PHYSICAL REVIEW E VOLUME 60, NUMBER 6 DECEMBER 1999

Simple model to study insertion of a protein into a membrane

Riccardo Bonaccini and Flavio Seno*
INFM, Dipartimento di Fisica, Università di Padova, Via Marzolo 8, 35131 Padova, Italy

(Received 21 January 1999; revised manuscript received 26 May 1999)

A simple coarse grained model on a two-dimensional lattice is presented to elucidate the main effects ruling the insertion of a protein into a polar environment such as a lipidic membrane. The amino acids are divided into two classes (hydrophobic or polar), and they behave differently according to their surroundings. In aqueous solution the hydrophobic amino acids are forced to minimize contacts with water, whereas in the apolar environment all the amino acids try to aggregate regardless to their specificity. The lattice is employed in order to perform exact calculations and to generate a fictitious protein data bank. Despite the simplicity of the model, some morphological features of the proteinlike lattice structures obtained by our model are compatible with the observed phenomenology of transmembrane proteins. These results seem to corroborate the hypothesis that the number of classes into which the amino acids can be divided that correctly describe the phenomena may be extremely low. [S1063-651X(99)08212-4]

PACS number(s): 87.15.By, 87.10.+e, 36.20.−r

I. INTRODUCTION

In the last decades a considerable effort has been expended into unraveling many of the mysteries behind the chemical and biological functionality of proteins. Most of this work has been devoted to water soluble globular proteins (WSP's), for which a large variety of three-dimensional native structures is experimentally known, and many theoretical aspects have been worked out [1,2].

On the other hand, much less is known about proteins (membrane proteins: MP's) that cross biological membranes and that rule solute transport, signal transmission, and energy conversion between the two isolated sides of the membrane. This lack of knowledge about MP's is related to the difficulty in experimental handling. Membranes consist of phospholipid bilayers with a hydrophobic interior: the surface of a MP that interacts with such an apolar environment is also hydrophobic and this property causes MP's to aggregate in aqueous solution, unless detergents are used [1,2]. This circumstance makes crystallization of MP's difficult and native structures have been determined only for nine of them [3,4].

The most important and studied class of MP's is that of transmembrane proteins (TMP's). These proteins span the membrane from one side to the other by one or more transmembrane segments (TMS's). They can have one or two functional globular domains outside the lipid bilayer, on the extracellular and/or the cytoplasmic side. As sketched in Fig. 1, TMP's are known to assume a rich variety of structures which in part are embedded in the hydrophobic membrane and in part in the polar solvent. Typical examples are the bacteriorhodopsin, made up of seven TMS's linked by loops external to the membrane, and the photosynthetic reaction center made up of four polipeptide chains with 12 TMS's and two globular domains external to the membrane.

A special class of TMP's are the pore-forming toxins, the most famous of which is Colicin A (so we will call them *Colicin A like* proteins): while soluble in aqueous media they nevertheless spontaneously insert themselves into lipid bilayers [1,2,4–6].

TMP's are characterized by the presence in the primary structure of long segments (20–30 amino acids) with a high degree of hydrophobicity [1,2], which correspond in the native structure to the TMS's. Another important feature is the stability inside the membrane of α helices and β sheets, since these structures allow the formation of hydrogen bonds between the backbone atoms, not possible with the surrounding apolar molecules [7]. This implies that TMS's are predominantly made up of α helices and β sheets.

The distribution of amino acids inside the membrane, and their mutual interactions, are less well understood. Early studies suggested that the distribution of hydrophobic and polar amino acids would be the opposite to that observed in WSP's, i.e., the polar amino acids are buried and the hydrophobic ones exposed to the lipid molecules, with the mean hydrophobic value of buried amino acids roughly conserved for both WSP's and TMP's [8,9]. More recent analysis has shown that such a scheme is too primitive, because some

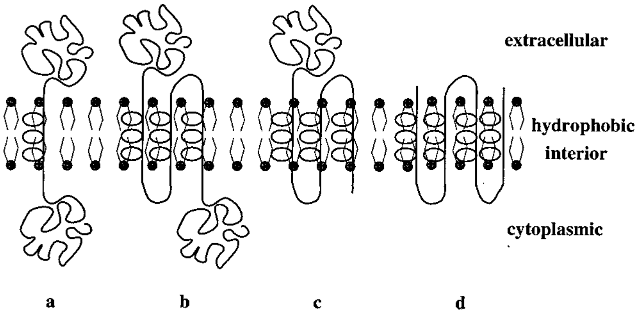

FIG. 1. Examples of TMP's with different morphological features: (a) A single transmembrane segment (TMS, for simplicity represented by a schematized helix) with a globular domain in both extracellular and cytoplasmic sides of the membrane. (b) Some TMS's with a globular domain on both sides of the membrane. (c) Some TMS's with a globular domain on only one side of the membrane (for example, the extracellular one). (d) Some TMS's without globular domains outside the membrane.

*Author to whom correspondence should be addressed. FAX: ++39-049-8277102. Electronic address: seno@padova.infn.it

1063-651X/99/60(6)/7290(9)/$15.00

amino acids, such as the aromatic and aliphatic ones, show different distributions [10,11]. Very little is known even now about the mutual interactions of amino acids inside the membrane, and between them and lipid molecules. It is clear only that TMS's aggregate to form a compact cluster of secondary structure units.

Given the difficulty in experimentally obtaining information about TMP's, theoretical studies are particularly important to better understand such complex phenomena. Very recently some efforts have been made in this direction, leading mainly to phenomenological models. The most successful of these is the so-called *two-stage model* [7], which states that the folding of TMP's occurs in a two-step process: in the first step α helices and β sheets are formed, while in the second step a native structure is reached.

In this paper we tackle a problem with a completely different strategy, namely, by using simple coarse grained exact models on a lattice [12–21]. The justification for such a method is that, despite the enormous number of degrees of freedom present in the phenomena, only a few parameters are relevant to describe the statistical behavior of folding. Once identified, these parameters enable calculation in a restricted configurational space (for example, on a lattice) where analytical or numerical calculations can be done to a high degree of precision. In this way it is possible to check the quality of the assumptions introduced in the model by comparing the properties of the model proteins to the real ones. Simple coarse grained lattice models have been so far applied to TMP's such as Monte Carlo simulations to study mesoscopic models in which sequences and conformations are treated in a very simplified way. Milik and Skolnik [22,23] studied the insertion of proteins inside the membrane by analyzing the dynamic behavior of few selected sequences starting from different initial configurations. Gersappe *et al.* [24] used a more schematic model on a cubic lattice to determine how the sequence distribution in amphiphilic chains affects their behavior with the membrane.

In this paper we develop a simple lattice model which takes into account only the hydrophobic effect, in order to verify whether this effect could be a mechanism sufficient to reproduce, at a coarse grained level, the phenomenology of TMP's. To this end we introduce a model on a square lattice that can be exactly analyzed through an exhaustive analysis of all possible sequences and conformations. As we will discuss in Sec. II, the choice of a two-dimensional lattice is not unrealistic because, for short chains, it correctly reproduces the exterior and interior ratio of proteins.

In the model, the amino acids are merely divided into two classes according to their affinity with water, and they interact through contact potentials that are different according to whether the amino acids are in aqueous solution or inside the membrane. External fields are also used to favor the presence of a $H(P)$ monomer inside (outside) the membrane.

Due to its simplicity, this model (as well as those used for WSP's) does not take into account secondary interactions such as hydrogen bonds. Therefore, our approach cannot be adequate to describe amino acid location inside the membrane where such interactions are driving the formation of helices, but it can be extremely useful to select the relevant forces that are stabilizing the protein across the membrane. For these reasons we decided to mimic the ''membrane'' in the simplest possible way, just as a hydrophobic line (in two dimensions) embedded in a polar environment, and to investigate whether this scheme is already sufficient to induce a competition among the hydrophobic and polar amino acids that stabilizes TMP-like configuration.

By looking at some phenomenological parameters, we see that at a coarse grained level the model reproduces the phenomenology of TMP's and the existence of Colicin A like proteins (sequences with different native states in water environment or in the presence of the hydrophobic line). These results seem to justify the approximations introduced in the model and the importance of the hydrophobic effect in assembling TMP's.

The paper is set out as follows. In Sec. II we summarize the main features of the hydrophobic-polar (HP) model used for WSP's. In Sec. III we introduce our model as an extension of it to represent some features of TMP's. In Sec. IV we present the numerical machinery used to take into account the symmetries of the model in the presence of a membrane and to define the proper space of conformations. Section V is devoted to a discussion of our results, and in Sec. VI possible further developments are presented.

II. SIMPLE EXACT MODELS FOR WSP's

Simple coarse grained exact models on a lattice have widely been used to study globular proteins [12–21,25–33]. They are recognized as extremely powerful tools to capture the essential features of the folding problem, and provide a versatile way to address questions of conformational change that are too complex to be treated with microscopic models.

As discussed in Sec. I the key idea is to eliminate redundant details. As a first step it is convenient to discretize space and to model the peptide chain by a self-avoiding random walk (SAW) on a lattice [34].

SAW's on a lattice can be exhaustively enumerated for relatively long chains [35] (from this fact the adjective exact comes from) and dynamically studied by standard Monte Carlo procedures [15,18]. Each step of such walks models the minimal unit of the model; it can represent a single amino acid (all the atoms forming it are substituted by a fictitious particle) or even clusters of them such as α helices or β sheets. An external parameter can be also introduced to discriminate between the sizes of these amino acids and their packing properties [21].

It is then important to find a suitable Hamiltonian that describes the interactions between amino acids and among amino acids and solvent molecules. An important simplification is generally introduced by considering only short range, two body effective interactions.

A sequence S^* can be a candidate for being a ''model protein'' (*good folder*) only when it has an unique ground state on some conformation Γ^* (*encodable structure*).

The existence of such peculiar sequences S^* and conformations Γ^* can be checked on exact models simply by counting how many sequences and conformations have the property that

$$\mathcal{H}_{\Gamma^*}(S^*) < \mathcal{H}_{\Gamma}(S^*) \quad \text{for any} \quad \Gamma \neq \Gamma^* \tag{1}$$

The ensemble of M sequences and conformations $[(S_1,\Gamma_1),(S_2,\Gamma_2),\ldots,(S_M,\Gamma_M)]$ with such a property can be seen as a model protein data bank, and its features have to be checked with those of real proteins in order to verify the quality and the robustness of the assumptions employed.

Perhaps the most popular lattice model employed to modelize the protein folding problem is the HP model of Lau and Dill [12,13,17] which consider two kinds of beads denoted by H (hydrophobic) and P (polar), and which is based on the hydrophobic effect [2,36]. In this model the formation of hydrophobic cores inside the native structures is induced by an effective two body attraction between H type beads, which is usually stronger than the attraction between two P's or between a P and an H. The effective nature of such interaction can be interpreted as a consequence of integration over the degrees of freedom of the solvent.

The virtues and drawbacks of the HP model are reviewed in detail in several papers [12,13,17,21], and we just recall that it has been used for at least two main kinds of problem: to deduce the main statistical features of the folding process [14–16] and to test the most advanced procedures to perform protein design and to extract statistical potentials among the amino acids [25–33]. The H and P classification in itself has been shown to be an opportune scheme for designing real proteins [36–38].

For the purpose of this paper it is worthwhile to notice that most of the work done with the HP model, as well as for analogous simplified models, has been done in two dimensions where exact enumerations are possible for longer chains. This, however, is not thought to be a major inconvenience because the surface-volume ratio is more important than the dimensionality of the space [17]. To correctly model the exterior-interior ratio of myoglobin in three dimensions requires simulations made up of around 150 monomers, but in two dimension simulations of only 16 step chains are needed [17,39]. Thus two-dimensional studies of short chains are regarded as models of longer three dimensional proteins.

III. MEMBRANE HP MODEL

In this section we derive a model for studying the insertion of a protein in a hydrophobic environment. This model is designed to capture the feature that TMP's have some parts which span the hydrophobic bilayer and others in contact with a polar solvent. We make the assumption that the classification of amino acids as hydrophobic or polar is able to roughly reproduce their position inside the membrane or in the external aqueous environment. The hydrophobic effect vanishes inside the membrane, so another type of effective interactions among the amino acids must be considered, however, at least in a first approximation the bare labels H and P could still be maintained.

As we discussed in Sec. I the bilayer is spanned from one side to the other by α helices and β sheets which make up the TMP's. By our HP coarse grained scheme we are not able to take into account the specificity of amino acid contacts inside the membrane, or the interactions with the polar heads of phospholipids, so we are not able to reproduce the details of amino acid position inside the bilayer. Moreover, the protein spans the whole bilayer whenever it is inserted into one of its surfaces. In order to model this aspect fully, we would have to introduce a sort of stiffness contribution into the Hamiltonian for the hydrophobic region, which leads to a complication of the computational task without a gain in the description we want to reach. Therefore, for our purposes, we can model the membrane by a penetrable hydrophobic surface in the lattice that separates two polar regions, that we call the *bulk* environment. This simplification will be *a posteriori* justified by the ability of the model to reproduce accurately the TMP phenomenology.

The model we present can be expressed in any dimensionality but, for the reasons presented in Sec. II and for computational convenience, we consider a two-dimensional square lattice where the membrane is merely represented by a line. The conformations a protein can adopt are represented by self-avoiding random walks made up of N monomers or beads (the considered fundamental unit of a protein), each of which is located on a lattice site $\mathbf{r_i}=(x_i,y_i)$. The line $y=0$ is our fictitious membrane. As in the HP model the monomers are divided in two classes, hydrophobic (H) and polar (P), and a sequence is represented by $S=(s_1,s_2,\ldots,s_N)$, where s_i is H or P depending on the class of the ith bead. When two monomers, both outside the membrane, not consecutive along the chain are on nearest neighbor sites they interact according to the HP scheme, i.e., they change energy by -1 if they are both H, or 0 otherwise [12,17]. H monomers in the bulk are forced to aggregate together in order to prevent contacts with the polar solvent (sites in the bulk not occupied by the SAW's). Obviously there is no interaction if one of the H beads is sitting on the hydrophobic line and the other one is off it. Clearly H monomers may prevent contacts with the solvent by staying inside the membrane. We therefore introduce a local field h_{HM} to favor the presence of H type beads on the hydrophobic line, and a second field h_{PM} to prevent the P type beads for sitting on the line.

For the interactions inside the lipid bilayer, we assume that we can integrate out the lipid degrees of freedom and consider short range two body effective interactions between the amino acids. In this way we assume that, at least in a first approximation, the entropic effects of packing amino acids with surrounding lipid molecules can be either neglected or represented by effective interactions.

Since the hydrophobic interactions do not exist in the membrane, we introduce a term in the Hamiltonian, which gives an energy gain ω for each monomer contact (regardless of type) inside the membrane line. The effect of this interaction is to favor the aggregation of monomers inserted in the lipid environment in compact domains, as observed in real TMP's, so ω cannot be positive. Notice that previous authors [22–24] who used coarse grained models, differentiated interactions inside the membrane according to the kind of amino acids. However, as discussed in Sec. I the HP classification is too schematic to reproduce the amino acid organization inside the membrane, so we believe that introducing differentiated interactions in such context is not relevant at this ''coarse grained'' level.

The complete Hamiltonian of the model can thus be expressed in the compact form

$$\mathcal{H}=\sum_{j>i,i=1}^{N}\Delta(\mathbf{r_i},\mathbf{r_j})[\epsilon(s_i,s_j)(1-\delta_{y_i,0})(1-\delta_{y_j,0})+\omega\delta_{y_i,0}\delta_{y_j,0}]+\sum_{i=1}^{N}\delta_{y_i,0}[h_{HM}f(s_i)+h_{PM}g(s_i)] \quad (2)$$

where (i) sums are over all beads forming the heteropolymer; (ii) $\Delta(\mathbf{r_i},\mathbf{r_j})$ is an adjacent matrix which entries a 1 if $\mathbf{r_i}$ and $\mathbf{r_j}$ are first neighbors in the lattice and not consecutive along the chain, and 0 otherwise; (iii) $\delta_{i,j}$ is the Kronecker delta function, which gives 1 if $i=j$ and 0 otherwise; (iv) f and g are two functions which give 1 when their argument is H or P, respectively, and 0 otherwise; (v) ϵ is the interaction matrix of the standard HP models [12] $\epsilon(H,H)=-1$ and $\epsilon(H,P)=\epsilon(P,P)=0$; (vi) h_{HM} and h_{PM} are the local fields which respectively favor the H monomers in the membrane line and the P monomers in the bulk, respectively; and (vii) ω is the parameter related to the interaction between amino acids inside the membrane.

We will name this model the membrane HP model (MHP model).

In our study h_{HM}, h_{PM}, and ω are free parameters . They cannot be known *a priori*, and they could strongly depend on the kind of biological membrane considered. We will show in Sec. IV that for a large spectrum of their values they correctly reproduce the TMP phenomenology, so we can conclude that our schematization is sufficiently precise and that the values of these parameters could be recovered by statistical analysis similar to those used to extract interaction potentials between amino acids in WSP's [30–33].

IV. CONFIGURATIONAL SPACE AND NUMERICAL ANALYSIS

Once the Hamiltonian of a lattice model for proteins is found, the next step consists of finding those sequences that could represent a protein and not merely a random heteropolymer. In Sec. II we showed that for the HP model this is accomplished by assuming the ground state structure of a sequence is its native one. This assumption is supported by the famous Anfinsen experiments [40].

Before extending this approach to TMP's it must be recognized that now the folding process is much more complex, because it requires a larger and more complex machinery represented by several molecules, necessary for the recognition, insertion, and translocation steps of the whole process [4,41–43]. Despite the complexity of such a biological process, it is most likely, from a statistical mechanics point of view, that the system is driven to a global minimum in its energy landscape. Such a minimum is required to take into account the thermodynamical stability observed in real TMP's, which diffuse along the membrane, interacting with many molecules but without significantly changing their three-dimensional structure.

An alternative possibility is a selection, driven by some interactions with the machinery and/or by related kinetical constraints, of a nonglobal, but local, minimum of the energy landscape. Even though this is in principle possible, it should require a very specific statistical or dynamical mechanism to select the native state in a minimum that is not global, so it seems unlikely to be generic. In conclusion, the assumption that the native state is the state of minimal energy is also the most plausible one for TMP's.

To find good folders one should check each of the q^N possible sequences (considering q different classes of amino acids), and determine whether it admits a unique ground state in the configurational space Ω, i.e., in the ensemble of all possible SAW conformations. Lattice symmetries should of course be exploited to eliminate redundant equivalent structures. This task is quite easy for WSP models: conformations that can be mapped one from the other by a set of symmetry operations should be considered as equivalent structures (see Fig. 2). Therefore, in a step by step construction of a SAW by the backtracking algorithm, the translational equivalence can be removed by fixing the starting point in the lattice, and the rotational and inversion ones by imposing the directions of the first step and of the first step that is not aligned [35].

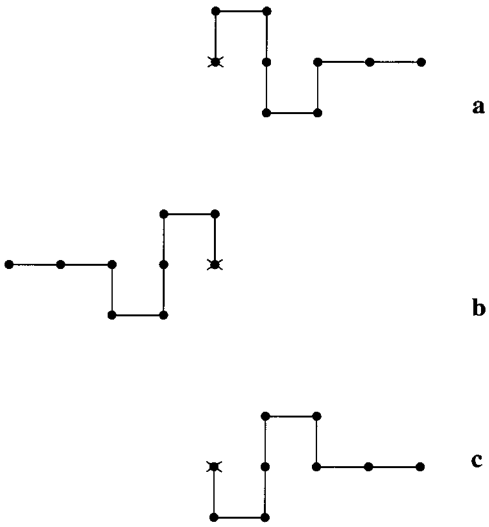

FIG. 2. Three conformations of WSP proteins on the square lattice ($N=9$) which are equivalent through symmetry operations. In determining the configurational space they are counted just as one conformation. Here and in the following figures, crosses indicate the starting point of the chain. It is kept fixed in performing enumerations to get rid of translation symmetries.

In the MHP model this becomes more tricky since the membrane breaks some of these symmetries. For example conformations are no longer equivalent through a translation along the y axis when at least one bead is touching the membrane (Fig. 3). For this reason we cannot fix the y coordinate of the starting point, so Ω now contains all the SAW that have any distance d between the starting point and the membrane line, and have at least one contact with the membrane. Moreover for $d\neq 0$ even mirror reflections along the axis passing through the starting point do not give equivalent conformations if the inversion axis is parallel to the hydrophobic line (Fig. 3). This requires, for each SAW generated by the standard backtracking algorithm, two different configurations for $d=0$ and four configurations for each distance $d\neq 0$ related to different orientations of the membrane line to respect the SAW. The determination of the configurational space consists in finding all these inequivalent conformations, the number $\|\Omega\|$ of which is clearly much larger than for the isotropic case. An exhaustive example of Ω is shown in Fig. 4 for a short chain of length 3 . At this point it is worthwhile to note that in Ω there are also present all the conformations without contacts with the membrane, i.e., the configurational space for the WSP case. In Table I, $\|\Omega_N\|$ is reported for different chain lengths N and compared with HP model results.

Our exact enumeration analysis for the MHP model has

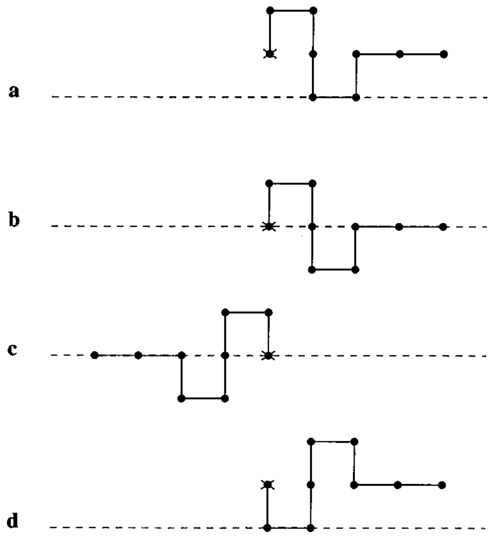

FIG. 3. Four configurations for the MHP model on the square lattice. The membrane is represented by the dashed line. Cases *a*, *b*, and *d* are inequivalent, and they model distinct structures, whereas cases *b* and *c* are symmetric and are counted just once in the configurational space. The starting point of the chain, marked by a cross, is not fixed because the translation symmetry is broken along the direction perpendicular to the membrane.

been carried out for chain lengths N ranging from 10 to 16 on a square lattice. Through a backtracking procedure we generated the complete set of 2^N sequences $S=(s_1,s_2,\ldots,s_N)$ made up with H and P monomers. Contrary to the isotropic case, the head to tail inversion is not an allowed symmetry operation because we are dealing with oriented walks. Given a set of values for the parameters h_{HM}, h_{PM}, and ω, we verified for each possible sequence whether it admits a unique ground state. The set $\{(S_1,\Gamma_1),(S_2,\Gamma_2),\ldots,(S_M,\Gamma_M)\}$ of good folders and encodable structures satisfying (1) determine the model protein data bank (MPDB) that strongly depends on the values of h_{HM}, h_{PM}, and ω. Obviously the M good folders S_i are all different, whereas some of the encodable structures Γ_i could be identical because dissimilar sequences can share the same ground state.

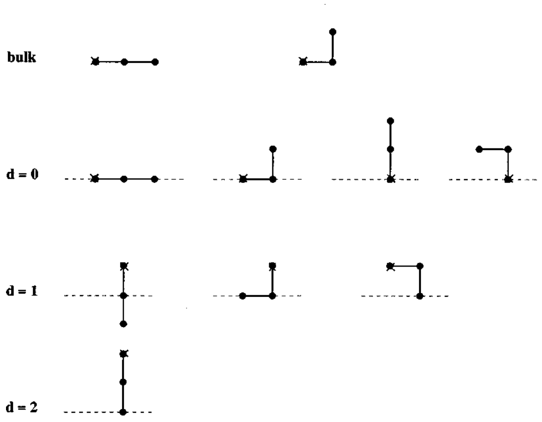

FIG. 4. Configurational space for the MHP model when $N=3$. d represents the Euclidean distance of the starting point of the chain from the membrane. The two structures in the first row are the only ones present in the bulk model.

TABLE I. Number of structures $\|\Omega_N\|$ present in the configurational space for the HP model in the bulk and for the MHP model for different values of the chain length *N*. The results are obtained on the square lattice

N	$\|\Omega_N\|$ (HP)	$\|\Omega_N\|$ (MHP)
10	2034	20 550
11	5513	59 345
12	15 037	171 224
13	40 617	488 155
14	110 188	1390 532
15	296 806	3926 032
16	802 075	11 076960

It is important to point out that some of the structures in the MPDB might not be in contact with the membrane: these structures do not gain from an interaction with the hydrophobic line, and they should not be considered as representing TMP's. More generally, good folders found in the HP model are strongly modified by the new interactions inserted in the MHP model Hamiltonian and they may lose the property of having a unique ground state. However in principle a sequence could be present both in the HP and MHP protein data bank, but with different ground state structures, resembling the behavior of Colicin A protein. In Sec. V we exploit the properties of the MPDB built by our model, and we compare it with real TMP's.

V. ANALYSIS OF THE RESULTS

In writing Hamiltonian (2) we stressed that, after assuming $\epsilon_{HH}=-1$, and $\epsilon_{HP}=\epsilon_{PP}=0$ as our choices for the bulk monomer interactions, we do not have any *a priori* knowledge of h_{HM}, h_{PM}, and ω. They are phenomenological parameters that might be extracted from statistical analysis [30–33] of the protein data bank of TMP once it has been shown they are sufficient conditions to model the problem. It is then crucial for the purpose of our analysis to show that there is at least a region in the physical part of the phase diagram spanned by h_{HM}, h_{PM}, and ω where the model match many experimental results on TMP's. The physical phase diagram can reasonably be limited by the conditions

$$0>h_{HM}\gtrsim -1=\epsilon_{HH}, \qquad |\epsilon_{HH}|=1\gtrsim h_{PM}>0,$$

$$\epsilon_{HH}/2=-1/2\lesssim\omega<0.$$

The first condition states that H beads ''prefer'' to stay on the membrane line, and that the strength of this effect should be comparable to the force segregating H monomers in the polar environment, because they are both directly related to the hydrophobic effect. On the other hand, the second condition implies that P monomers should prefer contacts with polar solvent with an interaction parameter not larger than $|\epsilon_{HH}|$, while the third condition stems from the fact that the attraction between the beads inside the membrane should be significantly less than the hydrophobic interactions, so we assume that its absolute value cannot be bigger than half $|\epsilon_{HH}|$.

TABLE II. Number of good folders, $N_{\rm GF}$; number of different encodable structures, $N_{\rm ES}$; and their ratio for different values of h_{HM}, h_{PM}, and ω. The results are obtained with $N=14$. In the first line the HP model results are reported.

h_{HM}	h_{PM}	ω	$N_{\rm GF}$	$N_{\rm ES}$	$N_{\rm ES}/N_{\rm GF}$
–	–	–	386	130	0.34
−0.2	0.0	0.0	819	251	0.31
−0.2	0.2	−0.1	1277	487	0.38
−0.2	0.4	−0.1	778	316	0.41
−0.2	0.6	−0.1	764	310	0.41
−0.2	0.8	−0.1	764	310	0.41
−0.2	1.0	−0.1	764	310	0.41
−0.4	0.0	0.0	915	259	0.28
−0.4	0.2	−0.1	2033	661	0.33
−0.4	0.4	−0.2	1473	549	0.37
−0.4	0.6	−0.2	963	448	0.47
−0.4	0.8	−0.2	962	432	0.45
−0.4	1.0	−0.2	948	432	0.46
−0.6	0.0	0.0	1132	316	0.28
−0.6	0.2	−0.1	2377	742	0.31
−0.6	0.4	−0.2	2236	705	0.32
−0.6	0.6	−0.3	2285	855	0.37
−0.6	0.8	−0.3	2191	901	0.41
−0.6	1.0	−0.3	1981	924	0.47
−0.8	0.0	0.0	878	238	0.27
−0.8	0.2	−0.1	2631	728	0.28
−0.8	0.4	−0.2	2076	821	0.40
−0.8	0.6	−0.3	2930	1210	0.41
−0.8	0.8	−0.4	2320	1039	0.45
−0.8	1.0	−0.4	2342	1059	0.45
−1.0	0.0	0.0	0	0	–
−1.0	0.2	−0.1	2645	298	0.11
−1.0	0.4	−0.2	2274	505	0.22
−1.0	0.6	−0.3	2427	702	0.29
−1.0	0.8	−0.4	1904	704	0.37
−1.0	1.0	−0.5	1556	590	0.38
0.0	0.2	0.0	0	0	–
0.0	0.4	0.0	0	0	–
0.0	0.6	0.0	0	0	–
0.0	0.8	0.0	0	0	–
0.0	1.0	0.0	0	0	–

TABLE III. $\langle C\rangle$, PM, PL, PMD, and PLD (defined in Sec. V) for different values of h_{HM}, h_{PM}, and ω. The listed results are obtained with $N=14$.

h_{HM}	h_{PM}	ω	$\langle C\rangle$	PM	PL	PMD	PLD
−0.2	0.0	0.0	2.5	38.3	33.0	1.7	99.0
−0.2	0.2	−0.1	2.2	23.8	10.0	11.9	90.6
−0.2	0.4	−0.1	1.7	1.1	0.3	21.6	85.9
−0.2	0.6	−0.1	1.7	0.0	0.0	22.0	85.6
−0.2	0.8	−0.1	1.7	0.0	0.0	22.0	85.6
−0.2	1.0	−0.1	1.7	0.0	0.0	22.0	85.6
−0.4	0.0	0.0	2.6	35.1	31.9	1.5	99.1
−0.4	0.2	−0.1	2.4	28.1	19.2	9.5	94.2
−0.4	0.4	−0.2	2.4	19.1	8.7	11.4	92.0
−0.4	0.6	−0.2	2.1	2.0	0.2	19.5	87.3
−0.4	0.8	−0.2	2.0	0.7	0.2	19.3	88.6
−0.4	1.0	−0.2	2.0	0.0	0.0	19.6	88.4
−0.6	0.0	0.0	2.8	30.7	29.5	1.9	97.9
−0.6	0.2	−0.1	2.6	24.2	17.6	9.4	92.6
−0.6	0.4	−0.2	2.7	16.9	11.0	10.3	88.2
−0.6	0.6	−0.3	2.8	9.2	5.0	11.4	80.3
−0.6	0.8	−0.3	2.8	5.3	2.1	15.1	77.2
−0.6	1.0	−0.3	2.8	0.7	0.2	17.8	73.2
−0.8	0.0	0.0	2.8	30.7	29.8	0.0	100.0
−0.8	0.2	−0.1	5.4	16.2	9.0	14.3	59.6
−0.8	0.4	−0.2	4.3	11.3	6.9	13.5	64.4
−0.8	0.6	−0.3	4.0	5.0	3.3	13.4	49.5
−0.8	0.8	−0.4	4.4	1.7	1.1	11.9	37.5
−0.8	1.0	−0.4	4.3	0.7	0.1	13.1	34.9
−1.0	0.2	−0.1	10.1	19.5	0.4	26.7	4.4
−1.0	0.4	−0.2	8.7	14.5	0.7	23.3	9.1
−1.0	0.6	−0.3	7.4	9.5	0.5	21.6	10.1
−1.0	0.8	−0.4	5.7	2.2	0.1	20.6	10.3
−1.0	1.0	−0.5	5.8	0.3	0.0	21.4	3.0

As a first test for the reliability of the model we report in Table II for $N=14$ and different values of ω, h_{HM}, and h_{PM} inside the previously mentioned ranges, the number N_{GF} of good folders, the number N_{ES} of different encodable structures, and the ratio of these two numbers, which gives the mean value of the number of sequences that fold in the same structure. In the first line of Table II these quantities are reported for the HP model.

For length $N=14$ the number of possible sequences is $2^{14}=16\,384$, and the number of structures with at least a contact with the surface (a proper membrane conformation) is 1 390 532. Except for the pathological cases ($h_{HM}=-1, h_{PM}=0$, and $\omega=0$ and $h_{HM}=0, 0.2\leqslant h_{PM}\leqslant 1.0$, and $\omega=0$), when we observe a strong adsorption of monomers in the membrane, for any point of the phase diagram the general physical criteria required for a protein model are fully satisfied. Only a few sequences are good folders, a reasonable number of structures are encodable, and on average more than one sequence selects the same folded state, guaranteeing the stability of the selected native conformations. All these results are corroborated by checks carried out at some specific values of the parameters for $N=16$, where a complete exploration of the phase diagram is computationally too demanding to be exhaustively performed.

In Table III we look for more particular properties of sequences and structures present in the MPDB. To do this in a proper way, we select four main properties which according to us better characterize the phenomenology of TMP's: (1) A significant portion of the protein has to be inserted in the membrane; (2) The large majority (but not the totality) of the transmembrane amino acids has to be of hydrophobic type; (3) The transmembrane polar amino acids should rarely be in contact with the lipid molecules; (4) The transmembrane amino acids should cluster together and form, at a coarse grained level, a single domain; (5) A significant number of proteins should have one or more large globular domains external to the lipid bilayer.

For the MPDB related to a given choice of parameters we focus our attention on these five quantities, all computed as average over all good folders: (1) The average number $\langle C\rangle$

TABLE IV. $\langle C\rangle$, PM, PL, PMD, and PLD obtained with $N=14$ for fixed values of h_{HM} and h_{PM}, but with different choices of ω.

h_{HM}	h_{PM}	ω	$\langle C\rangle$	PM	PL	PMD	PLD
-0.8	0.8	0.0	3.0	0.4	0.2	23.4	71.3
-0.8	0.8	-0.1	3.5	3.2	1.3	20.8	63.7
-0.8	0.8	-0.2	3.7	3.2	1.3	20.7	59.1
-0.8	0.8	-0.3	4.0	2.5	1.5	14.8	45.1
-0.8	0.8	-0.4	4.4	1.7	1.1	11.9	37.5

of monomers in contact with the membrane line; (2) The percentage PM of P type monomers among those present inside the membrane line; (3) The percentage PL of P type monomers among those in contact inside the membrane with a lipid bead, i.e., nearest neighbor to an empty site of the hydrophobic line; (4) The percentage PMD of good folders with multiple domains in the membrane, i.e., those that occupy more than one cluster of nearest neighbor sites for $y=0$; and (5) The percentage PLD of good folders that have native structures with large globular domains in the bulk, represented by at least four HH contacts in the bulk.

For $N=14$ and different values of parameters the quantities $\langle C\rangle$, PM, PL, PMD, and PLD are listed in Table III, where for each pair of h_{HM} and h_{PM} values $|\omega|$ is chosen equal to half of the minimum of $|h_{HM}|$ and $|h_{PM}|$.

With regard to $\langle C\rangle$ it is not possible to make any comparison with experimental results in view of the fact that our fictitious membrane is too schematic and not in scale with real ones; however, one could assume for $N=14$ that $\langle C\rangle \leqslant 8$ are physical allowable average values for this parameter, higher values being signals that most of the proteins are almost completely adsorbed by the membrane. From the results it seems that $\langle C\rangle$ is better tuned by h_{HM} than by h_{PM}, and that only for $h_{HM}=-1.0$, $0.0\leqslant h_{PM}\leqslant 0.4$ do we find unrealistic values for $\langle C\rangle$.

A physical estimation of PM for real TMP's exists, and it is around 30% [44]. Because of the coarse grained description of amino acids inside the membrane, we expect that the PM value in our model should be less than the one in real TMP's, so we require that PM should be less than an upper limit reasonably less than 30%, that we have chosen to be 10%. As shown in Table III, PM seems to depend mainly on h_{PM}, and it adopts values compatible with our requirements for $h_{PM}>0.5$. In this range of parameters values we also observe a satisfyingly low value of PL.

Moreover it turns out that in the restricted region fitting previous constraints for any nonvanishing value of ω the percentage PMD of multiple transmembrane domain structures is below the acceptable value of 25%. By decreasing ω, PMD can be significantly decreased, as shown in Table IV, confirming that the unspecific attraction of monomers inside the membrane line is able to guarantee that only a few good sequences are unrealistic with multiple intermembrane domains.

Finally, for PLD it is more difficult to determine a range of values which gives a better agreement with the TMP phenomenology; nevertheless we assume that it should be larger than 30%, so only for $h_{PM}=1.0$ are its values unacceptable. In any case, an important result is that increasing the absolute values of both local fields h_{PM} and h_{HM}, $\langle C\rangle$ increases, revealing that in these conditions the interactions of monomers with the membrane gives the largest contribution to the stabilization of native structures, while PLD decreases, revealing that the formation of hydrophobic cores in the bulk is now less important. This fact confirms that our simple model well reproduces the competition between two different types of interactions, those related to h_{HM} and h_{PM} contributions in the Hamiltonian, which favor H monomer insertion inside the membrane; that related to ϵ_{HH}, which favors the formation of hydrophobic cores in globular domains present in a polar environment. Nevertheless we stress again that $\langle C\rangle$ and PLD are average values, so for any choice of parameter values in the selected region we are able to find native structures either with many contacts with the membrane and with many HH contacts in the bulk, as shown in Fig. 5.

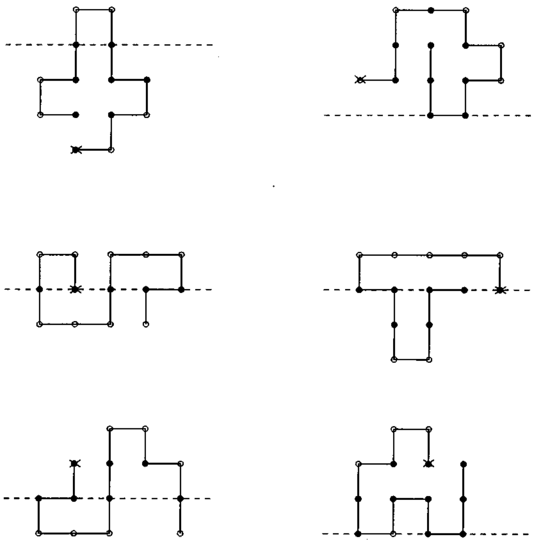

FIG. 5. Native structures for the MHP model obtained for $N=14$ and $h_{HM}=-0.6$, $h_{PM}=0.6$, and $\omega=-0.3$. Black circles indicate H monomers, white circles represent P monomers. In the first row: examples of good folders with many HH contacts in the bulk; in the middle row: good folders with many contacts with the membrane line; in the last row: intermediate cases. At the bottom left there is an example of a structure with more than one transmembrane domain.

From this analysis it is possible to conclude that, apart from the sharp position of its boundaries, a region in the phase diagram exists in which the proteins of MPDB roughly obey the main phenomenological features of TMP; this result supports *a posteriori* the crude approximations of our simplified model. It is also possible to look at the shape of these structures, and worthwhile to note that, despite the shortness of the chains and the regularization introduced by the lattice, qualitative analogies between this phenomenology and that of real TMP can be undoubtedly detected. In Fig. 5 examples of typical structures are illustrated.

We conclude this section by considering the Colicin A like behavior, i.e., sequences that have a unique ground state with both the standard HP model and the membrane HP model, considering some specified parameter values, but in different structures. Our work gave positive results: for many choices of parameter values in the selected range we found such classes of sequences; for instance, for $h_{HM}=-0.6$,

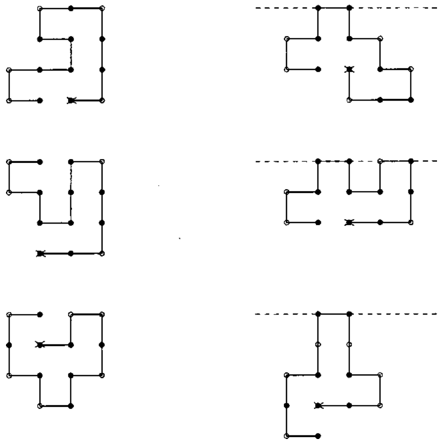

FIG. 6. Three examples of Colicin A like proteins found with the MHP model at $h_{HM}=-0.6$, $h_PM=0.6$, and $\omega=-0.3$ ($N=14$). On the left we show the native structure in the bulk, on the right the native structure in the presence of the membrane. The black (white) circles indicate H (P) monomers.

$h_{PM}=0.6$, and $\omega=-0.3$ we found 24 such sequences, three of which are shown in Fig. 6 as examples. Once again, the simple schematic strategy of our model is able to predict a phenomenon observed in nature, such as the possibility for a water soluble protein to change its native structure when it is in contact with a biological membrane. The investigation of the basic principles of such a mechanism, so far experimentally unknown, is of course outside the possibility and the purpose of the MHP model.

VI. CONCLUSIONS AND PERSPECTIVES

To summarize, in this paper we presented a simple and coarse grained model on a two-dimensional lattice to verify the importance of the hydrophobic effect in stabilizing proteins that can locate themselves either in a polar bulk environment or in a hydrophobic line. We used a procedure directly inspired by previous works on WSP's.

We emphasize that our approach cannot reproduce a detailed description of the microscopic behavior of proteins and membranes, but is aimed at understanding the main mechanism underlying such a complicated phenomena: this is the reason why simple mesoscopic models have been so widely and successfully studied in the literature. From our analysis it turns out that, even in a model where the amino acids are merely divided into two classes, the competition produced by perturbing a polar medium by the presence of an hydrophobic region (the membrane) is sufficient to generate unique ground states, i.e., proteinlike conformations, which have morphological features similar to those of TMP's (see Fig. 6 and the statistical analysis of Sec. V). It is also possible to verify the existence of sequences with different native states in the presence of different surrounding, in analogy with the behavior of Colicin A like proteins.

It is worthwhile to note that having obtained these results just for the simplest form of the hydrophobic membrane (a line) enforces the conclusion that the most relevant mechanism ruling the insertion of the protein into a membrane is really the bare competition induced by the different hydrophobicity between the inside and outside of the membrane. To use a thicker membrane is a necessary step to extract more detailed information about transmembrane segments, but the main physical features of the problem already seem to be caught by our simple model. It is important to point out how these results are not implicit in the model we have chosen because our analysis has been done by looking at the ensemble of unique ground states generated by all possible sequences. This ensemble can not be predicted *a priori*, and a first achievement of our work already consists of having found a small set of parameters for which such an ensemble could exist.

In conclusion, our results show that to reproduce the configurational arrangement of TMP's around the membrane, it is reasonable to look for simple models with a few classes of residues and a small set of effective interaction parameters and local fields. These interactions and fields should incorporate the hydrophobic effect as a main ingredient, and they must introduce an appropriate competition between the polar bulk and the hydrophobic membrane, that can even be mimicked by a line (in two dimension) in order to give physical results.

To obtain more complete and conclusive results regarding TMP's, it is necessary to consider a more precise representation of the membrane and a finer classification of the amino acids. In doing so one might also be able to determine the structure of the protein inside the membrane, and perform a closer comparison between the model and real membrane proteins.

On the other hand, the mere idea of using a few classes of amino acids can immediately [45] be employed to extract the statistical interaction potentials and the effective fields. At the same time the ultimate goal of designing new membrane proteins with a desired functionality can be seriously addressed. Recent theoretical results [33] for WSP's have indeed shown that, with a reduced number of amino acid classes, these goals can be simultaneously obtained by working with only a small set of known native structures as in the case of TMP's.

ACKNOWLEDGMENTS

We are indebted to Alessandro Desideri, Damien P. Foster, Amos Maritan, Cristian Micheletti, Tzachi Pilpel, Anna Tramontano, and Gunnar von Heijne for useful discussions. One of us (R.B.) is grateful to the *Istituto Nazionale di Ottica* of Firenze for kind hospitality.

[1] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991).

[2] T. E. Creighton, *Proteins: Structures and Molecular Properties*, (Freeman, New York, 1992).

[3] C. Ostermeier and H. Michel, Curr. Opin. Struct. Biol. **7**, 697 (1997).

[4] G. von Heijne, Prog. Biophys. Mol. Biol. **66**, 113 (1996).

[5] D.M. Engelman, Science **274**, 1850 (1996).

[6] F.G. van der Goot, J.M. Gonzalez-Manas, J.H. Lakey, and F. Pattus, Nature (London) **354**, 408 (1991).

[7] J.L. Popot and D.M. Engelman, Biochemistry **29**, 4031 (1990).

[8] D.C. Rees, L. De Antonio and D. Eisenberg, Science **245**, 510 (1989).

[9] W.A. Cramer, D.M. Engelman, G von Heijne, and D.C. Rees, FASEB J. **6**, 3397 (1992).

[10] F.A. Samatey, C. Xu, and J.L. Popot, Proc. Natl. Acad. Sci. USA **92**, 4577 (1995).

[11] T. Pilpel private communication.

[12] K.F. Lau, and K.A. Dill, Macromolecules **22**, 3986 (1989).

[13] H.S. Chan and K.A. Dill, Macromolecules **22**, 4559 (1989).

[14] H.S. Chan and K.A. Dill, Phys. Today **46**(2), 24 (1993).

[15] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[16] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).

[17] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, and P.D. Thomas, Protein Sci. **4**, 561 (1995).

[18] M.R. Betancourt and J.N. Onuchic, J. Chem. Phys. **103**, 773 (1995).

[19] H. Li, R. Helling, C. Tang, and N.S. Wingreen, Science **273**, 666 (1996).

[20] H. Li, C. Tang, and N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).

[21] C. Micheletti, J.R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **80**, 5683 (1998).

[22] M. Milik and J. Skolnick, Proc. Natl. Acad. Sci. USA **89**, 9391 (1992).

[23] M. Milik and J. Skolnick, Proteins **15**, 10 (1993).

[24] D. Gersappe, W. Li, and A.C. Balazs, J. Chem. Phys. **99**, 7209 (1993).

[25] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E. Shakhnovich, and K.A. Dill, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).

[26] J.M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).

[27] F. Seno, M. Vendruscolo, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).

[28] M.P. Morrisey and E. Shakhnovich, Folding Des. **1**, 391 (1996).

[29] C. Micheletti, F. Seno, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **80**, 2237 (1998).

[30] P.D. Thomas and K.A. Dill, J. Mol. Biol. **257**, 457 (1996).

[31] P.D. Thomas and K.A. Dill, Proc. Natl. Acad. Sci. USA **93**, 11 628 (1996).

[32] F. Seno, A. Maritan, and J.R. Banavar, Proteins: Struct., Funct., Genet. **30**, 244 (1998).

[33] F. Seno, C. Micheletti, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).

[34] B.H. Park and M. Levitt, J. Mol. Biol. **249**, 493 (1995).

[35] D.C. Rapaport, Comput. Phys. Rep. **5**, 265 (1987).

[36] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, and M.H. Hecht, Science **262**, 1680 (1993).

[37] M.H.J. Cordes, A.R. Davidson, and R.T Sauer, Curr. Opin. Struct. Biol. **6**, 3 (1996).

[38] C. Micheletti, F. Seno, A. Maritan, and J.R. Banavar, Proteins: Struct., Funct., Genet. **32**, 80 (1998).

[39] H.S. Chan and K.A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[40] C. Anfinsen, Science **181**, 223 (1973).

[41] A.C. Borel and S.M. Simon, Cell **85**, 379 (1996).

[42] W. Mothes, S.U. Heinrich, R. Graf, I.M. Nilsson, G. von Heijne, J. Brunner, and T.A. Rapoport, Cell **89**, 523 (1997).

[43] R.S. Hegde and V.R. Lingappa, Cell **91**, 575 (1997).

[44] G. von Heijne (private communication).

[45] R. Bonaccini and F. Seno (unpublished).