Protein folding in contact map space

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

Abstract

We represent a protein’s structure by its contact map. Our aim is to identify the unknown fold of a known sequence by minimizing a (free) energy defined in the space of contact maps. To this end, we developed an efficient method to search this space and to generate low energy maps that are also physical. We proved that the standard pairwise approximation to the free energy is unable to stabilize the native fold of a single protein against a set of carefully generated decoys.

1. Introduction

Nearly all the important biochemical tasks of organisms such as catalytic activity, molecular recognition and transmission of signals are performed by proteins [1]. The chemical composition of proteins is relatively simple: they consist of N covalently bound amino acids that form a linear chain. There are 20 different amino acids and hence a protein’s chemical composition is completely defined by its amino acid sequence, i.e., an N-letter “word”, in which each letter stands for a particular amino acid. The amino acid sequence of these macromolecules serves as the blueprint for their conformation, i.e., the specific shapes into which they fold under physiological conditions. This three dimensional structure determines the biological function of a protein. The protein folding problem, i.e., prediction of a protein’s (unknown) native structure from its (known) amino acid sequence is one of the most challenging open problems in computational physics, chemistry and biology. Since the number of sequenced proteins (about 200,000) is increasing steeply [2], while the structure has been determined only for a few thousand, theoretical attacks on the folding problem are a most timely undertaking. Solution of this problem will have far-reaching impact...
on our understanding the function of biologically active macromolecules, as well as on
the practical problems of central importance, like drug design.

As opposed to most problems studied by physicists, this problem is easy to state and
is obviously important. It is, however, an extremely complex problem whose solution
is not much closer today than, say, 20 years ago, two decades of intensive research
notwithstanding.

I review here the work done over a period of three years by myself and several
students and post-docs and limit my discussion to the approach we chose, the problems
we encountered, the way we solved some of them and, for some others, proved that
no solution exists. The main conclusion of our work can be viewed as pessimistic:

We have demonstrated that a certain widely used approximate energy function is
unable to stabilize the native fold of even a single protein.

I will explain this statement and try to outline future directions which may, nevertheless,
bring us closer to a solution of this complex and challenging problem.

2. The contact map representation of structure

A conceptually straightforward attempt to solve the problem is to construct, for
any given molecule, an energy function using the inter-atomic potentials and integrate
Newton’s equations at an energy corresponding to \( kT \). Such a direct attack on the
problem is doomed to fail for several reasons. First of all, the computational time
scales needed to follow the large number of atoms and ions that comprise a protein
lie beyond the possibilities of existing computers. Moreover, the exact potential is not
known (we are looking for a classical effective interaction between ions and atoms;
 furthermore, folding takes place in the presence of water and the water molecules must
be “integrated out”). This state of affairs points to a need for approximate, coarse
grounded or reduced representations of protein structure and derivation of corresponding
energy functions.

A minimalistic representation of a protein’s structure is given by its contact map
[3–7]. The contact map of a protein with \( N \) residues is a \( N \times N \) matrix \( S \), whose
elements are defined as

\[
S_{ij} = \begin{cases} 
1 & \text{if residues } i \text{ and } j \text{ are in contact,} \\
0 & \text{otherwise.}
\end{cases}
\]  

One can define contact between two residues in different ways. In this work, we will
consider two amino acids in contact when their two central carbon atoms (called \( C_\alpha \))
are closer than some threshold (8.5 or 9 Å ). The contact map corresponding to the
known structure of a short protein, 6pti, is shown in Fig. 1, together with a “ribbon
diagram” that depicts the three-dimensional fold of its backbone.

Clearly, the mapping from conformations to contact maps is many-to-one: many
conformations yield the same map. Nevertheless, given all the inter-residue contacts or
even a subset of them, it is possible to reconstruct quite well a protein’s structure, by means of either distance geometry [8], Molecular Dynamics [9] or Monte Carlo [7].

In contrast to detailed atomistic representations, in which the Cartesian coordinates of every ion are specified, the map representation of a protein’s structure is independent of the coordinate frame. This property made contact maps attractive for protein structure comparisons and for searching a limited database for similar structures [3–5]. A more challenging possibility was proposed recently [6]: 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 central premise of this program is that the contact map representation has an important computational advantage; changing a few contacts in a map induces rather significant large-scale coherent moves of the corresponding polypeptide chain [10]. The proposed program however, faces three considerable difficulties:

(1) In the course of a dynamic procedure one may produce non-physical contact maps, that do not correspond to physically realizable conformations of a chain. We ensure that our contact maps are physical by a reconstruction procedure.

(2) One needs an efficient procedure to execute non-local moves in the space of (physical) contact maps.

(3) One should construct a reliable (free) energy function, defined in contact map space, such that low-energy maps can be used to identify the native one.

We turn to discuss briefly these problems and their proposed solutions.

2.1. Identifying physical maps by reconstruction

We developed an efficient method to generate a three-dimensional polypeptide conformation from a contact map.
Assume we produced by some means a contact map $S$; our aim is to determine whether it is physical or not and, in any case, to produce another map, $S'$, which definitely is physical and close to the target map $S$.

To do this, we represent the protein as a “string of beads”, in which each bead stands for an amino acid – the coordinates of the center of a bead are identified with those of the corresponding $C_x$ atom. For any given conformation $c$ of our chain we can calculate the corresponding contact map $S_c$, which, by definition, is physical. Our reconstruction procedure moves the beads of the chain in a manner guided by the target map. That is, a new chain configuration $c'$ is accepted or rejected in a manner that depends on whether $S_{c'}$ is closer to the target than $S_c$. Using an annealing process we arrive at a chain configuration whose contact map is close to the target. The details of this procedure are described elsewhere [7]; here we quote only some of the results.

(1) Experimental contact maps were reproduced with high accuracy. For example, using the map of 6pti shown in Fig. 1 as our target, we arrived at a chain conformation whose contact map contained all the native ones and, in addition, two spurious contacts were generated. Repeating the reconstruction process many times we generated an ensemble of chain conformations, whose average rms distance from the native conformation was about 1.6 Å.

(2) Using non-physical maps as our target: we used several physical maps and corrupted them by randomly turning contacts into non-contacts and vice versa. For noise levels as high as 60% the reconstructed structure was fairly close to the original one, corresponding to the uncorrupted map. An example of such a reconstruction, using a noisy map as the target, is shown in Fig. 2.

2.2. Non-local moves in contact map space

We developed a stochastic method to perform dynamics in contact map space.

Our aim is to generate a large number of contact maps that can serve as candidates for the native structure. To take advantage of the contact map representation we need a procedure that yields very different, uncorrelated maps in reasonable time. Such maps are necessary for protein folding by means of energy minimization, as well as in order to generate decoys needed to test properties of various energy functions. Hence the requirements from any procedure that generates such maps are

• The generated maps should be physical.
• The maps should be “protein-like”; for example they should have secondary structure elements.
• The maps should have low values of the energy (defined in terms of the sequence and the contact map).
• Efficiency – in order to generate large numbers of independent maps in reasonable computing time.

The requirement of physicality is addressed by the method described in Section 2.1; whenever a new candidate map is generated, we use it as the target map of the
reconstruction procedure, and obtain, this way, a contact map which corresponds to a physical “chain of beads”.

In order to move efficiently in contact map space in a way that satisfies the requirements listed above, we introduced a four-step procedure which is described in detail elsewhere [10]. It involves non-local moves, which involve identifying, removing and creating blocks or clusters of contacts; modifying of the resulting map by local moves (adding and removing contacts near existing ones); reconstruction, to obtain a physical map, which then is refined by local moves executed in the conformation space of the chain of beads that has been generated.

These dynamic rules (non-local and local) generate uncorrelated starting points for the reconstruction process. In this way, after each four-step move, we obtain a good candidate map for the native state.

2.3. Pairwise contact approximation for the free energy

We introduce the exact free energy of a contact map and discuss a simple approximation to it.

As explained above, many microscopic configurations of a protein with sequence \( A = (a_1, a_2, a_3, \ldots, a_N) \) are characterized by the same contact map \( S \). We now show that one can define an exact free energy, \( \mathcal{H}(A, S) \), for the assignment of \( S \) to the sequence \( A \). Denote by \( \mathcal{C} \) a micro-state of the system, specified by the coordinates of all atoms of the protein (and of the solvent and any other relevant molecules). The
true, microscopic energy of this configuration is $E(\mathcal{C})$. In thermal equilibrium each micro-state appears with a probability proportional to the corresponding Boltzmann weight $e^{-(1/kT)E(\mathcal{C})}$.

The free energy $\mathcal{H}(A,S)$ (to which we refer simply energy) associated with sequence $A$ and map $S$ is defined as follows:

$$\text{Prob}(S) \propto e^{-\mathcal{H}(A,S)} = \sum_{\mathcal{C}} e^{-(1/kT)E(\mathcal{C})} \Delta(\mathcal{C},S),$$

(2)

where

$$\Delta(\mathcal{C},S) = \begin{cases} 1 & \text{if } S \text{ consistent with } \mathcal{C}, \\ 0 & \text{otherwise} \end{cases}$$

$\Delta(\mathcal{C},S)$ is a “projection operator” that ensures that only those configurations $\mathcal{C}$ whose contact map is $S$ contribute to the sum (2). In other words, only those micro-states whose contact map is $S$ contribute to the sum and hence to $\mathcal{H}(A,S)$.

This definition of the (free) energy of a map is exact; it is nothing but the negative log of the probability of observing the map $S$ for sequence $A$. Therefore $\mathcal{H}(A,S)$ has an important property; in as much as the native fold’s contact map, $S_0$, has the highest probability of appearing (in an equilibrium ensemble of configurations under physiological conditions), the corresponding (free) energy $\mathcal{H}(A,S_0)$ is the lowest among all possible $\mathcal{H}(A,S)$, i.e.,

$$\mathcal{H}(A,S_0) < \mathcal{H}(A,S), \quad \forall S \neq S_0. \quad (3)$$

The main problem with this exact energy is that the sum (2) is impossible to carry out. Therefore, one takes various phenomenologically motivated guesses for the form of $\mathcal{H}(A,S)$, that presumably would have been obtained had the sum been carried out. This approach is related in spirit to the phenomenological Landau–Ginzburg-type free energy used in several areas of condensed matter physics. We also start from the simplest approximate form for this complicated function – that of the pairwise contact energy:

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

(4)

That is, if there is a contact between residues $i$ and $j$, the parameter $w(a_i,a_j)$, which represents the free energy gained by bringing amino acids $a_i$ and $a_j$ in contact, is added to the energy.

The 210 parameters $w(a_i,a_j)$ were derived in different ways. The first idea was to derive a particular set of energy parameters from amino acid pairing frequencies observed in available crystallographic structures [11–13]. It was also realized that parameter derivation can be formulated as an optimization problem. For a fixed set of sequences with their native maps and for a fixed set of alternative contact maps, one defines a cost function in parameter space and looks for the set of parameters of “minimum cost”. Goldstein et al. [14] maximized the ratio $R$ between the width of the distribution of the energy and the average energy difference between the native state
and the unfolded ones. More recently, Mirny and Shakhnovich [15] expressed the same quantity analytically as a function of the energy parameters, which were then derived by optimization.

An alternative method was proposed later by Maiorov and Crippen [16]. Energy parameters were derived by requiring that the observed native structures should be the lowest in energy among an ensemble of alternative conformations (called decoys). This requirement, when translated to contact maps, means that the native map \( S_0 \) should have lower energy than every decoy map \( S \), i.e., inequalities of the form

\[ H^{\text{pair}}(A, S_0, w) < H^{\text{pair}}(A, S, w) \]  

should hold for each decoy. The same general idea was subsequently used by Maritan and collaborators [17,18] and also by our group [19–21].

### 3. Can one use the pairwise contact approximation to fold proteins?

It is important to emphasize from the outset the difference between the question we posed and that of other workers. The generally defined aim has been to propose ways to calculate a set of contact energy parameters that are, in some sense, optimal. The question we asked was different, e.g. *Is there a set of 210 contact parameters, that satisfies inequalities (5) even for a single protein, against a set of carefully selected decoys?*

This question is of paramount importance if one hopes to use the resulting energy function with the aim of identifying, for a given sequence, the native map as the one that minimizes the energy. If the answer to the question is negative, one will never be able to find the native map this way, no matter which method is used to derive the energy parameters.

As it turns out, the qualifying statement “carefully selected decoys” is of central importance. A very simple method of producing physical decoys is that of *gapless threading* [21]. In essence one reads off from the databank of known structures a set of proteins that are longer than the one whose structure is to be determined, and threads the shorter chain through all the longer structures. Each sub-chain generated this way is a piece of a real polypeptide chain and can serve as a decoy conformation. The decoys obtained this way are, however, fairly poor, and one can find contact energy parameters that stabilize the native map of the short protein against all of them.

If, however, the decoy conformations are generated by the search procedures outlined above, one can show that for a large enough set of these “hard” decoys *there is no set of contact energies that satisfies all the inequalities* of Eq. (5). We proved this statement by a perceptron learning procedure [22]: the inequalities are linear in the contact energies and hence each inequality is an “example” to be “learned” by a perceptron. In the course of learning, the procedure moves in the space of contact energies and calculates a monotonously increasing quantity, \( d \). If there is no solution to the inequalities, the \( d \) will exceed a critical value in a finite number of learing steps.
If a solution does exist, the procedure will find a solution before \( d \) has reached its critical value.

We have shown that for crambin no solution exists (i.e., the problem is unlearnable); hence it is not possible to find contact energy parameters that stabilize the native map of crambin against all the decoys that were generated. In order to prove this, we had to work hard and generate a large number of low-energy decoys. Had we used decoys obtained by gapless threading or by a less powerful search method, we could have easily been misled to believe that a solution does exist.

A similar result was obtained for a family of six immunoglobulins.

This finding indicates that the simple pairwise contact approximation to the free energy is not sufficient to fold proteins.

4. Discussion

There are several possible directions to explore.

(1) Controlled inclusion of additional energy terms, may help to attain foldability. We have already added hydrophobic (solvation) terms, and extended the number of contact energy parameters to 420 in two different ways [23]. The first assigns different energies to two different ranges of separation between the \( C_\alpha \) atoms. The second allows dependence of the energy of a given pair of residues on whether they are buried in the globule or on its surface. Both extensions improve significantly the ability of the energy function to stabilize native folds. Inclusion of hydrogen bonds, interactions between charges and multi-body interactions will probably provide additional improvement.

(2) Most importantly, we believe that the requirement that increased numbers of inequalities of the form (5) be satisfied will allow a step-by-step systematic improvement of the energy function.

(3) 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.

Acknowledgements

The contribution of Michele Vendruscolo to the work described here has been of central importance. I also benefited from collaborations with L. Mirny, E. Kussell, R. Najmanovich, J. Lebowitz, B. Subramanian, I. Kanter, K. Park, G. Getz and E. Shakhnovich. My research was supported by the Germany–Israel Science Foundation (GIF), the Minerva Foundation and the US–Israel Binational Science Foundation (BSF).
References