CHAPTER 13

MODELING OF MEMBRANE PROTEINS AND PEPTIDES

Jeffrey Skolnick and Mariusz Milik

INTRODUCTION

Even as the number of solved protein structures has increased exponentially over the last few years, the gap between the number of known protein sequences and structures has continued to grow. This disparity is particularly apparent in the area of membrane proteins, where only a few structures are known at satisfactory resolution.[1] Because of the paucity of solved structures, the use of computer modeling approaches for the prediction of three dimensional structures of proteins from sequence has become increasingly important. There have been a variety of approaches to the protein folding problem. These have ranged from inverse folding or threading approaches which attempt to match a given sequence to a known structure,[2] to de novo approaches that attempt to predict the tertiary structure from sequence information alone.[3-6] As might be expected, inverse folding approaches have met with greater success, but there still are problems with identifying folds of the same topology but random sequence homology.[7] These approaches have been mainly applied to water soluble, globular proteins. In contrast, in this chapter, we focus on preliminary attempts to predict the tertiary structure of membrane proteins where the field is less highly developed, but where the need is no less crucial.

Most modeling efforts have concentrated on the area of water-soluble proteins, where more structural information is known, and it is possible to use statistically derived potentials of mean force to evaluate the relative preference of a sequence for a given conformation.[6] Furthermore, since there is a plethora of structural information available, validation of the approach is straightforward, even if the derivation of suitable potentials is not.

Membrane Protein Assembly, edited by Gunnar von Heijne.

Much less information is available for membrane protein structures, and very few such structures are known at atomic resolution. Bacteriorhodopsin from *Halobacterium halobium* is the first membrane protein whose structure was solved[8,9] and carefully analyzed.[10] The final structure at 3.5 Å resolution became the model for the whole family of membrane proteins containing seven transmembrane helices. Many membrane protein 3-D structure prediction methods use the bacteriorhodopsin structure as a structural template on which the prediction is based. This is true despite the fact that sometimes there is no sequential homology between the bacteriorhodopsin and the protein which is modeled (for example, the models of G-protein coupled receptors[11]). Given that there are so few known membrane protein structures, this approach is very understandable. However, it must be kept in mind that it is equivalent to modeling all helical proteins on the basis of myoglobin, which was the first solved crystal structure of a water soluble protein.[12] While the number of topologies of helical membrane proteins may be more limited than in the case of water soluble proteins, there is no reason to believe a priori that most membrane proteins will be comprised of all helical structures, or that their topology must be that of BR. Alternatively, one may employ a combined method which attempts to build a model consistent with the known experimental restraints. One very promising approach that adopts this viewpoint has been proposed by Herzyk and Hubbard.[13]

A great step forward in the area of membrane protein structure analysis was the solution of the structure of photosynthetic reaction centers from *Rhodopseudomonas viridis*[14,15] and *Rhodobacter sphaeroides*.[16-18] These large complexes of membrane bounded proteins gave new insights into the rules of spatial organization of proteins in membranes. All the transmembrane fragments from protein structures mentioned above were α–helices, which lie more or less perpendicular to the membrane surface. The analysis of these examples led some researchers to the conclusion that most membrane protein helices have a simple, up-down helical topology. The idea that the structure of membrane buried protein fragments is predominantly helical was very popular until the structures of porins were discovered.[19,20] Transmembrane fragments of these proteins are formed by a 16-stranded, antiparallel β-barrel structure. The porin example very clearly shows how dangerous it is to draw general conclusions about membrane protein structure on the basis of very incomplete structural data.

A new type of membrane protein topology was found for the structure of light -harvesting complex from photosystem II, which was solved by Kuhlbrandt et al using electron crystallography.[21] In this predominately helical structure, one of the helices is positioned parallel to the membrane surface. This helix has the additional possibility of forming a membrane bound structure. Prostaglandin H2 synthase, whose X-ray structure was published in 1994,[22] exhibits yet another topology. It is the first example of a monotopic (non-spanning) membrane protein structure. Prostaglandin H2 synthetase is anchored to the membrane by surface adsorbed amphipathic α-helices. Furthermore, nobody can exclude the possibility that in the near future a structure of membrane protein with an α/β topology will be solved. In summary, the problem of prediction of topology and 3-D structure of membrane proteins turns out to be more difficult than one would expect on the basis of the initially solved structures.

Most work in the area of membrane protein structure prediction concerns the prediction of transmembrane helical fragments using sequence analysis methods.[23-26] This approach implicitly assumes a simple up-down helical topology for membrane proteins. Some researchers concentrate on the prediction of the point at which a transmembrane helix leaves the bilayer.[27] Other investigators use the information

about the asymmetry of distribution of charged residues between internal and external loops in membrane proteins[28-31] to predict not only the position of helices, but also their orientation (N-terminus inside or outside of the cell).[32-35]

Hofmann and Stoffel[32] have prepared a database of known membrane spanning protein segments (TMbase). The database not only contains sequences of putative transmembrane helices, but also has a lot of additional information concerning the type of membrane the protein is associated with, the orientation transmembrane fragments in membranes, etc. Statistical analysis of the information from this database was used by a number of authors as a basis for a sequence threading algorithm for the prediction of location and orientation of transmembrane fragments in helical membrane proteins.[33-36]

Using statistical information about preferences of the various amino acids for particular regions of the membrane, Jones and coworkers[33] have analyzed all possible topologies of membrane spanning protein of interest for the most probable topology. The method was tested on the set of 37 known multispanning helical membrane proteins and was shown to be very accurate, with 34 of 37 correctly predicted topologies.

Persson and Argos[34,35] made an analysis of the frequency of occurrence of amino acids in intra- and extracellular fragments of transmembrane proteins. The technique relies on the differences in amino acid composition in intra- and extracellular regions. They found that Asp occurs more often in extracellular regions, while Glu is indifferent, according to their statistics. They also find that Asn, Asp, Gly, Phe, Pro, Trp, Tyr and Val prefer the extracellular region. In contrast, Lys, Arg, Ala and Cys are more common on the intracellular side of the membrane. Additionally, they used the information from a multiple alignment method for known families of membrane proteins to improve the prediction from the analysis of the distribution of amino acids in transmembrane fragments and loops.[34] They showed that including this additional information results in about a 10% improvement in accuracy.

In the area of membrane protein modeling, Milik, Skolnick and coworkers have presented a model for membrane bound peptides and proteins.[37,38] The model represents the first stage of a larger project whose goal is to develop a method for simulating complex membrane-protein systems. In the initial stages of the project, separate models of the lipid membrane and the membrane protein were prepared and extensively explored. The lipid membrane originated from a simple model of chains anchored to an interface.[39] Subsequently, a united atom, Monte Carlo model of the lipid bilayer was developed.[40] The model reflects the internal molecular geometry of lipid molecules and the essential, physical properties of real lipid membranes. Using this model, the authors were able to reproduce the gel/liquid phase transitions and the lateral diffusion of phospholipids in the lipid bilayer. Additionally, the existence of a free-diffusion regime outside of the bilayer phase was confirmed by the results of the simulations.

The model of structure and dynamics of membrane proteins started from a simplified lattice representation.[41] The polypeptide molecule was represented by a diamond lattice chain, where every amino acid consisted of three consecutive diamond lattice points representing the NH, Cα plus side chain, and carbonyl group, respectively. The water and the hydrocarbon chain environments were taken into account via an effective, coordinate dependent hydrophobic potential. The model demonstrated that changes in the sequence of the model peptide chain may change the pathway of the helical hairpin insertion from "end first" to "turn first." Additionally, the results of the simulations support the idea of structural subassembly during the process of spontaneous insertion of a protein into a membrane. During the modeled insertion process, the fragments of

secondary structure were preassembled on the surface of the membrane, and then transported into the membrane. This effect is consistent with the conjecture that the interface may play a very important role in the spontaneous insertion process.[42]

The importance of the interface region was even more evident in the next, more detailed model of membrane proteins.[37] In this model, the protein chain was represented by a chain of balls with centers at the Cα carbon position. The internal structure of membranes was also schematically considered in this model. The lipid membrane phase was divided into head-group and hydrocarbon phases. The interactions between a protein chain and environment were modeled by a new hydropathy scale, based on experimental data. In spite of its simplicity, the model reproduced the experimental behavior of a number of systems. For example, it successfully distinguished between transbilayer (M2δ) and surface adsorbed (magainin2) helical peptides.[37] The model also gave a good prediction for the transbilayer and surface-adsorbed fragments of the bacteriophage coat proteins. The predicted structures of pf1 and fd coat proteins are in very good agreement with the experimental data for the positions and orientation of the residues in the membrane bounded system.[38]

These simulations confirmed the existence of a two stage insertion process for the spontaneous insertion of membrane peptides that was originally proposed by Jacobs and White[42] and is known as the "helix insertion" hypothesis. In the first stage, the peptide chain adsorbs on the membrane surface with accompanying formation of helical structure. The resulting preformed secondary structure fragment is then inserted into the membrane. This occurs because the energetic cost of not having any hydrogen bonds in the bilayer phase is very large. Thus, these simulations have provided new insights into the process of membrane peptide transport and stability.

Subsequently, the role of the internal structure of the membrane in the insertion process was explored by Baumgartner and Skolnick.[43,44] They used a dumbbell model for the molecules in each leaflet of the bilayer. In an initial study,[43] they examined the translocation of an inert polymer which is driven across the model membrane as a function of field strength. Below a characteristic field strength, they find that the membrane is practically impenetrable. Above this threshold, but at low fields, translocation can be described as a Kramers process associated with the escape of a Brownian particle over potential barriers. Next, they examined the effect of membrane curvature on the nature of the translocation process.[44] For flat membrane models, they find that a structureless polymer having attractive interactions with the bilayer spontaneously crosses the bilayer with equal probability in each direction. This is reasonable, since space is isotropic, and there are no field gradients across the bilayer. However, when the bilayer is highly curved, then there is essentially irreversible transport from the outside to the inside of the bilayer. This is due to an entropic effect arising from the fact that the bilayer is less dense in the interior leaflet than in the exterior leaflet. The importance of these studies resides in the fact that there may be situations where the internal structure of the membrane is important and important physical effects may be missed if it is simply treated as a structureless medium.

With the above caveat in mind, at the very least, the membrane peptide model must be extended to include a more realistic representation of amino acid side chains, a side chain-side chain interaction potential and a membrane-protein interaction potential. Here, we describe initial investigations along this direction and describe a method based on the approach of Jones et al[33] for extracting information about protein-membrane interactions from protein sequence data. The objective is to develop a mean-force asymmetric potential that includes spatial information which was absent in our original membrane protein models and to examine the consequence when such a potential is implemented in

the simplified protein model. Having such a potential in hand, we then explore whether the additional terms in the potential produce results which are consistent with experiments regarding to the location of the helices with respect to the bilayer. We also wish to establish that the inclusion of an amino acid specific, spatially anisotropic potential is compatible with other terms in the potential. These simulations offer the advantage that the helical locations are not encoded a priori, but may change. In fact, an amphipathic helix may lie on the surface of the membrane. They suffer from the disadvantage that the computational cost is substantially larger than more standard sequence based approaches.

The first problem which should be addressed in order to prepare a statistical potential that describes protein-membrane interaction is that the membrane itself is not an amorphous, isotropic system; rather, it has a complicated internal structure.[45,46] Lipids are asymmetrically distributed between the inside and outside leaflet of the membranes. For example, in mammalian erythrocyte membranes, phosphatidylcholine (PC) and sphingomyelin (Sph) molecules are preferentially concentrated in the outside leaflet of the membrane, while phosphatidylethanolamine (PE) and phosphatidylserine (PS) prefer the inside leaflet of the membrane.[47] An analogous situation also obtains in the case of plasma membranes, but PC is rather equally distributed between the leaflets.[47] This asymmetry in lipid composition is probably important for the cell function, because, according to experimental data, it is probably artificially maintained by a specialized apparatus.[45]

One possible explanation why membrane asymmetry is so essential for living cells is that this affects both the structure and function of membrane proteins. Some membrane proteins may be activated or inhibited by specific phospholipids. For example, according to experimental data, protein kinase C requires PS for its function, and b-hydroxybutyrate dehydrogenase requires PC.[46] A typical cell membrane has a ratio of 100 phospholipid molecules per membrane protein molecule.[46] Thus, every membrane bound protein is surrounded by several shells of phospholipid molecules. One can expect that the structure and orientation of a membrane protein must depend on the phospholipid type in the neighborhood. In particular, it may be sensitive to the chemical identity of the head-groups.

The asymmetry in the composition of loops exterior to the membrane in membrane protein sequences was discovered by von Heijne and co-workers.[48-51] Based on these observations, they formulated the "positive inside" rule. The rule reflects the observation that positively charged amino acids (Lys and Arg) are comparatively rare in periplasmic loops (about 5%) but are very common in cytoplasmic loops (about 15%).[31] This empirical rule has already been successfully used to predict the topology of membrane proteins.[31,33,35,52]

A suitable membrane-protein model should account for the internal membrane structure. However, the detailed treatment of the membrane makes the problem computationally intractable. Thus, we propose to develop a statistical or knowledge based potential that reflects the differential preferences of the various amino acids for the different regions of the membrane. In the case of globular protein structure prediction, many methods use information extracted from already known protein structures in the derivation of knowledge based potentials for protein folding.[6,53] Unfortunately, in the case of helical membrane proteins, there are too few known structures to make this practical. However, in the case of membrane proteins, we have the advantage that the helical transmembrane fragments are confined to the phospholipid bilayer boundaries and may be used to extract information about their environment. Using a set of transmembrane helices with known orientation, it should be possible to find some statistical regularities in the spatial distribution of amino acids.

STATISTICAL ANALYSIS OF TRANSMEMBRANE HELIX SEQUENCES

Preparation of the Database of Transmembrane Helices

The fragments of protein sequences containing the transmembrane helices were extracted from the "TMbase" which is a database of membrane spanning protein segments.[32] The database contains transmembrane sequences extracted from Swiss Prot release 25 and includes some additional information about the orientation of the helices and the type of membrane (when known). The orientation was deduced from Swiss Prot annotations or from the positions of glycosylation and phosphorylation sites.

In the present work, we concentrated on membrane proteins from mammalian plasma membranes and proceed in a fashion that is similar in spirit to that of Jones et al.[33] This greatly decreased the size of our data base, but it increases the coherence of our results. We expect that helical fragment spanning the same (or at least, very similar) membranes might give us some insight into specific protein-phospholipid interactions. Additionally, we filtered

Table 13.1. Names of proteins whose sequences were used in the transmembrane fragment database

Swiss Prot Name	Chosen Helices	Swiss Prot Name	Chosen Helices
5HT2_HUMAN	1 2 3 4 5 6 7	5HT3_HUMAN	1 2 3 4
5HTA_HUMAN	1 2 3 4 5 6 7	5HTB_MOUSE	1 2 3 4 5 7
5HTC_HUMAN	1 2 4 5 6 7	5HTD_HUMAN	1 2 3 4 5 6 7
5HTE_HUMAN	1 2 4 5 6 7	5HTE_MOUSE	1 3 6 7
5HTX_HUMAN	1 6 7	A1AA_HUMAN	1 2 4 5 6 7
A1AA_RAT	1	A1AB_RAT	1 2 3 4 5 6 7
A1AC_BOVIN	1 2 3 4 7	A2AA_HUMAN	1 2 3 4 5 6 7
A2AB_HUMAN	1 4 5 6 7	A2AB_MOUSE	1
A2AC_HUMAN	4	AA1R_BOVIN	1 2 3 4 5 6 7
AA2A_HUMAN	1 3 7	AA2B_HUMAN	1 4 7
ACH2_RAT	1 2 3 4	ACH3_RAT	4
ACH5_HUMAN	2 4	ACH5_RAT	4
ACHA_BOVIN	3 4	ACHB_BOVIN	2
ACHB_HUMAN	4	ACHD_BOVIN	2 4
ACHD_MOUSE	4	ACHE_BOVIN	2 3 4
ACHG_BOVIN	4	ACHG_HUMAN	2 3
ACHN_HUMAN	4	ACM1_HUMAN	1 2 3 4 5 6 7
ACM2_HUMAN	1	ACM_HUMAN	1
AG2R_BOVIN	1 2 3 4 5 6 7	B1AR_HUMAN	1 2 4 5 7
B2AR_HUMAN	4	B3AR_HUMAN	2 3 4 5 6 7
B3AR_MOUSE	1	BRB2_HUMAN	1 2 3 4 5 6 7
C5AR_HUMAN	1 2 3 4 5 6 7	CALR_PIG	1 2 3 4 5 6 7
CANR_HUMAN	1 2 3 4 5 6 7	D2DR_BOVIN	1 3 4 5 6
D3DR_MOUSE	2 6 7	D4DR_HUMAN	1 2 4 5 6 7
D5DR_HUMAN	1 2 4 5 6 7	DADR_HUMAN	1 3 4 5 6 7
DBDR_RAT	4 5 6 7	EDG1_HUMAN	1 2 3 4 5 6 7
ET1B_BOVIN	1 2 3 4 5 6 7	ET1R_BOVIN	1 2 4 5 6 7
FML1_HUMAN	1 2 3 4 5 6 7	FMLR_HUMAN	6
FSHR_HUMAN	1 2 3 4 5 6 7	GAA1_BOVIN	1 2 3 4
GAA4_BOVIN	1 3 4	GAA6_MOUSE	3
GAB1_BOVIN	2 3	GAC2_BOVIN	1

out the chosen transmembrane helices in order to remove very similar sequences. In the remaining set, no pair of transmembrane helix sequences is more than 50% identical. Table 13.1 presents the set of chosen mammal plasma membrane proteins with information about which transmembrane helices from these proteins are included in our data base. The resulting database consists of 484 transmembrane fragments extracted from 112 mammal proteins. Additionally, to all membrane spanning sequences, we added short (5 residues) flanking sequences from the N and C termini of the helices. These flanking sequences were also used in our statistical analysis.

Figure 13.1 presents a histogram of the length of transmembrane fragments in our database. Extremely long (greater than 26 residues) or short helices (shorter than 19 residues) were removed. After this procedure, the database consisted of 447 transmembrane fragments.

STATISTICAL ANALYSIS

As stated above, experiments indicate that the distribution of different types of phospholipids in cell membranes is asymmetric. The outside leaflet of the lipid bilayer from the plasma membrane contains more Sph, while the inside leaflet is richer in PE. The main difference between these phospholipids is in the shape and physicochemical features of their head groups. Thus, we expect that most of the information about the asymmetrical distribution of amino acids in membrane spanning fragments will be contained in the fragments of sequence closest to the head group fragment of the membrane—near the termini of the transmembrane fragments.

Table 13.1. (continued)

Swiss Prot Name	Chosen Helices	Swiss Prot Name	Chosen Helices
GAC2_BOVIN	1	GAD_MOUSE	3
GAR1_HUMAN	1	GCRB_RAT	1 2 3 4 5 6 7
GCRT_RAT	1 2 3 4 5 7	GLGP_RAT	1 2 3 4 5 6 7
GLPR_RAT	1 2 3 4 5 6 7	GRA1_HUMAN	1 2 3 4
GRA3_RAT	3	GRB_RAT	1 2 3
GRPR_HUMAN	1 2 4 5 6 7	HH2R_HUMAN	1 2 4 5 6 7
IL8A_HUMAN	1 2 3 4 5 6 7	IL8A_RABIT	1 2 3 4 5 6
IL8B_HUMAN	5 7	LSHR_HUMAN	1 2 3 4 5
LSHR_MOUSE	1	LSHR_PIG	7
MAS_HUMAN	1 2 3 4 5 6 7	NK1R_HUMAN	1 2 3 4 5 6 7
NK2R_BOVIN	1 2 5 6 7	NK2R_RAT	4
NMBR_HUMAN	1 4 5 7	NTR_RAT	1 2 3 4 5 6 7
NY1R_HUMAN	1 2 3 4 5 6 7	NY3R_BOVIN	1 2 4 5 6 7
OLF0_RAT	1 2 3 4 5 6 7	OLF2_RAT	4 5
OLF3_MOUSE	1 4 5 6 7	OLF3_RAT	1 4 5
OLF5_RAT	4 5	OLF6_RAT	1 2 4 5 6 7
OLF7_RAT	1 3 4 5 7	OPSB_HUMAN	1 2 3 4 5 6 7
OPSD_BOVIN	1 2 4 5 6 7	OPSG_HUMAN	1 2 3 4 5 6
OXYR_HUMAN	1 2 3 4 5 6 7	PAFR_HUMAN	1 2 3 4 5 6 7
PTRR_RAT	1 4 5 6 7	RDC1_HUMAN	1 2 3 4 5 6 7
RTA_RAT	1 2 3 4 5 6 7	SCRC_RAT	5
TA2R_HUMAN	1 2 3 4 5 6 7	THRR_HUMAN	1 2 3 4 5 6 7
THRR_MOUSE	7	TRFR_MOUSE	1 2 3 4 5 6 7
TSHR_HUMAN	5	V1AR_RAT	1 2 4 5 6 7
V2R_HUMAN	1 2 4 5 7	VIPR_RAT	4

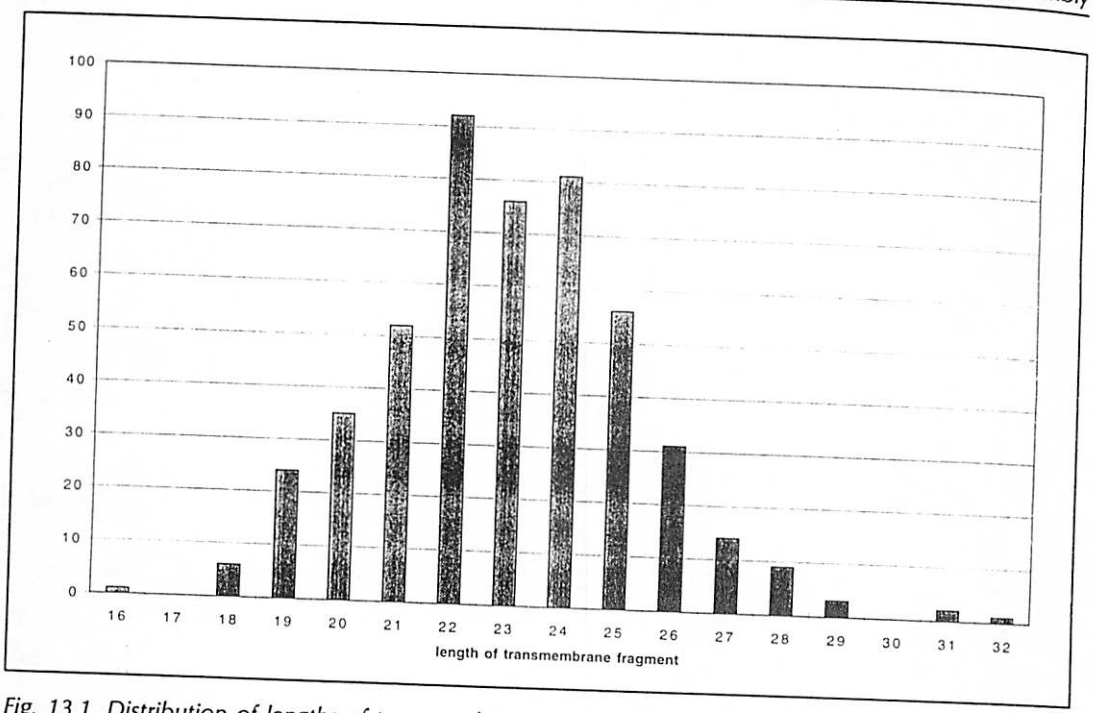


Fig. 13.1. Distribution of lengths of transmembrane helices used in the present work. The helices were extracted from mammal plasma membrane proteins.

From the sequence database, we extracted information about the interface regions and the central region of the cell membrane. Thus, as suggested by Jones et al,[33] we specify five regions: inside head group (HDI), inside hydrocarbon (HCI), central hydrocarbon (HCC), outside hydrocarbon (HCO) and outside head group (HDO). As shown in Figure 13.2, three residues comprise the inside and outside regions (about one turn of an α-helix) and five residues comprise the central region (see Fig. 13.2). With this spatial division, the question we asked was: is there any preference in the distribution of amino acids in the membrane spanning sequences chosen by us among these five regions?

To answer this question, we first calculated the frequencies of amino acids in the selected regions of the sequence. Figure 13.3 presents histograms of the calculated frequencies in comparison with the frequencies of amino acids in the whole Swiss Prot sequence data base. As expected, the transmembrane fragments are enriched in hydrophobic amino acids (Ile, Leu, Val) and contain less hydrophilic ones (Asp, Glu, Lys, Gln). Note that the frequencies of Asn and Arg in the transmembrane fragments are close to the average frequencies in the entire Swiss Prot data base. That is, in contrast to the other polar residues, their relative frequency is not substantially suppressed in membrane proteins. This may suggest that these residues play a specific role in transmembrane fragments.

Having in hand the expected values of occurrences of the individual amino acid in the transmembrane fragments, we calculated for all amino acids and all sequence fragments the ratio:

$$d^{AA}(r) = \frac{N^{AA}_{obs}(r)}{N^{AA}_{\exp}(r)} \qquad (1)$$

where: $d^{AA}(r)$ is the ratio for amino acid type AA in region r; $N^{AA}_{obs}(r)$ is the number of the observed occurrences of amino acid AA in region r; $N^{AA}_{\exp}(r)$ is the number of the expected occurrences of amino acid AA in region r and calculated from statistics for transmembrane fragments and size

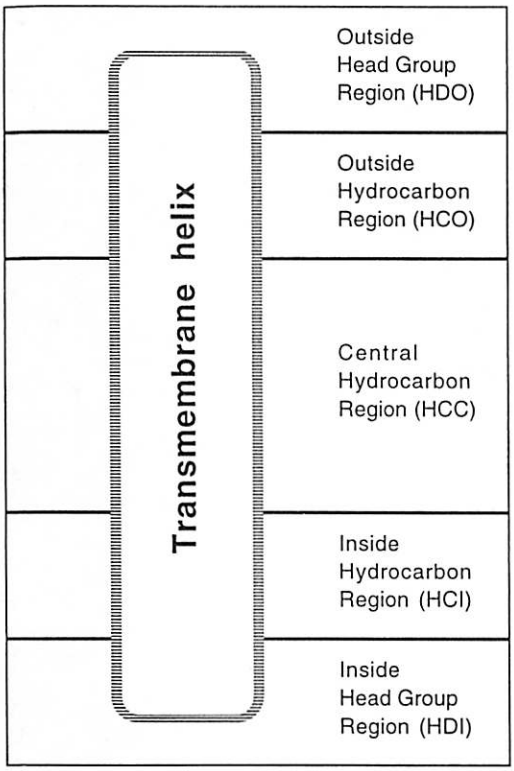


Fig. 13.2. Schematic view of the membrane divided into five regions.

of the region. That is, we simply calculate the ratio, q(r), of the length of region r to the total length of all five regions. $N^{AA}_{exp}(r)$ is just the total number of residues of type AA times q(r).

Values of $d^{AA}(r)$ for all regions and amino acid types are presented in Table 13.2. Graphically, a portion of the data are presented as histograms in Figure 13.4a-e. Figure 13.4a presents the ratios for the five different regions for the hydrophobic amino acids—Ile, Leu, Val and Phe. As one might expect, these amino acids are mostly concentrated in the hydrocarbon regions; the differences in ratios between inside and outside leaflets of the hydrocarbon portions of the membrane are on the level of the statistical error.

Figure 13.4b presents analogous histograms for positively charged amino acids: Lys, Arg and His. Again, as expected, these residues are concentrated in the head-group regions and the distribution is asymmetric, particularly for Arg, where ratio in the region HDI is almost two times larger than in the region HDO and to a lesser extent

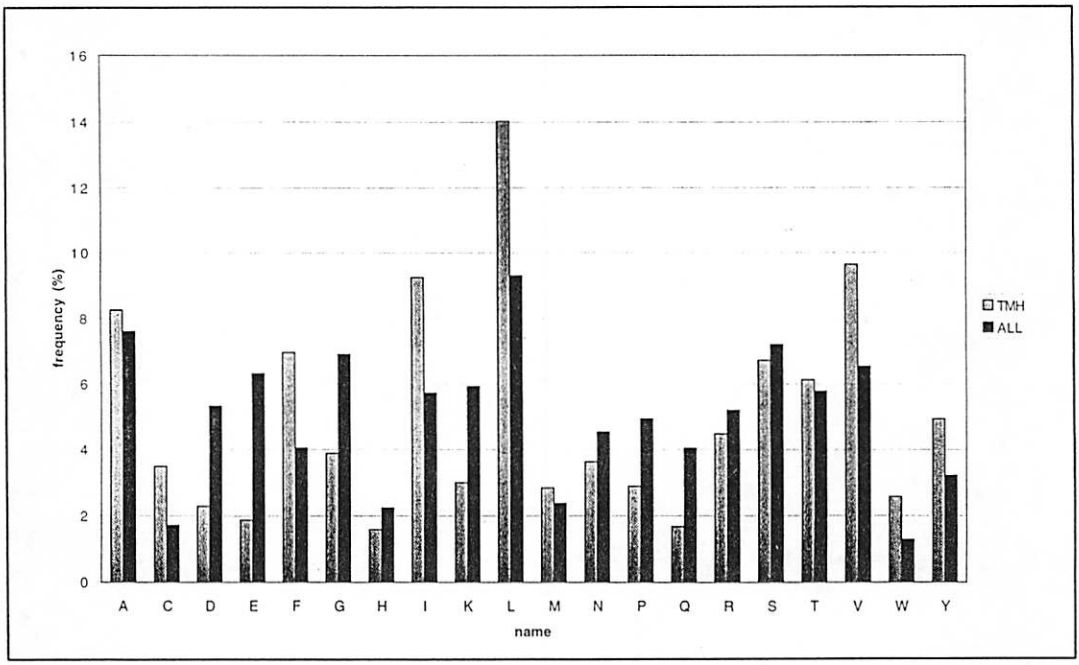
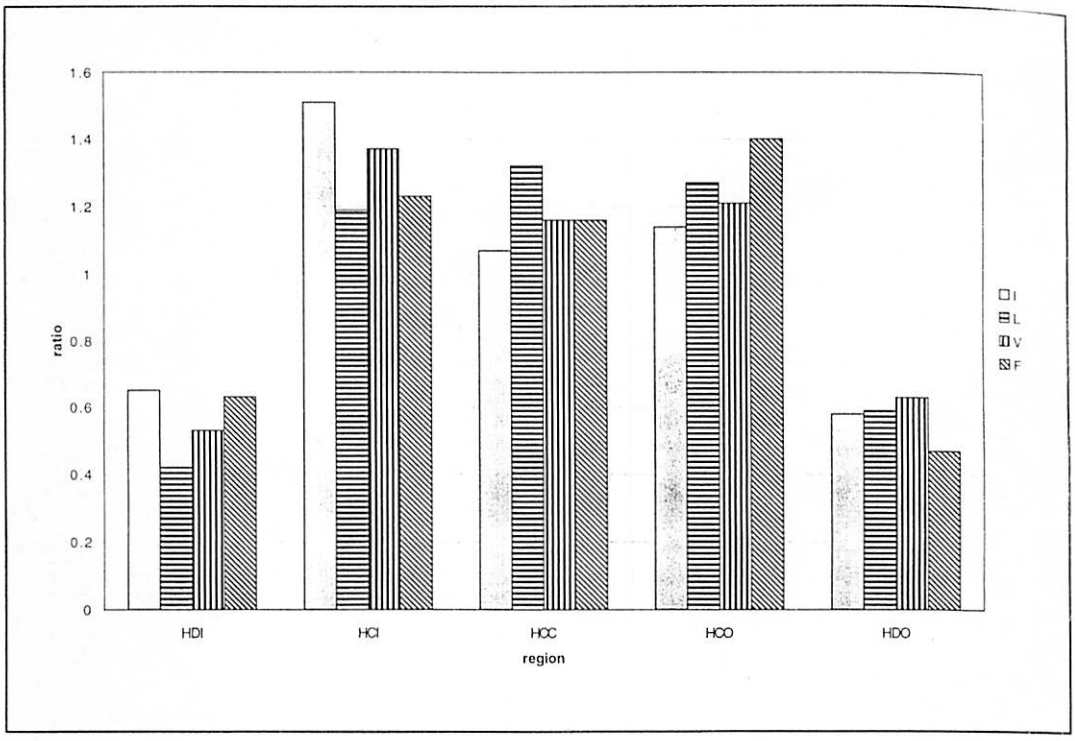
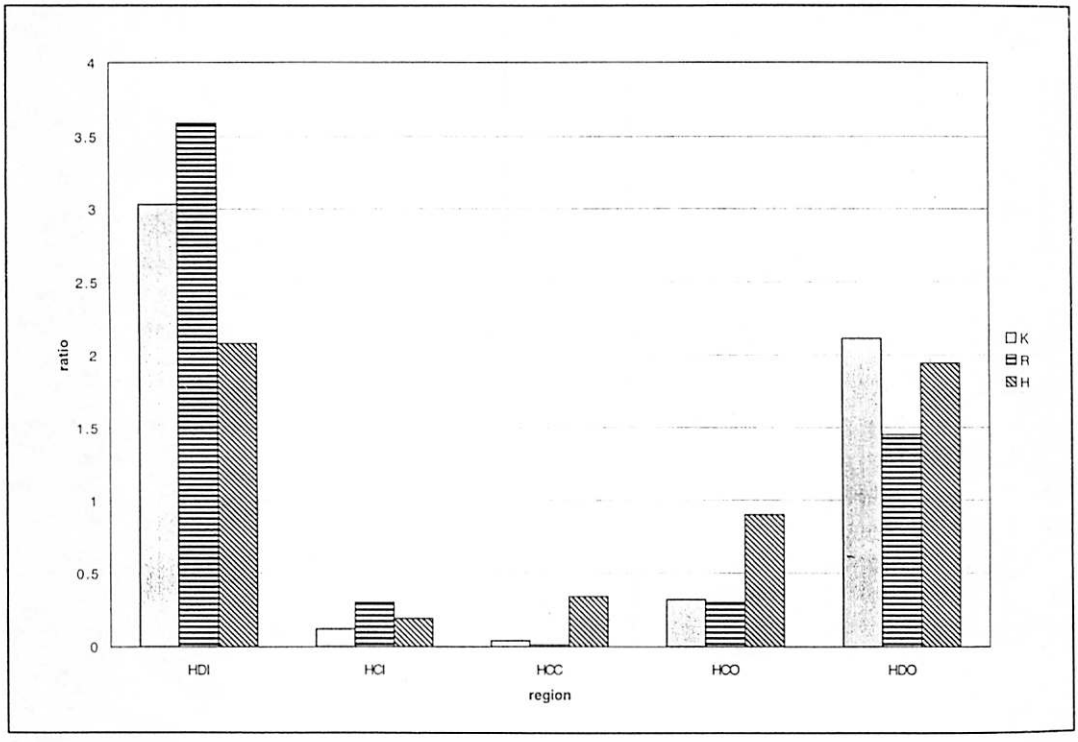


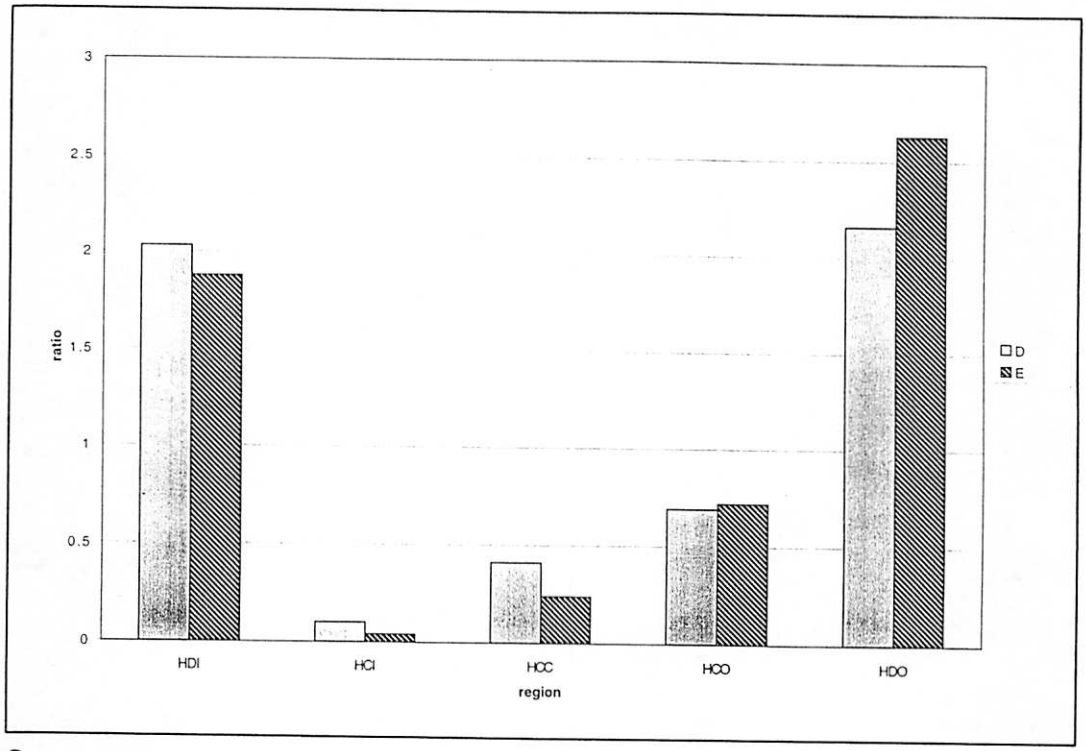
Fig. 13.3. Frequencies of amino acids in transmembrane helices in comparison to average frequencies in the Swiss Prot database.


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
ratio
HDI
HCI
HCC
HCO
HDO
region
I
L
V
F

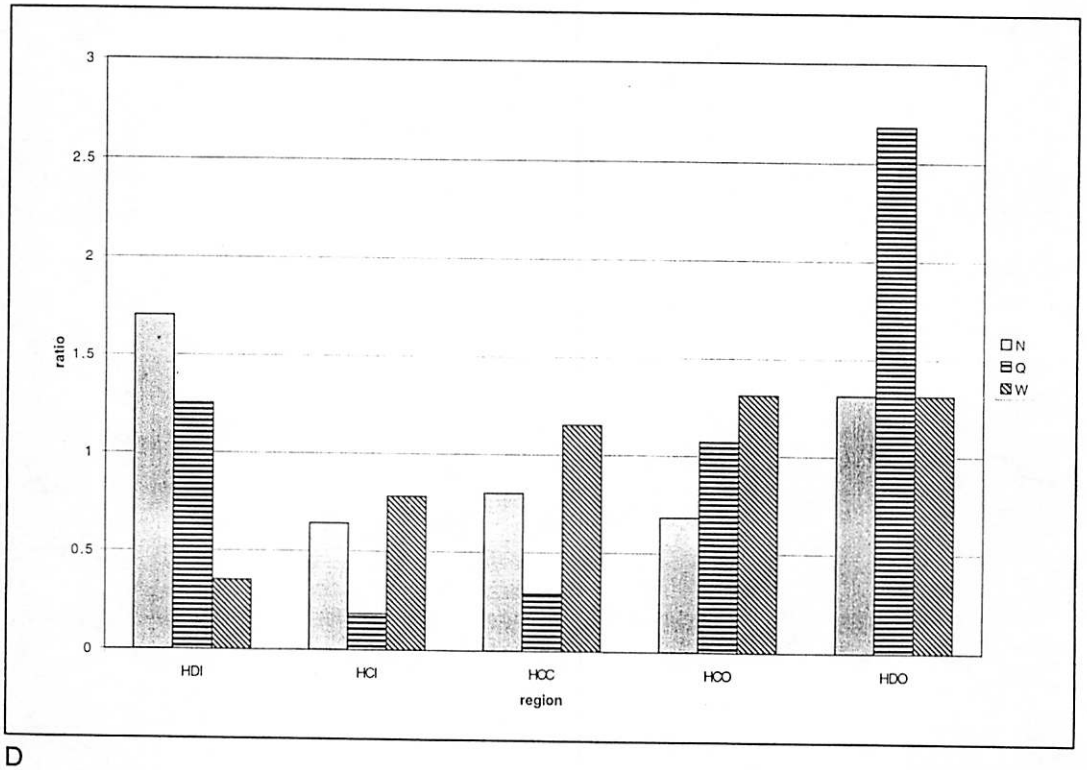
A


4
3.5
3
2.5
2
1.5
1
0.5
0
ratio
HDI
HCI
HCC
HCO
HDO
region
K
R
H

B

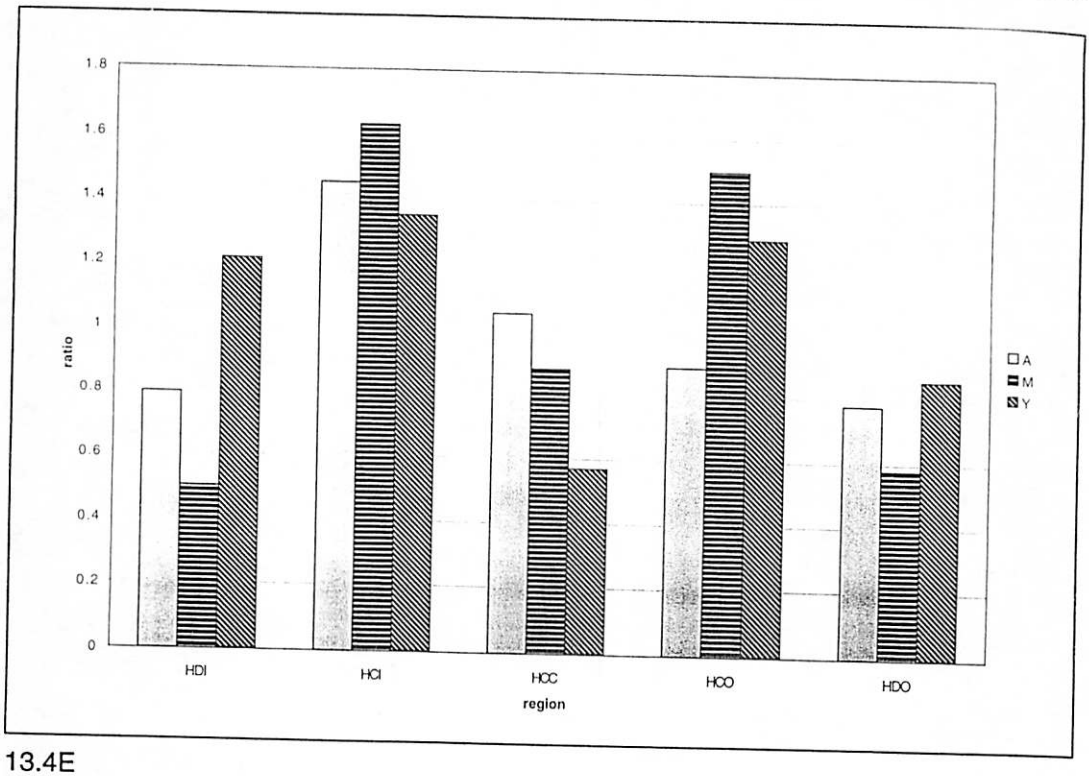


C



D

Fig. 13.4. Values of ratios $d^{M}(r)$ (see text for description) of different types of amino acids in membrane regions. (A) hydrophobic residues; (B) positively charged residues; (C) negatively charged residues; (D) noncharged, polar residues; (E, on following page) "membrane surface" residues.



13.4E

for Lys. In contrast, for His the differences are on the level of the statistical error.

Histograms for the negatively charged amino acids Asp and Glu are presented in Figure 13.4c. The distribution is again asymmetric for Glu, but this time with a preference for the outside part of the membrane. In the case of Asp, the differences are on the level of the error of the method. The overall trend for the positively and negatively charged residues supports the "positive inside" rule (discussed above), which was originally formulated for the outside of membrane loop fragments of membrane proteins. Our statistical data show that this rule is valid for the residues from the head-group region as well. However, in the case of His and Asp, the rule may hold weakly at best.

Figure 13.4d shows histograms for the polar, noncharged residues: Asn, Gln and Trp. The histograms show a distinct difference in the Gln and Asn distribution in transmembrane helices from plasma membranes. While the Asn distribution is almost symmetric, the Gln residues are strongly concentrated in the HDO region of the membrane. If this is not an artifact of the small size of our data base, this may provide an interesting insight into the specificity of protein-membrane interactions. This observation is consistent with that of Persson and Argos who used a set of transmembrane fragments in different environments.[35] Another interesting observation is the distribution of Trp residues. The statistics suggest that Trp has a preference for "outside" side of the plasma membrane. This tendency is also consistent with the work of Persson and Argos.[35]

Figure 13.4e shows histograms of ratios for residues with a tendency to concentrate in the hydrocarbon phase of membrane, but close to its border (regions HCI and HCO), namely Ala, Met, and Tyr. This effect is most evident for the Met and Tyr side chains.

Table 13.2. Numbers of observed ($N_{obs}^{AA}(r)$) and expected ($N_{exp}^{AA}(r)$) occurrences of amino acids in different regions of the membrane

	HDI			HCI			HCC			HCO			HDO		
	N_{exp}	N_{obs}	Ratio	N_{exp}	N_{obs}	Ratio	N_{exp}	N_{obs}	Ratio	N_{exp}	N_{obs}	Ratio	N_{exp}	N_{obs}	Ratio
A	110.8	88	0.79	110.8	161	1.45	184.7	194	1.05	110.8	99	0.89	110.8	86	0.78
C	46.8	21	0.45	46.8	53	1.13	77.9	121	1.55	46.8	22	0.47	46.8	48	1.03
D	30.5	62	2.03	30.5	3	0.10	50.9	21	0.41	30.5	21	0.69	30.5	66	2.16
E	25.1	47	1.88	25.1	1	0.04	41.8	10	0.24	25.1	18	0.72	25.1	66	2.63
F	93.2	59	0.63	93.2	115	1.23	155.3	180	1.16	93.2	130	1.40	93.2	44	0.47
G	52.1	33	0.63	52.1	34	0.65	86.8	120	1.38	52.1	56	1.08	52.1	52	1.00
H	21.2	44	2.08	21.2	4	0.19	35.3	12	0.34	21.2	19	0.90	21.2	41	1.94
I	124.1	81	0.65	124.1	187	1.51	206.8	222	1.07	124.1	141	1.14	124.1	72	0.58
K	40.2	122	3.03	40.2	5	0.12	67.1	3	0.04	40.2	13	0.32	40.2	85	2.11
L	187.9	79	0.42	187.9	223	1.19	313.2	413	1.32	187.9	239	1.27	187.9	111	0.59
M	38.1	19	0.50	38.1	62	1.63	63.5	56	0.88	38.1	57	1.50	38.1	22	0.58
N	48.7	83	1.70	48.7	31	0.64	81.2	65	0.80	48.7	33	0.68	48.7	64	1.31
P	38.6	31	0.80	38.6	8	0.21	64.4	70	1.09	38.6	35	0.91	38.6	75	1.94
Q	22.4	28	1.25	22.4	4	0.18	37.3	11	0.29	22.4	24	1.07	22.4	60	2.68
R	59.8	215	3.59	59.8	18	0.30	99.7	1	0.01	59.8	18	0.30	59.8	87	1.45
S	89.8	65	0.72	89.8	70	0.779	149.7	231	1.54	89.8	57	0.64	89.8	86	0.96
T	81.9	103	1.26	81.9	69	0.84	136.5	127	0.93	81.9	72	0.88	89.8	93	1.14
V	129.4	69	0.53	129.4	177	1.37	215.6	249	1.16	129.4	157	1.214	129.4	81	0.63
W	34.4	12	0.35	34.4	27	0.78	57.4	66	1.15	34.4	45	1.31	34.4	45	1.31
Y	66.0	80	1.21	66.0	89	1.35	110.0	63	0.57	66.0	85	1.29	66.0	57	0.86

See the text for explanation.

PREPARATION OF MEAN FORCE POTENTIALS

The mean-force potentials were prepared from the statistical data using the supposition of a Boltzmann distribution of energetic states. The values of the one-body potential for different regions and different amino acids were calculated according to the equation:

$$p^{AA}(r) = -kT\ln(d^{AA}(r)) = -kT\ln\frac{N^{AA}_{obs}(r)}{N^{AA}_{exp}(r)} \quad (2)$$

where: $p^{AA}(r)$ represents the value of the asymmetrical membrane potential for amino acid "AA" in the region "r", and $d^{AA}(r)$ is defined in Equation 1.

Table 13.3 presents the values of the potential calculated using equation 2. For most amino acids, the values of this potential are insignificant in comparison to the values of the hydrophobic potential used in the simulation procedures. For amino acids like Asp and Glu, the difference between the "inside" and "outside" leaflets of the membrane can be more than 1 kcal/mol. For Lys, Arg and Trp, these differences are about 0.5 kcal/mol. These numbers are indicative of the moderate to relatively strong preferences for these amino acids to be in the outside or inside leaflets of the membrane.

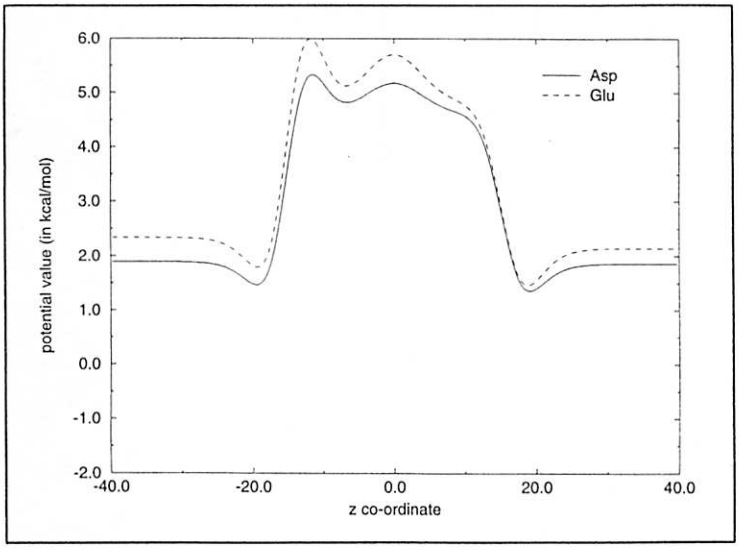
This coordinate dependent potential was then included as an additional energy factor into the simulation scheme previously employed in our previous work.[37] In effect, the z-dependent hydrophobic membrane potential is now asymmetrical, as is seen in the plots for the potential for Asp, Glu, Arg, Lys, Phe and Trp (Fig. 13.5).

SIMULATIONS

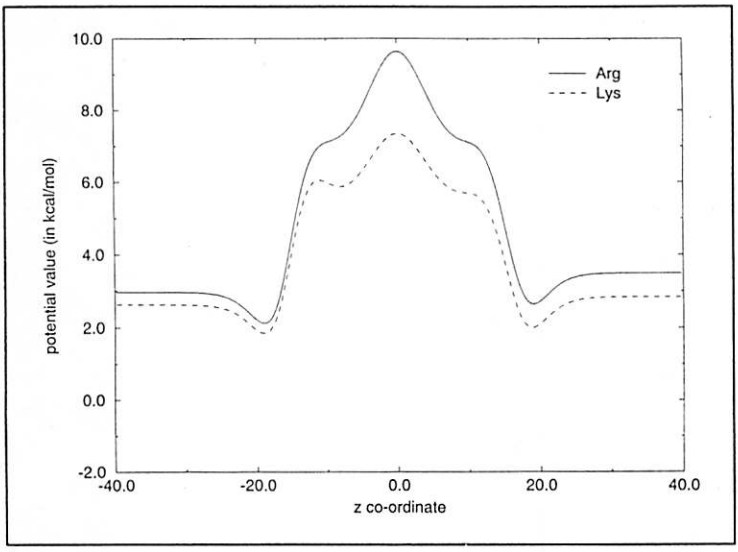
The resulting potential field was then tested on the example of a membrane protein from the human plasma membrane: interleukin-8 receptor A (IL8A_HUMAN) from the family of G-protein coupled receptors (GPCR).[54] The sequence was chosen because of the well-known overall membrane topology of GPCRs (positions of transmembrane helices and their orientations). In test runs, we have used transmembrane fragments from IL8A, extracted

Table 13.3. Values of asymmetric membrane potential for different membrane regions and different amino acids

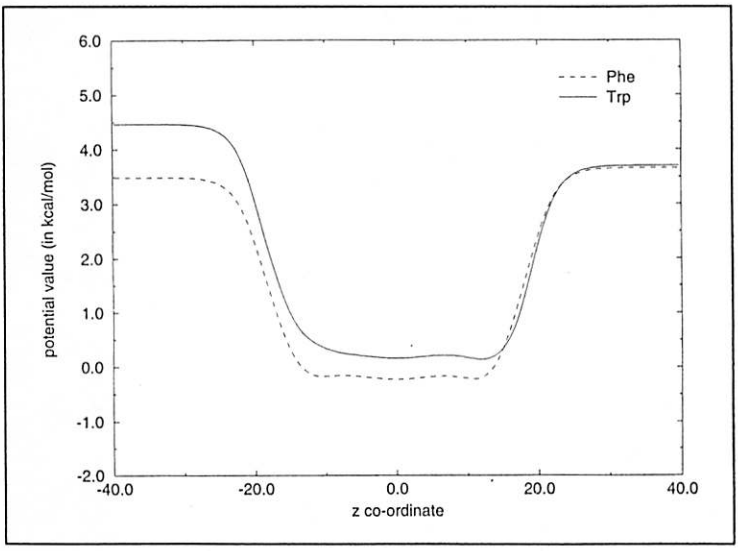
Name	HDI	HCI	HCC	HCO	HDO
A	0.13	-0.22	0.03	0.06	0.15
C	0.46	-0.07	-0.25	0.43	-0.02
D	-0.41	1.34	0.51	0.22	-0.44
E	-0.36	1.86	0.82	0.19	-0.56
F	0.26	-0.12	-0.08	-0.19	0.43
G	0.26	0.24	-0.19	-0.04	0.00
H	-0.42	0.96	0.62	0.06	-0.38
I	0.24	-0.24	-0.04	-0.07	0.31
K	-0.63	1.20	1.79	0.65	-0.43
L	0.50	-0.10	-0.16	-0.14	0.30
M	0.40	-0.28	0.07	-0.23	0.32
N	-0.31	0.26	0.13	0.22	-0.16
P	0.13	0.91	-0.05	0.06	-0.38
Q	-0.13	0.99	0.70	-0.04	-0.57
R	-0.73	0.69	2.65	0.69	-0.22
S	0.19	0.14	-0.25	0.26	0.02
T	-0.13	0.10	0.04	0.07	-0.07
V	o.36	-0.18	-0.08	-0.11	0.27
W	0.61	0.14	-0.08	-0.15	-0.15
Y	-0.11	-0.17	0.32	-0.14	0.08



A



B



C

Fig. 13.5. Asymmetric mean force membrane potential for representative amino acids

according to their annotations in the Swiss Prot database. Two flanking fragments, each 5 residues in length, were added to both ends of the helices. The sequences (with the flanking fragments) are presented in Table 13.4.

The simulation procedure consists of 28 independent simulated annealing runs for every helix, with a starting temperature of 500° K and a final temperature of 290° K. The geometrical parameters of membrane were the same in all simulations. The thickness of the hydrocarbon phase was equal to 33 Å and the thickness of both head group layers was equal to 4 Å. Uniform helical propensities of 1 kcal/mol per residue were used for all the chains. All the modeled chains started from random conformations whose centers of mass are about 30 Å from the membrane border. Half of the runs started from the chain situated on the "outer-cellular" side of the model membrane, and the other half started from the "cytoplasmic" side. For all tested transmembrane fragments, we observed topologies with "N-terminus inside," and "N-terminus outside" orientations. The average values of the energy for both the orientations and for all of the tested transmembrane helices are presented in Table 13.4.

According to the literature, all GPCRs have an overall "N-outside" topology.[32] This means that the first helix has its N-terminus situated on the outside leaflet of the plasma membrane, the second one is inside, etc. The expected result from the test of our potential is that the helical fragment in the native orientation should have lower average energy. This is the case for the first, second, fifth, and seventh helices. For the third and sixth helix, our method does not prefer any orientation. In the case of the helix number four, the predicted orientation is reversed. However, one can calculate the average total energy of the linked transmembrane helices, knowing that if the first helix is oriented with N terminus outside of the cell, the second one must be oriented in a N-inside topology, etc. In the case of the tested protein, we have two possible topologies for the seven linked helices. The difference in energies between the N-outside and N-inside topologies is 12 kcal/mol in favor of the native topology. This suggests that the preference for N-inside versus N-outside is result from the sum of differential preferences, some of which may be indifferent or weakly in conflict.

Table 13.5 contains the positions of beginning and end of the putative trans-

Table 13.4. Sequences of transmembrane fragments used in our testing simulations and minimal values obtained in simulated annealing procedures for native and inverse orientations

No. of Helix	Sequence	Minimal Energy for the "N- Inside" Topology	Minimal Energy for the "N- Outside" Topology
1	ETLNK YVVIIAYALVFLLSLLGNSLVMLVIL YSRVG	21.56	**16.30**
2	RSVTD VYLLNLALADLLFALTLPIW AASKV	**5.81**	9.47
3	LTFLC KVVSLLKEVNFYSGILLLACIS VDRYL	-1.30	**-1.33**
4	RHLVK FVCLGCWGLSMNLSLPFFL FRQAY	**-0.72**	-1.98
5	TAKWR MVLRILPHTFGFIVPLFVML FCYGF	-7.55	**-10.67**
6	HRAMR VIFAVVLIFLLCWLPYNLVLL ADTLM	**-8.45**	-8.36
7	NNIGR ARDATEILGFLHSCLNPIIYAF IGQNF	8.27	**6.29**

The bold numbers denote energies of transmembrane helices in native orientation[a,b]
a. Energy for 7 linked transmembrane fragments in the N-outside topology (native): 7.23 kcal/mol
b. Energy for 7 linked transmembrane fragments in the N-inside topology (inverted): 20.11 kcal/mol

Table 13.5. Positions of starts and ends of putative transmembrane helices in IL8A sequence, obtained by analysis of our simulation trajectories in the native and inverse orientations as well as the positions from the Swiss Prot annotations

Helix Number	Predicted, Native Topology	Predicted, Inverse Topology	Swiss Prot Annotation
1	42 - 63	43 - 63	40 - 66
2	77 - 94	78 - 94	76 - 96
3	121 - 136	120 - 134	112 - 133
4	153 - 175	152 - 174	155 - 174
5	202 - 224	201 - 221	200 - 220
6	244 - 266	244 - 262	243 - 264
7	286 - 309	288 - 311	286 - 308

membrane helices in the IL8A sequence, obtained by analysis of our simulation trajectories. The table presents results obtained for both native and inverse topologies. According to our simulations, we would predict that the differences in the location of the transmembrane helices in the two different topologies are minor. The only difference between the native and inverse topology is their energy, which is lower for the native one (see Table 13.4). For comparison purposes, the values from the Swiss Prot annotations are presented in the same table. The differences between our predictions and the Swiss Prot annotation are in most cases in the range of two residues (what is probably in the range of the error level of the method). The only significant difference is for the third helix, whose starting point is shifted by nine residues. It is possible that this is due to the placement in Swiss Prot of the residues "KEVN" in the central part of the putative transmembrane helix. Our method interprets this strongly hydrophilic fragment as a signal that it is outside of the membrane. This is suggestive that perhaps the positional assignment of the third helix needs to be reexamined.

CONCLUSIONS

At present, the prediction of membrane protein tertiary structure is still in the early stages of development. This is partly due to the paucity of known membrane protein structures against which the various prediction methods can be tested. The most powerful and general tools for the prediction of the helix location and overall topology (intra versus extra cellular handedness) are those which are amino acid composition based. In such cases, the assignment accuracy is quite high. Motivated by the success of such amino acid composition based methods, the potential in a reduced protein model developed by Milik and Skolnick[37] has been extended to include such terms. In contrast to sequence based approaches, this method starts from a protein fragment in an aqueous phase and simulates the conformational equilibria of the peptide between the aqueous and membrane phase. Such an approach is designed to predict the orientation and conformation of the peptide fragment with respect to the membrane bilayer. The effect of inclusion of the amino acid specific terms that reflects the environmental asymmetry for a given amino acid in the membrane has been investigated in a preliminary set of simulations on the interleukin-8 receptor. The simulations predict that the majority of the helices favor the native topology, but some may be indifferent to or even favor the alternative topology. Interestingly, the Swiss Prot annotation for the location of the third helix is inconsistent with the simulation results, whose

starting point is shifted by 9 residues. This suggests that the method may be used to refine the putative location of the transmembrane fragments.

The next stage in membrane tertiary structure prediction will require the development of potentials that describe the relative lipophilicity of a given amino acid in the bilayer. For example, does a leucine prefer to interact with the lipid or does it prefer to be buried in the membrane interior? Then, having such potentials in hand, terms which describe differential preferences for interacting pairs of amino acids are required. A major problem with the development of such potentials is the lack of solved membrane protein structures, but here multiple sequence alignment information may be of some help. Such potentials will be required for the de novo prediction of membrane protein tertiary structure, a capability which does not yet exist.

These simplified models can also provide a number of insights into the mechanism of membrane protein insertion into the bilayer, and the detailed role played by the membrane in determining membrane protein structure. At the very least, information such as a generalized positive inside rule needs to be encoded into such models. Indeed, the differential preference for amino acids to locate in the various regions of the membrane can provide a thermodynamic driving force for protein insertion. More generally, effects of membrane curvature may need to be considered, and there may well be situations where the detailed membrane structure needs to be accounted for. However, computational tractability will require that these models be kept as simple as possible. Nevertheless, simplified models have proven to provide a number of insights into the water soluble, globular protein folding problem, and this is reason to hope that they will prove to be as powerful when applied to the prediction of membrane protein tertiary structure. This will constitute the direction of research taken by our group in the next few years.

Acknowledgment

Valuable discussions with Dr. Andrzej Kolinski are gratefully acknowledged. This research was supported in part by grant No. GM-38794 of the Division of General Medical Sciences of the National Institutes of Health.

References

1. Stowell MHB, Rees DC. Structure and stability of membrane proteins. Adv Prot Chem 1995; 46:279-311.
2. Wodak SJ, Rooman MJ. Generating and testing protein folds. Curr Opin Struct Biol 1993; 3:247-259.
3. Jernigan RL. Protein folds. Curr Opin Struct Biol 1992; 2:248-256.
4. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. Proteins 1994; 18:353-366.
5. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994; 18:338-352.
6. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996; 6:195-209.
7. Miller RT, Jones DT, Thornton JM. Protein fold recognition by sequence threading: tools and assessment techniques. FASEB 1996; 10:171-178.
8. Henderson R. The structure of the purple membrane from Halobacterium hallobium: analysis of the X-ray diffraction pattern. J Mol Biol 1975; 93:123- 38.
9. Henderson R, Unwin PNT. Three-dimensional model of purple membrane obtained by electron microscopy. Nature 1975; 257:28-32.
10. Henderson R, Baldwin JM, Ceska TA, Zemlin F, Beckmann E, Downing KH. Model for the structure of Bacteriorhodopsin based on high-resolution electron cryo-microscopy. J Mol Biol 1990; 213:899-929.
11. Hibert MF, Trump-Kallmeyer S, Hoflack J, Bruinvels A. This is not a G protein-coupled receptor. Trends Pharm Sci 1993; 14(1):7-12.
12. Kendrew JC. The three-dimensional struc-

ture of a protein molecule. Sci Am 1961; 205:662-666.
13. Herzyk P, Hubbard RE. Automated method for modelling seven-helix transmembrane receptors from experimental data. Biophys J 1995; 69:2419-2442.
14. Deisenhofer J, Epp O, Miki K, Huber R, Michel H. X-ray structure analysis of a membrane protein complex. Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. J Mol Biol 1984; 180:385-98.
15. Deisenhofer J, Michel H. Nobel lecture. The photosynthetic reaction centre from the purple bacterium *Rhodopseudomonas viridis*. EMBO Journal 1989; 8:2149-70.
16. Allen JP, Feher G, Yeates TO et al. Structural homology of reaction centers from *Rhodopseudomonas sphaeroides* and *Rhodopseudomonas viridis* as determined by X-ray diffraction. Proc Natl Acad Sci USA 1986; 83:8589-93.
17. Allen JP, Feher G, Yeates TO, Komiya H, Rees DC. Structure of the reaction center from Rhodobacter sphaeroides R-26: The protein subunits Proc Natl Acad Sci USA 1987; 84:6162-6.
18. Chang CH, el-Kabbani O, Tiede D, Norris J, Schiffer M. Structure of the membrane bound photosynthetic reaction center from *Rhodobacter sphaeroides*. Biochem 1991; 30:5352-60.
19. Weiss MS, Wacker T, Wackesser J, Welte W, Schultz GE. The three-dimensional structure of porin from *Rhodobacter capsulatus* at 3 Å resolution. FEBS Letters 1990; 267:268-72.
20. Cowan SW, Schirmer T, Rummel G et al. Crystal structure explain functional properties of two *E. coli* porins. Nature 1992; 358:727-33.
21. Kuhlbrandt W, Wang DN, Fujiyoshi Y. Atomic model of plant light-harvesting complex by electron crystallography. Nature 1994; 367:614-621.
22. Picot D, Loll PJ, Gavarito RM. The X-ray crystal structure of the membrame protein prostaglandin H2 synthase 1. Nature 1994; 367:243-249.
23. Rao JKM, Argos P. A conformational preference parameter to predict helices in integral membrane proteins. Biochim Biophys Acta 1986; 869:197-214.
24. Fasman GD, Gilbert WA. The prediction of transmembrane protein sequences and their conformation: an evaluation. Trends Biochem Sci 1990; 15:89-92.
25. Ponnuswamy PK, Gromiha MM. Prediction of transmembrane helices from hydrophobic characteristics of proteins. Int J Pept Prot Res 1993; 42:326-341.
26. Casadio R, Fariselli P, Taroni C, Compiani M. A predictor of transmembrane α- helix domains of proteins based on neural networks. Eur Biophys J 1996; 24:165-178.
27. Donnelly D, Cogdell RJ. Predicting the point at which transmembrane helices protrude from the bilayer: a model of the antenna complexes from photosynthetic bacteria. Prot Engn 1993; 6:629-635.
28. Nilsson I, von Heijne G. Fine-tuning the topology of polytopic membrane protein: role of positively and negatively charged amino acids. Cell 1990; 62(6):1135-41.
29. Dalbey RE. Positively charged residues are important determinants of membrane protein topology. Trends Biochem Sci 1990; 15:253-257.
30. von Heijne G Membrane protein structure prediction: Hydrophobicity analysis and positive-inside rule. J Mol Biol 1992; 225:487-494.
31. von Heijne G Membrane proteins: from sequence to structure. Annu Rev Biophys Biomol Struct 1994; 23:167-92.
32. Hofmann K, Stoffel W. TMBASE—A database of membrane spanning protein segments. Biol Chem Hope-Seyler 1993; 374:166.
33. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical mmebrane protein structure and topology. Biochem 1994; 33: 3038-3049.
34. Persson B, Argos P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 1994; 237(2):182-192.
35. Persson B, Argos P. Topology prediction of

membrane proteins. Prot Sci 1996; 5: 363-371.
36. Reithmeier RAF. Characterization and modelling of membrane proteins using sequence analysis. Curr Biol 1995; 5:491-500.
37. Milik M. Insertion of peptide chains into lipid membranes: an off-lattice Monte Carlo dynamics model. Proteins 1993; 15:10-25.
38. Milik M, Skolnick J. A Monte Carlo model of fd and pf1 coat proteins in lipid membranes. Biophys J 1995; 69:1382-1386.
39. Milik M, Kolinski A, Skolnick J. Monte Carlo dynamics of a dense system of chain molecules constrained to lie near an interface. A simplified membrane model. J Chem Phys 1990; 93:4440-4446.
40. Milik M, Skolnick J, Kolinski A. Monte Carlo studies of an idealized model of lipid - water system. J Phys Chem 1992; 96:4015-4022.
41. Milik M, Skolnick J. Spontaneous insertion of polypeptide chains into membranes: A Monte Carlo model. Proc Natl Acad Sci USA 1992; 89:9391-9395.
42. Jacobs RE, White SH. The nature of the hydrophobic binding of small peptides at the bilayer interface: Implications for the insertion of transbilayer helices. Biochem 1989; 28:3421-3437.
43. Baumgartner A, Skolnick J. Polymer electrophoresis across a model membrane. J Phys Chem 1994; 98:10655-10658.
44. Baumgartner A, Skolnick J. Spontaneoous translocation of a polymer across a curved membrane. Phys Rev Lett 1995; 74: 2142-2145.
45. Rothman JE, Lenard J. Membrane asymmetry. Science 1977; 195:743-753.
46. Newton AC. Interaction of proteins with lipid head groups: lessons from protein kinase C. Ann Rev Biophys Biomol Struct 1993; 22:1-25.
47. Op den Kamp JAF. Lipid asymmetry in membranes. Ann Rev Biochem 1979; 48:47-71.
48. von Heijne G. The cytoplasmic domain of Escherichia coli leader peptidase is a translocation poison sequence. Proc Natl Sci USA 1988; 85:3363-3366.
49. Sipos L, von Heijne G. Predicting the topology of eukaryotic membrane proteins. E. J Biochem 1993; 213(3):1333-40.
50. Andersson H, von Heijne G. Membrane protein topology: effects of delta mu H+ on the translation of charged residues explain the "positive inside" rule. EMBO J 1994; 13(10):2267-72.
51. von Heijne G. Membrane protein assembly: rules of the game. Bioessays 1995; 17(1): 25-30.
52. Claros MG, von Heijne G. TopPredII: an improved software for membrane protein structure prediction. CABIOS 1994; 10(6):685-6.
53. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995; 4:2107-2117.
54. Lee J, Horuk R, Rice GC, Bennett GL, Camerato T, Wood WI. Characterization of two high affinity human interleukin-8 receptors. J Biol Chem 1992; 267(23): 16283-7.