, and the maximal preference for the TM helix too low. Numerical values for polar residues are based on the observed frequency of these residues in TM segments and neighboring extramembrane segments from the training database. The minimal value of the TMH elimination parameter, (predicted TMH length-9)(maximal TMH preference)/(polar residue score), was 1.5 and 2.5 for less than four and four or more predicted TMH, respectively. The higher value is applied only in the case when the ends of a predicted TM helix were more than 90 residues away from sequence start or from the C-terminus of the nearest downstream predicted helix and from the sequence end or from the N-terminus of the nearest upstream predicted helix. All segments, with that parameter less than the minimal value, are eliminated. Potential voltage sensor segments with regularly spaced arginines[38] are generally spared by omitting the score due to such arginines from the total score for polar residues. The FORTRAN source code of the algorithm is freely available at http://pref.etfos.hr.

Reported Performance Parameters. Chosen performance parameters punish underprediction and overprediction. The prediction accuracy parameter A_{TM} for residues in the TMH structure takes into account overpredicted o_{TM}, underpredicted u_{TM}, and observed number N_{TM} of residues found in the TMH structure:

$$A_{TM} = (N_{TM} - o_{TM} - u_{TM})/N_{TM} \quad (2)$$

Per-segment prediction accuracy A_{TMH} is also estimated by using eq 2 when a number of overpredicted and underpredicted TM segments is known. To compare the results from I. Simon[8] and S. White[16] laboratory we also used the Q parameter:

$$Q = 100*N_{cor}*(N_{obs}*N_{pred})^{-1/2} \quad (3)$$

Here N_{cor}, N_{obs}, and N_{pred} are respectively correctly predicted, observed, and predicted numbers of TM segments.

Protein specific performance is estimated by counting (a) the number of sequences N_{CT} in which for all predicted TM segments at least nine residues overlap observed and predicted segment without any underpredicted or overpredicted segment and (b) the number of sequences N_{CN} in which both protein N-terminus location and sequence location of predicted TM segments (according to the condition (a)) are completely accurate. Protein specific prediction accuracy is then the ratio of N_{CT} and N_{CN} with the total number N_{MP} of tested sequences (multiplied by 100 to express percentage).

Predictions of Helical Amphipathicity, Extramembrane Helices, and Helix Initiation Sites. We described earlier how helix initiation sites can be found in soluble proteins using preference functions.[39] The same technique can be used for membrane proteins as well. All reported results about helix-initiation sites are obtained with the Kumar and Bansal middle helix preferences extracted from long helices in soluble proteins[40] and corresponding preference functions.[39] Extramembrane helices are predicted by using the Richardson scale of middle-helix preferences extracted from soluble proteins,[37] also through corresponding preference functions.[27] The Eisenberg scale[41] is used through the hydrophobic moment threshold function[27] in all reported results for predicted amphipathic helices. The earlier SPLIT versions were used with the input choice of the Kyte-Doolittle hydropathy scale.[25]

RESULTS

Runs of Positive Charges Are More Frequent in Membrane Than in Soluble Proteins. We used 100 soluble and 361 membrane proteins all less than 25% identical to each other. Basic charges and basic motifs were counted throughout each sequence. The frequency of positive charges (Arg and Lys only) was very similar: 9.78% for soluble and 9.45% for membrane proteins. There were 113 overlapping basic triplets in soluble proteins and 2018 such triplets in membrane proteins. Random distribution of R and K in the sequence created 147 ± 10 triplets in soluble proteins and 1378 ± 43 triplets in membrane proteins. As observed earlier, positive charge clusters are scarce in soluble protein sequences from the PDB data bank.[42] For instance the BBB triplet appears only nine times in these 100 proteins, while the expected number is 21 ± 4 and 17 from random and binomial distribution, respectively. The same triplet appears 353 times in 361 membrane proteins, while the expected number is 193 ± 14 and 174 from random and binomial distribution, respectively. It appears that runs of positive charges are more frequent in membrane than in soluble proteins. Is there any preferential sequence location for positive charge motifs in membrane proteins?

Transmembrane Distribution of Basic Residues and Basic Clusters. Nonhomogeneous and asymmetric basic charge distribution is expected according to the positive inside rule,[3] with charges depleted in putative transmembrane segments and enriched in cytoplasmic (inside) more than in extracytoplasmic (outside) loops. This is indeed found both for single charges and for basic charge clusters (Figure 1). The number of all residues, in sequence segments where motifs are counted (see Methods), is about the same in cytoplasmic (30949) and in extracytoplasmic loops (27913), but the number of Arg/Lys residues is considerably greater in cytoplasmic (4224) than in outside loops (1783). Isolated basic residues of the type XXXBXXX are only slightly more frequent in cytoplasmic loops, because the majority of such residues in cytoplasmic loops are engaged in creating sequence motifs and clusters (Table 1).

Our choice of motifs to be counted was based on desire to compare all simple motifs in cytoplasmic and extracytoplasmic loops when enough motifs could be collected among 361 nonhomologous membrane proteins. We set an arbitrary

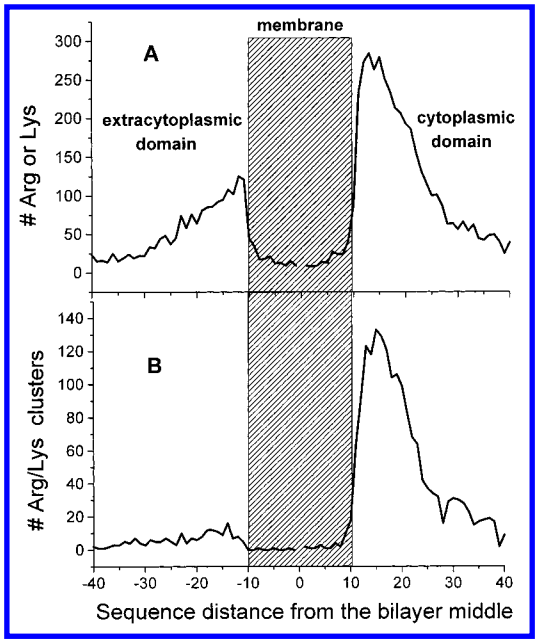

Figure 1. Transmembrane profile of basic residues (Figure 1A) and clusters with three or more positive amino acids (Figure 1B) in 361 nonhomologous integral membrane proteins. Positive charges are summed for each sequence position starting from the bilayer middle toward membrane surface and up to 30 extramembrane residues. The same range is used for cluster counting in which for each sequence position we summed only those basic residues that are part of clusters. Basic residues (B) separated from other such residues by more than two nonbasic residues (X) were not counted in Figure 1B.

lower limit of about 50 identical motifs (in terms of X, B combinations) for triplets and of about 20 identical motifs for quadruplets (Table 1). The motifs ratio in favor of cytoplasmic loops is about 5, 11, and 38 for charge pairs, triplets, and quadruplets, respectively. For instance, out of 906 triplets 833 are found in cytoplasmic loops and 73 in extracytoplasmic loops. Positive charge bias is about three. Therefore, the predominance of triplet motifs near the cytoplasmic surface is a stronger bias than the K/R charge bias (Figure 1). In certain classes of membrane proteins particular triplet motifs, such as the BBB motif in 84 and 68 proteins with OUT topology having one and seven TM segments, respectively, appear exclusively in the cytoplasmic position (Table 2). Different counting boundaries inside membrane spanning segments, that is collecting motifs from 12 boundary residues or from whole TM segment, produced very little or no change in the Table 2 results.

Does the Positive Inside Bias Explain the Motif Bias as Well? We asked the question if the positive inside bias is responsible for motifs bias as well. Higher frequency of basic charges in cytoplasmic loops (0.1365) than in outside loops (0.0639) is expected to create more inside motifs in purely random fashion. Randomization of basic charges in each protein in all inside or outside loops created expected number of motifs as calculated from binomial distribution for the frequency of Arg/Lys residues (Table 1). In extracytoplasmic loops there is no significant difference in the number of real and randomized or expected motifs. However, in cytoplasmic loops, the situation with Arg/Lys motifs is quite different. While for minority of motifs, such as the BXB and BXBXB, the numbers found in real proteins are similar to virtual proteins with randomized charges, for most motifs there is a significant excess in their cytoplasmic loops numbers over the expected or random numbers (Table 1). For instance, in 68 proteins with seven TM segments motifs BXBB and BBXXB appear in cytoplasmic location 48 times each, while from random distribution of Arg/Lys near cytoplasmic membrane surface expected numbers are 26 ± 4 and 21 ± 5, respectively.

We also tried different counting boundaries outside membrane spanning segments when positive charges and clusters are collected from real or randomized extramembrane segments long 15, 20, 25, 35, and 40 residues. The evidence (enclosed in Supporting Information) for specific enrichment of triplet motifs such as BBB, BBXXB, and BXXBB in these cytoplasmic segments, way above expected random distribution, was as strong as for the 30 residues long extramembrane segments (Table 1).

Predicting the N-Terminus Location When Sequence Location of TM Segments Is Given. Three topological determinants are taken into account when combined predictor is used: charge bias,[3] charge pattern bias (this work), and charge difference bias.[11] The combined predictor can be constructed in two ways. The first is the priority predictor (COMB1). It accepts the motifs bias prediction if such bias is different from zero. If it is zero, then positive charge bias prediction is accepted. The charge difference prediction is accepted only if the first two biases are equal to zero. The second possibility is to use the overall bias as defined in the Methods section (COMB2).

For 36 sequences of known X-ray structure, not from inner mitochondrial membrane, the motif bias predictor produces perfect IN/OUT prediction in 18 sequences where motifs are found and where motif bias is different from zero (Table 3). All of 30 positive charge clusters and 37 overlapping triplets are located near cytoplasmic membrane surface. Charge bias is different from zero for 35 sequences. It fails in the prediction of the IN/OUT location of protein N-terminus in three cases. Charge difference across the first transmembrane segment is different from zero for 31 sequences, and it fails as the IN/OUT predictor in nine sequences, eight of which are bacterial.

The prediction accuracy for the N-terminus location is quite low for 16 sequences of known X-ray structure from mitochondrial inner membrane. Only five sequences are associated with motif bias different from zero, and the triplet bias produces correct IN/OUT prediction only in two cases. Out of nine clusters of positive charges five are IN (matrix location) and four are in OUT location. When positive charges are counted individually, prediction accuracy is also low: four out of 12 sequences (having charge bias different from zero) are predicted with wrong N-terminus location. The charge difference across the first transmembrane segment fails as IN/OUT predictor in four out of 14 sequences where it is different from zero.

Overall accuracy of the combined predictor for all of 361 nonhomologous membrane proteins is 90.3% for priority predictor and 91.7% for overall bias predictor (Table 4). We choose to use in the following sections only the overall bias predictor as the combined predictor (COMB2). The motif bias is the most accurate predictor for protein N-terminus location (Table 4), but it cannot be used in 26% proteins, where it is equal to zero. The positive inside rule alone is 91.2% accurate for the subset of 340 proteins having the

Table 2. Distribution of Triplet Motifs in Protein Data Sets

data set[a]	1 TM (IN)	1 TM (OUT)	2 TM seg.	3 TM seg.	4 TM seg.	6 TM seg.	7 TM seg.	12 TMS	36 X-p.	96 prot.	265 prot.
(A) Total Number of Proteins with Transmembrane (TM) Segments											
no. of proteins	36	84	44	16	55	18	68	29	36	96	265
no. of proteins with triplets	22	57	30	9	39	15	56	28	18	65	201
no. of triplets	47	153	63	24	170	52	290	125	37	179	727
no. of clusters	24	72	38	14	94	33	170	74	30	105	417
(B) Total Number of Triplets versus Triplets in OUT Loops											
BBB	13:0	49:0	12:1	4:0	37:8	17:1	43:0	21:2	10:0	36:4	156:7
BXBB	7:0	19:0	11:1	3:0	31:4	5:1	48:0	18:1	1:0	27:0	101:5
BBXB	4:0	25:2	8:0	5:0	26:2	5:2	35:2	20:3	7:0	24:1	95:9
BXXBB	5:1	13:0	10:0	3:0	7:2	5:0	28:0	18:1	6:0	23:2	83:4
BBXXB	7:1	15:2	7:0	3:2	15:1	3:0	50:2	8:1	4:0	22:2	87:8
BXBXB	3:0	10:2	6:1	0:0	15:3	3:0	20:0	7:2	2:0	17:2	44:4
BXXBXB	3:2	6:1	3:0	1:1	14:0	5:0	22:0	15:1	1:0	13:1	50:5
BXBXXB	2:0	12:1	3:0	3:1	12:2	4:0	21:0	8:2	3:0	9:1	54:5
BXXBXXB	3:1	4:1	3:2	2:0	13:2	5:3	23:1	10:0	3:0	9:1	57:12

[a] Protein data sets include proteins with increasing number of TM segments (from 1 to 12) from training (265 proteins) and testing data sets (158 proteins), proteins of known structure not from inner mitochondrial membrane (36 X-p.), and data sets of 96 and 265 nonhomologous proteins.

charge bias different from zero. The charge difference across the first TM segment is the least accurate (87.8%). It cannot be used (it is zero) in 9% of proteins, and it is more accurate for eukaryotic (90.5%) than for prokaryotic proteins (84.2%). The triplet bias (in sequences where it is different from zero) gave similar 95% accuracy in eukaryotic and prokaryotic proteins.

Fair comparison of performance measures would use only such proteins where all three of the biases are different from zero. Putting together nonhomologous proteins from training and testing set, and removing all proteins with zero biases, leaves 242 nonhomologous proteins. The rank list of predictors in terms of their performance is then motif bias (95.9%), positive charge bias (95.0%), and charge difference bias (90.9%) (Table 4). The overall bias predictor is the most accurate for this data set of proteins (96.3%).

The prediction accuracy depends on the number of transmembrane segments expected to be in the helix conformation (Table 5). Proteins with seven TM segments (mainly G-protein coupled receptors) are predicted with correct transmembrane orientation with the highest accuracy (97.1%). For some membrane proteins the triplet bias alone never fails when triplets are found in a sequence. This is the case for 56 proteins with seven TMS, for 57 type I single span proteins, and for 18 membrane protein sequences of known crystal structure (Tables 2 and 3). There is only one prediction error for 18 proteins with six TMS. It is for the SecD part of the protein export Sec system from *Escherichia coli*, which is strongly predicted with the IN orientation, but its expected orientation is OUT.[8] However, this is an assignment error instead of a prediction error. The SecD protein contains six transmembrane stretches (not four as adopted in the version 38 of the Swiss-Prot database) and a large periplasmic domain, which implies its IN orientation.[43] The current Swiss-Prot model for the SECD_ECOLI (release 40) has indeed six TMS and cytoplasmic N-terminal location. Topological determinants found in two, three, and four TMS proteins produce less accurate IN/OUT predictions.

The Topology Predictor SPLIT 4.0. In the previous section our IN/OUT predictor (COMB2) used the assigned sequence location of TM segments. A much more interesting goal is to produce different topological models, for a given membrane protein sequence, and to select the optimal model corresponding to the highest overall bias factor. In our earlier work with the SPLIT predictor we noticed that different sequence location and number of potential TM helices are predicted in some cases even when very similar hydropathy scales are used as an input to the preference functions method. Therefore, a selection of different scales of amino acid attributes can be used to produce different topological models for a tested sequence, and an overall bias factor will select the most likely model (see Methods). The prediction results can be seen in Tables 3−7.

The training data set was used to see how important are different program routines mentioned in the Methods. For instance, the prediction accuracy for transmembrane segments, as measured by the A_{TMH} parameter (Methods), decreases from its Table 6 value to 0.878 when Richardson preference functions are not used at all. It also decreases to $A_{TMH} = 0.856$ when these functions are always used. Seemingly small accuracy gain, when Richardson functions are used, as described in Methods, is due to correct transmembrane topology prediction for important receptors (ionotropic and muscarinic acetylcholine receptors) and channels (voltage-gated potassium channel). In another example, the prediction accuracy decreases to $A_{TMH} = 0.850$ when polar residue score routine (see Methods) is omitted. As expected, the prediction accuracy also decreases when only one amino acid scale (the Kyte-Doolittle[25]) is used, instead of 15 different scales ($A_{TMH} = 0.861$).

Several Examples How SPLIT 4.0 Creates and Selects Topological Models for a Given Sequence. Let us see in one example how SPLIT 4.0 works. The yeast glucose transporter HXT5_YEAST is assigned in the Swiss-Prot database with IN topology but with 11 TM segments. Other glucose transporters are assigned with IN topology and 12 TM segments. In the case of the HXT5, only the Edelman optimal parameter[36] scales 51, 52, 53, and 54 produce 12 potential TM helices and IN topology, while scales 19, 27, 17, 12, 3, 85, 43, 4, 82, 81, and 1 produce 11 potential TM helices and OUT topology (Table 7). The predictor easily decides that the last to be used of Edelman scales (#52) is the optimal scale, because all Edelman scales produce the overall bias of 23, while all other scales produce negative

Table 3. Topology Predictors for Membrane Proteins of Known Crystallographic Structure

		observed biases					SPLIT 4.0 predictions	
protein[a] no./name/code	no. of TMH	obs. topol.	motif bias	charge bias	charge diff.	overall bias	overall bias predicted	no. of TMH correctly predicted (scale)
#1−6 Photosynthetic Reaction Center								
1PRC_H	1	OUT	−2	−5	−2	−13	−11	1 (1)
2RCR_H	1	OUT	0	−1	2	1	2 (IN)	1 (1)
1PRC_L	5	IN	2	6	3	15	18	5 (1)
2RCR_L	5	IN	2	6	1	13	15	5 (3)
1PRC_M	5	IN	0	6	0	6	10	5 (85)
2RCR_M	5	IN	0	6	1	7	9	5 (51)
#7−13 Cytochrome c Oxidase								
1QLE_A	12	IN	0	7	2	9	7	12 (1)
1QLE_B	2	OUT	−1	−1	−5	−9	−9	2 (1)
1QLE_C	7	IN	0	3	0	3	5	7 (52)
COX4_PARDE	1	IN	0	2	−2	0	4	1 (1)
1EHK−I	13	IN	3	17	0	26	27	13 (1)
1EHK−II	1	IN	0	−1	−2	−3	2	1 (1)
1EHK−IIa	1	IN	0	1	−1	0	0	1 (1)
#14 bacteriorhodopsin 1BRX	7	OUT	0	−2	0	−2	−6	7 (1)
#15 halorhodopsin 1E12	7	OUT	−2	−2	−5	−13	−15	7 (27)
#16 rhodopsin 1F88	7	OUT	−1	−9	−4	−16	−16	7 (82)
#17 potassium channel 1BL8	2	IN	0	2	4	6	6	2 (1)
#18 mechanosensitive ion channel 1MSL	2	IN	1	5	1	9	8	2 (1)
#19 glycophorin A 1MSR	1	OUT	−2	−4	−6	−16	−19	1 (54)
#20 Ca2+ transporting ATPase 1EUL	10	IN	7	30	1	52	47	10 (3)
#21 F0 ATPsynthase C E. coli 1A91_C	2	OUT	0	−2	−1	−3	−6	2 (81)
#22−24 Coat Proteins								
COAB_BPFP	1	OUT	−1	−3	−5	−11	−13	1 (4)
COAB_BPPF1	1	OUT	0	−1	−4	−5	−5	1 (1)
COAT_BPPF3	1	OUT	0	−2	−4	−6	−6	1 (1)
#25−27 Fumarate Reductase Sequences								
1FUM_C	3	IN	1	9	8	20	23	3 (1)
1FUM_D	3	IN	0	2	0	2	5	3 (1)
1QLA_C	5	IN	3	4	3	16	18	5 (1)
#28 glycerol channel 1FX8	6	IN	1	5	1	9	10	6 (43)
#29 aquaporin 1 1FQY	6	IN	3	7	1	17	19	6 (85)
#30 lipid flipase MSBA_ECOLI	6	IN	3	21	2	32	33	6 (1)
#31 histoc. antigen 1B14_HUMAN	1	OUT	−1	−3	−4	−10	−8	1 (1)
#32−36 Light Harvesting Complex								
CB21_PEA	3	IN	1	−3	−6	−6	8	3 (1)
1KZU_A	1	IN	0	−1	−1	−2	−2 (OUT)	1 (1)
1KZU_B	1	IN	0	1	−2	−1	−2 (OUT)	1 (1)
1LGH_A	1	IN	0	1	−1	0	−2 (OUT)	1 (1)
1LGH_B	1	IN	0	0	−2	−2	−2 (OUT)	1 (1)

[a] Sequences from mitochondrial inner membrane are not included.

Table 4. Prediction Accuracy of Topology Predictors in Protein Data Sets

protein data set[a] (no. of proteins)	motif bias (no. of proteins)	charge bias (no. of proteins)	charge difference (no. of proteins)	overall bias predictor	SPLIT 4.0 N_{CN}/N_{MP} (percentage)
nonhomologous (361)	94.7 (266)	91.2 (340)	87.8 (329)	91.7	71.2
Tusnády and Simon nonhomologous (96)	93.8 (65)	90.2 (92)	86.7 (83)	90.6	81.3
our nonhomologous (265)	94.5 (201)	91.1 (248)	87.8 (245)	91.7	68.3
bacterial (125)	95.3 (86)	89.0 (118)	84.2 (114)	90.4	68.0
eukaryotic (230)	94.4 (179)	92.2 (217)	90.5 (210)	92.6	73.0
with all topolog. determinants (242)	95.9 (242)	95.0 (242)	90.9 (242)	96.3	74.8
with known X-ray structure (36)	100 (18)	91.4 (35)	71.0 (31)	83.3	86.1

[a] See Materials and Methods section for description.

overall bias less than 10 in the absolute value. There are three triplet motifs in this sequence: BBXB, BXXBB, and BXBXB. The BXXBB motif is the R(404)FGRR signature of sugar transporters, which must be located in the 8−9 cytoplasmic loop. For Edelman's scales it is indeed in the cytoplasmic (IN) positions as well as other two motifs from this sequence, so that high positive bias is obtained. With preference functions derived from Edelman's scale #52, SPLIT predicts seventh TM segment at the sequence location 341−361. Models based on hydrophobicity do not see that TM segment and accordingly locate the RFGRR motif into the extracytoplasmic space, which produces negative overall bias. The RR(345) positive doublet at the cytoplasmic looking N-terminus of the TM VII is the cause of its decreased hydrophobicity.

The examples of potassium voltage-gated channels CIQ1_HUMAN and CIQ4_HUMAN are also instructive. Their topology is now correct in the latest Swiss-Prot anno-

Table 5. Prediction Accuracy of Topology Predictors in Protein Classes

class[a] (#TMS)	COMB2	SPLIT4.0 A_{TM}	SPLIT4.0 A_{TMH}	SPLIT4.0 N_{CN}/N_{MP}
1 OUT	0.940	0.745	0.929	0.869
1 IN	0.806	0.747	0.944	0.750
2	0.886	0.609	0.852	0.636
3	0.688	0.576	0.833	0.500
4	0.745	0.728	0.927	0.527
6	0.944	0.536	0.824	0.500
7	0.971	0.689	0.931	0.731
12	0.931	0.637	0.937	0.655

[a] See Table 2 and its legend with respect to proteins included in each class.

Table 6. SPLIT Prediction Results with Versions 3.1,[26] 3.5,[27] and 4.0 (This Work)

protein list	performance parameter	SPLIT 3.1	SPLIT 3.5	SPLIT 4.0
265 nonhom.	A_{TM}/Q	0.661	0.614	0.626/79.0
265 nonhom.	A_{TMH}/Q	0.831	0.840	0.883/94.2
265 nonhom.	N_{MP}	263	265	265
265 nonhom.	N_{CT}	159	163	198
265 nonhom.	N_{CN}			181
96 nonhom. Simon	A_{TM}/Q	0.710	0.684	0.682/81.9
96 nonhom. Simon	A_{TMH}/Q	0.908	0.920	0.948/97.4
96 nonhom. Simon	N_{MP}	96	96	96
96 nonhom. Simon	N_{CT}	72	72	85
96 nonhom. Simon	N_{CN}			78
36 X-ray membrane	A_{TM}/Q	0.626	0.670	0.720/77.8
36 X-ray membrane	A_{TMH}/Q	0.949	0.942	1.00/1.00
36 X-ray membrane	N_{MP}	36	36	36
36 X-ray membrane	N_{CT}	29	30	36
36 X-ray membrane	N_{CN}			31
100 X-ray soluble	N_{MP}	11	11	10

Table 7. SPLIT 4.0 Run through Amino Acid Attributes in the Case of the HXT5_YEAST Glucose Transporter

scale no.	no. of TMH predicted	motif bias	charge bias	charge diff.	overall bias	predicted topology
1	11	−1	1	−5	−7	OUT
19	11	−1	4	−5	−4	OUT
27	11	−1	3	−5	−5	OUT
17	11	−1	3	−5	−5	OUT
12	11	−1	3	−5	−5	OUT
3	11	−1	3	−5	−5	OUT
85	11	−1	3	−5	−5	OUT
43	6	1	8	−5	6	IN
4	11	−1	4	−5	−4	OUT
54	12	3	19	−5	23	IN
53	12	3	19	−5	23	IN
51	12	3	19	−5	23	IN
52	12	3	19	−5	23	IN
82	11	−1	1	−5	−7	OUT
81	11	−1	1	−5	−7	OUT
1	11	−1	1	−5	−7	OUT
52 (selected)	12	3	19	−5	23	IN

tation update (release 40, August 2001), but it was not so in previous releases. The potassium channel CIQ1_HUMAN was assigned in the Swiss-Prot (release 38) without transmembrane voltage sensor 226−248 and with the pore P-segment (H) 300−320 wrongly annotated as the TM segment S5. Both mistakes are easy to make because the S4 voltage sensor segment has low hydrophobicity due to arginines positioned at each fourth position, while the membrane buried helix of the P-segment is often hydrophobic enough to be mistaken for the TM segment. The SPLIT 4.0 predictor selects the PRIFT scale #27 for optimal amphipathic helices.[30] It is the only scale that predicts the correct number of six potential TM helices with the S4 segment included at the sequence location 221−244 and the P-segment excluded as the TM segment. The selection of the #27 scale occurred due to large positive overall bias of 30 that was twice as high as biases produced by all other scales predicting five TM helices.

The potassium channel CIQ4_HUMAN was assigned in the Swiss-Prot (release 38) with misplaced first four TM segments. Its first potential TM segment was L(45)-GLLGSPLPPGAPGPGSGS, which contained too many prolines, glycines, and serines, and the fourth potential TM segment at the 173−193 sequence location was annotated as the voltage sensor S4, which, however, did not contain a single arginine. With such TM segments assignments both first cluster K(78)RYRR and second cluster R(166)FRFARK of basic residues would be in the extracytoplasmic location if protein amino terminus is cytosolic. The prediction of five, instead of six, transmembrane segments in this sequence, due to the underprediction of the S4 segment, is a common error made by online Web predictors such as DAS,[15] SOSUI,[44] TMHMM,[1] TopPred II,[45] PHDhtm,[7] and TMAP.[6,46] It leads to extracytoplasmic location of the last two positively charged clusters R(213)MVRMDRR and K(337)RR and cytoplasmic location of the pore segment, when N-terminus is assumed to be cytoplasmic. Other online predictors, such as HMMTOP,[8] predict six TM segments so that S4 is omitted, but the P-segment (H) is added as the potential TM segment. This leads to wrong membrane topology for that protein too.

The SPLIT 4.0 predictor has an easy choice when working on the CIQ4_HUMAN sequence. Only the PRIFT scale (#27),[30] for optimal amphipathic helices, produces a large positive bias of 52, with six potential TM helices predicted. The S4 segment at predicted sequence location 194−218 has regularly spaced arginines, and it agrees well with what is now thought to be the correct sequence position 202−224 for the S4 segment (the Swiss-Prot version 40 has now the correct model for the CIQ4). All other scales miss the S4 segment and produce topological models with five potential TM helices. These models are associated with a much smaller overall bias of 12−14. None of the automatically generated topological models by the SPLIT 4.0 is as bad as the Swiss-Prot model in the release 38, which was associated with negative overall bias of −10 and accordingly the OUT prediction for the N-terminus location. Therefore, at least in the case of these potassium channels, the choice of SPLIT 4.0 for automatic annotation would be advantageous.

The Performance on Protein Data Sets. When different protein data sets are compared, the protein based accuracy increases for better known proteins (Table 4). With respect to earlier SPLIT versions there is about a 15% increase in the number of proteins with all TM segments correctly predicted (the N_{CT} parameter in Table 6). The number of underpredicted + overpredicted TM segments is considerably reduced, which is seen as an increase in the A_{TMH} and Q parameters for segment prediction (Table 6).

To compare different procedures how charged amino acids could be engaged for topology prediction we used again 242 nonhomologous sequences having all biases different from zero. When only one topological determinant is used, the

protein based prediction accuracy (the N_{CN}/N_{MP} parameter) is 72% for the bias in triplet motifs of basic residues, and 70% for the positive charge bias as the selection parameter. The predictor with all three biases combined in (1) finds 75% proteins in which both protein N-terminus location and sequence location of predicted TM segments are completely accurate (Table 4).

Overprediction of TM segments in longer sequences, having only one such segment, and underprediction of TM segments in multispanning membrane proteins is quite common. This was partially corrected during the training procedure, but the price paid is the introduction of several exceptions to the general rule that the optimal scale for choosing topological model is associated with the largest absolute bias (see Methods). As expected, on the basis of the IN/OUT predictor performance, potential TM helices are easier to predict in some protein classes than in another (Table 5). Difficult cases are sequences with two, three, four, and six TM segments, while sequences with one or seven TM segments are much easier to predict correctly. Even in the worst case the A_{TMH} parameter is higher than 0.81, which means that on average less than one in ten TM segments is underpredicted or overpredicted. For proteins having 12 TM segments such an error rate would be too high, because every second such protein would be predicted with wrong topology. The observed per protein accuracy measure (last column in the Table 5) of 65.5% in the case of 29 proteins with 12 TMS is not much better than that. Notice that for 348 observed, 342 predicted, and 334 correctly predicted TM segments, the per-segment accuracy measures, $A_{TMH} = 93.7$ and $Q = 96.8$, still appear very high.

Also expected on the basis of the IN/OUT predictor results is poor performance for 16 inner mitochondrial sequences of known secondary structure. Only nine are predicted both with correct sequence location of all TM helices and with correct location for the N-terminus. Out of 41 observed TM helices two are underpredicted and one overpredicted.

All of 137 TM helices belonging to 36 sequences from X-ray determined proteins that are not from inner mitochondrial membrane (Table 3) are correctly predicted without overpredictions (Table 6). For 11 additional sequences of known structure (see Methods), out of 28 observed TM helices, 27 are correctly predicted by the SPLIT 4.0 and one is overpredicted. When 41 sequences with known X-ray structure from the MPtopo database[16] are combined with 11 additional sequences of known structure (see Methods) the prediction accuracy for TM helices is still very high $Q = 99.1$, because out of 178 observed helices only two are underpredicted and one is overpredicted. The Rieske iron−sulfur protein contributes one underpredicted and one overpredicted TM helix.

The Comparison with Other Methods. A recently created database of membrane protein topology, MPtopo,[16] can be used to compare the accuracy of our predictor with PHDhtm,[7] HMMTOP,[8] TopPred II,[45] and TMAP.[46] The prediction accuracy Q (see Methods) is used for that comparison. For the 3D-helix subset of 41 proteins with verified structure by means of crystallography the Q factor is equal to 99, 97, 96, 95, 95 for SPLIT, PHDhtm, TMAP, HMMTOP, and TopPred II algorithms, respectively. Out of 150 observed helices 149 are correctly predicted, one is underpredicted, and another one overpredicted by the SPLIT algorithm. The required overlap among observed and predicted TM helix is nine residues in the SPLIT method,[26] which is a more strict requirement than, for instance, five residue overlap used by the PHDhtm [7] and three residue overlap used by the HMMTOP[8] algorithm. For the 1D_helix subset of 38 proteins whose structure was verified by means of gene fusion and other methods the Q factor is 95, 94, 93, 92, and 89 for the HMMTOP, SPLIT, PHDhtm, TMAP, and TopPred II, respectively. Out of 242 observed TM helices 10 are underpredicted and 18 overpredicted by the SPLIT algorithm.

Earlier SPLIT versions do not use automatic selection of optimal topological model and are less accurate. For the 3D-helix subset the Q factor is respectively 97 and 96.6 for SPLIT 3.5 and SPLIT 3.1. Corresponding values for the 1D-helix subset are 93.5 and 93. Using the optimal combination of amino acid attributes in the SPLIT 3.5 version[27] (#60[37] together with #52[36]), instead of the Kyte-Doolittle hydropathy scale,[25] results in $Q = 98$ for the 3D-helix subset. Four TM helices are underpredicted, two are overpredicted, and two sequences are not predicted as membrane protein sequences.

DISCUSSION

Positive Charge Clusters as Topological Determinants for All Membrane Proteins? The first aim of this work was to find accurate predictors of protein N-terminus location (cytoplasmic-IN or extracytoplasmic-OUT) for a given sequence assignation of TM segments. The positive-inside rule[3] and charge difference across the first TM segment[11] are combined with the bias in basic charge clusters to achieve that goal. Our choice of triplet motifs, as basic charge motifs to be counted, is certainly not unique. Positive charge pairs and quadruplets of positive charges are also topological determinants (Table 1). Triplets can be defined with more flexibility, allowing for three, four, or even five consecutive X amino acids, between two neighboring positive charges in a sequence. Our choice of standard nine triplets for motif bias calculation (Table 2b) is a compromise between having the highest IN/OUT motif ratio and high abundance of motifs. The advantage is that it is the simplest choice with the smallest number of different motifs to be counted, which still results in high prediction accuracy, when such triplets are present. The disadvantage is that such triplets are absent in about one-fourth of all membrane protein sequences.

The initial assumption, that triplets of positive charges can be used as topological determinants in all sequences where such charge clusters are found, irrespective of their origin, is suspect in the case of proteins from inner mitochondrial membrane. For 16 of these sequences, with known structure, the prediction of protein N-terminus location is only slightly better than random. This may well be because we did not attempt to distinguish nuclear from plastid encoded sequences with respect to predictor's performance.[47] Another tacit assumption, about the absence of membrane-spanning basic clusters, is certainly wrong in the case of the S4 TM segments with the BXXBXXB motifs from voltage-gated channels.[38] The IN/OUT classification is questionable for such clusters, which suggests one possible reason for relatively low BXXBXXB IN/OUT ratio (Table 2, last column).

At the present time only a small number of membrane proteins have been characterized by high-resolution structural

techniques.[16] When 36 of these sequences, that are not from inner mitochondrial membrane, are separately examined, then all of the 37 Arg/Lys triplet motifs are found near putative cytoplasmic membrane surface. In other words, triplets of positive charges are perfect topological predictors and one is tempted to regard this finding as the triplet inside rule. The positive inside rule is also accurate (91%) but far from a perfect predictor for these sequences (Table 4).

Functional Importance of Basic Charge Clusters. Positive charge clusters near cytoplasmic membrane surface may have functional and structural roles. Such clusters can serve to create obligatory cytoplasmic motifs for signal transduction in the cytoplasm through interaction with other cytoplasmic proteins.[48−50] Basic charge clusters can interact with negatively charged polar headgroups of phospholipids from the cytoplasmic layer of membrane bilayer[51] or can be important in the binding of negatively charged ligands. For instance, the CD44 bitopic membrane protein contains two conserved triplets of the BBB type in its cytoplasmic domain that are involved in the binding of CD44 to hyaluronic acid.[52]

Some basic clusters form cytoplasmic anchors for transmembrane domains and can be described as topological signals.[53−55] There are two conserved pentameric BXGBB motifs in the Glut1 glucose transporter (GTR1_HUMAN, Swiss-Prot code), in the second (loop 2−3), and in the fifth cytoplasmic loop (loop 8−9).[55] Mutating the three arginines into glycines in either motif caused each loop to translocate into the exoplasmic (OUT) compartment along with its two flanking TM segments.[55] This substitution produced aberrant 10 TMS (instead of 12 TMS) topology and completely abolished transport function. Another conserved motif is the GXXX(D/E)BXGBB motif in lactose permease that also appears in the loop 2−3 and in the loop 8−9.[56,57] These authors suggested that the basic residues in conserved cytoplasmic motifs of such residues play a role in protein insertion or stability. For instance, mutations transforming basic to neutral residues in the loop 8−9 motif reduce the level of permease expression.[56]

Many other transporters thought to have 12 TM helices have (D/E)BXGBB motif in the 8−9 cytoplasmic loop. Some of these are yeast glucose transporters HGT1, HXT1, HXT4, HXTC, RGT2, plant glucose transporter STP1, bacterial proline and α-ketoglutarate transporters PROP and KGTP, respectively, and yeast myo-inositol transporter ITR1. The DRFGRR motif in the second cytoplasmic loop of bacterial tetracycline transporter contains three amino acid residues, D(66), G(69), and R(70), that are functionally and/or structurally important, one of them essential (D(66)).[58]

A conserved BBXXB motifs of basic residues appears often at the carboxyl-terminal third intracellular loop (the 5−6 cytoplasmic loop) in 7 TMH G-protein coupled receptors.[48−50,59,60] The G protein-coupled receptors make 88% of our seven transmembrane domain sequences. High accuracy of IN/OUT prediction (Table 5) may have arisen due to conserved cytoplasmic triplet motifs even in its members that are no more than 25% identical. Indeed, the G-protein-activating domain is heterogeneous in amino acid sequence but conserved in secondary structure, possibly as the amphipathic α-helical extension in the cytoplasm of the sixth membrane-spanning segment.[61] Receptors for serotonin, adenosine, acetylcholine (muscarinic), bombesin, gastrin, cholecystokinin, corticotropin, dopamine, growth hormone, histamine, melanotropin, tachykinin, neuropeptide Y, prostaglandin, and somatostatin have the BBXXB motif in the 5−6 cytoplasmic loop as well as adrenergic, angiotensin, chemokine, dopamine, somatostatin, chemoatractant, opioid, N-formyl peptide, and other structurally diverse G protein-coupled receptors. When one of two X amino acids is serine or threonine, the BBXXB motif is a potential phosphorylation site for protein kinase C or calcium activated protein kinase. For instance, the KRTPR(140) motif in the second cytoplasmic loop of muscarinic acetylcholine receptors is a conserved BBXXB triplet motif found in all mAChR subtypes as a consensus sequence for phosphorylation by protein kinase C.[60] Therefore, this motif may be coupled to phospholipase C/protein kinase C pathway in addition to G protein pathway and may act as a potential switch point between these pathways. The KKAAR(365) motif in the third cytoplasmic loop of that protein is found to be obligatory for stimulation of phosphoinositide hydrolysis and cAMP accumulation.[60]

The ratio of 37 IN to zero OUT for all triplet motifs (Table 3) or 48,49 IN to zero OUT for some triplets (BBB) in some classes of membrane proteins (Table 2) also argues for structural and functional role for such triplets at the cytoplasmic membrane side. More direct evidence is found for point mutations and deletions in triplet motifs that are causing human genetic diseases. For example, the vasopressin V2R receptor contains two similar basic charge clusters in cytoplasmic loops 1−2 and 5−6 (RRGRR and RRRGRR, respectively). The receptor with deletions in these motifs fails to activate the adenylyl cyclase system causing congenital nephrogenic diabetes insipidus.[62] Both motifs are predicted by us as helix initiation sites in amphipathic helix conformation.[27,39] A mutation transforming alanine 568 to valine in the KIAKK motif from LSHR_HUMAN receptor and of alanine 623 to isoleucine in the KIAKR motif from TSHR_HUMAN receptor cause constitutive activation of the adenylyl cyclase system resulting in male precocious puberty and thyroid adenomas, respectively.[63,64] The KIAK(K/R) motifs are also located at the C-terminal of the 5−6 cytoplasmic loop and are also predicted by SPLIT to be in the amphipathic helix secondary conformation.

Point mutations in the C-terminal motif G(189)RLRFARK of the 2−3 cytoplasmic loop of the potassium CIQ1 human channel (kcnq1 gene) cause deafness and long QT syndrome type 1 (a congenital heart disease).[65] This motif is also predicted as an amphipathic helix in its C-terminal which is fused with the N-terminal of the S3 TM helix. Another motif of the BXXBB type appears in the 4−5 cytoplasmic loop of the cystic fibrosis chloride channel. Its second B is arginine 297. Cystic fibrosis develops when it becomes Q. This motif is part of a longer predicted amphipathic helix segment and helix initiation site just next to the cytoplasmic beginning of the fifth TM segment. It may act as a topogenic signal during protein membrane insertion.[66] The KPQ deletion in the SKKPQK(1508) triplet from the C-terminal of the cytoplasmic III−IV domain in heart sodium channel (scn5a gene or CIN5_HUMAN in the Swiss-Prot code) causes long QT syndrome type 3 (sudden cardiac arrest). This triplet motif is part of a longer run of basic residues next to hydrophobic IFM (1487) triplet that is required for fast Na^+ channel inactivation.[67] It is part of the inactivation loop between repeats III and IV and part of predicted amphipathic helix. The SKK part of the triplet is predicted PKC

phosphorylation site (according to Prosite database). In another sodium channel, CIN1_HUMAN (scn1a gene), the R1648H mutation in the voltage sensor S4 helix of the repeat IV (the BXXBXXB(1648)XXBXXBXXB motif), causes generalized epilepsy with febrile seizures. As in other voltage sensor segments, regularly spaced positive charges of the S4 segment are thought to move through the membrane in response to electric field change.[38] The calcium channel CCAF_HUMAN (cacna1f gene) has an even larger basic charge cluster BBXBBXXBXXB(508)XBXBBXXB in the C-terminal of the cytoplasmic loop connecting domains I and II. The R508Q point mutation causes stationary night blindness. It is part of the predicted interaction site with the $G_{\beta\gamma}$ subunits of GTP-binding proteins,[68] predicted helix initiation site, and predicted amphipathic helix. The coincidence of triplet motifs and high potential for amphipathic helix formation near membrane cytoplasmic surface is often observed (manuscript in preparation) and may be simply due to requirement for the conserved basic motif to assume amphipathic helix conformation during protein maturation.

Assignment Errors and Predictor's Performance. Assignment errors are a common problem for a large number of topological models of membrane proteins collected in databases, even in a well-curated database such as the Swiss-Prot (we indeed found some, see Methods). Decreased prediction accuracy in topology models adapted by the Swiss-Prot database in comparison to models based on more experimental evidence (Tables 4 and 6) is likely to be caused in part by systematic and random errors in the Swiss-Prot database.

The assignment errors are risky when overtraining can be achieved and assignment errors memorized during procedures that do not maintain strict separation between training and testing data sets. We were careful in using the Swiss-Prot topological models only during our unrefined training procedure (see Methods). Such a training procedure is considerably less sensitive to assignment mistakes than much more intensive training associated with the neural network or Hidden Markov model procedures. Also, we did not use homologous sequences to the tested one to improve the prediction, and, accordingly, the assignment errors could not become the cause for wrong alignment of homologous sequences and for erroneous prediction.

There is, however, one case when assignment errors are systematic rather than occasional. The TM segment assignments, by different authors using different methods, are optimized to isolate only the membrane-spanning part of a potential TM helix. Transmembrane helices in solved crystal structures are, however, often extended into cytoplasmic space.[27] Our predictor of TM segments is trained on shorter probable TM helices, and because of that reason it produces the underprediction in 52 X-ray solved sequences (987 TM residues underpredicted and 420 overpredicted). On another hand, it can extend potential TM helix, if there is a high potential for creating soluble or amphipathic helix next to membrane-spanning hydrophobic helix.[27] This produces overprediction in the predicted TM helix residues when Swiss-Prot assignments are used as the "standard of truth" (the decrease in the A_{TM} parameter in the Table 6, with respect to SPLIT 3.1 version, is due to overprediction).

The Selection of an "Optimal" 2D Topological Model. In order of decreasing prediction accuracy, topological determinants considered in this work can be listed as triplet motifs bias, positive charge bias, and charge difference across the first TM segment. In order of decreasing frequency the list of determinants is reversed: positive charges, charge difference across the first TM segment, and triplet motifs. Therefore, the best topological predictors are the most specific and the least frequent. One must simply use what is available for topology prediction in each sequence, without relying only on one class of topological determinants. An overall bias factor (see Methods) takes into account all three topological determinants considered in this work, as the selection parameter used to select among different topological models.

Different topological models for the same sequence are created by using 15 different amino acid scales (see Methods) through the preference functions method.[26] We do not claim that our choice of 15 amino acid attributes is optimal. For instance, the dehydration cost of peptide bonds is not accounted for in any of the older hydrophobicity scales that we used. The prediction accuracy may well increase when novel, experiment-based scales[69] are included in the automatic selection process.

In general, functionally important TM helices have characteristic polar residues inside their span, may be partially shielded from lipids by other helices,[70] and are difficult to predict with any hydropathy scale. One possible approach to this problem is a direct recognition of TM segments, such as voltage sensors in voltage-gated channels, with characteristic distribution of polar or charged residues. The second is the usage of amino acid attributes such as optimal amphipathicity (the PRIFT scale[30]), or statistically derived optimal predictors (the Edelman scales[36]), that are not directly related to hydrophobicity. The third is to use Richardson middle helix preferences for helices in soluble proteins[37] and corresponding preference functions[27] when possible underprediction is detected in multispanning proteins based on the prediction with the Kyte-Doolittle scale, as the first scale (see Methods). In multispanning membrane proteins, with some rare exceptions, potential or observed TM helices are found in groups close to one another in a sequence or close to protein N or C terminal. The predictor performs better when it recognizes this clustering tendency.[8]

Overprediction of TM helices is a common problem with sliding-window analysis based only on hydrophobicity.[69] Hydrophobic, but not transmembrane, segments are likely to be found in almost any long sequence. The absolute bias in positive charges/triplets will favor models having more than one hydrophobic TM segment. We corrected for such an overprediction tendency by using the relative bias (bias divided by the number of predicted TM helices) as the selection parameter in the case when a small number of potential TM helices are predicted.

Another approach to decrease the overprediction is to take explicitly into account the avoidance of polar residues, such as the Asn, inside potential TM helix. Asparagine promotes self-association of TM helices,[71] and, possibly because of that reason, it is much less frequent in TM helices than in water-soluble helices. Therefore, some frequency bias between residues in observed TM helices and residues in extramembrane loops or water-soluble helices can also serve, as in our case (see Methods), to lower the overprediction of TM helices.

We did not attempt to optimize the predictor in distinguishing membrane from soluble proteins. The prediction of a single TM helix by only two out of 15 scales (see Methods) is enough to consider the tested sequence as the membrane protein sequence. This stipulation leads to false prediction of TM helices in 10% of soluble proteins, but all membrane proteins from our data sets are correctly identified as such. This error rate can be significantly reduced quite easily, for instance, by requiring that for membrane protein recognition more than half of amino acid attributes used in the automatic run must predict at least one TM segment. Also, we did not train the predictor to distinguish signal sequences from TM segments (work in progress). SPLIT should be used only after the consultation of signal sequence prediction methods[72] and eventual removal of a signal sequence.[73]

The selection based on the highest absolute value of the overall bias factor is usually working well when that bias factor value is considerably greater than the second best value. However, situations when quite different topological models are associated with similar or same bias factor are not so rare, and some other selection mechanisms have to be devised. One is the majority rule when an optimal scale is selected among the majority of scales predicting the same number of potential TM helices. Another is a different definition of a selection parameter. It may well turn out that judicious use of a compositional bias[12] instead of a charge difference,[11] in a definition of an overall-bias, is a better way to approach this problem. Charge difference is less sensitive to predicted topology than other topological determinants (Table 7). Extracting preference functions from selected membrane proteins with novel amino acid scales also leaves room to improve the method.

CONCLUSION

Cytoplasmic location of certain charge clusters in membrane proteins is undoubtedly conserved for a good reason. Basic charge clusters near the cytoplasmic membrane surface are ideally positioned to be involved in protein−protein interactions and signal transmission routes in addition to being the topological determinants. Our predictor selects those topological models of membrane protein that are in accord not only with sliding-window hydrophobicity analysis but also with the obligatory cytoplasmic location of certain basic charge clusters. It can be regarded as an additional tool for analysis and automatic annotation of integral membrane proteins with potential helical TM segments, whose performance will not decrease when sequences homologous to tested sequence are small in number or absent. When tested on the recently collected MPtopo database of the best-characterized membrane proteins[16] SPLIT 4.0 performs better than do four other topology predictors. Cluster counting in SPLIT, combined with the preference functions method,[26] results in the perfect prediction score for 137 TM helices from 36 sequences of exactly known structure (mitochondrial sequences omitted). With mitochondrial sequences included in the complete MPtopo database the SPLIT prediction accuracy for TM helices is about five percentage points better than the TopPred II[45] accuracy. The TopPred II is not the best topology predictor today, but it is still widely used as a hydrophobicity-analysis algorithm with transparent rules, simpler but similar to SPLIT in design.

ACKNOWLEDGMENT

We thank Ana Jerončić for homology tests with the CLUSTALW program. Bono Lučić kindly provided some articles from journals not available at the University of Split. Thanks are also due to Istvan Simon and Alex Tossi for insightful comments and helpful corrections. This work was supported by Grant No. 177060 to Davor Juretić from the Croatian Ministry of Science.

Supporting Information Available: Basic charge motifs near expected membrane surface in 361 nonhomologous membrane proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Krogh, A.; Larsson, B.; von Heijne G.; Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **2001**, *305*, 567−580.
(2) Tusnády, G. E.; Simon, I. Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 364−368.
(3) von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and positive-inside rule. *J. Mol. Biol.* **1992**, *225*, 487−494.
(4) Claros, M. G.; von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **1994**, *10*, 685−686.
(5) Jones, D. T.; Taylor, W. R.; Thornton, J. M. A model recognition approach to the prediction of all- helical membrane protein structure and topology. *Biochemistry* **1994**, *33*, 3038−3049.
(6) Persson, B.; Argos, P. Topology prediction of membrane proteins. *Protein Sci.* **1996**, *5*, 363−371.
(7) Rost, B.; Fariselli, P.; Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **1996**, *5*, 1704−1718.
(8) Tusnády, G. E.; Simon, I. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.* **1998**, *283*, 489−506.
(9) von Heijne, G.; Gavel, Y. Topogenic signals in integral membrane proteins. *Eur. J. Biochem*. **1988**, *174*, 671−678.
(10) Sipos, L.; von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem*. **1993**, *213*, 1333−1340.
(11) Hartmann, E.; Rapaport, T. A.; Lodish, H. F. Predicting the orientation of eukaryotic membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5786−5790.
(12) Nakashima, H.; Nishikawa, K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett.* **1992**, *303*, 141−146.
(13) Juretić, D.; Lee, B. K.; Trinajstić, N.; Williams, R. W. Conformational preference functions for predicting helices in membrane proteins. *Biopolymers* **1993**, *33*, 255−273.
(14) Juretić, D.; Lučić, B.; Zucić, D.; Trinajstić, N. In *Theoretical and Computational Chemistry;* Párkányi, C., Ed.; Elsevier Science: Amsterdam, 1998; Vol. 5, *Theoretical Organic Chemistry*, Chapter 13, p 405−445.
(15) Cserzö, M.; Wallin, E.; Simon, I.; von Heijne, G.; Elofsson, A. Prediction of transmembrane α-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.* **1997**, *10*, 673−676.
(16) Jayasinghe, S.; Hristova, K.; White, S. H. MPtopo: A database of membrane protein topology. *Protein Sci.* **2001**, *10*, 455−458.
(17) Chang, G.; Roth, C. B. Structure of MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **2001**, *293*, 1793−1800.
(18) Möller, S.; Kriventseva, E. V.; Apweiler, R. A collection of well characterised integral membrane proteins. *Bioinformatics* **2000**, *16*, 1159−1160.
(19) Bairoch, A.; Boeckmann, B. The SWISS−PROT proteins sequence bank. *Nucleic Acids Res.* **1991**, *19*, 2247−2249.
(20) Bairoch, A.; Apweiler, R. The SWISS−PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*. **1997**, *25*, 31−36.
(21) Wo, Z. G.; Oswald, R. E. Transmembrane topology of two kainate receptor subunits revealed by N-glycosilation. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 7154−7158.
(22) Bennett, J. A.; Dingledine, R. Topology profile for a glutamate receptor: three transmembrane domains and a channel-lining reentrant membrane loop. *Neuron.* **1995**, *14*, 373−384.

(23) Higgins, D. G.; Thompson, J. D.; Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **1996**, *266*, 383−402.
(24) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23*, 566−579.
(25) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.
(26) Juretić, D.; Zucić, D.; Lučić, B.; Trinajstić, N. Preference functions for prediction of membrane-buried helices in integral membrane proteins. *Computers Chem.* **1998**, *22*, 279−294.
(27) Juretić, D.; Lučin, A. The preference functions method for predicting protein helical turns with membrane propensity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 575−585.
(28) Juretić, D.; Jerončić, A.; Zucić, D. Sequence analysis of membrane proteins with the Web server SPLIT. *Croat. Chem. Acta* **1999**, *72*, 975−997.
(29) Ponnuswamy, P. K.; Gromiha, M. M., Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int. J. Peptide Protein Res.* **1993**, *42*, 326−341.
(30) Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659−685.
(31) Garnier, R.; Robson, B. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York, 1989; Chapter 10, p 417.
(32) Ponnuswamy, P. K.; Prabhakaran, M.; Manavalan, P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* **1980**, *623*, 301−316.
(33) Juretić, D.; Lučić, B.; Trinajstić, N. Predicting membrane protein secondary structure: preference functions method for finding optimal conformational parameters. *Croat. Chem. Acta* **1993**, *66*, 201−208.
(34) Kuhn, L. A.; Leigh, J. S., A statistical technique for predicting membrane protein structure. *Biochim. Biophys. Acta* **1985**, *828*, 351−361.
(35) Engelman, D. M.; Steitz, T. A.; Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **1986**, *15*, 321−353.
(36) Edelman, J. Quadratic minimization of predictors for protein secondary structure: Application to transmembrane α-helices. *J. Mol. Biol.* **1993**, *232*, 165−191.
(37) Richardson, J. S.; Richardson, D. C. Amino acid preferences for specific locations at the ends of α helices. *Science* **1988**, *240*, 1648−1652.
(38) Larsson, H. P.; Baker, O. S.; Dhillon, D. S.; Isacoff, E. Y. Transmembrane movement of the Shaker K^+ channel S4. *Neuron* **1996**, *16*, 387−397.
(39) Juretić, D.; Jerončić, A.; Zucić, D. Prediction of initiation sites for protein folding with α-helix preferences. *Periodicum Biologorum* **1999**, *101*, 339−347.
(40) Kumar, S.; Bansal, M. Geometrical and sequence characteristics of α-helices in globular proteins. *Biophysical J.* **1998**, *75*, 1935−1944.
(41) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot. *J. Mol. Biol.* **1984**, *179*, 125−142.
(42) Zhu, Z.-Y.; Karlin, S. Clusters of charged residues in protein three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8350−8355.
(43) Pogliano, K. J.; Beckwith, J. Genetic and molecular characterization of the *Escherichia coli* secD operon and its products. *J. Bacteriol.* **1994**, *176*, 804−814.
(44) Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **1998**, *14*, 378−379.
(45) Claros, M. G.; von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **1994**, *10*, 685−686.
(46) Milpetz, F.; Argos, P.; Persson, B. TMAP: A new email and WWW services for membrane-protein structural predictions. *Trends Biochem. Sci.* **1995**, *20*, 204−205.
(47) Gavel, Y.; von Heijne, G. The distribution of charged amino acids in mitochondrial inner- membrane proteins suggests different modes of membrane integration for nuclearly and mitochondrially encoded proteins. *Eur. J. Biochem.* **1992**, *205*, 1207−1215.
(48) Obosi, L. A.; Hen, R.; Beadle, D. J.; Bermudez, I.; King, L. A. Mutational analysis of the mouse 5- HT7 receptor: importance of the third intracellular loop for receptor-G-protein interaction. *FEBS Lett.* **1997**, *412*, 321−324.
(49) Wang, H. L. Basic amino acids at the C-terminus of the third intracellular loop are required for the activation of phospholipase C by cholecystokinin-B receptors. *J. Neurochem.* **1997**, *68*, 1728−1735.
(50) Wang, H. L. A conserved arginine in the distal third intracellular loop of the mu-opioid receptor is required for G protein activation. *J. Neurochem.* **1999**, *72*, 1307−1314.
(51) Van Klompenburg, W.; Nilssen, I.; von Heijne, G., De Kruijff, B. Anionic phospholipids are determinants of membrane protein topology. *EMBO J.* **1997**, *16*, 4261−4266.
(52) Liu, D.; Liu, T.; Sy, M. S. Identification of two regions in the cytoplasmic domain of CD44 through which PMA, calcium, and forskolin differentially regulate the binding of CD44 to hyaluronic acid. *Cell Immunol.* **1998**, *190*, 132−140.
(53) Yamane, K.; Akiyama, Y.; Ito, K.; Mizushima, S. A positively charged region is a determinant of the orientation of cytoplasmic membrane proteins in *Escherichia coli*. *J. Biol. Chem.* **1990**, *265*, 21166−21171.
(54) Krishtalik, L. I.; Cramer, W. A. On the physical basis for the cis-positive rule describing protein interaction in biological membranes. *FEBS Lett.* **1995**, *369*, 140−143.
(55) Sato, M.; Mueckler, M. A conserved amino acid motif (R-X-G-R-R) in the Glut1 glucose transporter is an important determinant of membrane topology. *J. Biol. Chem.* **1999**, *274*, 24721−24725.
(56) Pazdernik, N. J.; Jessen-Marshall, A. E.; Brooker, R. J. Role of conserved residues in hydrophilic loop 8−9 of the lactose permease. *J. Bacteriol.* **1997**, *179*, 735−741.
(57) Pazdernik, N. J.; Matzke, E. A.; Jessen-Marshall, A. E.; Brooker, R. J. Roles of charged residues in the conserved motif, G-X-X-X-D/E-R/K-X-G-[X]-R/K-R/K, of the lactose permease of *Escherichia coli*. *J. Membr. Biol.* **2000**, *174*, 31−40.
(58) Tamura, N.; Konishi, S.; Iwaki, S.; Kimura-Someya, T.; Nada, S.; Yamaguchi, A. Complete cysteine-scanning mutagenesis and site-directed chemical modification of the Tn10-encoded metal-tetracycline/H^+ antiporter. *J. Biol. Chem.* **2001**, *276*, 20330−20339.
(59) Okamoto, T.; Nishimoto, I. Detection of G protein-activator regions in M_4 subtype muscarinic cholinergic and α_2-adrenergic receptors based upon characteristics in primary structure. *J. Biol. Chem.* **1992**, *267*, 8342−8346.
(60) Lee, N. H.; Geoghagen, N. S. M.; Cheng, E.; Cline, R. T.; Fraser, C. M. Alanine scanning mutagenesis of conserved arginine/lysine-arginine/lysine-X-X-arginine/lysine G protein-acivating motifs on m1 muscarinic acetylcholine receptors. *Molecular Pharmacol.* **1996**, *50*, 140−148.
(61) Strader, C. D.; Fong, T. M.; Michael, R. T.; Underwood, D.; Dixon, R. A. F. Structure and function of G protein-coupled receptors. *Annu. Rev. Biochem.* **1994**, *63*, 101−132.
(62) Pan, Y.; Metzenberg, A.; Das, S.; Jing, B.; Gitschier, J. Mutations in the V2 vasopressin receptor gene are associated with X-linked nephrogenic diabetes insipidus. *Nat. Genet.* **1992**, *2*, 103−106.
(63) Latronico, A. C.; Anasti, J.; Arnhold, I. J.; Mendonca, B. B.; Domenice, S.; Albano, M. C.; Zachman, K.; Wajchenberg, B. L.; Tsigos, C. A. Novel mutation of the luteinizing hormone receptor gene causing male gonadotropin-independent precocious puberty. *J. Clin. Endocrinol. Metab.* **1995**, *80*, 2490−2494.
(64) Parma, J.; Duprez, L.; Van Sande, J.; Cochaux, P.; Gervy, C.; Mockel, J.; Dumont, J. E.; Vassart, G. Somatic mutations in the thyrotropin receptor gene cause hyperfunctioning thyroid adenomas. *Nature* **1993**, *365*, 649−651.
(65) Splawski, I.; Timothy, K. W.; Vincent, G. M.; Atkinson, D. L.; Keating, M. T. Molecular basis of the long-QT syndrome associated with deafness. *N. Engl. J. Med.* **1997**, *336*, 1562−1567.
(66) Tector, M.; Hartl, F. U. An unstable transmembrane segment in the cystic fibrosis transmembrane conductance regulator. *EMBO J.* **1999**, *18*, 6290−6298.
(67) West, J. W.; Patton, D. E.; Scheuer, T.; Wang, Y.; Goldin, A. L.; Catterall, W. A. A cluster of hydrophobic amino acid residues required for fast Na^+-channel inactivation. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10910−10914.
(68) Herlitze, S.; Hockerman, G. H.; Scheuer, T.; Catterall, W. A. Molecular determinants of inactivation and G protein modulation in the intracellular loop connecting domains I and II of the calcium channel alpha 1A subunit. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 1512−1516.
(69) Jayasinghe, S.; Hristova, K.; White, S. H. Energetics, stability and prediction of transmembrane helices. *J. Mol. Biol.* **2001**, *312*, 927−934.
(70) Lodish, H. F. Multi-spanning membrane proteins: how accurate are the models? *Trends Biochem. Sci.* **1988**, *13*, 332−334.
(71) Choma, C.; Gratkowski, H.; Lear, J. D.; DeGrado, W. F. Asparagine-mediated self-association of a model transmembrane helix. *Nature Struct. Biol.* **2000**, *7*, 161−166.
(72) Nielsen, H.; Brunak, S.; von Heijne, G. Machine learning approaches to the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **1999**, *12*, 3−9.
(73) Möller, S.; Croning, M. D. R.; Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **2001**, *17*, 646−653.

CI010263S