Protein Folding in Contact Map Space

M. Vendruscolo, R. Najmanovich, and E. Domany

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

(Received 25 June 1998)

Changing a few contacts in a contact map corresponds to a large scale move in confrontation space; hence, one gains a lot by using the contact map representation for protein folding. We developed an efficient search procedure in the space of physical contact maps, which could identify the native fold as of the lowest free energy, provided on had a free energy function whose ground state is the native map. We prove rigorously that the widely used pairwise contact approximation to the free energy cannot stabilize even a single protein's native map. Testing the native map against a set of decoys obtained by gapless threading, one may be misled to the opposite conclusion. [S0031-9007(98)08231-3]

PACS numbers: 87.15.By, 87.10.+e

One of the most challenging open problems in computational physics, chemistry, and biology is that of *protein folding;* e.g., to predict the conformation of a polypeptide chain from its amino acid sequence. The solution of this problem will have a far reaching impact on our understanding the function of biologically active macromolecules, as well as on practical problems of central importance, such as drug design. The number of sequenced proteins (about 170 000) is increasing steeply [1], while the structure has been determined only for a few thousand. Since this gap is expected to widen considerably as a result of concentrated sequencing efforts, theoretical attack on the folding problem is a most timely undertaking. A conceptually straightforward attempt to solve the protein folding problem is to construct, for any given molecule, an energy function using the interatomic potentials and look for its minima, or use molecular dynamics, integrating Newton's equations at an energy corresponding to kT. Such a direct attack on the problem is unrealistic, partly because solving the dynamics for large molecules lies beyond the possibilities of existing computers and partly because the exact potential is not known. Interest among physicists [2] covers various aspects of the problem, ranging from generally theoretical issues, addressed by a variety of field-theoretical methods [3], to developing numerical methodologies [4], suitable for folding specific real proteins [5]. The issues raised include dynamic aspects of the folding process [6,7], determining factors that govern thermodynamic stability of the native fold [8] and its stability against mutations [9], etc. Every realistic effort aimed at answering any specific question must use some energy function. The exact potential that governs the folding process is not known [10]. One is looking for classical effective interactions between atoms; furthermore, folding takes place in the presence of water and the water molecules must be "integrated out." Hence, one needs approximate, coarse grained, or reduced representations of protein structure and of the corresponding *free* energy.

A large fraction of recent numerical work in the physics literature uses pairwise *contact potentials* to represent the energy of a particular conformation [5,11]:

$$\mathcal{H}^{\text{pair}}(\boldsymbol{S}, \boldsymbol{A}) = \sum_{i<j}^{N} \boldsymbol{S}_{ij} w(a_i, a_j) . \tag{1}$$

Here, $\boldsymbol{A} = (a_1, a_2, a_3, \ldots, a_N)$ denotes the sequence of a protein of N residues; a_i is one of the letters of a 20-letter alphabet, identifying the amino acid at position i along the chain. The *contact map* $\boldsymbol{S}$ is an $N \times N$ matrix whose elements are either 0 or 1; $S_{i,j} = 1$ if residues i and j are in contact or 0 if not. Since we identify [12] physical structures by their backbone conformation, we define two residues to be in contact by the distance between their C_α atoms (we used the threshold R_c of 8.5 Å).

The choice of the 210 contact energy parameters $w(a, b)$ varies from the simplest binary valued HP model [13], through values selected from random distributions [6], to knowledge-based determination that uses data available from the Protein Data Bank (PDB) [14]. These methods employ either quasichemical approximations [5,15,16] or optimization of, for example, the Z scores of the native folds with respect to an ensemble of decoys [17–19]. Attempts were made to fit the contact parameters obtained this way to simple forms from which various conclusions about the potential can be drawn [20].

Since the contact map of a protein is independent of the coordinate frame used, contact maps are convenient for protein structure comparisons and for *searching a limited database* for similar structures. A more challenging possibility was proposed recently [5]: to use the contact map representation for *folding, e.g., to search the space of contact maps* for the map that corresponds to the native fold. The main computational advantage of this strategy is that changing a few contacts in a map induces rather significant large-scale coherent moves of the corresponding polypeptide chain [21]. Given all the inter-residue contacts or even a subset of them, it is possible to reconstruct quite well a protein's structure [12,22].

In this work, we reinterpret $\mathcal{H}^{\text{pair}}$; rather than viewing it as the energy, it plays the role of a simplest phenomenological approximation to the "true" *free* energy of a protein with a given contact map. With this in mind, the native map must be identified as the one with the highest probability of appearance and, hence, of the lowest possible free energy. Clearly, if the global minimum of an energy function is not at (or near) the native fold, an efficient energy minimization will mislead us to a wrong structure.

In this work, we take the first step in an attempt to determine *which potentials are capable to be tuned so that they have their minimum at the native map, S_0.* Whether $\mathcal{H}^{\mathrm{pair}}$ is a good enough approximation to play the role of the function whose minimization yields the native fold of proteins is not clear at all. Hence, it is natural to pose a well defined question regarding the ability of the pairwise contact approximation (1) to predict correctly the native fold of even a single protein. In other words, we ask the following: Is it possible to find a set of 210 contact parameters $w(a,b)$ such that

$$\mathcal{H}^{\mathrm{pair}}(\mathbf{S}_0, \mathbf{A}, \mathbf{w}) < \mathcal{H}^{\mathrm{pair}}(\mathbf{S}_\mu, \mathbf{A}, \mathbf{w}) \quad \forall\ \mu\,, \qquad (2)$$

i.e., the energy of $\mathbf{S}_0$, the native contact map of a protein is lower than all other maps $\mathbf{S}_\mu$? We proved [23] that the answer to this question is *negative.* This means that simple approximations to the contact free energy *cannot be used to identify the native map of a protein.*

It is impractical to construct all of the $O(e^{aN})$ physical contact maps of a chain of length N. Therefore the native map can be tested only against a relatively small number of decoys, and the answer to our question will depend on the competing conformations that were generated. For example, for "low quality" decoys obtained by *threading* the answer is positive not just for a single protein; rather, we can stabilize simultaneously the native maps $\mathbf{S}_0^p$ of $p = 1, \ldots, M_p$ proteins, each against all of its decoys $\mathbf{S}_\mu^p$ obtained by threading, i.e.,

$$\mathcal{H}^{\mathrm{pair}}(\mathbf{S}_0^p, \mathbf{A}^p, \mathbf{w}) < \mathcal{H}^{\mathrm{pair}}(\mathbf{S}_\mu^p, \mathbf{A}, \mathbf{w}) \quad \forall\ p, \mu\,. \qquad (3)$$

This result holds [24] for (typically) $M_p < 100$ and for $7.5 < R_c < 15$ Å. To avoid being misled to a positive answer to the question (2), it is essential to test the native fold against "hard" candidate contact maps. Generating such decoys is far from easy. To produce them, we developed an efficient way to explore the space of contact maps. Executing moves in the 3D conformation space of chains is inefficient; on the other hand, moves made in the space of contact maps give rise to a major difficulty, in that such moves usually lead to nonphysical maps. To make sure that our candidate maps are *physical,* e.g., correspond to real C_α chain conformations, we used a previously developed algorithm to project maps generated by our search procedure onto the subspace of physical maps [12].

Once a large set of hard candidate maps is assembled for a single protein, we answer the question by searching for contact energies $w(a,b)$, for which (2) holds for all decoys. This search is done by *perceptron learning.* The version of the perceptron learning rule that we use signals when the training set is unlearnable, meaning, in the present context, that such a set of $w(a,b)$ does not exist. This was found to be the case when we tried to "learn" contact parameters that stabilize the native fold of *crambin* against a large set of decoys. We present now our work in more detail.

Free energy.—Denote by C a microstate of the system (of the chain, the water molecules, etc.); since many microscopic conformations share the same contact map $\mathbf{S}$, it is appropriate to define a *free energy* $\mathcal{H}(\mathbf{S}, \mathbf{A})$ associated with this sequence and map:

$$\mathrm{Prob}(\mathbf{S}) \propto e^{-\mathcal{H}(\mathbf{S},\mathbf{A})} = \sum_C e^{-(1/kT)E(C)} \Delta(C, \mathbf{S})\,, \qquad (4)$$

where $\Delta(C, \mathbf{S}) = 1$ if $\mathbf{S}$ is consistent with C and $\Delta = 0$ otherwise; the "projection operator" Δ ensures that only those configurations whose contact map is $\mathbf{S}$ contribute to the sum (4) (note that summation over positions of water and other solvent molecules is also implicit). $E(C)$ is the unknown true microscopic energy.

Since it is impossible to evaluate this *exact* definition of the free energy of a map, we resort to a phenomenological approach, guessing the form of $\mathcal{H}(\mathbf{S}, \mathbf{A})$ that would have been obtained had the sum (4) has been carried out. $\mathcal{H}^{\mathrm{pair}}$ of Eq. (1) is a simplest approximation to the true free energy. To test the extent to which this approximate form is capable of stabilizing the native map of a protein against other non-native maps, we must assemble a set of such decoys.

Generating decoys by threading.—The simplest way to generate candidate contact maps is by "threading" a chain of length N_p through the (known) configuration of a longer chain, of length N_q. In the context of contact maps, this amounts to cutting out $N_p \times N_p$ submatrices that lie along the diagonal of the larger map, yielding $N_q - N_p + 1$ decoys for the shorter chain. Since each of these submatrices is a map of a segment of length N_p of a real protein, it is guaranteed to represent a physical C_α chain. We assembled a set of 153 proteins [24]; for every protein p of the set, we generated from the PDB its native contact map $\mathbf{S}_0^p$, as well as a set of candidate maps (decoys), generated by *threading* it into every member of the set whose length N exceeds N_p. This simple method has an obvious deficiency; the candidate map uses a structure which was "tailored" for the sequence of a segment of the longer protein and may not fit at all that of the shorter one. Hence, in general, the resulting map, albeit physical, will not yield a low energy when used in Eq. (1).

Generating low energy decoys.—We have presented elsewhere [12,21] a *three-step* Monte Carlo method to generate physical maps of low energy, which we summarize briefly.

In the first step, we perform nonlocal moves, updating "clusters" of contacts in an existing map. These clusters represent either α helices or β sheets (parallel or antiparallel), or small groups of contacts between amino acids that are well separated along the chain. The "energy" of the resulting coarse map is evaluated and a low energy map is retained. This map is refined in the second step by local moves in which contacts that are in the vicinity of existing ones are turned on or off (mostly one at a time). Only moves that lower (or do not significantly raise) the energy are accepted. In the third step, which takes most of the computer time, we deal with the major problem of ensuring that we stay in the subspace of physical maps [12]. A string of beads that represents the backbone of the polypeptide chain is moved around without tearing the chain and without allowing one bead to invade the space of another. The motion of this string is controlled by a "cost function" which vanishes when the contact map of the string coincides with that of the target map that was produced by the second step. The cost increases when the difference between the two maps increases. This procedure ends up with a chain configuration whose contact map is physical by definition and close to the target map. Thus we are able to efficiently "project" any map that we have generated in the first two steps onto the subspace of physical maps.

The contact maps obtained by this procedure have nativelike average radius of gyration and number of contacts. A serious shortcoming of the C_α representation, which applies both to gapless threading and to low energy decoys, is that, even for a physical C_α trace, an attempt to fit in side chains may result in violations of steric constraints and bond angle. Thus, our definition of physicality means only that the C_α chains are realizable.

For the reason explained above, the maps obtained by our search procedure are of much lower energy than those obtained by threading. This can be seen in Fig. 1, which presents histograms of the energies of two families of decoys for *crambin.* One family was obtained by threading and the other by searching for low energy maps, using in (1) different contact potentials from the literature. Evidently the decoys obtained by threading are of much higher energy, with only a small fraction below the native one. On the other hand, all maps obtained by our search have significantly lower energies than the native one. Therefore, finding a set of contact parameters for which (2) holds for all maps of this set constitutes a much more difficult challenge than doing the same for the threading ensemble.

Learning the contact energy parameters.—The answers quoted above regarding the existence of contact energy parameters that satisfy the set of conditions (2) or (3) were derived by perceptron learning.

Note that for any map $\boldsymbol{S}_\mu$ the energy [Eq. (1)] is *linear* in the parameters $\mathbf{w}$; therefore the conditions (3) and (2) are linear as well. The difference between the energy of a decoy map and the native one can be written as

$$\Delta \mathcal{H}_\mu = \sum_{c=1}^{210} [N_c(\boldsymbol{S}_\mu) - N_c(\boldsymbol{S}_0)]\mathbf{w}_c \propto \mathbf{w} \cdot \mathbf{x}_\mu , \quad (5)$$

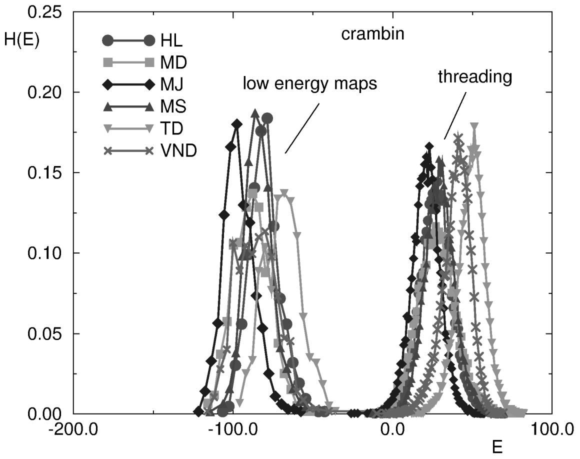

FIG. 1. Histograms that demonstrate the difference in energy between ensembles of contact maps obtained by threading and by energy minimization, shown for different contact energy parameters: VND: as obtained in this work, by finding a solution for a threading ensemble; HL: Ref. [16]; MD: [5]; MJ: [15]; MS: [18]; TD: [19]. The energy parameter sets were shifted and rescaled to obtain $\langle \mathbf{w} \rangle = 0$ and $\langle \mathbf{w}^2 \rangle - \langle \mathbf{w} \rangle^2 = 1$ (averages are over the 210 energy parameters). Energies were shifted to set the native state to $E = 0$.

where $N_c(\boldsymbol{S}_\mu)$ $[N_c(\boldsymbol{S}_0)]$ is the total number of contacts of type $c = 1, 2, \dots, 210$ that actually appear in map $\boldsymbol{S}_\mu$ $[\boldsymbol{S}_0]$, and we denoted by $\mathbf{x}_\mu$ the normalized vector of contact differences. Thus the conditions of Eq. (2) take the form

$$\mathbf{w} \cdot \mathbf{x}_\mu > 0 \quad \forall\ \mu . \quad (6)$$

Perceptron learning is a standard procedure to look for a set of $\mathbf{w}$ that satisfies such linear inequalities for all $\mu = 1, 2, \dots, P$ "examples" that constitute the "training set." The examples μ are presented sequentially; after presentation of example μ for which $\mathbf{w} \cdot \mathbf{x}_\mu < 0$, the following update takes place:

$$\mathbf{w}' = (\mathbf{w} + \eta \mathbf{x}_\mu)/|\mathbf{w} + \eta \mathbf{x}_\mu| . \quad (7)$$

The perceptron learning procedure is guaranteed to converge [25] to a solution $\mathbf{w}^*$ of (6), if one exists. By setting the value of the parameter η according to a learning rule that was introduced in Ref. [26], we are able to detect that the *training set is unlearnable;* e.g., there are no contact parameters that stabilize the native map against all of the decoys. This is done by evaluating, as we learn, a monotonously increasing quantity d called *despair;* if d exceeds a critical value d_c [24] before a solution is reached, the problem is unlearnable.

Results for threading.—Our database contained 153 proteins. Using the perceptron learning technique presented above, we proved that there is no set of contact energy parameters that can satisfy Eq. (3) when all of the possible 1 248 667 decoys obtained by gapless threading are used simultaneously. On the other hand, for a typical randomly drawn subset of 100 proteins, a set of contact parameters *can* be learned.

Results for low energy maps.—Crambin is a protein of length $N = 46$, composed of 15 kinds of amino acids, of which three appear only once. Hence, only 117 of the total number of 210 possible contacts can actually appear for any fold of crambin. In an unlearnable case, there exist sets of examples for which no solution can be found; for large enough P the training set will include, with nonvanishing probability, such an unlearnable subset. For $M = 117$, the critical despair is [26] $d_c \simeq 10^{163}$. We had to generate, by our search procedure discussed above, $p = 298\,710$ examples to obtain an unlearnable set. Out of these, we identified a hard subset of 10 000 examples, which we tried to learn. d_c was reached after about 37 500 learning cycles; the problem is unlearnable. Repeating the learning procedure using only decoys with less contacts than native crambin does not change the conclusion. We have repeated the same procedure for a set of six immunoglobulines (8fab, 1baf, 1cbv, 1dba, 2f19, 2fdl) which we attempted to stabilize simultaneously. For this problem $M = 210$, and again we proved unlearnability (details will be presented elsewhere).

We demonstrated that it is impossible to parametrize a simple potential in a way that guarantees that the *native fold is the state of lowest energy even for a single protein;* a good search procedure identifies lower energy decoys which are *very different* from the native map. One should not be mislead by similar work [27] on model proteins, in which a database of foldable sequences is designed using a contact potential and subsequently a set of contact energy parameters is recovered. Success in this case is possible because the contact energy of Eq. (1) is the *exact* form of the free energy of the model.

There are at least two possible directions to explore. (i) Controlled inclusion of additional energy terms, such as hydrophobic (solvation), hydrogen bond, or multibody interactions may help to attain foldability. We believe that the optimization scheme presented here will allow a step-by-step improvement of the energy function. (ii) We have also found evidence that, although the overall fold of crambin remains elusive with our optimized contact potential, partial success is obtained on a smaller scale, either identifying long range contacts or local structural features. Whether we will be able to use these advantages to improve the predictive power of some novel method of identifying structure remains to be seen.

We are grateful to Ron Elber for discussing with us a similar approach, based on Eq. (5) (unpublished). This research was supported by grants from the Minerva Foundation, the Germany-Israel Science Foundation (GIF), and by a grant from the Israeli Ministry of Science.

[1] A. Bairoch and R. Apweiler, Nucl. Acids Res. **25**, 31 (1997).

[2] H. Frauenfelder and P.G. Wolynes, Phys. Today **47**, No. 2, 58 (1994); V.S. Pande, A.Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. (to be published).

[3] T. Garel, H. Orland, and D. Thirumalai, in *New Developments in Theoretical Studies of Proteins,* edited by R. Elber (World Scientific, Singapore, 1997).

[4] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, Phys. Rev. Lett. **80**, 3149 (1998); T. Garel and H. Orland, J. Phys. A **23**, L621 (1990).

[5] L. Mirny and E. Domany, Proteins **26**, 391 (1996).

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

[7] C.J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).

[8] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994); H. Li, R. Helling, C. Tang, and N.S. Wingreen, Science **273**, 666 (1996); F. Seno, M. Vendruscolo, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).

[9] M. Vendruscolo, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **78**, 3967 (1997); V.S. Pande, A.Yu. Grosberg, and T. Tanaka, Fold. Des. **2**, 109 (1997).

[10] A.V. Finkelstein, Curr. Opin. Struct. Biol. **7**, 60 (1997).

[11] C. Micheletti, F. Seno, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **80**, 2237 (1998); A.M. Gutin, V. Abkevich, and E.I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996); K.D. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996); V.S. Pande, A.Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).

[12] M. Vendruscolo, E. Kussel, and E. Domany, Fold. Des. **2**, 295 (1997).

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

[14] For a recent review, see R.L. Jernigan and I. Bahar, Curr. Opin. Struct. Biol. **6**, 195 (1997).

[15] S. Miyazawa and R.L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[16] D.A. Hinds and M. Levitt, J. Mol. Biol. **243**, 668 (1994).

[17] M.H. Hao and H.A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **93**, 4984 (1996).

[18] L. Mirny and E.I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).

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

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

[21] M. Vendruscolo and E. Domany, Fold. Des. **3**, 329 (1998).

[22] A.T. Brünger, P.D. Adams, and L.M. Rice, Curr. Opin. Struct. Biol. **5**, 325 (1997).

[23] We allow all decoy maps whose backbone satisfies our definition (see below) of being physical.

[24] M. Vendruscolo, R. Najmanovich, and E. Domany (unpublished).

[25] M.L. Minsky and S.A. Papert, *Perceptrons* (MIT, Cambridge, MA, 1969).

[26] D. Nabutovsky and E. Domany, Neural Comput. **3**, 604 (1991).

[27] J. Van Mourik, C. Clementi, A. Maritan, A.F. Seno, and J.R. Banavar, cond-mat/9801137.