Recovery of protein structure from contact maps

Michele Vendruscolo[1], Edo Kussell[2] and Eytan Domany[1]

Background: Prediction of a protein's structure from its amino acid sequence is a key issue in molecular biology. While dynamics, performed in the space of two-dimensional contact maps, eases the necessary conformational search, it may also lead to maps that do not correspond to any real three-dimensional structure. To remedy this, an efficient procedure is needed to reconstruct three-dimensional conformations from their contact maps.

Results: We present an efficient algorithm to recover the three-dimensional structure of a protein from its contact map representation. We show that when a physically realizable map is used as target, our method generates a structure whose contact map is essentially similar to the target. Furthermore, the reconstructed and original structures are similar up to the resolution of the contact map representation. Next, we use nonphysical target maps, obtained by corrupting a physical one; in this case, our method essentially recovers the underlying physical map and structure. Hence, our algorithm will help to fold proteins, using dynamics in the space of contact maps. Finally, we investigate the manner in which the quality of the recovered structure degrades when the number of contacts is reduced.

Conclusions: The procedure is capable of assigning quickly and reliably a three-dimensional structure to a given contact map. It is well suited for use in parallel with dynamics in contact map space to project a contact map onto its closest physically allowed structural counterpart.

Addresses: [1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Chemistry, Harvard University, Cambridge, MA 02138, USA.

Correspondence: Michele Vendruscolo
E-mail: femichel@wicc.weizmann.ac.il

Key words: contact, distance, dynamics, map, protein, reconstruction

Received: **08 May 1997**
Revisions requested: **27 May 1997**
Revisions received **07 July 1997**
Accepted **08 July 1997**

Published: **10 September 1997**
http://biomednet.com/elecref/1359027800200295

Folding & Design 10 September 1997, **2**:295–306

Introduction

Considerable effort has been devoted to finding ways to predict a protein's structure from its known amino acid sequence $\mathbf{A} = (a_1, a_2, \ldots a_N)$. The contact map of a protein is a particularly useful representation of its structure [1,2]. For a protein of N residues, the contact map is an $N \times N$ matrix $\mathbf{S}$, whose elements are $S_{i,j} = 1$ if residues i and j are in contact and $S_{i,j} = 0$ otherwise. 'Contact' can be defined in various ways; for example, in a recent publication Mirny and Domany [3] defined contact $S_{i,j} = 1$ when a pair of heavy (all but hydrogen) atoms, one from amino acid i and one from j, whose distance is < 4.5 Å can be found. Secondary structures are easily detected from the contact map. α-Helices appear as thick bands along the main diagonal since they involve contacts between one amino acid and its four successors. The signatures of parallel or anti-parallel β-sheets are thin bands, parallel or anti-parallel to the main diagonal. On the other hand, the overall tertiary structure is not easily discerned. The main idea of Mirny and Domany [3] was to use this representation to perform a search, executed in the space of possible contact maps $\mathbf{S}$, for a fixed sequence $\mathbf{A}$, to identify maps of low 'energy' $\mathcal{H}(\mathbf{A}, \mathbf{S})$. They defined the energy $\mathcal{H}(\mathbf{A}, \mathbf{S})$ as the negative logarithm of the probability that structures whose contact map is $\mathbf{S}$ occur for a protein with the sequence $\mathbf{A}$; therefore, a map of low energy corresponds to a highly probable structure.

One of the most problematic aspects of their work was that by performing an unconstrained search in the space of contact maps, i.e. freely flipping matrix elements from 1 to 0 and vice versa, one obtains maps of very low energy which have no physical meaning, since they do not correspond to realizable conformations of a polypeptide chain. To overcome this problem, Mirny and Domany introduced heuristic restrictions on the possible changes one is allowed to make to a contact map, arguing that if one starts with a physically realizable map, the moves allowed by these restrictive dynamic rules will generate maps that are also physically realizable. Even though their heuristic rules did seem to modify the dynamics in the desired way, there is no rigorous proof that indeed one is always left in the physical subspace, there is no clear evidence that the resulting rules are not too restrictive and, finally, the need to start with a physical fold, copied from a protein of known structure, may bias the ensuing search and get it trapped in some local minimum of the energy.

The aim of the present publication is to present a method to overcome these difficulties. The idea is to provide a test that can be performed 'online' and in parallel with the dynamics in the space of contact maps which will 'project' any map onto a nearby one that is guaranteed to be in the subspace of physically realizable maps. That is, for a given target contact map $\mathbf{S}$, we search for a conformation that a 'string of beads' can take, such that the contact map **S**′ of our string is similar (or close) to **S**. Needless to say, the contact map associated with a string of beads is, by definition, physical.

This particular aim highlights the difference between what we are trying to accomplish and the goals of existing methods [2,4–11]. These methods use various forms of distance geometry [12,13], supplemented by restricted molecular dynamics [14] or simulated annealing [15], to construct three-dimensional structures from distance information. Our method addresses a different problem: how to convert a possibly ill-defined nonphysical set of contacts to a legitimate one. We should emphasize here the distinction between a contact map and a distance map. In a contact map, a minimal amount of information is available — given a pair of amino acids, we know only whether they are in contact or not, i.e. only lower and upper bounds on their separation are given. A distance matrix, on the other hand, presents real-valued distances between pairs of amino acids. Therefore, it is considerably harder to reconstruct a structure from a contact map than from a distance matrix. Rather than being concerned with obtaining a structure that is close to a real experimental one, we mainly want to check whether a contact map **S** is physically possible or not, and if not, to propose some **S**′ that is physical and, at the same time, is not too different from **S**. The three-dimensional structure is in our case a means, rather than an end. The method has to be fast enough to run in parallel with the search routine (which uses $\mathcal{H}(\mathbf{A}, \mathbf{S})$ to identify candidate maps **S** of low energy). Another important requirement is to be able to recover contacts that do not belong to secondary structure elements and may be located far from the map's diagonal. Such contacts are important to nail down the elusive global fold of the protein. We believe that the main advantage of performing a dynamic search in the space of contact maps is the ease with which such contacts can be introduced, whereas creating them in a molecular dynamics or in a Monte Carlo procedure of a real polypeptide chain involves coherent moves of large sections of the molecule — moves that take a very long time to perform. To make sure that this advantage is preserved, our method must be able to efficiently find such conformations, if they are possible, once a new target contact has been proposed.

Existing methods are capable of dealing with the noise that can arise from experimental errors in the distances derived from NOESY spectra, from uncertainty in the identification of the pair of atoms that gave rise to a particular distance signal, or from distances that are assigned to wrong atom pairs. Typically, in a distance geometry approach, distances are first filtered by applying triangle inequalities, and then by some iterative embedding procedure [12,13]. Till now, to our knowledge, no work has been specifically aimed at assessing quickly and reliably whether a given set of contacts is physically realizable or not. The main conclusion of the work of Havel *et al.* [2] is that it is possible to reconstruct a structure from the knowledge of the correct set of contacts. Successive work within the distance geometry framework, for the reasons explained above, has focused mainly on the treatment of noise in distance assignment [7,10]. A comparison with the work of Bohr and co-workers [6] shows that our method is considerably more robust against inconsistencies in the assignment of the contacts in the map. This is to be expected, since we use a stochastic method whereas Bohr and co-workers use a deterministic (gradient descent) one; inconsistencies and noise give rise to multiple minima of the cost function, which are overcome efficiently by stochastic methods.

We are currently working on combining the method presented here with dynamics in the space of contact maps. The results of the combined procedure will be presented in a future publication. In this paper, we give a detailed explanation of the method. We show how it works on native maps of proteins with the number of residues ranging from $N = 56$ to $N = 581$. Success of the algorithm is measured in terms of the number of contacts recovered and the root mean square displacements of the recovered three-dimensional structures from the native ones. We study the answers given by our algorithm when it faces the task of finding a structure, using a nonphysical contact map as its target. As the first test, we added and removed contacts at random in a physical map and found that the reconstructed structure did not change by much, i.e. we could still reconstruct the underlying physical structure. As a second test, we used the constrained dynamic rules proposed by Mirny and Domany [3]: starting from an experimental contact map, we obtained a new map by a denaturation/renaturation computer experiment. Since the rules are heuristic, this map is not guaranteed to be physical. Our reconstruction method projects a nonphysical map onto one that is close to it and physically allowed.

We then discuss the extent to which the quality of the structure obtained from a contact map gets degraded when the number of given contacts is reduced. This issue has considerable importance beyond the scope of our present study, since experimental data (e.g. from disulfide bridge determination, crosslinking studies and NMR) are often available for only a small number of distance restraints. Clearly, the more restraints one has the smaller is the number of possible conformations of a chain that are consistent with the constraints contained in the contact map. The issue we address is when this reduction of the number of possible conformations suffices to define the corresponding structure with satisfactory accuracy.

Results and discussion

Methodology

In this work, we adopt a widely used definition of contact: two amino acids, a_i and a_j, are in contact if their distance

$d(a_i, a_j)$ is less than a certain threshold d_t. The distance $d(a_i, a_j)$ is defined as:

$$d(a_i, a_j) = \left|\mathbf{r}_i - \mathbf{r}_j\right| \tag{1}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the coordinates of the C_α atoms of amino acids i and j.

The algorithm is divided into two parts. The first part, growth, consists of adding one monomer at a time, i.e. a step-by-step growth of the chain. The second part, adaptation, is a refinement of the structure, obtained as a result of the growth stage, by local moves. In both stages, to bias the dynamics, we introduce cost functions defined on the basis of the contact map. Such cost functions contain only geometric constraints and do not resemble the true energetics of the polypeptide chain.

Growth

Single monomer addition

Suppose we have grown $i - 1$ monomers and we want to add point i to the chain. To place it, we generate at random N_t trial positions (typically $N_t = 10$):

$$\mathbf{r}_i^{(j)} = \mathbf{r}_{i-1} + \mathbf{r}^{(j)} \tag{2}$$

where $j = 1,\ldots,N_t$. The direction of the vector $\mathbf{r}^{(j)}$ is selected from a uniform distribution in the region of the space allowed by the stiffness of the C_α chain. From a statistical analysis of several proteins in the PDB, we derived the lower bound for the angle between two successive C_α, which is expressed by the condition $\mathbf{r}_{i-1}{\cdot}\mathbf{r}_i^{(j)}/\left|\mathbf{r}_{i-1}\right|\left|\mathbf{r}_i^{(j)}\right| <$ -0.3. The length of $\mathbf{r}^{(j)}$ is distributed normally with average r_a and variance σ. Since in our representation monomers identify the C_α positions, we took $r_a = 3.79$ and $\sigma = 0.04$. We assign a probability $p^{(j)}$ to each trial in the following way. For each trial point $\mathbf{r}_i^{(j)}$, we calculate the contacts that it has (see equation 1) with the previously positioned points $\mathbf{r}_1,\ldots,\mathbf{r}_{i-1}$. Contacts that should be present, according to the given contact map, are encouraged and contacts that should not be there are discouraged according to a cost function E_g that will be specified below. One out of the N_t trials is chosen according to the probability:

$$p^{(j)} = \frac{e^{-E_g^{(j)}/T_g}}{Z} \tag{3}$$

where the normalization factor is given by:

$$Z = \sum_{j=1}^{N_t} e^{-E_g^{(j)}/T_g} \tag{4}$$

The notation for the cost function E_g and for the parameter T_g that guide the growth are chosen in the spirit of the Rosenbluth method [16] to suggest their reminiscence to energy and temperature, respectively.

Chain growth

The step-by-step growth presented in the previous section optimizes the position of successive amino acids along the sequence. The main difficulty in the present method is that the single step of the growing chain has no information on the contacts that should be realized many steps (or monomers) ahead. To solve this problem, we carry out several attempts (typically 10) to reconstruct the structure, choosing the best one. In practice, this is done as follows.

For each attempt, when position $\mathbf{r}^{(j)}$ is chosen for monomer i according to equation 3, its probability is accumulated in the weight:

$$W_i = \prod_{k=1}^{i} p_k^{(j)} \tag{5}$$

When we have reached the end of the chain, we store the weight W_N. The trial chain with the highest W_N is chosen.

Cost function

The probabilities in equation 3 are calculated using the following cost function:

$$E_g^{(j)} = \sum_{k=1}^{i-1} f_g\left(r_{ik}^{(j)}\right) \tag{6}$$

where $r_{ik}^{(j)} = \left|\mathbf{r}_i^{(j)} - \mathbf{r}_k\right|$ and:

$$f_g(r_{ik}) = d \cdot \mathrm{a}_g(S_{ik}) \cdot \vartheta(d_t - r_{ik}) \tag{7}$$

The enhancing factor $d = i - k$ is introduced to guide the growth towards contacts that are long ranged along the chain; ϑ is the Heaviside step function and the constant a_g can take two values: $\mathrm{a}_g(S_{ik} = 1) \geq 0$ and $\mathrm{a}_g(S_{ik} = 1) \leq 0$. That is, when a contact is identified in the chain, i.e. $r_{ik} < d_t$, it is either 'rewarded' (when the target map has a contact between i and k) or penalized. In this work, we have grown chains with $\mathrm{a}_g(0) = 0$. In this case, for a given contact map S, the function f_g rewards only those contacts that are realized and should be present. No cost is paid if contacts that are not in the map are realized by the chain (false positive contacts). Typically, we chose the values $\mathrm{a}_g(1) = -1.0$ and $T_g = 1$.

Adaptation

When we have grown the entire chain of N points, we refine the structure according to the following scheme. We choose a point i at random and try, using a crankshaft move [17], to displace it to $\mathbf{r}'_i$, keeping fixed the distances from both points i–1 and i+1. We use a local cost function $E_a^{(i)}$:

$$E_a^{(i)} = \sum_{k=1}^{i-1} f_a(r'_{ik}) \tag{8}$$

where $r'_{ik} = |\mathbf{r}' - \mathbf{r}_k|$ and:

$$f_a(r_{ik}) = \mathrm{a}_a(S_{ik}) \cdot \vartheta(d_t - r_{ik}) \tag{9}$$

Note that the enhancing factor d has been omitted, so that f_a does not favor contacts between monomers that are distant along the chain. The displacement is accepted with probability π, according to the standard Metropolis prescription:

$$\pi = \min(1, \exp(-\Delta E_a / T_a) \tag{10}$$

where ΔE_a is the change in the cost function E_a induced by the move and T_a is a temperature-like parameter, used to control the acceptance ratio of the adaptation scheme. A key ingredient of our method is annealing [18]. As in all annealing procedures, the temperature-like parameter T_a is decreased slowly during the simulation to help the system find the groundstate in a rugged energy landscape.

In our method, however, instead of using simulation time as a control parameter on the temperature, we chose the number n of missing contacts. Two regimes were roughly distinguished. In the first regime, many contacts are missed and the map is very different from the target one. In the second regime, few contacts are missed and the map is close to the target. The parameters a_a and T_a are interpolated smoothly between values suitable for these two limiting cases. In the first regime, we strongly favor the recovery of contacts that should be realized, whereas in the second regime, we strongly disfavor contacts that are realized but should not be present. We set:

$$T_a^{(n)} = T_a^f + (T_a^i - T_a^f)\sigma(n) \tag{11}$$

The function $\sigma(n)$ interpolates between the initial value a^i and the final value a^f:

$$\sigma(n) = \frac{2}{1 + e^{-\alpha_g n}} - 1 \tag{12}$$

By choosing a^i, a^f, T_a^i, T_a^f and α_g we define the two regimes, far from and close to the target map.

Chirality

A contact map contains no information about chirality. When an overall structure is reconstructed, the mirror image conformation is equally legitimate, having the same contact map. Since existing proteins do have a definite chirality, we are allowed to supply this information.

The C_α–C_α contact constraints allow a local refinement of the reconstructed structure, with no loss in our geometrical cost function. α-Helices can be detected as a thick band along the main diagonal of a contact map. A preliminary scan of the map identifies the sections that should be reconstructed as α-helices. Next, we push the C_αs in the α-helices to positions that give the correct chirality, which is formally defined as the normalized triple product:

$$c_i = \frac{\mathbf{v}_i \times \mathbf{v}_{i+1} \cdot \mathbf{v}_{i+2}}{|\mathbf{v}_i| \cdot |\mathbf{v}_{i+1}| \cdot |\mathbf{v}_{i+2}|} \tag{13}$$

where $\mathbf{v}_i = \mathbf{r}_i - \mathbf{r}_{i-1}$.

In a typical α-helix, $c_i = c_o = 0.778$ [8]. To refine the chirality of the preliminary chain obtained from the map by growth and adaptation as described above, we perform an additional Monte Carlo procedure. This procedure uses as 'energy' a function that strongly favors the value quoted above for c:

$$E_c = a_c\left[\frac{2}{1 + exp[-\alpha_c(c - c_o)^2]} - 1\right] \tag{14}$$

Since our Monte Carlo moves do not conserve the C_α–C_α bond length, we added a term E_b to the energy function:

$$E_b = a_b\left[\frac{2}{1 + exp[-\alpha_b(r - r_a)^2]} - 1\right] \tag{15}$$

At each step, a monomer i is selected randomly and its position displaced to:

$$\mathbf{r}'_i = \mathbf{r}_i + \delta \tag{16}$$

where δ is a small random vector. The total variation in the cost function, $E_a + E_c$, is evaluated with:

$$c = \frac{c_{i-1} + c_i + c_{i+1}}{3} \tag{17}$$

used in equation 14.

Growth and adaptation yield a particular recovered structure, **C**. We first create $\bar{\mathbf{C}}$, the mirror image of **C**, and use both structures as initial states for the final refinement procedure. Usually either **C** or $\bar{\mathbf{C}}$ evolves to a structure with the correct value of the average chirality rather quickly by our Monte Carlo process, while the mirror image does not, due to the lower compatibility of the latter structure with the correct chirality.

All-β structures deserve a special treatment, since we cannot use chirality to filter out mirror images. Proteins 1acx and 1tlk in Table 1 are all-β. We used the following empirical procedure. We generated several hundreds of conformations and computed the relative distances D (as from equation 20). We found that reconstructed structures cluster in two sets. Data presented in Figure 1 actually refer to the reconstructed structures in the set that has a closer distance with the known experimental conformation.

Table 1

List of PDB proteins used to test the reconstruction procedure.

Protein	N	N_c	Protein	N	N_c
6pti	56	342	2sodO	151	1066
2ci2	65	350	1bmv1	185	1084
1tlk	103	610	1akeA	214	1348
5cytR	103	644	1trmA	223	1595
1ltsD	103	571	1abe	305	2179
9rnt	104	623	1pii	452	3070
1acx	108	652	3gly	470	3383
2trxA	108	628	3cox	500	3680
1f3g	150	1049	1gal	581	4369
1aak	150	922			

Alternative strategies

We devote the rest of this section to the discussion of alternative strategies that we have tried. Some of these may prove to be useful in future applications for more complex problems, but we have found that they are not necessary for the specific task dealt with in this work. We present these experiments because they underscore some nontrivial aspects of the problem.

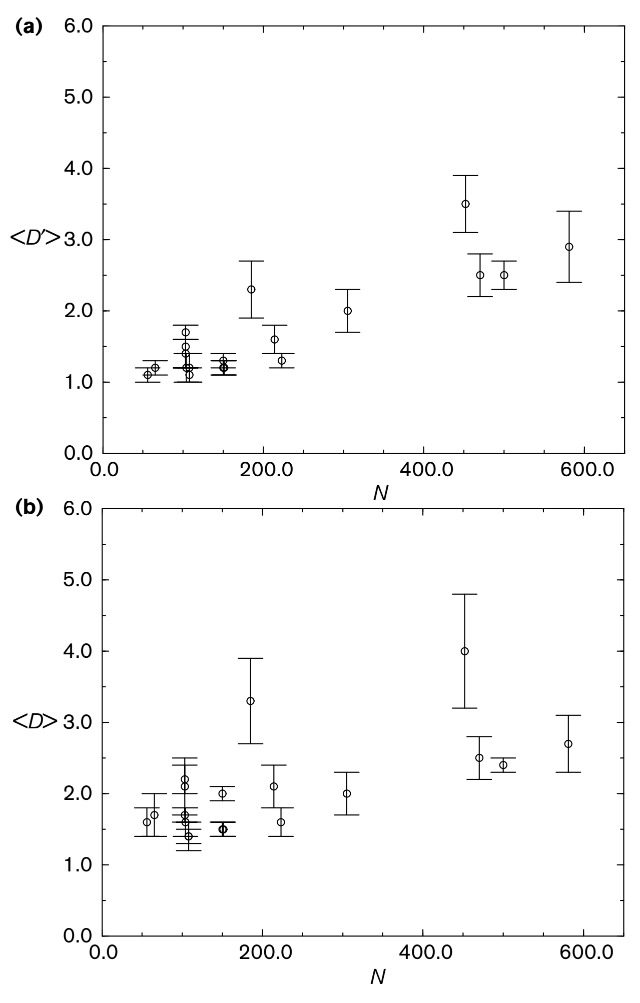
Adaptation alone

It is interesting to note that for short chains ($N < 200$) we can skip the growth stage; starting from a random structure, the adaptation procedure alone suffices to recover the correct set of contacts. The computer time needed for recovery, however, increases very fast with N. Since we are interested in recovering the structure in as short a time as possible, growth must be used, especially for long chains. We found that starting the adaptation stage from a grown (versus random) initial chain speeds up the procedure by a factor of about 10 for proteins of length $N \simeq 100$. Moreover, for longer chains ($200 < N < 1000$), the cost function landscape is rougher and reconstruction by adaptation alone becomes unfeasible.

Piecewise growth

The importance of local contacts (i.e. contacts that involve amino acids nearby along the chain) versus nonlocal ones has been discussed recently in the literature [19,20]. In these works, evidence is given in support of the idea that nonlocal interactions are decisive in stabilizing the folded structure. A long-standing alternative hypothesis [21] is that the folded structure is stabilized mainly by local interactions. We can test in the present work whether the purely geometrical (versus energetic) part of the reconstruction can or cannot be helped much by emphasizing the role of local contacts. To this end, we used secondary structure elements as guidelines for the step-by-step growth. To implement this kind of growth instead of growing the entire chain of N amino acids, a section of M steps is built, with M ranging from 4 to 10 to match the size of a turn in an α-helix or in a β-sheet. A set of sections is generated and the one with the best weight, according to equation 5, is chosen. Consistent with the findings in [19,20], we found that this secondary structure driven growth does not help much in the recovery.

Figure 1



Average distances **(a)** ⟨D'⟩ and **(b)** ⟨D⟩ versus chain length N for the proteins listed in Table 1.

A related idea is to optimize the relative positions of successive secondary structure elements. To realize this, we have tried the following method (similar to that above). Sections of chain of M steps are grown, but now M is chosen randomly from 20 to 50. In this way, we explore the space forward on the length scale of secondary structures to hook important contacts, i.e. those that fix the positions of secondary structures relative to each other. This scheme biases the growth to build a bridge to the next important contact, which usually is either inside a secondary structure or between different secondary structures. This method also allows one to go back if too many mistakes are detected. As was the case for the previous attempted method, our experience suggests that this forward exploration is not necessary for solving the present task.

For multidomain proteins, we tried growing one domain at a time and then refining the structure by an adaptation cycle. The overall results were, however, only slightly affected. By allowing the growth to start alternatively from either end, we have verified that no bias is introduced if the growth is started always from the same end as discussed above.

An alternative idea that we tried is to bias the growth towards reaching a particular 'fixed point' [22]. For example, if it is known that two amino acids i and $j = i + k$ should be in contact, then it is possible to bias the formation of a loop of length k. This method is well suited for very sparse contact maps, where it is easy to identify target points for the growing chain. We have verified that in dense maps the reconstruction speed is not increased by this scheme, due to the cumbersome identification of the target points.

Different cost functions

As mentioned above, we have used $a_g(0) = 0$. In general, this could lead to an overcompaction of the final structure. To assign unfavorable weight to false positive contacts, we should set $a_g(0) > 0$. This would introduce frustration to the growth process, however, since it is guided by positive and negative energies. We discuss here the results of a possible method that we have tested to bias the growth away from conformations that contain 'spurious' contacts, i.e. contacts not present in the map. We have assigned a positive cost $a_g(0) > 0$ to generating a spurious contact (i,j) if the closest existing contact (as measured on the map) is more than a distance of R units away, e.g.:

$$\min_{(h,k)} \sqrt{(i-h)^2 + (j-k)^2} > R \tag{18}$$

where (h,k) runs over all the existing contacts $S_{h,k} = 1$ in the given map. For the proteins we have analyzed (see Table 1), we have extensively scanned possible values for R and $a_g(0)$. We found that that there is a strongly frustrated regime for small R and large $a_g(0)$ where reconstruction is hindered, and a weakly frustrated regime for large R and small $a_g(0)$ where the efficiency of the reconstruction is only slightly improved with respect to using $a_g(0) = 0$. The intermediate regime (typically $R = 5$–10 and $a_g(0) = 0.1$–0.01 for the values $a_g(1) = -1.0$ and $T_g = 1$ given above) may prove to be useful for proteins longer than those tested is the present work.

As for the functional form of the cost function, another possible choice, following Bohr and co-workers [6,8], is to smooth the step function that defines a contact with a sigmoid. Again, we did not find this necessary to achieve fast reconstruction.

In principle, it would be possible to add to the function f_g of [7] a hard core repulsion:

$$h(r) = \sigma_0 (r - r_0)^{-\alpha} \tag{19}$$

to try to overcome a general problem that arises when working with distance inequalities: an overcompaction of the globule, as measured e.g. by the gyration radius. In practice, a good recovery prevents the overlap between C_αs automatically and the addition of such a term is not necessary.

Experimental contact maps

In this section, we present results concerning the reconstruction of experimental contact maps as taken from the PDB. Since our purpose, as explained in the Introduction, is to use the reconstruction in connection with dynamics, we chose $d_t = 9$ Å to obtain the most faithful representation of the energy of the protein [3]. Such a threshold is determined by the requirement that the average number of C_α–C_α contacts for each amino acid is roughly equal to the respective numbers obtained with the all-atom definition of contacts.

Two dissimilarity measures between structures are widely used. The most commonly used [23–25] is the root mean square distance D:

$$D = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{r}'_i)^2} \tag{20}$$

where one structure is translated and rotated to get a minimal D. Another possible choice is the distance D':

$$D' = \sqrt{\frac{1}{N^2}\sum_{i,j=1}^{N} (\mathbf{r}_{i,j} - \mathbf{r}'_{i,j})^2} \tag{21}$$

The relation between D and D' is derived by Cohen and Sternberg [26]. The dissimilarity measure between contact maps is defined as the Hamming distance:

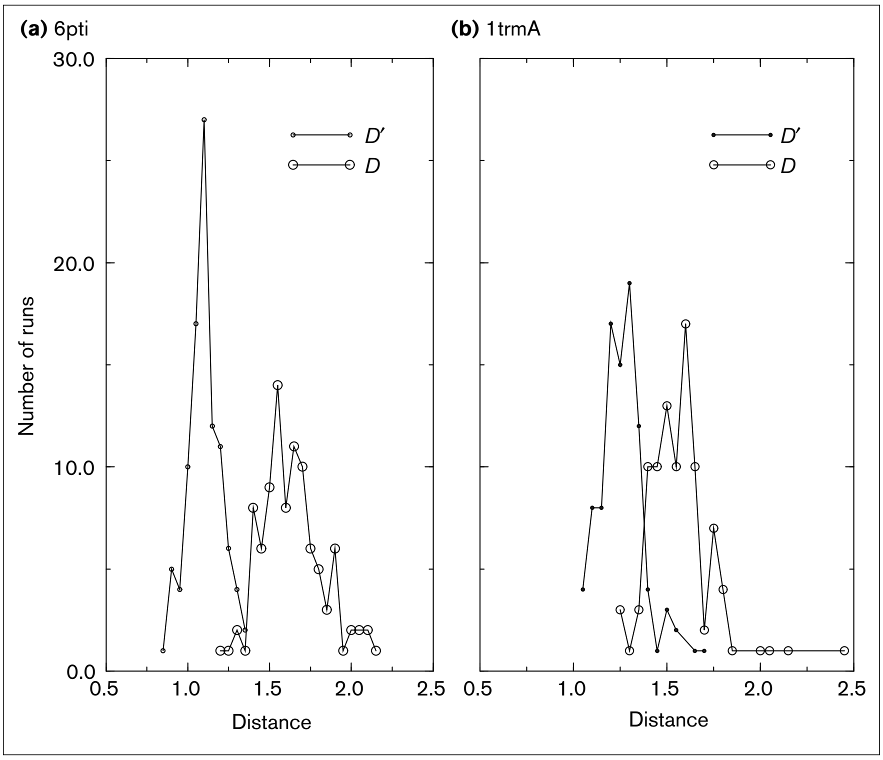
$$D^{map} = \sum_{j>i} \left| S_{ij} - S'_{ij} \right| \tag{22}$$

which counts the number of mismatches between maps **S** and **S′**.

For several proteins, we present in Figure 1 the distances D and D' plotted versus the chain length N. The proteins considered (with their respective lengths N and number of contacts N_c) are reported in Table 1.

The values of D and D' presented in Figure 1 were obtained by averaging over 100 reconstruction runs for chains up to $N = 223$ and over 10 runs for longer chains.

Figure 2



Distances *D′* and *D* for the 100 runs used to test the reconstruction procedure. Data are presented for proteins **(a)** 6pti and **(b)** 1trmA.

Error bars represent the variances as obtained from the corresponding sets of runs, as shown, for example, for proteins 6pti (bovine pancreatic trypsin inhibitor) and 1trmA (rat trypsin, chain A) in Figure 2.

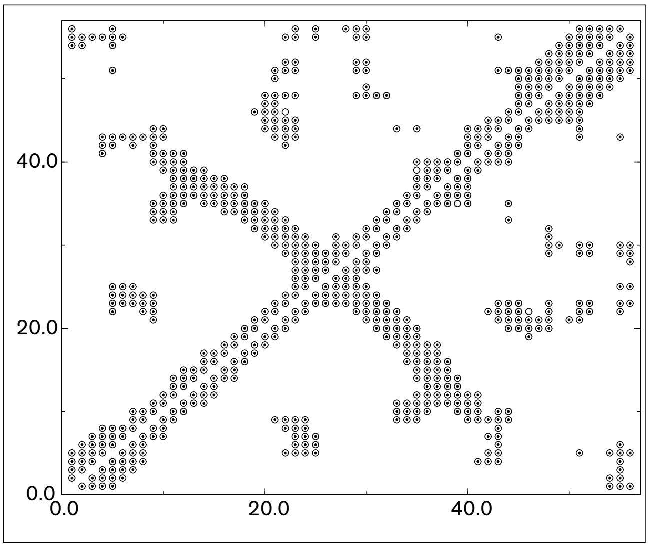
In Figure 3, we show the contact map for the protein 6pti, N = 56, as taken from PDB, that was used as a target to construct a chain. The contact map of a typical reconstructed chain is also shown. In this particular case, none of the 342 original contacts was missed and only two false positive contacts were added. These are close to clusters of correct contacts, indicating slight local differences with the crystallographic structure. The distances recorded in this case were D' = 1.06 and D = 1.56.

In Figure 4, we show similar results for the larger protein 1trmA, with N = 223 and 1595 contacts. For clarity, we have separated the experimental contact map from the reconstructed one. In the particular case shown, there are nine missing contacts and 84 false positives, and the corresponding distances are D' = 1.34 and D = 1.59. On average, in the 100 runs, six contacts were missed and 75 false positives were spuriously added. As in the case of 6pti, wrong contacts are mostly neighboring correct ones. Averages distances are $\langle D'\rangle$ = 1.3 ± 0.1 Å and $\langle D\rangle$ = 1.6 ± 0.2 Å (see also Figure 2). The corresponding conformations for both 6pti and 1trmA are shown in Figure 5. (These superpositions can be compared with those in Figure 2 of [8].)

Using the distances D and D' to assess the quality of our results is misleading, since we are searching only for a chain that reproduces the contacts of a given map, whereas D and D' measure similarity between structures. Information that is all-important to obtain low values for D and D', such as the positions of amino acids that belong to loops or slight rotations of secondary structures, is not contained at all in the map. For example, for the two-domain protein 1pii (phosphoribosylanthranilate isomerase), which has the largest distance in Figure 1, only two out of 3070 contacts were missed, on average, in the 10 reconstruction runs. However, changes in the relative orientations of the two domains lead to large distances. In fact, the target native maps were nearly perfectly reconstructed for all proteins tested.

We turn now to estimating the range of expected values for D and D'. The lower limit of our resolution for the chain, imposed by the geometrical constraints contained in the contact map, is about 1 Å. To support this statement, we present the results of the following test. We subjected the native maps of 6pti, 1acx (actinoxanthin) and 1trmA from the PDB to an adaptation cycle at low T_a. No native contacts were lost and no spurious ones were generated throughout the simulation, even though the structure (i.e. the positions of the beads) did vary; the most probable value for the distance D' between the generated structures was found to be around 1 Å. This

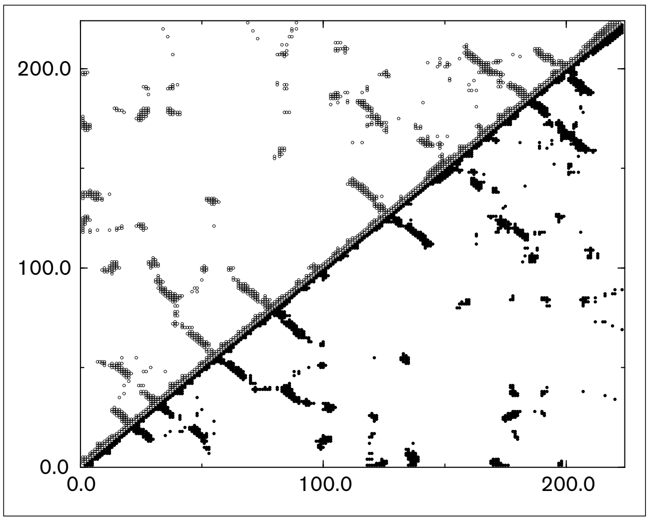
Figure 3



Contact map for protein 6pti for a threshold $d_t = 9$ Å. Dots are the PDB data, open circles the output of the reconstruction procedure. None of the target contacts is missed and two spurious ones are added.

result clearly indicates the extent to which the contact map representation does not allow us to nail down one specific structure to arbitrary precision. This ambiguity is compatible with the usual experimental resolution of PDB structures and hence the contact map representation is useful. Moreover, from low temperature flash photolysis experiments [27], X-ray diffraction result analyses [28] and molecular dynamics simulations [29], the native fold of a protein is believed to consist of a set of conformational substates rather than of a unique structure [30]. The upper limit of the range of expected distances in our reconstruction is that between two completely unrelated structures, which can be as large as 15 Å.

Figure 4



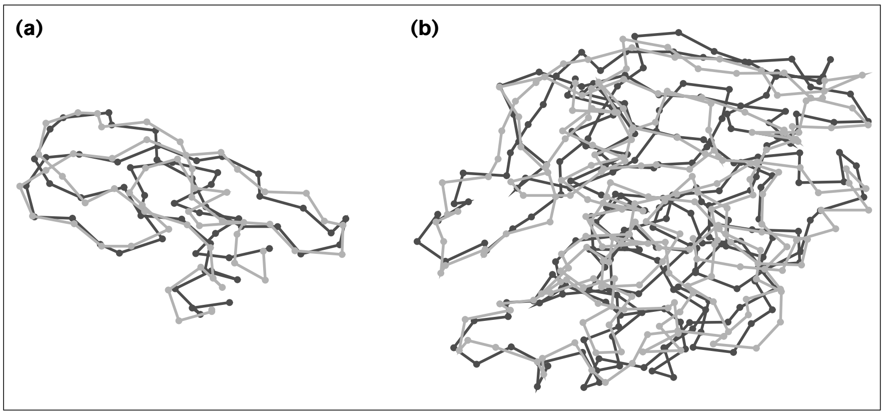
Contact map of protein 1trmA. Experimental contact map (above diagonal) and reconstructed one (below diagonal).

The conclusion of our studies is that our method produces, using a native contact map as target, a structure whose contact map is in nearly perfect agreement with the target. Furthermore, the distance of this reconstructed chain from the native structure is quite close to the resolution that can be obtained from the information contained in contact maps.

Nonphysical contact maps

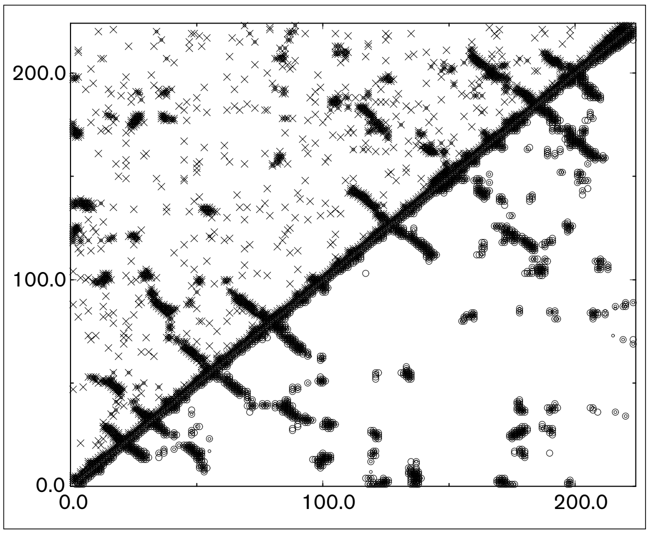
As stated in the Introduction, our main purpose is to develop a strategy to construct a three-dimensional structure, starting from a given set of contacts, even if these contacts are not physical, i.e. not compatible with any conformation allowed by a chain's topology. In such a case, we require our procedure to yield a chain whose conformation is as 'close' as possible to the contact map we started with.

Figure 5



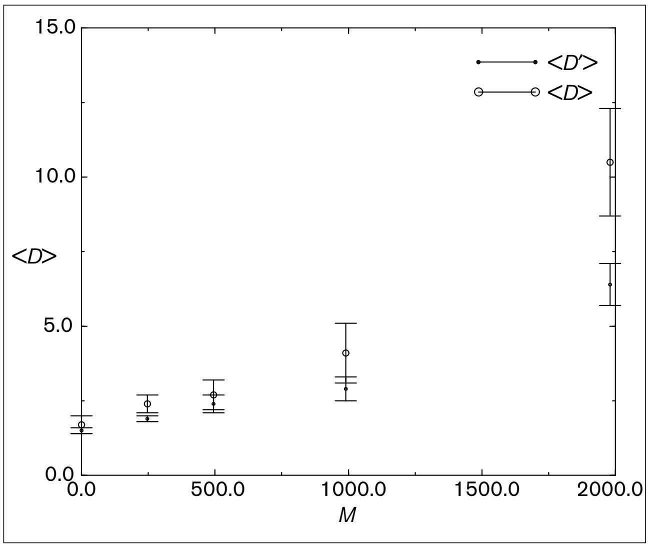
Backbone conformations as generated from a typical reconstruction run with threshold 9 Å for proteins **(a)** 6pti and **(b)** 1trmA. The experimental crystallographic structures are also shown for comparison.

Figure 6



Contact map for protein 1trmA. Above diagonal: reference map (crosses) obtained by randomizing the underlying physical map (dots). Below diagonal: reconstructed contact map (open circles) obtained using the noise-corrupted map as target.

Figure 7



Average distances ⟨*D*′⟩ and ⟨*D*⟩ versus noise *M* for protein 1trmA.

The exact measure of such closeness depends on the source of nonphysicality, as will be demonstrated in two examples described below.

Our first examples of nonphysical contact maps were obtained by randomizing a native contact map; this was done by flipping M randomly chosen entries. Contacts between consecutive amino acids (neighbors along the chain) were conserved.

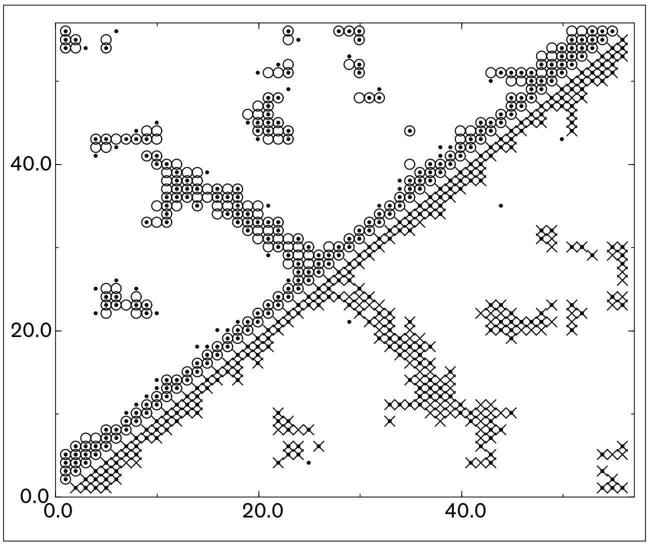
A typical contact map with noise is shown in Figure 6. The protein is 1trmA, whose contact map has 1595 contacts when the threshold is set to 9 Å. We show the native map and the target map obtained by flipping at random $M = 400$ entries of the native map, together with the map produced by our method. For the particular case shown, distances to the crystallographic structure are $D' = 2.1$ Å and $D = 2.4$ Å. The most important conclusion that can be drawn from Figure 6 is that isolated nonphysical contacts are identified as such and ignored and the underlying physical contact map is recovered.

The dependence of this recovery on the noise level is shown in Figure 7, where we present the average distance of the final structure from the uncorrupted 1trmA contact map for various values of M. Averages were taken over 10 different realizations of the noise, and over 10 reconstruction runs for each realization. The distance for totally unrelated structures for 1trmA is around 15 Å. It is remarkable that up to $M < 1000$, a fair resemblance to the experimental structure is still found. Even with the addition of a noise, which is around 60% of the signal, the reconstruction procedure works. We have found similar results for the smaller protein 6pti, which has 342 contacts and can be fairly well reconstructed with a noise of up to 200 flipped contacts.

We can to a certain extent compare these results with those of Bohr *et al.* [6]. They found, using a threshold $d_t = 16$ Å, that their method is robust against noise up to 3%. In their case, the noise is normalized by the total number of possible contacts. For example, their 3% noise corresponds to flipping about 750 entries of the map. The threshold of 16 Å used by Bohr *et al.* is too large to be appropriate for our parametrization of the contact energy in terms of a single number. On the other hand, with such a high value of the threshold, the contact map contains much more information, i.e. 7336 contacts out of the 24,753 entries of the contact map matrix. The information in the contact map is no longer minimal, and it becomes possible to add a noise that is of the order of N^2. We used their threshold to generate our maps, introduced noise as before and repeated the averaging procedure described above. We found that with up to $M = 5000$ flipped contacts (which corresponds to about 20% of the total number of possible contacts), our method achieved the same quality of reconstruction ($\langle D\rangle = 4.1 \pm 0.4$ Å and $\langle D'\rangle = 3.4 \pm 0.2$ Å) as was obtained by Bohr *et al.* for a noise level of 3–5%.

The family of possible nonphysical contact maps, which is most relevant to our program, is produced by using the heuristic constrained dynamic rules in contact map space that were introduced by Mirny and Domany [3]. Following them, we started with a native map of 6pti; when using

Figure 8



Reconstruction of a contact map of protein 6pti obtained by the constrained dynamics introduced by Mirny and Domany [3]. Above diagonal: reference map (dots) obtained by a denaturation/renaturation cycle starting from the experimental map (open circles). Below diagonal: reconstructed contact map (crosses) obtained by using as a target the contact map (dots) that was obtained by constrained dynamics.

a threshold d_t = 8.5 Å, the map derived from the PDB structure has 289 contacts, represented by open circles above the diagonal of Figure 8. We have recomputed the energy parameters for the present definition of contacts (which involves C_α atoms only) and for a threshold of 8.5 Å (energy parameters are available from the authors on request). The energy of the native 6pti map obtained in this way is 36.81. This contact map is subjected to repeated denaturation/renaturation cycles, using the constrained dynamics introduced by Mirny and Domany [3]. We first heat the protein, inducing its unfolding, which is signaled by melting of secondary structure elements in the contact map. For moderate temperature shocks, the protein is generally able to refold upon annealing [3].

In this work, we add a second step to this experiment, by subjecting the contact map obtained by constrained dynamics to our reconstruction procedure. To discuss in some detail the result of this combined scheme, we introduce three classes of contacts: we denote by A the contacts present in $\mathbf{S}_A$, the experimental contact map of protein 6pti (native contacts); by B the contacts present in $\mathbf{S}_B$, the contact map obtained by constrained dynamics in contact map space, starting from $\mathbf{S}_A$; and by C the contacts in $\mathbf{S}_C$, the reconstructed contact map obtained using $\mathbf{S}_B$ as a target. We present in Table 2 the total number of contacts in each class and the number of contact in common between two classes. Map $\mathbf{S}_B$ has 255 contacts, 215 of which are in common with map $\mathbf{S}_A$. This difference is due to 74 missing contacts and 40 spurious ones (see also Figure 8, above the diagonal). The energy of $\mathbf{S}_B$ is 13.35. The reconstructed map $\mathbf{S}_C$ has 310 contacts, 251 of which are in common with the target map $\mathbf{S}_B$. The difference arises from four missing contacts and 61 false positives. This reconstruction score is significantly larger than those typically obtained by applying directly our reconstruction procedure to native maps of proteins of similar size (see e.g. our discussion about Figure 3). This suggests, although without proving it, that the first step of the experiment, when we apply the rules of constrained dynamics introduced by Mirny and Domany [3] is not guaranteed to yield a physically realizable map. From a closer inspection, however, we can derive an overall consistency argument that implies that the Mirny and Domany rules do not drive the system very far away from the physical region in contact map space. The map $\mathbf{S}_C$ and the native map $\mathbf{S}_A$ have 249 common contacts. $\mathbf{S}_C$ has 40 missing and 61 false positives with respect to $\mathbf{S}_A$. Nevertheless, the distances in the three-dimensional structures are D' = 2.14 and D = 2.97, respectively, indicating a rather successful refolding.

From these results, we argue that the Mirny and Domany rules alone are possibly not restrictive enough to keep the trajectory of the system in the physical region of the contact map space during a denaturation/renaturation experiment. This problem can be corrected by the reconstruction procedure discussed in this work. Our procedure projects the contact map obtained by the Mirny and Domany rules onto a contact map that is admissible by construction. Rather consistently, the projected map is quite close to the target one.

Reducing the number of contacts

In this section, we address a very important issue: the effect of reducing the number of contacts on the accuracy of the reconstructed structure. Even though resolving this problem is not essential for our goals, its resolution is an interesting spinoff obtained from our algorithm. The issue is relevant to a number of problem areas where contacts are of importance, such as protein structure determination from NMR data [31] and studies of DNA and crosslinked polymers. The latter are known to undergo a vulcanization transition from a liquid phase to a frozen amorphous phase if the number of contacts exceeds a critical value [32,33].

Table 2

Number of contacts in classes A, B and C and the number of common contacts between classes.

A	B	C	AB	AC	BC
289	255	310	215	249	251

The stochastic reconstruction method described in this work is rather general and potentially applicable to these systems as well.

In real proteins, the number of contacts scales with the chain length N, as shown in Figure 9 for a representative set of 246 proteins taken from the PDB. Fitting the data with a single power law:

$$M = aN^{\nu} \tag{23}$$

yields best fit for $\nu = 1.07$, but the data are also compatible with linear scaling (e.g. $\nu = 1$) as well as with a combination of linear scaling and surface corrections:

$$M = aN + bN^{2/3} \tag{24}$$

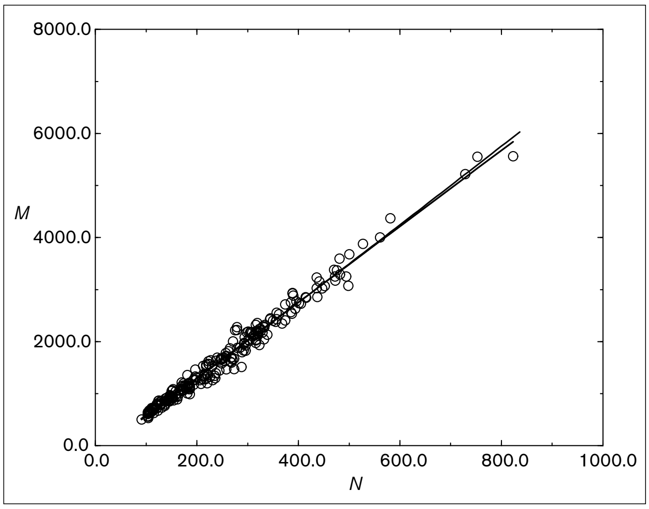
All three fits are shown in the figure and are nearly indistinguishable on the scale used. Figure 9 was obtained using the definition of contacts as given by equation 1 with a threshold of d_t = 9 Å. In the range 5 Å < d_t < 9 Å, only the prefactor a changes, while the exponent ν remains the same. This result holds also for the Mirny and Domany definition of a contact. It has been proposed that in order to have a compact structure, the minimum number of contacts of a random heteropolymer should scale linearly with N [34–36] or with N/ln N [37]. These findings suggest that in proteins the number of contacts required to determine the native fold also cannot scale with a power that is much less than linear with N. The relevant issue, to which considerable effort has recently been devoted [9,11], concerns how small the prefactor a can be, in order to achieve reasonable reconstruction of protein structure from incomplete experimental distance informations. We address this point by analyzing the feasibility of the reconstruction as the threshold d_t is decreased. The smaller d_t, the smaller is the number of contacts present in the contact map.

We now present detailed studies of protein 1acx, with N = 108. For d_t = 9 Å, the number of contacts was 652; for d_t = 6 Å, this number becomes 253 and 154 for d_t = 5 Å. Note that the optimal parameters used for annealing depend on d_t; for d_t = 5 Å, for example, the values $a^f(1) = -5$, $a^i(1) = -20$, $a^f(0) = 0.1$ and $a^i(0) = 0.5$ were used.

In all cases, our method produced chains whose contact maps were in nearly perfect agreement with the respective target maps (deviating by one or two spurious contacts). The distances of the corresponding structures from the native one are, however, very different as d_t decreases. As shown in Figure 10, the values of the average distances D' and D (obtained from 100 runs for each d_t) increase from 1–2 Å for d_t = 9 Å to 5–8 Å for d_t = 5 Å.

This striking increase of D with decreasing d_t shows that even when the target contact map is essentially perfectly recovered, the corresponding structure can be very different from the true one. This suggests that for low values of d_t more information than that contained in the contact map should be provided to get acceptable resemblance to the experimental structure.

Figure 9

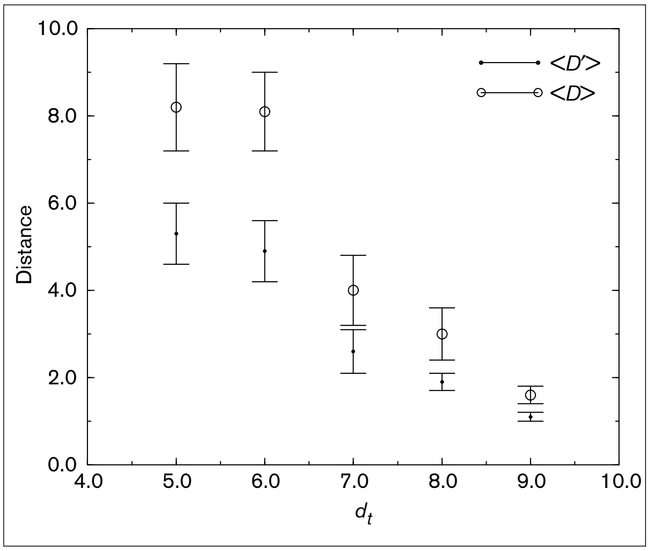


Scaling of M, the number of contacts, with the length N of the proteins. Data refer to 246 proteins taken from the PDB, and to a threshold d_t = 9 Å.

Summary and conclusions

We have presented a stochastic method to derive a three-dimensional structure from a contact map representation. We have shown that for physically realizable target contact

Figure 10



Average distances ⟨D'⟩ and ⟨D⟩ as a function of the threshold d_t for protein 1acx.

maps our method is very fast and reliable to find a chain conformation whose contact map is nearly identical to the target. Moreover, the method is able to find a good candidate structure even when the target map has been corrupted with nonphysical contacts.

The information contained in a known native contact map suffices to reconstruct a conformation that is relatively close to that of the original structure, as was already observed by Havel *et al.* [2]. There is, however, an intrinsic limit in the resolution of a contact map. We used a threshold of 9 Å between C_α atoms to define contact; for this threshold, the distance between two typical structures that are both compatible with the contact map is about 1 Å. The threshold of 9 Å is relevant for our purpose of working with contact energies in a scheme to derive structure from sequence. The present work is instrumental in achieving this long-term goal and it is within this context that it should be viewed. That is, whereas existing methods aim at obtaining a structure of 'high quality' (as measured, for example, by the distance *D* from the true structure), we are interested mainly in starting from a possibly nonphysical contact map and producing from it one that is guaranteed to be physical. Reconstructing structure from a (possibly noisy) contact map is also an important problem and we believe that our work has contributed to its solution.

The contact map representation is intimately connected with the parametrization of every contact energy by a single number. We are currently studying the dynamics in contact map space, controlled by such a simple energy function. We believe that such a study will reveal whether the contact map representation, together with the assumptions implicit in working with a pairwise contact-based approximation for the energy, suffice to single out the native state of a protein.

Acknowledgements

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. E Domany thanks the members of the Laboratoire de Physique Statistique at the Ecole Normale Superieure, Paris for their hospitality. E Koussell thanks the K Kupcinet fellowship program for partial support of his stay as a summer student at the Weizmann Institute.

References

1. Lifson, S. & Sander, C. (1979). Antiparallel and parallel beta strands differ in amino acids residue preferences. *Nature* **282**, 109-111.
2. Havel, T.F., Crippen, G.M. & Kuntz, I.D. (1979). Effect of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* **18**, 73-81.
3. Mirny, L. & Domany, E. (1996). Protein fold recognition and dynamics in the space of contact maps. *Proteins* **26**, 391-410.
4. Havel, T.F. & Wüthrich, K. (1985). An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution. *J. Mol. Biol.* **182**, 281-294.
5. Brünger, A.T., Clore, G.M., Gronenborn, A.M. & Karplus, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc. Natl Acad. Sci. USA* **83**, 3801-3805.
6. Bohr, J., *et al.*, & Petersen, E.F. (1993). Protein structures from distance inequalities. *J. Mol. Biol.* **231**, 861-869.
7. Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulfide connectivities. *J. Mol. Biol.* **245,** 645-650.
8. Lund, O., Hansen, J., Brunak, S. & Bohr, J. (1996). Relationship between protein structure and geometrical constraints. *Protein Sci.* **5**, 2217-2225.
9. Aszódi, A. & Taylor, W R. (1996). Homology modelling by distance geometry. *Fold. Des.* **1**, 325-334.
10. Mumenthaler, C. & Braun, W. (1996). Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465-480.
11. Skolnick, J., Kolinski, A. & Ortiz, A.R. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.
12. Crippen, G. & Havel, T.F. (1988). *Distance Geometry and Molecular Conformation*. Wiley, New York.
13. Kuntz, I.D., Thomason, J.F. & Oshiro, C.M. (1989). Distance geometry. *Methods Enzymol.* **177**, 159-204.
14. Scheek, R.M., Van Gunsteren, W.F. & Kaptein, R. (1989). Molecular dynamics simulation techniques fro determination of molecular structures from nuclear magnetic resonance data. *Methods Enzymol.* **177**, 204-218.
15. Brünger, A.T., Adams, P.D. & Rice, L.M. (1997). New applications of simulated annealing in X-ray crystallography and solution NMR. *Curr. Opin. Struct. Biol.* **5**, 325-336.
16. Rosenbluth, M.N. & Rosenbluth, A.W. (1955). Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.* **23**, 356-359.
17. Šali, A., Shakhnovich, E.I. & Karplus, M. (1994). Kinetics of protein folding. *J. Mol. Biol.* **235**, 1614-1636.
18. Kirkpatrick, S., Gelatt, C.D. Jr, & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.
19. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460-471.
20. Govindarajan, S. & Goldstein, R. (1995). Why are some protein structures so common? *Proc. Natl Acad. Sci. USA* **93**, 3341-3345.
21. Anfinsen, C. & Scheraga, H. (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* **29**, 205-300.
22. Vendruscolo, M. (1997). Modified configurational bias Monte Carlo for simulation of polymer systems. *J. Chem. Phys.* **106**, 2970-2976.
23. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **34**, 827-828.
24. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **32**, 922-923.
25. McLachlan, A.D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49-79.
26. Cohen, F.E. & Sternberg, M.J. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.* **163**, 613-621.
27. Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I.C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* **14**, 5355-5360.
28. Frauenfelder, H., Parak, F. & Young, R.D. (1988). Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.* **17**, 451-479.
29. Elber, R. & Karplus, M. (1987). Multiple conformational states of proteins: Molecular dynamics of myoglobin. *Science* **235**, 318-321.
30. Frauenfelder, H., Sligar, S. & Wolynes, P.G. (1991). The energy landscapes and motions in proteins. *Science* **254**, 1598-1603.
31. Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids.* Wiley, New York.
32. Goldbart, P.M. & Zippelius, A. (1993). Amorphous solid state of vulcanized macromolecules: a variational approach. *Phys. Rev. Lett.* **71**, 2256-2259.
33. Castillo, H.E., Goldbart, P.M. & Zippelius, A. (1994). Distribution of localization lengths in randomly crosslinked macromolecular networks. Europhys. *Lett.* **28**, 519-524.
34. Camacho, C.J. (1996). Entropic barriers, frustration and order: basic ingredients in protein folding. *Phys. Rev. Lett.* **77**, 2324-2328.
35. Camacho, C.J. & Schanke, T. (1997). From collapse to freezing in random heteropolymers. *Europhys. Lett.* **37**, 603-608.
36. Gutin, A.M. & Shakhnovich, E.I. (1994). Statistical mechanics of polymers with distance constraints. *J. Chem. Phys.* **100**, 5290-5293.
37. Bryngelson, J.D. & Thirumalai, D. Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.* **76**, 542-546.