Algorithms **2009**, *2*, 76-92; DOI: 10.3390/a2010076

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Review

On the Reconstruction of Three-dimensional Protein Structures from Contact Maps

Pietro Di Lena [1,⋆], Marco Vassura [1], Luciano Margara [1], Piero Fariselli [2] and Rita Casadio [2]

[1] Department of Computer Science, University of Bologna, Via Mura Anteo Zamboni, 7, 40127 Bologna, Italy

[2] Biocomputing Group, Department of Biology, University of Bologna, Via Irnerio, 42, 40127 Bologna, Italy

E-mails: {dilena, vassura, margara}@cs.unibo.it; piero@biocomp.unibo.it; casadio@alma.unibo.it

⋆ Author to whom correspondence should be addressed.

Received: 30 November 2008; in revised form: 8 January 2009 / Accepted: 20 January 2009 / Published: 22 January 2009

Abstract: The problem of protein structure prediction is one of the long-standing goals of Computational Biology. Although we are still not able to provide first principle solutions, several shortcuts have been discovered to compute the protein three-dimensional structure when similar protein sequences are available (by means of comparative modeling and remote homology detection). Nonetheless, these approaches can assign structures only to a fraction of proteins in genomes and ab-initio methods are still needed. One relevant step of ab-initio prediction methods is the reconstruction of the protein structures starting from inter-protein residue contacts. In this paper we review the methods developed so far to accomplish the reconstruction task in order to highlight their differences and similarities. The different approaches are fully described and their reported performances, together with their computational complexity, are also discussed.

Keywords: Protein folding; contact map; molecular modeling.

---

1. Introduction

Proteins are polymers of *amino acids* (also referred to as *residues*). Every protein is uniquely identified by its one-dimensional (1D) sequence of amino acids (*primary structure*), which are covalently bonded together to create a continuous and non-overlapping chain called the *backbone* of the protein. The Anfinsen's dogma [1] states that, under the same environmental conditions, every protein *folds* spontaneously into a characteristic three-dimensional (3D) structure (*tertiary structure*). Protein biological functions are strictly related to the folded state and several biological diseases are believed to be the consequence of misfolded proteins. The process by which proteins fold into their native states is still a matter of debate. At the coarsest level it seems that often the folding process involves first the establishment of regular *secondary structure* elements, such as *alpha helices* and *beta sheets*, and afterwards the establishment of the tertiary structure. The most important aspect is that protein primary sequences seem to contain all information needed to drive the fold to the native structures. One of the grand challenges in Bioinformatics is to understand the rules which drive the folding process by starting from only the residue sequence.

Currently, protein structures are typically solved experimentally by time-consuming X-ray crystallography or NMR spectroscopy. In the last 25 years the amount of electronic information contained in the protein data bank repository (PDB*) has been growing enormously. The parallel increase of computational power motivated the development of computational approaches to the folding recognition problem, being building by *homology* the most effective one. This method is based on the notion that proteins which share high sequence similarity fold into the same native structure (with few exceptions). Protein evolutionary process tends to preserve more the structure than the sequence and a target protein can be usually modeled with good accuracy on a related template protein (for which the fold is known) by performing a *sequence alignment*, i.e. by mapping residues of the target protein into residues of the template protein. The overall quality of the fold recognition depends on the quality of the alignment. Homology modeling is no more effective when the set of template proteins are distantly related or not related at all to the target protein. Then other approaches must be explored. We can partition current approaches into two main classes: *ab-initio* methods (see, for example, [2]) and methods based on the *remote homolog detection* (see, for example, [11]).

An interesting subproblem in protein structure prediction is the problem of predicting residue intramolecular contacts. Two residues are said to be in *contact* if (in the protein structure) their distance is below some given threshold. Contact prediction methods have been developed and assessed in CASP (Critical Assessment of techniques for protein Structure Prediction) experiments [8]. The interest in contact prediction is justified by the fact that the knowledge of the whole map of residue-to-residue contacts (*contact map*) or the knowledge of just few correct contacts can be of great help for discriminating among all possible folds for a target protein. After a contact map is predicted, the problem of reconstructing the protein 3D structure can be tackled [3]. Notwithstanding the fact that the reconstruction problem from native contact maps is computationally intractable, the heuristic algorithms developed so far have very good performances. The overall effectiveness of this approach is essentially affected by the accuracy of contact map prediction, which remains very poor to date. In this paper we focus and review in detail the

*Freely available at *http://www.pdb.org*.

computational methods proposed so far in literature to solve the reconstruction problem from (predicted) contact maps.

The paper is organized as follows. In Section 2. we define formally the reconstruction problem from contact maps and introduce the notation. In Section 3. we provide a general discussion about the computational complexity of the reconstruction problem. In Section 4. we review in detail the best known reconstructing algorithms and their computational complexity. Section 5. is devoted to a final discussion about the effectiveness of these methods.

2. Protein contact maps

There are several definitions of *residue-to-residue contacts* in literature. The two most widely accepted ones are those which define two residues to be in contact when (in the protein native structure) the distance between their carbon-alpha (C_α) or carbon-beta (C_β) atoms is below some given threshold. While there is some difference between C_α and C_β contact maps, the representation adopted is not crucial for the algorithms we will present here and throughout the rest of the paper we will only consider C_α contact maps. An important property of C_α contacts is that the distance between the C_α atoms of two consecutive residues is always an almost constant value close to 3.8Å. This is no more true if we consider C_β distances of consecutive residues.

Given some protein P consisting of N residues we can describe its structure by means of a set of three-dimensional coordinates $c_i = (x_i, y_i, z_i), 1 \leq i \leq N$ which correspond to the positions of the C_α atoms of the protein residues. The *contact map* $M \in \{0,1\}^{N \times N}$ of *threshold* t for protein P is a two-dimensional approximation of the protein structure and it is defined as

$$M_{ij} = \begin{cases} 1 & if\ |c_i - c_j| \leq t \\ 0 & if\ |c_i - c_j| > t \end{cases}$$

where

$$|c_i - c_j| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

is the Euclidean distance between coordinates c_i and c_j. From now on we will keep the notation consistent in the rest of the paper and we will refer to the length of a protein with the symbol N; with c_i we will refer to some i-th coordinate and we will use the symbols M and t to denote a contact map and its threshold, respectively.

The threshold value determines the amount of information about the protein structure contained in a contact map. For low threshold values the number of contacts observed in a protein structure is relatively sparse if compared to the number of non-contacts. The typical threshold values adopted are 7-8Å. These threshold values are the ones which minimize the distance between residue physical contacts and C_α, C_β contacts (see, for example, [4]). The C_β representation with threshold 8Å is the definition of contact currently adopted to evaluate the performances of contact predictors in CASP experiments (see, for example, [8]). With these definitions of threshold, the number of contacts between residues which are distant in the protein primary sequence (long range contacts) is very small. This is an important point since long-range contacts are the most informative about protein structures and they are also the most difficult to predict. Increasing the threshold value usually implies better performances for the

reconstructing algorithms but, on the contrary, this does not imply an increase in the accuracy of the contact predictions, which is currently very low.

A reconstructing algorithm from a contact map M must be able to recover a set of coordinates which satisfies as much as possible the constraint imposed by M. We say that a set of coordinates $c_1, .., c_N$ is consistent with M if $\forall i, j, |c_i - c_j| \leq t$ if and only if $M_{ij} = 1$. Very likely a predicted contact map M is not *physical* which implies that no set of consistent coordinates with M exist at all.

The success of a reconstructing algorithm from contact maps is measured in terms of the number of contacts correctly recovered while a qualitative measure of the success of the algorithm is provided by the *root mean square deviation* (RMSD) between the reconstructed three-dimensional structure and the native one. The RMSD measure is commonly used to compare two molecular structures described by some set of coordinates $C = (c_1, .., c_N)$ and $C' = (c'_1, .., c'_N)$. It is defined as the smallest distance

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{c}'_i - c_i)^2}$$

where $\tilde{C}' = (\tilde{c}'_1, .., \tilde{c}'_N)$ is obtained by rotating and translating the coordinate set C'.

3. Computational complexity of the reconstruction problem

In Computational Geometry the problem to determine a n-dimensional representation of a graph is usually known as the *graph realization problem*. A *realization* of a graph in the n-dimensional Euclidean space is a mapping from the vertices of the graph into n-dimensional coordinates such that the spheres of unit radius centered in every pair of such coordinates intersect if and only if the related vertices of the graph are connected. The *unit* of the sphere is not crucial here since a realization remains valid also if we scale the distances in the realized structure. Note that any undirected graph can be represented by means of an *adjacency matrix*, i.e. a binary symmetric matrix which contains 1 in the entry i, j if and only if the i-th vertex and the j-th vertex are connected. The graph realization problem is thus a generalization of the reconstruction problem from native protein contact maps (i.e. contact maps which contains no errors).

The computational complexity of the graph realization problem was first addressed in the mid nineties by H. Breu and D. G. Kirkpatrick [6]. The authors of [6] show that the problem to determine the *sphericity* of a graph is *NP-complete* (i.e. it cannot be solved by any polynomial-time algorithm unless P=NP, see for example [9]). The *sphericity* of a graph is the smallest dimension in which the graph can be realized. Graphs of sphericity 1 are called *unit interval graphs* and they can be recognized in polynomial time. Breu and Kirkpatrick show that the problem to determine whether a graph can be realized in dimension 2 is no less difficult than the problem to determine whether there is an assignment to the variables of a boolean formula which makes the formula evaluate to true (i.e. the proof is by reduction from the SAT problem which is historically the first decision problem proved to be NP-complete). While the sphericity problem is NP-complete this does not imply that the problem cannot be solved in polynomial-time for some fixed dimension greater than 2 (as it happens for unit interval graphs). The authors of [6] argue that their construction can be easily modified to show that the sphericity problem is NP-complete also for dimension 3 (which is the most interesting case from our point of view) and, moreover, they conjecture that this is also true for every dimension greater that 1.

At first glance it can seem that the problem to determine the sphericity of a graph has few to do with the graph realization problem. It is not difficult to see that if the graph realization problem could be solved in polynomial-time then the sphericity problem would be solved in polynomial-time, yielding to a contradiction. Assume that there is some polynomial-time algorithm which can compute a 2D representation of a graph if such graph has sphericity 2, on the contrary the algorithm returns some not consistent structure. Then we could use such algorithm to solve in polynomial-time the sphericity problem for dimension 2: given a graph we compute a 2D realization in polynomial-time and check (also in polynomial-time) if such realization is consistent with the graph; if the structure is not consistent then the graph has not sphericity 2, on the contrary the answer is positive. Actually, the negative result of Breu and Kirkpatrick has also been extended in a recent work [17], in which the authors show that the graph realization problem cannot be even approximated in polynomial-time unless P=NP.

While the graph realization problem is intractable, the reconstruction problem from protein contact maps is not so general. From a biological point of view a realization of a contact map which does not satisfies the constraint of a non-overlapping backbone is useless. With the additional backbone constraint the realization problem could be solved in polynomial-time. Nothing has been proved so far in this direction, anyway the construction in [6] seems to be enough versatile to admit the backbone constraint without lowering the computational complexity of the general problem.

4. Approaches to the reconstruction problem from protein contact maps

In this section we review the approaches reported in literature for the contact map reconstruction problem. To be useful, a reconstructing algorithm from contact maps must satisfy some non-trivial requirements. First of all it must be fast enough to be used for large scale reconstructions while we showed in the previous section that the problem to realize a contact map in the three-dimensional space is generally intractable. The difficulty of the problem is moreover increased by the fact that a reconstructing algorithm must be designed in order to deal also with highly blurred contact maps. Given the impossibility to find an optimal solution to this problem in reasonable time, all algorithms developed so far are heuristic. An heuristic algorithm is not required to produce the best possible solution but only a good solution in the average case. One necessary requirement of a reasonable solution is that it must satisfy at least some minimal biological constraints such as the *backbone constraint*.

Existing algorithms proceed in two phases: in the first phase they generate an initial set of 3D coordinates which are refined in the second phase in order to obtain a 3D structure as consistent as possible with the input contact map. We can coarsely identify three approaches to the reconstruction problem. In the first approach (Section 4.1.) the realization problem is defined as the search in the three-dimensional space of the set of coordinates which minimize an opportune cost function. The search in the coordinates space is based on *gradient descent* methods. The main problem of gradient descent minimization approach is that, during the computation, the procedure can be trapped in local minima of the cost function from which it cannot escape. This is much more likely to happen when the contact map is not physical (i.e. when the constraints imposed are contradictory). For this reason, the other methods developed so far adopt a *local search heuristic* approach. A local search algorithm starts from a candidate solution and then iteratively moves to a neighbor solution until the best solution found has not been improved in a given number of steps or until the time bound given to the procedure is elapsed. The local search avoids

that the procedure gets trapped in a local minimum. Our second class of methods (Section 4.2.) is based on a local search approach known as *simulated annealing*. To finish, a third and recent method, uses a Distance Geometry (Section 4.3.) approach to generate an initial solution and then it uses a local search technique to refine such an initial solution. After a detailed description, we discuss the computational complexity of these methods (Section 4.4.).

4.1. Methods based on the gradient minimization of a cost function

The general idea of these methods is that starting from an initial set of coordinates the reconstruction problem can be formulated as the minimization of the squared deviation between the current set of distances and the expected distances. In detail, given a $N \times N$ contact map M the reconstructing procedure tries to compute a set of coordinates which minimize some energy cost function of the form

$$E = \sum_{i=1}^{N} \sum_{j=i}^{N} (|c_i - c_j| - d_{ij})^2 \tag{1}$$

where, as usual, c_i is the Cartesian coordinate of the i-th residue, the real number $|c_i - c_j|$ is the Euclidean distance between the i-th and the j-th coordinates and d_{ij} is the *expected* distance between the i-th and j-th residues. The minimization of the energy function is carried out by standard gradient descent based methods (see, for example, [20]).

A first difficulty with the minimization approach is that the expected distances are usually difficult to compute with high accuracy since a contact map contains only information about coarse lower and upper bounds. A second problem is that the cost function (1) does not embody the backbone constraint and the minimization procedure can eventually lead to a structure which has a broken backbone, i.e. the distances between two consecutive residues can be much larger than those in real proteins.

In the sequel we review two methods which use this approach and which differ essentially on the way by which some constraints are added to equation (1) in order to solve the two problems described above.

Bohr et al. (1993)

The authors of [7] overcome the backbone and the expected distances problems by adding extra terms to the cost function (1) which becomes

$$E = \sum_{i=1}^{N} \sum_{j=i}^{N} w_{ij} \cdot [S(|c_i - c_j|) - d_{ij}]^2 + \sum_{i=1}^{N-1} E^i_{bond} \tag{2}$$

The distance $|c_i - c_j|$ is projected into the interval $[0, 1]$ by the sigmoidal function

$$S(x) = \frac{1}{1+e^{-s(x-t)}}$$

where s is a slope constant (reported as 1.5^{-1} Å) which determines the steepness of the sigmoidal function in proximity of $x = t$. The idea to project the current distances into the interval $[0, 1]$ solves the problem to find a good approximation of the expected distances, thus the expected distance d_{ij} can be simply defined by

$$d_{ij} = \begin{cases} 0 & if\ M_{ij} = 1 \\ 1 & if\ M_{ij} = 0 \end{cases}$$

avoiding any sort of approximation. A counter-effect of this approach is that the energy function has a local minimum when the distances diverge to infinity. This is a consequence of the fact that when a low threshold value t is chosen (recall that a typical value is around 8Å) the distances in real proteins below t are much more sparse compared to those above t. This problem is solved by adding the weighting term

$$w_{ij} = \begin{cases} \frac{N_L}{N_L+N_S} & if\ d_{ij} = 0 \\ \frac{N_S}{N_L+N_S} & if\ d_{ij} = 1 \end{cases}$$

which has the effect to balance the weight of shorter distances with respect to larger distances (N_L and N_S are the number of distances above and below the threshold t, respectively). The backbone constraint is introduced by adding the term $\sum_{i=1}^{N-1} E^i_{bond}$ with

$$E^i_{bond} = \begin{cases} 3.8 - \epsilon - |c_i - c_{i+1}| & if\ |c_i - c_{i+1}| < 3.8 - \epsilon \\ 0 & if\ 3.8 - \epsilon < |c_i - c_{i+1}| < 3.8 + \epsilon \\ |c_i - c_{i+1}| - 3.8 - \epsilon & if\ |c_i - c_{i+1}| > 3.8 + \epsilon \end{cases}$$

where the parameter ϵ is a small quantity (reported as 0.2 Å). The initial configuration is chosen at random in a cubic box of total volume $100N$Å^3.

Galaktionov and Marshall (1994)

The main difference between this method [12] and the previous one is in the approach adopted to deal with the approximation of the expected distances. The distance between two residues i, j in the protein structure can be upper bounded by just considering the successive powers ($M^2 = M \cdot M, M^3 = M \cdot M \cdot M, ...$) of the protein contact map. In fact, if for some power n of the contact map M the entry M^n_{ij} is greater than 0 then we have that in the protein structure there must exist at least one path of length at least n connecting the i-th and the j-th residue which imposes the constraint

$$|c_i - c_j| \leq nt.$$

The authors describe a method to obtain more accurate measures. We can partition the set of distances on the protein structure into N disjoint sets. For $1 \leq n \leq N$, we denote

$$d_{ij} \in D_n \text{ if } M^1_{ij} = 0, ..., M^{n-1}_{ij} = 0 \text{ and } M^n_{ij} > 0.$$

If a couple of residues i, j has expected distance $d_{ij} \in D_n$ then there are exactly M^n_{ij} shortest paths of length n connecting i and j. By experimental tests, the authors report the existence of a linear dependence between the values M^n_{ij} and the $d_{ij} \in D_n$ distances which can be approximated by the polynomials

$$d_{ij} = g^{(n)}_0 + \sum_k g^{(n)}_k (M^n_{ij})^k, d_{ij} \in D_n \tag{3}$$

where the coefficients $g^{(n)}_k$ are those which assure the best fitting to equation 3 and have to be determined experimentally. The energy cost function becomes

$$E = \sum_{i=1}^{N} \sum_{j=i}^{N} \gamma_{ij} \left[|c_i - c_j| - d_{ij} \right]^2 + \sum_{i=1}^{N-1} E^i_{bond} \tag{4}$$

where the coefficients γ_{ij} depends on the class of the d_{ij} distance, i.e. if $d_{ij}, d_{kl} \in D_n$ then $\gamma_{ij} = \gamma_{kl}$. The backbone constraint term $\sum_{i=1}^{N-1} E^i_{bond}$ is defined by

$$E^i_{bond} = \gamma_0(|c_i - c_{i+1}| - 3.8)^2.$$

The constants γ_0, γ_{ij} represent the weights of each relative constraint in the energy function. Also in this case the initial solution is computed at random.

4.2. Methods based on simulated annealing

The *simulated annealing* is a generic technique used to find an approximation of the global minimum of a given function when the search space is very large (see, for example, [13]). This technique is inspired from annealing in metallurgy where the microstructure of a material is altered by heating to a predetermined temperature and then cooling down to room temperature to improve ductility. By analogy with the physical process, at each step of the algorithm the current solution is replaced by a random nearby solution, chosen with a probability which depends on the current solution and on a global parameter T, called *temperature*, that is gradually decreased during the computing process. The current solution changes almost randomly when T is large while the randomness decreases as T goes to zero. This prevents the method from becoming stuck at local minima, as it can happen with gradient descent methods.

To our knowledge there are two methods that use a simulated annealing approach for the reconstruction problem. Both methods consist of two phases. A *growth* phase, in which an almost random initial solution is built by adding step-by-step one coordinate at a time which satisfies local constraints. An *adaptation* phase which uses simulated annealing to iteratively refine the initial solution.

Vendruscolo et al. (1997)

This method has been described in [24]. The key ingredients are the definition adopted for the temperature-like parameter and the energy cost function used to evaluate the computed structure.

During the growing phase one coordinate c_i at a time is added to the growing chain. Assume that the first $c_1, ..., c_{i-1}$ coordinates have been chosen already. A number k (reported as 10) of random direction vectors $c^{(1)}, ..., c^{(k)}$ are generated obtaining a set of possible candidates for c_i

$$c_i^{(j)} = c_{i-1} + c^{(j)}.$$

The set of direction vectors $c^{(j)}$ is randomly chosen from a uniform distribution. The set of distances $|c_i^{(j)} - c_{i-1}|$ has average distribution 3.8 and small variance 0.04. Among the k possible candidates the best one is chosen to grow the chain. The best candidate is the one which maximizes the number of true positive contacts and minimizes the number of false positive contacts consistently with the contact map. In detail, every candidate $c_i^{(j)}$ is scored according to the probability

$$p^{(j)} = \frac{e^{-\frac{E_g(c_i^{(j)})}{T_g}}}{Z} \tag{5}$$

where

$$Z = \sum_{j=1}^{k} e^{-\frac{E_g\left(c_i^{(j)}\right)}{T_g}}$$

is a normalization factor and T_g is a temperature-like parameter which is initialized to 1 in this phase. The energy cost function E_g (note that it is computed only with respect to the newly selected coordinate $c_i^{(j)}$) is defined by

$$E_g\left(c_i^{(j)}\right) = \sum_{k=1}^{i-1}(i-k)\cdot K_g(M_{ik})\cdot\vartheta(t-|c_i^{(j)}-c_k|) \tag{6}$$

where the factor $(i-k)$ is introduced to give more importance to long range contacts, ϑ is the Heaviside step function and $K_g(M_{ij})$ is a two-valued function defined by $K_g(M_{ij}) \leq 0$ if $M_{ij} = 1$ and $K_g(M_{ij}) \geq 0$ if $M_{ij} = 0$. This term has the effect to penalize false positive contacts and to reward true positive contacts. That is, when the distance $|c_i^{(j)} - c_k|$ is below the contact map threshold t and $M_{ij} = 1$ then the candidate $c_i^{(j)}$ is rewarded otherwise it is penalized. During the growth phase the values $K_g(0) = 0$ and $K_g(1) = -1$ are chosen, i.e. false positive contacts are not penalized. Every chain grown in this way has a score defined by the product of the probabilities of its coordinates (i.e. the probabilities computed with equation (5)). The growth phase is computed several times from scratch (typically ten times) and the structure with the best score is selected as the starting point for the adaptation phase.

In the adaptation phase, coordinates are displaced at random by using crankshaft moves (which avoid the backbone to be broken). The displacement of coordinate c_i into c_i' is accepted with probability

$$\pi = \min\{1, e^{\frac{\Delta E_a}{T_a}}\}$$

according to the standard Metropolis prescription. The term $\Delta E_a = |E_a(c_i) - E_a(c_i')|$ is the change in the cost function induced by the move. The energy function of the adaption phase, defined by

$$E_a(c_i) = \sum_{k=1}^{N} K_a(M_{ik})\cdot\vartheta(t-|c_i-c_k|), \tag{7}$$

apparently differs from the energy function of the growth phase only for the missing factor $(i-k)$ which means that in this phase long range contacts are not favored against short range contacts. Note that, also in this case, the energy function is computed only for the displaced coordinate since the move doesn't affect the rest of the structure. The temperature-like parameter T_a is not constant as in the previous phase but it is decreased slowly during the computation according to the number of missing contacts. Two regimes are identified. In the first one, the structure contains many missing contacts and it is very far from being consistent with the contact map. In the second regime, the structure contains few missing contacts and it is very close to be completely consistent with the contact map. The function K_a and the temperature-like parameter T_a are interpolated between these two limiting cases:

$$K_a^{(n)}(M_{ij}) = K^f(M_{ij}) + [K^s(M_{ij}) - K^f(M_{ij})]\cdot\sigma(n)$$

and

$$T_a^{(n)} = T_a^f + (T_a^s - T_a^f)\cdot\sigma(n)$$

where the function

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

has the scope to interpolate between the two regimes (here n is the current number of iterations). The distance between the two regimes is obtained by opportune choices of $K_a^s, K_a^f, T_a^s, K_a^f$ and α.

The authors describe another refinement stage after the adaption phase which fixes the chirality of the recovered structure. For more details refer to [24].

Pollastri et al. (2006)

This algorithm is currently used as the final step in the *Distill*[†] architecture, a machine learning approach to ab initio protein structure prediction [18]. Distill resembles a set of machine learning predictors for 1D and 2D protein features and a 3D reconstructing algorithm which uses a simulated annealing procedure to recover the 3D structures from contact maps predicted in the previous stages. One notable difference with the other methods described here is that the Distill reconstructing algorithm takes into account some 1D constraints (for instance, whether some fragment should fold to alpha helices) which cannot be inferred directly from contact maps without having any knowledge about the protein primary sequences. This is an example of how more constraints can be added to the reconstruction problem in order to improve structure prediction.

Also in this case the reconstructing algorithm consists of a growth phase and an adaptation phase. The growing procedure is essentially the one described in the previous section except for one difference. If the fragment going from the i-th to the j-th residue of the chain is predicted as an alpha helix, then the coordinates from i to j are chosen in order to model an ideal alpha helix with random orientation. Also the adaptation procedure is similar to the one described in the previous section. The coordinates are displaced at random by using crankshaft moves. One difference from the previous method is that the secondary structure elements (helices) are displaced as a whole without modifying their shape. The main difference here consists in the way the energy function encodes the constraints on the whole structure. In detail, a displacement is accepted with probability

$$\pi = \min\{1, e^{\frac{\Delta E}{T^{(n)}}}\}$$

where the temperature-like parameter $T^{(n)}$ is initialized to a value proportional to the protein length and it is decreased by some inverse exponential function on the number of steps n. The energy function E contains both contact map constraints and geometrical constraints imposed by the protein structure topology. Formally it is defined by

$$E = \alpha_0 \left\{ 1 + \sum_{(i,j)\in\mathcal{F}_0} \left(\frac{|c_i - c_j|}{t} \right)^2 + \sum_{(i,i+1)\in\mathcal{B}} (|c_i - c_{i+1}| - 3.8)^2 \right\} + \alpha_1 F + \alpha_2 \sum_{(i,j)\in\mathcal{C}} e^{5-|c_i-c_j|} \quad (8)$$

where

$$\mathcal{F}_0 = \{(i,j) \mid |c_i - c_j| > t \text{ and } M_{ij} = 1\}$$

is the set of coordinate pairs which define false negative contacts,

[†]Freely available at *http://distill.ucd.ie/distill/*.

$$\mathcal{F}_1 = \{(i,j) \mid |c_i - c_j| \leq t \text{ and } M_{ij} = 0\}$$

is the set of coordinate pairs which define false positive contacts and F is the cardinality of $\mathcal{F}_1$,

$$\mathcal{B} = \{(i,j) \mid |i-j| = 1 \wedge |c_i - c_j| \notin [3.733, 3.873]\}$$

is the set of successive coordinate pairs which do not satisfy correctly the backbone constraint and

$$\mathcal{C} = \{(i,j) \mid |i-j| > 1 \wedge |c_i - c_j| < 5\}$$

is the set of non-consecutive coordinate pairs which do not satisfy the minimal observed distance threshold for hard core repulsion (5Å). The constants $\alpha_0 = 0.2$ (false non-contacts), $\alpha_1 = 0.02$ (false contacts), $\alpha_2 = 0.05$ (clashes) represent the weights of each constraint term.

Note that, also in this case, the energy function can be computed only for the displaced coordinates (which are more than one if a whole alpha helix is moved).

4.3. Methods based on distance geometry and local search

Distance Geometry deals with the characterization of mathematical properties which can be derived from distance values between pairs of points. Distance Geometry can be seen as the mathematical foundation for a geometric theory of molecular conformation. One of the most relevant results in this context is that, given a consistent set of distances in the three-dimensional space, the problem to find a set of coordinates which satisfy such exact distance constraints can be solved by a polynomial-time algorithm [5] (the problem becomes NP-complete when the given set of distances is sparse [19]). One of the best known application of this theory is the determination of molecular conformations from experimental data such as NMR spectroscopy and X-ray crystallography. Because of experimental errors, we can usually obtain only a set of lower and upper bounds to inter-atomic distances rather than exact values. Even with such relaxed distance constraints, the problem to compute a set of consistent coordinates is intractable [16]. While the problem is intractable, in the eighties Havel and Crippen [14, 15] developed a polynomial-time recovering algorithm from a sparse set of lower and upper bounds. This is the first use of a distance geometry based approach for protein structure recovering. Their algorithm first uses some bound smoothing technique to estimate bound values for the missing distances. Then it uses an algebraic technique known as the Embed algorithm to generate an approximate set of 3D coordinates. The 3D structure recovered is, in some sense, the best three-dimensional fit for the set of distances considered. One important feature of Embed is that it provides a reasonable solution also when the set of distances in not embeddable in the three-dimensional space.

The Embed algorithm cannot be directly used with great success for the contact map reconstruction problem because, even in the presence of maps without errors, the set of lower and upper bounds which can be inferred from contact maps is usually coarse and not that accurate. Anyway, the Embed algorithm can be still used to compute a good initial solution better than a random one to be used as a starting point for a successive refinement procedure. To our knowledge there is only one method reported in literature which uses this approach.

Vassura et al. (2007)

This algorithm[‡] has been developed recently [22, 23]. A first version of this algorithm which works only on native contact maps has been described in [21]. The algorithm is in three phases. In the first phase an initial set of coordinates is generated. The initial set of coordinates is iteratively refined in the second phase of the algorithm up to a final solution. The third phase fixes, if necessary, the backbone constraint.

In the first phase, the procedure tries to guess a possible initial solution. This first phase relies essentially on the Embed algorithm. In detail, the contact map is scanned to find independent subcomponents. Two disjoint fragments of the protein are independent if there are no medium-long range contacts between them. Once independent fragments (if any) are identified the algorithm tries to guess a possible 3D structure for any such component. A set of distances is guessed according to the following general scheme

$$D_{ij} = \begin{cases} 0 & if \;\; i = j \\ 3.8 & if \;\; |i-j| = 1 \\ 6 \pm \epsilon & if \;\; |i-j| = 2 \; and \; M_{ij} = 1 \\ 7.5 \pm \epsilon & if \;\; |i-j| = 3 \; and \; M_{ij} = 1 \\ t \pm \epsilon & if \;\; |i-j| \geq 4 \; and \; M_{ij} = 1 \\ \infty & if \;\; M_{ij} = 0 \end{cases}$$

where ϵ is some small random error (in function of the threshold t). The missing distances (i.e. those initialized with ∞) are upper bounded by using the standard Shortest Paths algorithm (see, for example, [9]). The initial 3D structure of each component can be then recovered by using the Embed algorithm. The recovered structures are finally merged to obtain a unique initial solution. The merge procedure adds one component at a time by choosing a random orientation. For every component a number (reported as 50) of random tries are done in order to choose the best possible orientation i.e. the orientation which minimizes the errors with respect to the contact map.

In the second phase a correction/perturbation procedure is applied iteratively in order to refine the structure obtained in the first phase. The algorithm stops when either all constraints of the contact map are satisfied or when a control parameter $\delta > 0$ decreases to 0. The correction procedure tries to displace one coordinate at a time in order to satisfy more contact map constraints without introducing new errors. A coordinate c_i is *well placed* if $\forall j \in [1, N]$ if $M_{ij} = 1$ then $|c_i - c_j| \leq t$ and if $M_{ij} = 0$ then $|c_i - c_j| > t$. Every not well placed coordinate c_i is displaced by moving it on the surface of a sphere centered in c_i of radius (of mobility)

$$r_i = \min\{D_0 - t, t - D_1\} \tag{9}$$

where

$$D_0 = \min\{|c_i - c_j| \mid M_{ij} = 0 \text{ and } |c_i - c_j| > t\}$$

[‡]Freely available at *http://bioinformatics.cs.unibo.it/FT-COMAR*.

and

$$D_1 = \min\{|c_i - c_j| \mid M_{ij} = 1 \text{ and } |c_i - c_j| \leq t\}.$$

Note that, by definition, every point in the surface of the sphere of radius r_i is *safe* for c_i in the sense that it does not introduce new errors. A good position c_i' for c_i can be approximated by

$$c_i' = \mathbf{F}\frac{r_i}{|\mathbf{F}|}$$

where the pseudo-force $\mathbf{F}$ is defined by

$$\mathbf{F} = \sum_{j=0}^{N} sgn_{ij}\frac{c_i - c_j}{|c_i - c_j|} \tag{10}$$

$$sgn_{ij} = \begin{cases} -1 & if \;\; M_{ij} = 1 \\ +1 & if \;\; M_{ij} = 0 \end{cases}$$

The perturbation procedure introduces small changes in the set of coordinates in the following way. For every couple of coordinates c_i, c_j if $t - \delta < |c_i - c_j| \leq t$ and $M_{ij} = 1$ then the coordinates c_i, c_j are moved in order to be $\delta/10$ closer. If $t < |c_i - c_j| \leq t + \delta$ and $M_{ij} = 0$ then they are moved in order to be $\delta/10$ more distant. The perturbation can introduce new errors but it has the effect to avoid that the radius of mobility of not well placed coordinates becomes too small thus preventing the structure from getting stuck. The δ parameter acts like the temperature-like parameter in simulated annealing approaches. It is initialized to a positive value at the beginning of the refinement phase and it decreased slightly after every refinement. If it reaches 0 then the procedure stops.

In the third phase the reconstructed structure is scanned in order to fix the eventually broken backbone. The δ parameter is set to a positive value and exactly the same correction/perturbation procedure of the second phase is applied. The only difference is that a coordinate c_i is displaced only if the conditions $|c_i - c_{i+1}| < 3.5$ or $|c_i - c_{i+1}| > 4$ hold. If δ reaches 0 and again the backbone constraint is not satisfied then the bad pair of consecutive coordinates are moved closer or apart in order to correctly satisfy the backbone constraint without taking care of the other constraints imposed by the contact map. This is justified by the fact that backbone constraints are always satisfied in native protein structures, while the contact map may contain errors.

The convergence speed of the algorithm is highly influenced by the quality of the initial structure guessed in the first phase. Note that the first phase is flexible enough to accommodate specific modifications e.g. if the reconstruction is applied to a family of proteins for which there is additional knowledge of intra-residue distances available. Another feature of this algorithm is that it works also if some entries of the contact map (typically, the most *unsafe* entries) are not specified (this is actually the behavior of the algorithm as described in [22]). The algorithm simply considers the unspecified entries of the contact map as if they were 0 in the first phase and it does not consider them at all in the second phase.

4.4. Computational complexity of the methods

The analysis of the computational complexity of the heuristic algorithms presented here is difficult because all such algorithms are iterative and there is no simple way to determine good upper bounds to

the speed of convergence, which typically depends of how well the initial solution is chosen. Anyway, it is still possible to provide an accurate analysis of the cost needed to generate an initial solution and to execute an iterative step.

The initial solution for the methods described in Section 4.1. is randomly chosen then it costs $O(N)$ (there are just N coordinates which describe the structure). Assume to use a gradient descent method to minimize the energy cost function. The cost of an iterative step for the first (Bohr et al.) and the second (Galaktionov and Marshall) methods can be upper bounded by the time needed to compute the partial derivatives of equations (2) and (4), respectively. Since in both cases the summation term contains exactly $N^2 + (N-1)$ distances, the computational complexity of an iterative step for both algorithms is upper bounded by $O(N^2)$.

In the simulated annealing approaches of Section 4.2., the cost of the growth phase is dominated by the cost to compute the score for every newly added coordinate to the growing chain. Since for every newly added i-th coordinate we have to compute the score by considering the previous $(i-1)$ coordinates (see, for example, equation (6)), the total cost is on the order of $N(N+1)/2$. Then growing the chain costs $O(N^2)$. In the adaptation phase the cost of the displacement is bounded by the time needed to compute the energy functions (7) for the first method (Vendruscolo et al.) and (8) for the second method (Pollastri et al.). Recall that for the second method, it can happen to displace a whole alpha helix. Then the cost of a displacement is $O(N)$ for the first method and $O(dN)$ for the second method, where d is the length of the largest alpha helix. If we assume that an iterative step of these methods consists of the displacement of all coordinates then for both algorithms an iterative step costs $O(N^2)$ in the worst case.

In the distance geometry based algorithm of Section 4.3. (Vassura et al.), the cost to generate an initial solution is bounded by the cost of the Embed algorithm which is $O(N^3)$. In fact, in the worst case, the contact map contains no independent components and the Embed algorithm runs on the entire map. During the iterative correction phase (both in the second and third phase) the two more time-consuming operations are the calculation of the radius of mobility (9) and of the pseudo-force (10). The cost to correct the position of one coordinate is then $O(N)$. In the same way, the cost to perturbate one coordinate is $O(N)$ which implies that the total cost of a single displacement is $O(N)$. As in the previous case, if one iteration step terminates when all coordinates have been eventually displaced, then an iterative step costs $O(N^2)$ in the worst case.

In conclusion, the iterative step is quadratic in the dimension of the input for all reviewed algorithms. The speed of convergence is very likely to depend on the quality of the initial solution. In this sense, although more costly, the choice to build an initial solution by means of a growing process or by using a distance geometry approach can be more effective than the random approach. The only reports about the running time of the algorithms on modern processors refer to the two more recent methods (i.e. Pollastri et al, and Vassura et al.). In both cases the authors report their methods to have total running time on the order of a few seconds for a medium-size protein.

5. Discussion and perspectives

A direct comparison of the performances of the algorithms described here is not possible since there are not currently available implementations of the older ones. Moreover, due to the limited computational capabilities of the mid nineties, the older methods have been tested only on a limited set of targets. Thus

here we try to discuss the effectiveness of these approaches according to what is reported in literature

In [21] the performances of the reconstructing algorithms have been described when the input is a native contact map of threshold 8 Å (the review does not include the Pollastri et al. algorithm, whose performance was reported in [18]). In the presence of no errors in the contact map, all such methods seem to behave quite well. The reconstructed structure, in almost all cases, matches well with the native one and its RMSD distance is quite small (the average RMSD reported is around 4Å for all the methods). A more extensive set of tests reported in [21] shows that native contact maps of threshold 8Å do not always contain enough information to assure a good reconstruction. For instance, there are native contact maps for which there are completely consistent realizations that are actually really distant from the native structures (in terms of RMSD). This happens when the contact map of threshold 8Å contains few or no long range contacts which actually impose the most relevant constraints to the protein structure. From this point of view, contact maps of higher thresholds (from 12Å to 18Å) seem to be the most informative in the sense that the reconstruction from these maps is always very accurate (the mean RMSD reported in [21] is below 2Å which is very close to the experimental error of NMR spectroscopy). Another interesting result is that the complete knowledge of the contact map is not necessary to obtain a good reconstruction. In [21], the authors show that with their algorithm just 25% of random entries of a contact map is sufficient to obtain good performances (assuming that the map has high threshold and that 25% of the entries are correct). The scenario changes dramatically when contact maps are allowed to contain errors. A simple way to add errors is to randomly flip the entries of the native maps. With this simple and *artificial* error model the quality of the reconstructions rapidly decreases. The method of Galaktionov and Marshall scores poorly when the amount of random error is around the 5% of the entire map. The methods of Vendruscolo et al. and Vassura et al. seem to be more robust since they can tolerate random errors up to 5% quite well (the mean RMSD obtained is below 5Å which usually implies high similarity with the native structure). These results are quite unsatisfactory since the amount of error contained in predicted contact maps is routinely much higher.

This does not mean that the methods are completely useless for reconstructing the 3D structure from predicted contact maps. One notable example is the method of Pollastri et al. which obtained encouraging reconstruction results on a large set of proteins [18]. The average RMSD obtained for the reconstructed structures is around 13Å. The method of Vassura et al. has been tested on a set of 100 predicted contact maps [22] (the maps have been predicted with CORNET predictor [10]). In this case, the average RMSD of the reconstructions is around 16Å. These results seem promising and suggest that new protein reconstruction algorithms that incorporate more information in the form of constraints and/or predictions will contribute towards the solution of the protein folding problem.

Acknowledgements

The authors thank MIUR for the PNR 2003 (FIRB art.8) on Bioinformatics for Genomics and Proteomics and Laboratorio Internazionale di BioInformatica (LIBI). This work was also supported by the Biosapiens Network of Excellence Project LSHG-CT-2003-503265 (a Grant of the European Unions VI Framework Programme).

References and Notes

1. Anfinsen, C. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223-230.
2. Baker, D. Prediction and design of macromolecular structures and interactions. *Phil. Trans. R. Soc. London B Biol. Sci.* **2006**, *361*, 495-463.
3. Bartoli, L.; Capriotti, E.; Fariselli, P.; Martelli, P.L.; Casadio, R. The pros and cons of predicting protein contact maps. In *Protein Structure Prediction*. Zaki, M.J., Bystroff, C. Eds.; Humana Press: New York, NY, USA, 2008; pp. 199-217.
4. Bartoli, L.; Fariselli, P.; Casadio, R. LETTERS AND COMMENTS: The effect of backbone on the small-world properties of protein contact maps. *Physical Biology* **2007**, 1-5.
5. Blumental, L.M. *Theory and applications of distance geometry*. Clarendon Press, Oxford, United Kingdom, 1953.
6. Breu, H.; Kirkpatrick, D. Unit disk graph recognition is NP-hard. *Computational Geometry* **1998**, *9*, 3-24.
7. Bohr, J.; Bohr, H.; Brunak, S.; Cotterill, R.M.; Fredholm, H.; Lautrup, B.; Petersen, S.B. Protein structures from distance inequalities. *J. Molecular Biology* **1993**, *231*, 861-869.
8. Izarzugaza, J.M.; Grana, O. ; Tress, M.L.; Valencia, A.; Clarke, N.D. Assessment of intramolecular contact predictions for CASP7. *Proteins* **2007**, *69*, 152-158.
9. Cormen, T.H.; Leiserson, C.E., Rivest, R.L.; Stein, C. *Introduction to Algorithms*, 2^{nd} Ed.; MIT Press, Cambridge, MA; McGraw-Hill Book Co., Boston, MA, 2001.
10. Fariselli, P.; Osvaldo, O.; Valencia, A.; Rita, C. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* **2001**, *14*, 835-843.
11. Fariselli, P.; Rossi, I.; Capriotti, E.; Rita, C. The WWWH of remote homolog detection: The state of the art. *Briefings in Bioinformatics* **2007**, *8*, 78-87.
12. Galaktionov, S.G.; Marshall, G.R. Properties of intraglobular contacts in proteins: an approach to prediction of tertiary structure. In Proceedings of the 27th Hawaii International Conference on System Sciences, IEEE Society Press, USA, 1994; pp. 326-335.
13. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671-680.
14. Crippen, G.M.; Havel, T.F. *Distance geometry and molecular conformation*. Research Studies Press Ltd, Taunton, England, 1988.
15. Havel, T.F. Distance Geometry: Theory, Algorithms, and Chemical Applications. In *Encyclopedia of Computational Chemistry*. Ragué, V., Schreiner, P.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer, H.F., Eds.; J. Wiley & Sons: New York, 1998; pp. 723-742.
16. Moré, J.; Wu, Z. ϵ-Optimal solutions to distance geometry problems via global continuation. In *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*. Pardalos, P. M., Shalloway, D., Xue, G. Eds.; American Mathematical Society, 1995; pp. 151-168.
17. Kuhn, F.; Moscibroda, T.; Wattenhofer, R. Unit disk graph approximation. In Proceedings of 2nd ACM Joint Workshop on Foundations of Mobile Computing, Philadelphia, Pennsylvania, USA, October 2004.

18. Baú, D.; Pollastri, G.; Vullo, A. Distill: a machine learning approach to ab initio protein structure prediction. In *Analysis of Biological Data: A Soft Computing Approach*. Bandyopadhyay, S., Maulik U., Wang J. T. L. Eds.; World Scientific, 2007.
19. Saxe, J. B. Embeddability of weighted graphs in k-space is strongly NP-hard. In Proceedings of the 17th Allerton Conference on Communications, Control, and Computing, 1979; pp. 480-489.
20. Snyman, J.A. *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*. Springer-Verlag: New York, NY, USA, 2005.
21. Vassura, M.; Margara, L.; Di Lena, P.; Medri, F.; Fariselli, P.; Casadio, R. Reconstruction of 3D Structures From Protein Contact Maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2008**, *5*, 357-367.
22. Vassura, M.; Margara, L.; Di Lena, P.; Medri, F.; Fariselli, P.; Casadio, R. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics* **2008**, *24*, 1313-1315.
23. Vassura, M.; Margara, L.; Di Lena, P.; Medri, F.; Fariselli, P.; Casadio, R. Fault Tolerance for Large Scale Protein 3D Reconstruction from Contact Maps. Springer Verlag Lecture Notes in Bioinformatics **2007**, *4645*, 25-37.
24. Vendruscolo, M.; Kussell, E.; Domany, E. Recovery of protein structure from contact maps. *Folding and Design* **1997**, *2*, 295-306.