Reconstruction of 3D Structures from Protein Contact Maps

Marco Vassura, Luciano Margara, Pietro Di Lena, Filippo Medri, Piero Fariselli, and Rita Casadio

Abstract—The prediction of the protein tertiary structure from solely its residue sequence (the so-called Protein Folding Problem) is one of the most challenging problems in Structural Bioinformatics. We focus on the protein residue contact map. When this map is assigned, it is possible to reconstruct the 3D structure of the protein backbone. The general problem of recovering a set of 3D coordinates consistent with some given contact map is known as a unit-disk-graph realization problem, and it has been recently proven to be NP-hard. In this paper, we describe a heuristic method (COMAR) that is able to reconstruct with an unprecedented rate (3-15 seconds) a 3D model that exactly matches the target contact map of a protein. Working with a nonredundant set of 1,760 proteins, we find that the scoring efficiency of finding a 3D model very close to the protein native structure depends on the threshold value adopted to compute the protein residue contact map. Contact maps whose threshold values range from 10 to 18 Ångstroms allow reconstructing 3D models that are very similar to the proteins' native structure.

Index Terms—Combinatorial algorithms, contact map, molecular modeling, protein structure prediction.

1 Introduction

Protein folding is the process by which a protein assumes its 3D structure. All protein molecules are endowed with a primary structure consisting of the polypeptide chain. Folding of this chain in the solvent space is constrained to a different and yet unsolved extent by the protein's different residue composition, and this results into the so-called 3D protein structure. The biological functional unity is strictly dependent on the protein 3D structure. At the coarsest level, it is believed that folding involves first the local establishment of secondary structures, specifically alpha helices and beta sheets, and that only after the hydrophobic collapse is the 3D protein structure formed. Actually, the greatest yet open problem in Structural Bioinformatics is the 3D protein structure prediction from its primary structure [14]. The ab initio solution of the folding problem is still lacking; a typical alternative approach is to identify a set of subproblems, such as the prediction of protein secondary structures, solvent accessibility and/or prediction of residue contacts, and/or design of heuristic solutions. Among different possibilities, the prediction of protein contact maps starting from the protein chain is particularly promising, since even a partial solution of it can significantly help the prediction of the protein 3D structure [9].

A contact map of a given protein is a binary matrix M such that $M[i, j]$ has value of one if and only if the distance between residues i and j in the native structure is less than or equal to a preassigned threshold. Having at hand a contact map, a reliable and fast reconstruction procedure of the 3D structure is needed. The problem is equivalent to the unit-disk-graph realization, which has been proved to be NP-hard [6]. Other well-studied similar problems are NMR structure determination [12], [16] and protein conformational freedom [11]. However the different nature of distance constraints induced by the protein contact map requires the implementation of other methods and tools. A series of heuristic algorithms have been developed to solve the problem. Galaktionov and Marshall [10] reconstructed the structures of five small proteins by adopting information relative to the residue coordination numbers. Vendruscolo et al. [20] described a method based on simulated annealing with the contact map as a target potential. They achieved an average Root-Mean-Square Deviation (RMSD) of 2.5 Ångstroms (Å) on some 20 protein structures. Other approaches rely on steepest descent with inequality distance constraints [5] and, alternatively, on an algorithm that minimizes a continuous cost function that embodies constraints associated with contact and angle maps [17], respectively. On average, these methods reconstruct the protein structures without completely satisfying the contact map in that the reconstructed protein structures may have contact maps that slightly differ from the native ones. However, these methods can deal also with nonphysical contact maps, allowing in principle the 3D reconstruction from blurred contact maps.

In this paper, we face the problem of 3D protein reconstruction starting from its native contact map. The main contribution of this paper is a heuristic technique called COntact MAp Reconstruction (COMAR), which is able to compute in few seconds of runtime a 3D structure that exactly matches all the entries of the input contact map.

The motivations for our work can be summarized as follows:

- *M. Vassura, L. Margara, P. Di Lena, and F. Medri are with the Department of Computer Science, University of Bologna, Via Mura Anteo Zamboni, 7, 40127 Bologna, Italy. E-mail: {vassura, margara, dilena, medri}@cs.unibo.it.*
- *P. Fariselli and R. Casadio are with the Biocomputing Group, Department of Biology, University of Bologna, Via Irnerio, 42, 40127 Bologna, Italy. E-mail: piero@biocomp.unibo.it, casadio@alma.unibo.it.*

Manuscript received 7 Aug. 2007; revised 21 Jan. 2008; accepted 7 Feb. 2008; published online 15 Feb. 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2007-08-0096.
Digital Object Identifier no. 10.1109/TCBB.2008.27.

- The problem of computing 3D structures that exactly match a given contact map is a well-known NP-hard problem in computational geometry. The problem remains NP-hard under some additional biological assumptions, e.g., fixed distances between adjacent amino acids. An efficient heuristic technique for solving this problem is of interest in itself.
- Our technique aims to reduce the (quite large) gap existing between the world of contact predictors and that of reconstruction techniques. In order to further reduce this gap, we are currently testing our technique using noisy and incomplete contact maps. For some preliminary (encouraging) experimental results, see Section 5.
- Restricting attention to native contact maps provides some insight on contact-map-related problems. As an example, experimental results show that in many cases, contact maps do not contain enough information for selecting the native structure among all the feasible ones we have computed (see Section 4.5). Such evidence heavily depends, as expected, on the contact threshold: small thresholds produce higher uncertainty.

In detail, in this paper, we show that COMAR is reliable in the sense that it is able to produce a reconstructed structure that has the same contact map as the native structure. This is so for our nonredundant data set consisting of 1,760 complete protein structures and irrespective of the threshold value adopted in computing the contact map. Although the unit-disk-graph realization problem is in general intractable, the average execution times of COMAR vary from 3 to 15 seconds, depending on the contact map threshold value. The relation between contact map threshold, protein size, and protein 3D structure is analyzed, showing that on the average, contact maps computed at thresholds ranging from 10 to 18 Å allow a better 3D structure recovery than those computed at lower values (ranging from 7 to 9 Å). These results and a partial description of our algorithm already appeared in [19].

Here, we analyze also the generalization capability of COMAR, i.e., the ability of finding a 3D structure completely consistent with a given native contact map as a function of the different steps of the algorithm. We also show that COMAR capability is increased by introducing a randomization procedure that increases the probability of finding a 3D model consistent with the input contact map.

2 Protein Structure Reconstruction

2.1 Protein Representation and Contact Maps

Proteins structures are described by the coordinates of the atoms that collectively constitute the macromolecule. For a protein with n atoms, we need $3n$ numbers to specify its 3D structure. An alternative view is to consider the distance matrix. The distance matrix is a symmetric matrix that contains in its cells the euclidean distance between each pair of atoms. If the number of atoms is n, we need n^2 elements; since the matrix is symmetric (the distance between atoms i and j is the same of that between j and i), the effective number of needed elements is only $n^*(n-1)/2$. In order to simplify the protein representation, not all protein atoms are taken into account, and residues are considered as unique entities. In this case, the distance matrix has a number of rows (and columns) equal to the residue number. Each distance matrix entry is then the distance between residue i and j. The distance between two residues can be defined in different ways:

- the distance between a specific pair of atoms (i.e., Cα-Cα or Cβ-Cβ),
- the shortest distance among the atoms belonging to residue i and those belonging to residue j, and
- the distance between the centers of mass of the two residues.

Starting from the protein distance matrix and selecting an arbitrary distance cutoff (threshold), a further simplified representation can be obtained: the protein contact map. Residues are in contact if their distance is less than or equal to the preassigned threshold. Contact maps are binary symmetric matrices, whose elements different from zero (and set to one) represent the contacts between residues.

In this paper, we use the Cα representation of the protein backbone, and for sake of simplicity, we refer to the protein Cα trace as the "protein structure" or 3D protein structure.

2.2 Distance Geometry and Protein Structure Reconstruction

Distance geometry (see [4] for an introduction) deals with the characterization of mathematical properties that can be derived from distance values between pairs of points. The mathematical foundation of distance geometry is essentially due to Cayley (1841) and Menger (1928), who showed how some basic geometry properties such as convexity could be defined in terms of distance values. One fundamental problem in distance geometry is to find a correct set of 3D euclidean coordinates that satisfy a set of distance constraints. In general, a set of points in the 3D space that satisfies some given constraints does not exist. However, Cayley and Menger gave necessary and sufficient conditions for a set of positive values to be the exact distances between pairs of 3D coordinates. Thus, given a consistent set of distances in the 3D space, the problem of finding coordinates that satisfy such exact distance constraints can be solved by a polynomial-time algorithm [4], while the problem is NP-hard when the given set of distances is sparse [18].

NMR spectroscopy and X-ray crystallography are the most widely used experimental techniques to obtain bounds to the interatomic distances between residues. Because of experimental errors, we can usually obtain only a set of lower and upper bounds to such interatomic distances rather than exact values. The distance-geometry-based approach to the protein structure reconstruction problem aims at developing techniques to recover the 3D protein structure, given a set of lower and upper bounds to residue interatomic distances. The problem of computing a set of consistent coordinates is generally intractable [15]. Crippen and Havel developed a recovering algorithm from a sparse set of lower and upper bounds to the interatomic distances [8], [12]. Their algorithm first uses some bound smoothing techniques 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 adopted as a starting solution for an optimization procedure.

While the problem of recovering protein structures from a set of distances is known to have a polynomial-time solution, the same problem from contact maps is NP-hard [6]. However, empirical developed applications seem to suggest that the approach of predicting the protein structure from contact maps can be fruitful (see, for example, [5], [20], and [21]). An introduction to such approach can be found elsewhere [3].

3 Algorithm Description

In this section, we describe COMAR, a heuristic algorithm to find a set of 3D coordinates consistent with some given native contact map CM of threshold t.

```
COMAR($CM \in \{0,1\}^{n \times n}$, $t \in N$)
repeat

    //Phase 1: initial solution
    $C \leftarrow$ RANDOM-PREDICT($CM, t$)

    //Phase 2: refinement
    $C \leftarrow$ CORRECT($CM, C, t$)
    set $\varepsilon$ to a strictly positive value
    while $C$ is not consistent with $CM$ and $\varepsilon > 0$ do
        $C \leftarrow$ PERTURBATE($CM, C, t, \varepsilon$)
        $C \leftarrow$ CORRECT($CM, C, t$)
        decrement slightly $\varepsilon$
    endwhile

until $C$ is consistent with $CM$

return $C$
```

The algorithm consists of two phases (see the pseudocode above). In the first phase (Phase 1), it generates a random initial set of 3D coordinates $C \in \boldsymbol{R}^{3 \times n}$ (**RANDOM-PREDICT**) that will be the starting point for the refinement procedure in the second phase. A detailed description of **RANDOM-PREDICT** can be found in Section 3.1.

In the second phase (Phase 2), the algorithm iteratively applies two local correction/perturbation techniques to the current set of coordinates: **CORRECT** and **PERTURBATE**. This is performed in order obtain a new set of coordinates "more consistent" with the given contact map. A detailed description of **CORRECT** and **PERTURBATE** can be found in Section 3.2. The refinement continues until the set of coordinates is consistent with the given contact map or until a control parameter ε becomes zero. The control parameter ε has initially a positive value, and it is decremented every some refinement steps. If it reaches the zero value before a consistent set of coordinates is found, then a new random initial set of coordinates is generated; ε is initialized again to a strictly positive value, and the refinement procedure restarts on the new set.

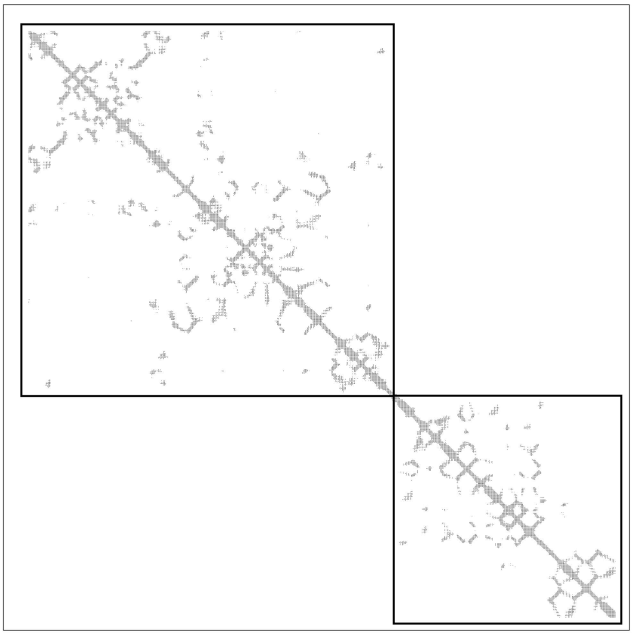


Fig. 1. Example of a contact map, computed with a threshold value of 14 A, with two splittable components (protein: phenylalanyl-tRNA synthetase complexed with phenylalanine and a phenylalanyl-adenylate analog, PDB code 1b7y chain B). The two submatrices corresponding to each component are shown.

3.1 First Phase: Finding the Initial Solution

The first step of the algorithm consists of the partially random generation of a starting coordinate set that will be refined in the second phase of the algorithm.

```
RANDOM-PREDICT($CM \in \{0,1\}^{n \times n}$, $t \in N$)
$\{CM_1, \ldots, CM_k\} \leftarrow$ SPLIT($CM$)
for $i \leftarrow 1$ to $k$ do
    $C_i \leftarrow$ EMBED(GUESS-DIST($CM_i$, t))
endfor
$C \leftarrow$ MERGE($C_i, \ldots, C_k, CM$)
return $C$
```

The computation of the initial solution is preceded by a scanning of the contact map for the existence of *splittable* components (**SPLIT**). Splitting the initial contact map in submatrices is done to locate those fragments of protein that demonstrate a high degree of independence with respect to mutual interactions (Fig. 1). The submatrices are then separately used to create sets of coordinates (**EMBED**) to be merged (**MERGE**) in an initial solution. The merging procedure is managed by selecting, between a set of equally distributed 3D angles, the best rotation of the coordinates corresponding to each component with respect to the lower number of errors generated in the contact map. The pseudocodes and detailed descriptions of **SPLIT** and **MERGE** procedures are in Appendix A.

A fast and reliable way to obtain good starting coordinates for the splittable components is provided by the metric matrix embedding (**EMBED**) algorithm [12]. The **EMBED** algorithm can be used to compute a set of 3D coordinates that is, in a certain sense, the best 3D fit for some distance matrix D. By using some a priori knowledge about the physical conformation of the proteins, the **GUESS-DIST** procedure tries to guess a possible set of distances $D \in R^{n \times n}$ consistent with some native contact map $CM \in \{0,1\}^{n \times n}$. Generally, no set of 3D points is consistent with some distance matrix D. However, **EMBED** uses standard numerical linear algebra methods to find the least distorted projection of D in the 3D euclidean space [10]. The pseudocode and a detailed description of the **GUEST-DIST** procedure are in Appendix B.

3.2 Second Phase: Refinement of the Coordinates

The second step of the algorithm applies iteratively a local correction/perturbation heuristic technique to the randomly predicted set of coordinates to obtain a new set of coordinates closer to the native contact map.

We call *not well placed* those residues whose coordinates are not consistent (according to the contact map) with the coordinates of all other residues. The local correction technique **CORRECT** attempts to change the coordinates of every not well-placed residue i as soon as the change does not introduce new errors in the coordinate set.

Let us consider a contact map $CM \in \{0,1\}^{n \times n}$ of threshold t and a set of coordinates $C \in \boldsymbol{R}^{3 \times n}$. Let us denote with $d_{ij} = |C[i] - C[j]|$ the distance between residues i and j of coordinates $C[i]$ and $C[j]$, respectively. Formally, we say that a residue j is *well placed* with respect to residue i whether either $(CM[i,j] = 1$ and $|C[i] - C[j]| \leq t)$ or $(CM[i,j] = 0$ and $|C[i] - C[j]| > t)$. A residue i is *well placed* if every other residue $1 \leq j \leq n$ is well placed with respect to i.

CORRECT attempts to change the coordinate of every not well-placed residue i in a new coordinate that does not affect the old set of well-placed residues with respect to i.

```
CORRECT($CM \in \{0,1\}^{n \times n}$, $C \in \boldsymbol{R}^{3 \times n}$, $t \in \boldsymbol{N}$)
for $i \leftarrow 1$ to $n$ do
    if $i$ is not well placed then
        $C[i] \leftarrow \mathbf{MOVE}(CM, C, t, i)$
    endif
endfor
return $C$
```

The procedure to approximate a good and safe coordinate for some residue i is described in **MOVE**. It changes the coordinate $C[i]$ to the coordinate of a point on the surface of the sphere of radius r_i centered in $C[i]$. The point is chosen in a region of the surface that is supposed to be as distant as possible from the whole set of residues j not well placed with respect to i such that $CM[i,j] = 0$ and as close as possible to the whole set of residues k not well placed with respect to i such that $CM[i,k] = 1$. The radius of mobility r_i of the residue i is defined as

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

where

- $D_0 = \min\{d_{ij} | d_{ij} > t \text{ and } CM[i,j] = 0\}$ and
- $D_1 = \max\{d_{ij} | d_{ij} \leq t \text{ and } CM[i,j] = 1\}$.

Then, by definition, the coordinate $C[i]$ of the residue i can be safely changed in any coordinate $c \in \boldsymbol{R}^{3 \times n}$ such that $|C[i] - c| \leq r_i$ without decreasing and eventually increasing the cardinality of the set of residues well placed with respect to i. The pseudocode and a detailed description of the **MOVE** procedure are in Appendix C.

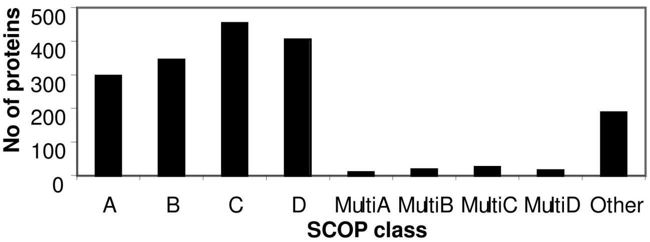


Fig. 2. Distribution of our protein set according to the SCOP classes. $\mathrm{A} = \mathrm{all\ alpha}$. $\mathrm{B} = \mathrm{all\ beta}$. $\mathrm{C} = \mathrm{Alpha/Beta}$. $\mathrm{D} = \mathrm{Alpha + Beta}$. $\mathrm{Multi}\text{-}\{\mathrm{A,B,C,D}\}$ and Other contain multidomain proteins.

A run of the correction procedure may reduce the radius of mobility for not well-placed residues. In order to maintain as large as possible the radius of mobility for such residues, after a correction procedure, we apply small perturbations to the coordinate set using the **PERTURBATE** procedure.

```
PERTURBATE($CM \in \{0,1\}^{n \times n}$, $C \in \boldsymbol{R}^{3 \times n}$, $t \in \boldsymbol{N}$, $\varepsilon \in \boldsymbol{R}$)
1:  for $i \leftarrow 1$ to $n$ do
2:      for $j \leftarrow 1$ to $n$ do
3:          if $t - \varepsilon < |\mathrm{C[i]} - \mathrm{C[j]}| \leq t$ and $CM[i,j] = 1$ then
4:              bring closer $C[i]$ and $C[j]$ of $\varepsilon/10$
5:          endif
6:          if $t < |\mathrm{C[i]} - \mathrm{C[j]}| < t + \varepsilon$ and $CM[i,j] = 0$ then
7:              move away $C[i]$ and $C[j]$ of $\varepsilon/10$
8:          endif
9:      endfor
10: endfor
11: return $C$
```

For every residue i and every residue j well placed with respect to i, if their distance d_{ij} is under the given threshold $(CM[i,j] = 1)$ but close to the threshold, then **PERTURBATE** changes the coordinates of i and j in order to make them a bit more closer (lines 3-5). If d_{ij} is above the given threshold $(CM[i,j] = 0)$ but close to the threshold, then **PERTURBATE** changes the coordinates of i and j in order to make them a bit more distant (lines 6-8). A perturbation can introduce new errors to the coordinate set, but conversely, it avoids not well-placed residues from getting stuck.

4 Experimental Results

4.1 Protein Set

We selected the list of proteins with their relative structural classifications from SCOP [2] release 1.67. We then downloaded the corresponding protein structures from the PDB, and we retained only those files with coordinates obtained with x-ray experiments, with resolution < 2.5 A, and without missed internal residues. Finally, using BLAST [1], we removed sequence redundancies, ending up with a data set of 1,760 protein chains with sequence similarity lower than 25 percent. The distribution of the 1,760 protein chains according to the SCOP classification is shown in Fig. 2. Our protein set contains 1,502 one-domain and 258 multidomain chains. The complete list is available at the website http://vassura.web.cs.unibo.it/protlist.tgz.

4.2 Hardware Configuration

All the test runs are executed on personal computers equipped with an Intel Pentium 4 processor with a clock rate of 2.8 GHz and 1 Gbit of RAM memory. Times reported are measured using the `time()` C library function. During each run, the program collects time information before reading the input and, again, after computing the result; the CPU time actually elapsed is computed as the difference between the two figures.

4.3 Distance Measures

To measure the difference between contact maps, we use the simple Hamming distance that counts the number of different bits; this distance is also the target function of the problem. When we deal with two protein structures, the classical RMSD is computed between the native and the reconstructed structure. RMSD is commonly used to compare two molecular structures described by some set of coordinates $C, C' \in R^{3\times n}$. It is defined as the smallest distance;

$$D_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C'[i] - C_k[i])^2},$$

where $C_k \in R^{3\times n}$ is obtained by rotating and translating the coordinates set C.

4.4 Highlighting COMAR Capability

In this section, we experimentally test the convergence capability of COMAR, namely, the ability of COMAR to rapidly find a 3D structure that matches the contact map taken as input.

The termination conditions let the algorithm run until a set of coordinates consistent with a given input contact map is found (Section 3). Formally, given a contact map CM and a set of coordinates C, we say that COMAR is *convergent* when the refinement of C (see Section 3.2) leads to a new set of coordinates consistent with CM. COMAR refinement (Phase 2) is more likely to converge as soon as the 3D structure described by the initially guessed coordinate set is sufficiently similar to the native one (see also Section 4.4.2). COMAR capability of finding a 3D structure with a given input contact map depends therefore on the interplay between the quality of the initial guessed solution (Phase 1) and the result after the refinement procedure (Phase 2).

To prove this, we test independently the robustness of the prediction phase (Section 3.1) and that of the refinement phase (Section 3.2) by evaluating the RMSD value of each 3D model to the corresponding native structure before and after refinement (see below). All tests have been performed on the data set described in Section 4.1, adopting a Cα protein representation and computing the contact map with a threshold value of 12 Å. As discussed in Section 4.5, our choice is consistent with the observation that contact maps computed at lower thresholds are found to admit 3D structures that, in spite of being completely consistent with the input contact maps, are largely different from the native structures.

It is interesting to discuss how much the random perturbation of our statistical information is relevant in order to obtain convergent computations. We tested what is the percentage of convergence of COMAR when the initial solution is not randomly perturbed. For instance, COMAR converges for 98.69 percent for contact maps of threshold 12 Å. In Fig. 3, the percentage of convergence of COMAR for different values of contact map threshold is reported. We obtain that the convergence rate is above 90 percent for all thresholds over 7 Å. The reconstruction quality is higher for thresholds ranging between 10 and 18 Å (see Section 4.5).

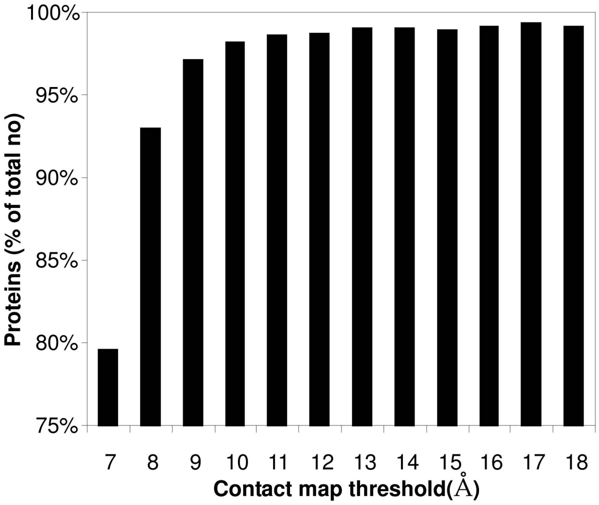


Fig. 3. Proteins within our set for which COMAR finds a 3D structure consistent with the corresponding input contact map computed at different threshold values without random perturbations.

4.4.1 Quality of the Prediction Phase

The RANDOM-PREDICT procedure (Phase 1, see Section 3.1) tries to guess a possible set of coordinates for a given contact map by using available statistical information on contact distribution distances in real proteins. The prediction is partially random in the sense that the predicted set of distances is actually obtained by introducing random perturbations on a set of distances recovered from statistical information (see procedure DISTANCE in Appendix B).

The quality of the prediction phase can be measured in terms of the RMSD from the native structure. We performed a series of tests for both the sets of distances generated with and without randomness. In Fig. 4, we show how the proteins in our data set are distributed according to the RMSD between the native structure and the nonrandom initial structure. The maximum RMSD value reached is 19.4 Å with an average RMSD of about 3.1 Å.

In Fig. 5, we show the results of the same test when the initial guessed solutions are randomly perturbed as detailed in Section 3.1. For each protein, the RMSD value considered is the average RMSD value obtained after 50 different runs. The maximum RMSD value reached is 25.5 Å at an average RMSD of 4.7 Å.

From the test results, it appears that initial structures guessed without randomization have on the average better quality when compared to the native ones. However, as shown in Fig. 3, when randomization is omitted, the

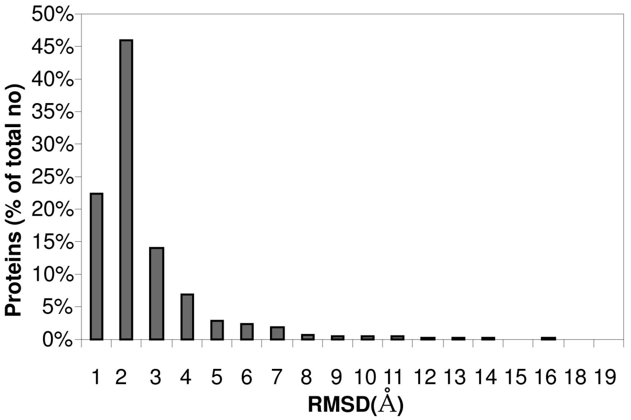


Fig. 4. Distribution of proteins according to the RMSD value between the native 3D structure and the guessed initial structure as evaluated after the RANDOM-PREDICT phase (Section 3.1) without random perturbation (Appendix B).

algorithm procedure fails in recovering all 3D models as a function of the threshold value. As a test case, the initial nonrandom solution of the protein of the Cricket Paralysis Virus (1b35, chain B) has a very high RMSD value (19.4 Å) from the native structure. Alternatively, when randomization is introduced for the same protein, among the 50 random initial structures generated, at least some have an RMSD value lower than 5 Å from the native structure. This is an example of how randomizing on the initial set of coordinates can effectively improve the performance of COMAR when the nonrandom initial solution leads to nonconvergence.

4.4.2 Error Tolerance of the Refinement Phase

In this section, we test the convergence of the refinement (Phase 2) in terms of RMSD of the initial solution to the native structures. This is done randomly, generating structures with RMSD values ranging between 1 Å to 32 Å. Such structures are generated by perturbing the native ones. A native set of coordinates C is perturbed with maximum error n Å, $1 \leq n \leq 32$, by randomly moving every coordinate in C of at most n Å. We verified that this procedure leads to a 3D structure whose RMSD from the original one is around n Å (data not shown).

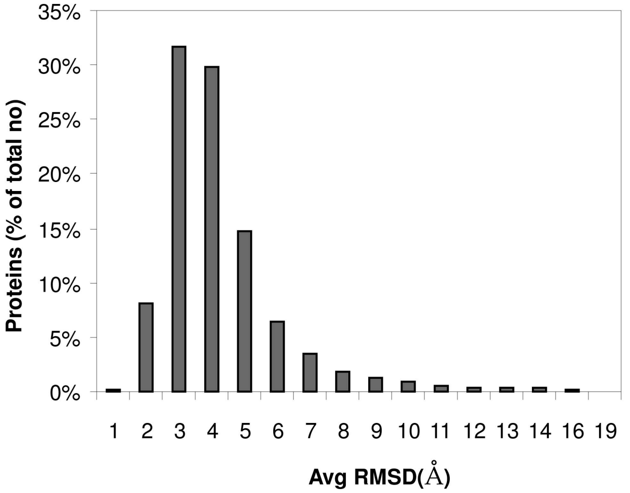


Fig. 5. Distribution of proteins according to the RMSD value between the native 3D structure and the guessed initial structure as evaluated after the RANDOM-PREDICT phase (Section 3.1) with random perturbation (Appendix B).

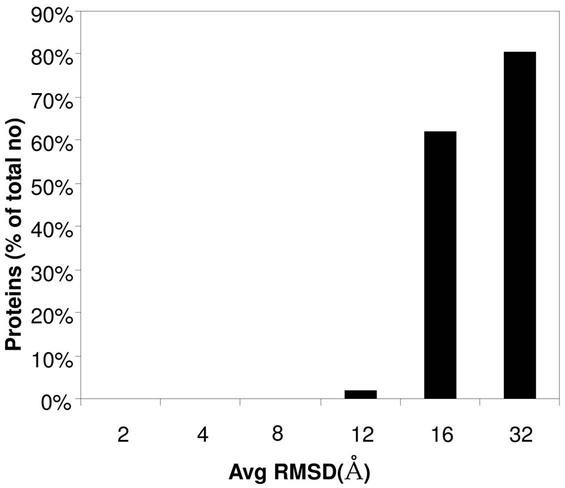


Fig. 6. Percentage of proteins of our set for which Phase 2 of COMAR is not able to converge as function of the RMSD value of the initial structure to the native structure. See text for details.

For each native structure, we perform a series of 10 random tests. The percentage of convergence in terms of the class of error is shown in Fig. 6. All native structures perturbed up to 8-Å RMSD are refined to structures exactly matching the native contact maps. The number of nonconverging structures is rapidly increasing when the RMSD value from the native structure is above 12 Å. This indicates that COMAR has good convergence capability: in nearly all tested cases, Phase 1 generates an initial structure having RMSD that is at the most 8 Å from the native one (see Fig. 5). This is further corroborated by the fact that Phase 2 can greatly reduce the RMSD of the given structure from the native one even when convergence is not obtained (Fig. 7). For example, for initial structures perturbed up to 16 Å, the average RMSD obtained after the refinement procedure is 3.2 Å; 61.9 percent of these structures are, however, not consistent with the corresponding native contact map.

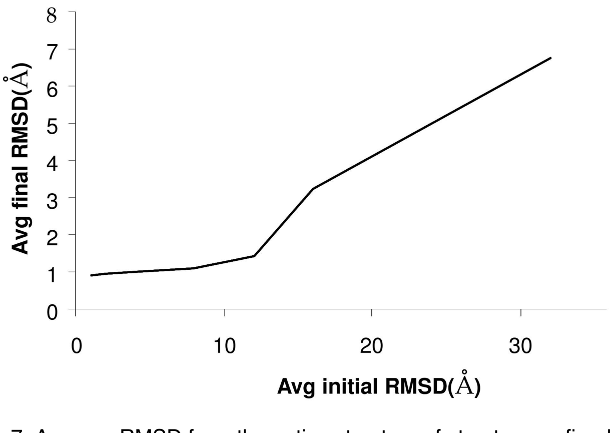


Fig. 7. Average RMSD from the native structure of structures refined by Phase 2 of COMAR as a function of the RMSD of the initial structure from the native structure. See text for details.

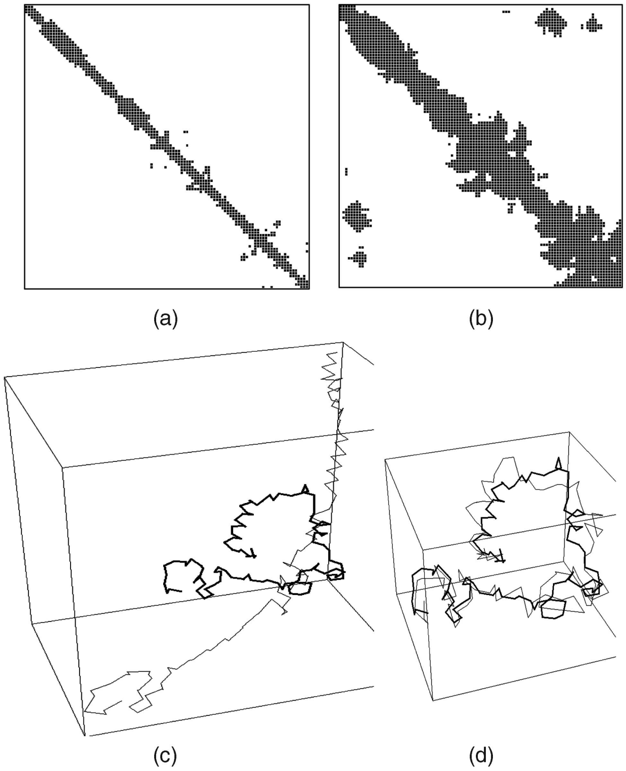


Fig. 8. Contact map degeneracy: a test case. The recovery of the 3D structure of Human Myeloperoxidase Isoform C (1cxp chain B,104 residues, all-alpha). (a) 1cxp contact map computed at a threshold of 7 Å. (b) 1cxp contact map computed at a threshold of 16 Å. (c) 1cxp native structure (thick line) compared to a recovered structure with the same contact map (a) $(\mathrm{RMSD} = 41.31\ \text{Å})$. (d) 1cxp native structure (thick line) compared to a recovered structure with the same contact map (b) $(\mathrm{RMSD} = 4.95\ \text{Å})$.

4.5 Three-Dimensional Structure Recovery

For each protein of our selected nonredundant data set, containing 1,760 protein structures (see Section 4.1), we generate 12 different contact maps by changing the contact threshold from 7 to 18 Å with a 1-Å step, and then, we run our procedure for all the 12*1,760 generated contact maps.

The most relevant result of our procedure is the fact that all the reconstructed protein structures satisfy the native contact maps. This means that the Hamming distance between the native and the reconstructed contact maps is zero, or in other words, that given the contact map of a protein, our algorithm finds a 3D structure that has the same contact map as the native protein. In spite of this, in some cases, the RMSD of the reconstructed protein with respect to the native structure can be very large (Fig. 8).

This indicates that some contact maps can represent a huge ensemble of protein conformations. Usually, this means that the map contains only a broad central band of local contacts, and no constraints are posed on the global bending of the protein. The reconstruction ambiguity is more evident when the contact map is generated using low values of contact thresholds (ranging from 7 to 9 Å) and decreases as the contact threshold increases (Table 1). Our results indicate that at increasing contact map threshold, both average RMSD and standard deviation values decrease over the all-protein set (Table 1). At increasing threshold values, global features in the contact map help in finding the 3D structure likely to be more similar/close to the native one.

TABLE 1
Scoring the Recovery of 3D Structure from the Contact Maps of 1,760 Proteins

Threshold (Å)	Cmap dist (Å)	Avg RMSD (Å)	AvgSD RMSD (Å)	Avg Time (s)	AvgSD Time (s)
7	0	6.11	4.09	15	136
8	0	4.58	3.86	9	110
9	0	3.37	3.42	9	155
10	0	2.62	2.98	10	157
11	0	2.21	2.69	5	71
12	0	1.97	2.51	3	15
13	0	1.75	2.29	2	13
14	0	1.58	2.09	3	16
15	0	1.47	2.01	10	274
16	0	1.39	1.90	2	9
17	0	1.36	1.75	5	94
18	0	1.35	1.79	3	17

Threshold = the threshold used to compute the input contact map; Cmap dist = the Hamming distance between the contact map of the native structure and the contact map of the recovered structure; Avg RMSD = the average, over all proteins, RMSD between the native structure and the recovered structure; AvgSD = the average standard deviation over all proteins; and Avg Time = the average, over all proteins, time needed to recover the 3D structure.

A typical example is shown in Fig. 8 for the protein Human Myeloperoxidase Isoform C (1cxp, chain B). The contact map computed with a threshold equal to 7 Å (Fig. 8a) does not contain enough global information of the protein structure, and a large number of protein structures are represented by that map. For instance, a possible reconstruction is reported in Fig. 8c where the RMSD to the native structure is 41.3 Å. When the contact map is computed at a threshold of 16 Å (Fig. 8b), more features are available off of the main diagonal, and the recovered 3D structure is closer to the native one. Indeed, RMSD decreases now to 4.9 Å (Fig. 8d).

This finding prompted us to do a search in the threshold space to optimize the RMSD values. We find that a better 3D reconstruction is obtained when a high threshold value is adopted (10 Å or higher), while the average runtime (over 1,760 proteins) does not depend on the threshold adopted (Table 1). RMSD values between the reconstructed and the corresponding native 3D protein structures are analyzed as a function of the four main SCOP classes, clustered in monodomain and multidomain proteins. The results are shown in Fig. 9. As a general trend, we find that multidomain proteins are more easily reconstructed with our procedure than monodomain proteins. This is so rather independent of the threshold value adopted. One possible explanation is that the contact map of multidomain proteins carries information about the interdomain residue contacts that poses more constraints to the reconstruction of the 3D protein structure. Another interesting point that emerges from Fig. 9 is the fact that the contact maps of monodomain all-alpha proteins (a SCOP label) tend, on average, to be more ambiguous in their reconstruction. This is in agreement

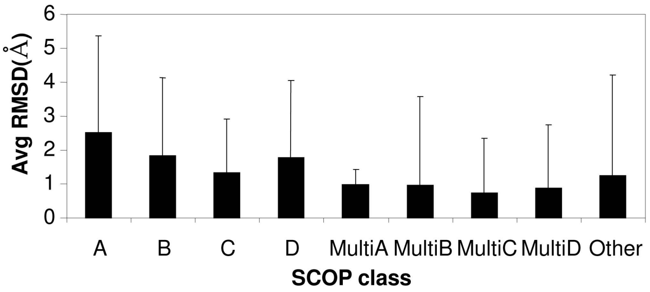


Fig. 9. Average RMSD values on the different SCOP classes as obtained using contact maps computed with a threshold of 13 Å.

with the fact that all-alpha proteins are characterized by contact maps with a great number of contacts made by sequence nearest neighbor residues, and this hampers global 3D reconstruction.

An analysis of our procedure as a function of the protein length shows that the method works independently of the protein size and that long proteins are on average reconstructed as well as short ones (Fig. 10).

4.6 Comparison with Previous Methods

To our knowledge, only four methods have been introduced so far to reconstruct the protein 3D structures starting from the contact map information [5], [10], [17], [20]. The approach developed by Vendruscolo et al. [20] was tested on some 20 proteins. Unlike our results, their findings indicate that RMSD, on the average, increases when the protein length increases.

This effect may be due to the adopted simulated annealing procedure that require more optimization steps for large than for short proteins; furthermore, they stop the search without a complete satisfaction of the contact maps ($\text{Cmap distance} = 0$). On the contrary, our method runs till the satisfaction of the contact map (Table 1).

When our method is tested on the Vendruscolo et al. set [20], it is worth noticing that even when a comparable threshold of 9 Å is used, the reconstructed RMSD is lower, on the average, than that previously obtained (Fig. 11). At higher contact map threshold values (for instance, 13 Å, as shown in Fig. 11), all the proteins of the Vendruscolo et al. set [20] are reconstructed with RMSD values lower than 2 Å and again with zero errors in the contact map. The average execution time on this set is less than 1 second.

Galaktionov and Marshall [10] report values for only five proteins, with RMSD values lower than 1 Å. In Table 2, we show that our method performs similarly on their data set (results are obtained with a contact map threshold of 13 Å).

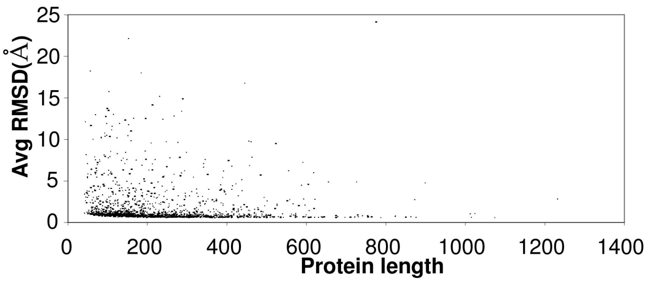


Fig. 10. Actual RMSD distribution as a function of the protein length (number of residues) when contact maps are computed with a contact threshold of 13 Å.

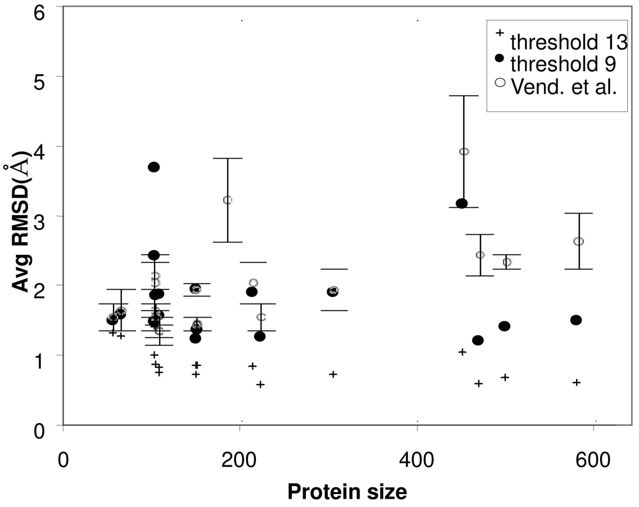


Fig. 11. Reconstruction accuracy (RMSD) of our method on the set of Vendruscolo et al. [20]. The results correspond to a contact threshold of 9 Å (for a direct comparison with [20]) and of 13 Å, respectively. The error associated with the Vendruscolo et al. reported data is due to the fact that the complete satisfaction of the contact map is not a constraint for their search.

Two other papers [5], [17] describe reconstruction procedures; however, they adopt predicted constraints or predicted contacts to fold the proteins so that a direct comparison with them is not possible.

5 Conclusions and Perspectives

In this paper, we address the problem of reconstructing protein structures from the native contact maps. We introduce the general problem (which has been shown to be computationally intractable), and we describe an efficient and very fast procedure (COMAR) to solve it. We show that contact maps computed using threshold values greater than those commonly used for $C\alpha$-$C\alpha$ distances allow better 3D structure recovery than those computed at lower thresholds (7-9 Å). This is mainly due to the fact that for some proteins (in particular but not exclusively, the all-alpha monodomain), there exist a large number of

TABLE 2
Comparison of Our Method with that of Galaktionov and Marshall [10]

PDB code	Size	Galaktionov Marshall [10] RMSD(Å)	Our method(*) RMSD(Å)	Time(s)
1rdg	52	0.66	1.08	0.01
1pcy	99	0.88	0.90	0.08
4fd1	106	0.86	0.74	0.14
1acx	108	0.96	0.83	0.13
1cpv	108	0.89	0.80	0.12
Avg		0.85	0.87	0.096

The protein set is the same of Galaktinov and Marshall [10].

(*) In this specific case we used a cut-off threshold of 13 Å; the results with other thresholds are similar.

different conformations that satisfy the native contact map. When the threshold of the contact map computation is increased, the ensemble of possible different solutions is reduced by increasing the number of structural constraints. Our finding indicates that the best cutoff threshold is in the range of 10-18 Å.

We are currently testing the robustness of COMAR in the case of noisy or incomplete contact maps. Preliminary results obtained on a subset of 120 proteins with lengths ranging between 50 and 1,100 residues show that COMAR can reconstruct 3D structures having $\mathrm{RMSD} < 4$ Å from the corresponding native structures with up to 5-10 percent random errors in the input contact map. In the case of incomplete contact maps, the same reconstruction accuracy is obtained when up to 75 percent of the input (i.e., entries of the map) is not considered. In other words, COMAR provides much better solutions if it receives as input a small number of well-predicted entries rather than a large number of predicted entries with a small number of errors.

In summary, in this paper, we show the following:

- Our method converges with high probability to a correct solution, and it is sufficiently robust to generate a solution close to the correct one also in those cases in which it is not convergent.
- Our method can reconstruct with zero contact map errors all the protein structures of our data set, and to our knowledge, this result has not been achieved before by other authors.
- The required computational time is in the range of 3-15 seconds when a normal personal computer is available, making the program a useful tool also for wide-scale applications.
- Our results are obtained on a nonredundant data set comprising 1,760 proteins, and this is the largest data set used so far for this specific task.

Finally, even if COMAR seems to tolerate many more errors in the contact map than all the others techniques proposed in the literature so far, the best available predictors are far from producing enough accurate contact maps. Part of this gap might be filled by using error filters for predicted contact maps or, equivalently, by posing more emphasis on the reliability of predicted contacts.

Appendix A

The **SPLIT** procedure splits a native contact map into submatrices in relation to those fragments of the protein that show a high degree of independence with respect to mutual interactions. In other words, we identify submatrices of the contact map such that their residues have no contacts outside the submatrix itself (Fig. 1). In searching these submatrices, we ignore contacts near the main diagonal, since each residue is in contact with the residues close to it in the protein chain. Therefore, we call *thickness* the minimum distance from the main diagonal of a contact to be considered in the splitting procedure.

Formally, we say that a contact map matrix $CM \in \{0,1\}^{n \times n}$ is *splittable* with thickness T in the two submatrices $CM_{1,j} \in \{0,1\}^{j \times j}$ and $CM_{j,n} \in \{0,1\}^{n-j+1 \times n-j+1}$ if and only if $CM[h,k] = 0 \; \forall \; h \in [1,j]$, $k \in [j,n]$, such that $|h-k| \geq T$.

Given a contact map CM, the **SPLIT** function determines if it is splittable and returns its submatrices. First, it computes the number of contacts shared by residues before and after each position in the sequence of residues (**SPLICE-CREATION**). Then, it divides CM into submatrices of size at least *AcceptedSize*, sharing no contacts with other submatrices. We have two types of submatrices:

- one having no contacts besides the ones near the main diagonal, allowed by the thickness T and denoted by a sequence of zeros in the array of shared residues V (line 7);
- one sharing no contacts with neighbor submatrices, denoted by a sequence of values preceded and followed by a zero in the array of shared residues V (line 8).

```
SPLIT($CM \in \{0,1\}^{n \times n}$)
1: $V \leftarrow$ SPLICE-CREATION($CM$)
2: $AcceptedSize \leftarrow 13$
3: $s \leftarrow 1$
4: $D \leftarrow \{\}$
5: foreach $i \leftarrow 1$ to $n$ do
6:     if ($i - s > AcceptedSize$) then
7:         if ($V[k] = 0 \; \forall \; k \in [s+1, i-1]$ and $V[s], V[i] \neq 0$) or
8:            ($V[s] = 0$ and $V[i] = 0$) then
9:             $D \leftarrow D \cup \{submatrix\ of\ CM\ from\ s\ to\ i\}$
10:            $s \leftarrow i$
11:        endif
11:    endif
12: endfor
13: return $D$
```

For each position $i \in [1,n]$, **SPLICE-CREATION** counts the number of contacts in the rectangular submatrix of CM having a lower left corner at position i on the main diagonal (line $i-1$, row $i+1$) and the same upper right corner as CM (line 0, row n). Contacts in the lower left corner of this rectangular submatrix, having position j, k such that $|j-k| \leq T$, are not considered (line 6). The *thickness* parameter T is initialized as the mean over all residues of the column of the first zero found starting from the main diagonal.

```
SPLICE_CREATION($CM \in \{0,1\}^{n \times n}$)
1: $V \leftarrow \{\}$
2: for $i \leftarrow 1$ to $n$ do
3:     $V[i] \leftarrow 0$
4:     for $j \leftarrow 1$ to $i-1$ do
5:         for $k = i+1$ to $n$ do
6:             if $|j-k| > T$
7:                 $V[i] \leftarrow V[i] + CM[j,k]$
8:             endif
9:         endfor
10:    endfor
11: endfor
12: return $V$
```

The **MERGE** procedure tries to merge coordinates $C_1, \ldots, C_k$, [each one constructed by the corresponding submatrix splitted from contact map CM (Section 3.1)] into a structure consistent with the whole contact map CM. The merging process is performed incrementally (lines 3-17), adding at each step i the set of coordinates C_i to the resulting structure C. The **TRANSLATE** procedure (line 4) translates the coordinates in C_i to superimpose the common residue between C_i and the already built structure. Then, 50 random rotations of C_i over the common residue are generated. The best rotation is selected as the one for which the contact map of the current structure has the minimum number of differences with the corresponding submatrix of the original contact map (lines 6-16). The **RANDOM** procedure generates three random numbers in the intervals specified. The **ROTATE**(C_i, $\{x, y, z\}$) function returns the rotation of the set of coordinates C_i over the three principal axes by angles $\{x, y, z\}$.

```
MERGE($C_1 \in R_1^{3\times n}, \ldots, C_k \in R_k^{3\times n}$, $CM \in \{0,1\}^{n\times n}$)
Require: $n_1 + \ldots + n_k = n$
1: $C \leftarrow \{\}$
2: $e \leftarrow \infty$
3: for $i \leftarrow 1$ to $k$ do
4:     $C_{old} \leftarrow C$
5:     $C_i \leftarrow$ TRANSLATE$(C_i, C)$
6:     for $j \leftarrow 1$ to 50 do
7:         $\{x,y,z\} \leftarrow$ RANDOM$([0,\pi],[-\pi,\pi],[\pi,\pi])$
8:         $C'_i \leftarrow$ ROTATE$(C_i, \{x,y,z\})$
9:         $C' \leftarrow$ append $C'_i$ to $C_{old}$
10:        $CM' \leftarrow$ contact map of $C'$
11:        $e' \leftarrow$ differences between $CM'$ and $CM$
12:        if $e > e'$ then
13:            $e \leftarrow e'$
14:            $C \leftarrow C'$
15:        endif
16:    endfor
17: endfor
18: return $C$
```

Appendix B

The **GUESS-DIST** procedure tries to guess a possible set of distances $D \in R^{n\times n}$ consistent with some contact map $CM \in \{0,1\}^{n\times n}$ of threshold t by using some a priori knowledge about the physical conformation of the proteins. For instance, residues that form the backbone of a protein are usually placed according to the typical distance value of 3.8 Å (the Cα-Cα distance). Other typical distance values can be obtained experimentally from the real proteins. The set of experimental typical values used by the **GUESS-DIST** procedure are collected in **DISTANCE**, which returns a random typical value for every couple of residues i and j and threshold t. The **RANDOM** procedure generates a random number in the interval specified.

```
DISTANCE($t \in N$, $i \in N$, $j \in N$)
Require: $1 \le i, j \le n$
if $i = j$ then return $0$
if $|i-j| = 1$ then return $3.8$
if $|i-j| = 2$ then return $6 +$ RANDOM$([-1.5, 1.5])$
if $|i-j| = 3$ then return $\mathrm{Max}\{0, 7.5 +$ RANDOM$([7.5-t, t-7.5])\}$
if $|i-j| > 3$ then return $(0.91 - t/100)t +$
    RANDOM$([-t + (0.91 - t/100)t, t - (0.91 - t/100)t])$
```

Any set of distances D must satisfy the triangle inequality (i.e., for all: $1 \le i, j, k \le n$, $D[i,j] \le D[i,k] + D[k,j]$) in order to be 3D consistent. To obtain from D a set of guessed distances that satisfy the triangle inequality, we run the (standard) **SHORTEST-PATH** algorithm (see, for example, [7]) on the weighted graph identified by the D matrix.

```
GUESS-DIST($CM \in \{0,1\}^{n\times n}$, $t \in N$)
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow i$ to $n$ do
        if $CM[i,j] = 1$ then
            $D[i,j] \leftarrow$ DISTANCE$(t, i, j)$
        else
            $D[i,j] \leftarrow \infty$
        endif
        $D[j,i] \leftarrow D[i,j]$
    endfor
endfor
return SHORTEST-PATH$(D)$
```

Appendix C

The **MOVE** procedure projects some $C[i]$ coordinate on the surface of the sphere of radius (of mobility) r_i and centered in $C[i]$ (see Section 3.2). The direction of the projection is described by a vectorial pseudoforce $\boldsymbol{F}$ applied to i. For every residue j not well placed with respect to i, let us consider the (vectorial) pseudoforce $\boldsymbol{F}_j = (C[i] - C[j])/d_{ij}$ of magnitude one and direction ij. The point on the surface of the sphere (line 12) is then identified by the pseudoforce $\boldsymbol{F}$ resulting from the (vectorial) addition of forces $\boldsymbol{F}'_j$, where $\boldsymbol{F}'_j = \boldsymbol{F}_j$ when $CM[i,j] = 1$, and $\boldsymbol{F}'_j = -\boldsymbol{F}_j$ has opposite direction to $\boldsymbol{F}_j$ when $CM[i,j] = 0$ (lines 2-11).

```
MOVE($CM \in \{0,1\}^{n\times n}$, $C \in R^{3\times n}$, $t \in N$, $i \in [1, n]$)
1: $r_i \leftarrow$ radius of mobility at threshold $t$ of residue $i$
2: $F \leftarrow \{0,0,0\}$
3: for $j \leftarrow 1$ to $n$ do
4:     if $j$ is not well placed with respect to $i$ then
5:         if $CM[i,j] = 1$ then
6:             $F \leftarrow F - (C[i] - C[j])/d_{ij}$
7:         else
8:             $F \leftarrow F + (C[i] - C[j])/d_{ij}$
9:         endif
10:    endif
11: endfor
12: return $C[i] + F(r_i/|F|)$
```

Acknowledgments

The authors thank MIUR for the following Grants: Projects PNR 2001-2003 (FIRB art.8) and PNR 2003 (FIRB art.8) on Bioinformatics for Genomics and Proteomics and Laboratorio Internazionale di BioInformatica (LIBI), both delivered to RC. This work was also supported by the Biosapiens Network of Excellence Project LSHG-CT-2003-503265 (a Grant of the European Union's VI Framework Programme).

REFERENCES

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.-J. Lipman, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research,* vol. 25, no. 17, pp. 3389-3402, Sept. 1997.

[2] D. Andreeva, S.E. Howorth, T.J. Brenner, C. Hubbard, and A.G. Chothia, "SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data," *Nucleic Acids Research,* database issue, vol. 32, pp. D226-D229, Jan. 2004.

[3] L. Bartoli, E. Capriotti, P. Fariselli, P.L. Martelli, and R. Casadio, "The Pros and Cons of Predicting Protein Contact Maps, in Protein Structure Prediction," *Methods and Protocols Humana Press,* in press.

[4] L.M. Blumental, *Theory and Applications of Distance Geometry.* Chelsea House Publishers, 1970.

[5] J. Bohr et al., "Protein Structures from Distance Inequalities," *J. Molecular Biology,* vol. 231, pp. 861-869, 1993.

[6] H. Breu and D.G. Kirkpatrick, "Unit Disk Graph Recognition Is NP-Hard," *Computational Geometry,* vol. 9, pp. 3-24, 1998.

[7] T. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms,* second ed. MIT Press, McGraw-Hill, 2001.

[8] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation.* John Wiley & Sons, 1988.

[9] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Progress in Predicting Inter-Residue Contacts of Proteins with Neural Networks and Correlated Mutations," *Proteins,* vol. 45, Suppl 5. pp. 157-162, 2001.

[10] S.G. Galaktionov and G.R. Marshall, "Properties of Intraglobular Contacts in Proteins: An Approach to Prediction of Tertiary Structure," *System Sciences,* Proc. 27th Hawaii Int'l Conf. Biotechnology Computing, vol. 5, nos. 4-7, pp. 326-335, Jan. 1994.

[11] B.L. de Groot, D.M.F. van Aalten, R.M. Scheek, A. Amadei, G. Vriend, and H.J.C. Berendsen, "Prediction of Protein Conformational Freedom from Distance Constraints," *Proteins,* vol. 29, pp. 240-251, 1997.

[12] T.F. Havel, *Distance Geometry: Theory, Algorithms, and Chemical Applications in the Encyclopedia of Computational Chemistry,* 1998.

[13] D.A. Hinds and M.A. Levitt, *Proc. Nat'l Academy of Sciences of the USA,* vol. 89, p. 2536, 1992.

[14] A. Lesk, *Introduction to Bioinformatics.* Oxford Univ. Press, 2006.

[15] J. Moré and Z. Wu, "[Epsilon]-Optimal Solutions to Distance Geometry Problems via Global Continuation," *Global Minimization of Non-Convex Energy Functions: Molecular Conformation and Protein Folding,* P. M. Pardalos, D. Shalloway, and G. Xue, eds., pp. 151-168, Am. Math. Soc., 1995.

[16] J. Moré and Z. Wu, "Distance Geometry Optimization for Protein Structures," *J. Global Optimization,* vol. 15, pp. 219-234, 1999.

[17] G. Pollastri, A. Vullo, P. Frasconi, and P. Baldi, "Modular DAG-RNN Architectures for Assembling Coarse Protein Structures," *J. Computational Biology,* vol. 13, no. 3, pp. 631-650, 2006.

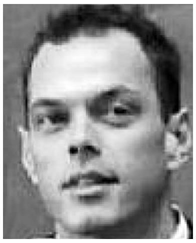
[18] J.B. Saxe, "Embeddability of Weighted Graphs in K-Space Is Strongly NP-Hard," *Proc. 17th Allerton Conf. Comm. Control, and Computing,* pp. 480-489, 1979.

[19] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio, "Reconstruction of 3D Structures from Protein Contact Maps," *Proc. Third Int'l Symp. Bioinformatics Research and Applications (ISBRA '07),* pp. 578-589, 2007.

[20] M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of Protein Structure from Contact Maps," *Folding and Design,* vol. 2, no. 5, pp. 295-306, Sept. 1997.

[21] M. Vendruscolo and E. Domany, "Protein Folding Using Contact Maps," *Vitamins and Hormones,* vol. 58, pp. 171-212, 2000.

Marco Vassura received the Laurea degree in informatica and the Dottorato di Ricerca in informatica (PhD degree in computer science) from the University of Bologna, Italy, in 2001 and 2005, respectively. He is a research assistant of computer science at the University of Bologna. His research supervisors are Professors Luciano Margara and Rita Casadio. His research interests include combinatorial optimization, optical networks, and, recently, bioinformatics.



Luciano Margara received the Laurea degree in scienze dell'informazione and the Dottorato di Ricerca in informatica (PhD degree in computer science) from the University of Pisa in 1991 and 1995, respectively. Since 1995, he has been with the University of Bologna, Italy, where he was a research associate from 1995 to 2000, was an associate professor from 2000 to 2005, and is currently a full professor of computer science. He has been a visiting scientist at the International Computer Science Institute, Berkeley, and a visiting professor in the Department of Computer Science, Cornell University. His research interests include discrete-time dynamical systems, optical networks, computational complexity, and, recently, bioinformatics.

Pietro Di Lena received the Laurea degree in informatica and the Dottorato di Ricerca in informatica (PhD degree in computer science) from the University of Bologna, Italy, in 2003 and 2007,respectively. He is a research assistant of computer science at the University of Bologna. His research supervisors are Professors Luciano Margara and Rita Casadio. His research interests include combinatorial optimization, cellular automata, and, recently, bioinformatics.

Filippo Medri received the Laurea degree in informatica from the University of Bologna, Italy, in 2003. He is currently a PhD student in computer science at the University of Bologna. His research supervisor is Professor Luciano Margara. His research interests include combinatorial optimization, probability and statistics, and bioinformatics.

Piero Fariselli has a PhD degree in biophysics and a Laurea degree in physics. He is a permanent researcher in the Biocomputing Group, University of Bologna. His main research interests include computational biology and machine learning. He is the author of more than 80 publications.

Rita Casadio is a full professor of biochemistry/bioinformatics at the University of Bologna, Italy, and is the group leader of the Biocomputing Group (www.biocomp.unibo.it). Her research interests include bioinformatics, computational biology, machine learning, and molecular theoretical biophysics. She is the author of more than 200 publications. She is president of the Bologna International Master Degree in Bioinformatics.