FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps

M. Vassura¹,*, L. Margara¹, P. Di Lena¹, F. Medri¹, P. Fariselli² and R. Casadio²
¹Department of Computer Science and ²Biocomputing group, Department of Biology, University of Bologna, Italy.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: FT-COMAR (Fault Tolerant Contact Map Reconstruction) is a heuristic algorithm for the reconstruction of the protein three-dimensional structure from (possibly) incomplete (i.e. containing unknown entries) and noisy contact maps. FT-COMAR runs within minutes, allowing its application to a large-scale number of predictions.

Availability: http://bioinformatics.cs.unibo.it/FT-COMAR
Contact: vassura@cs.unibo.it

1 INTRODUCTION

The knowledge of the protein tertiary structure may help in understanding its biological function. However, 3D protein structure prediction directly from its primary structure is still a really complex and unsolved problem (Lesk, 2006). A possible approach is the prediction of the protein contact map as an intermediate step between primary and tertiary structure. While different methods have been described for predicting protein contact maps, the problem of an efficient 3D reconstruction has been poorly addressed. A contact map of a given (known) protein 3D structure is a binary matrix M such that M_{ij} = 1 iff the Euclidean distance between residue Cα atoms i and j in the native structure is less than or equal to a pre-assigned threshold value.

The general problem of recovering a set of 3D coordinates having a specific contact map has been proved to be NP-hard (Breu et al., 1998). Other well studied similar problems are NMR structure determination (Havel, 1998; Moré and Wu, 1999) and protein conformational freedom (Groot et al., 1997). However, the different nature of distance constraints available in such settings requires the use of methods and tools which are different from the techniques needed for Cα-trace contact map reconstruction. A series of heuristic algorithms have been proposed to solve the contact map reconstruction problem (Galaktionov et al., 1994; Vendruscolo et al., 1997; Bohr et al., 1993; Pollastri et al., 2006). Most of them have been tested only on the contest of a specific predictor performance and none of them is available to the scientific community. The FT-COMAR implementation described here has at least the same reconstruction quality of the above quoted methods with a running time short enough to allow a large scale number of predictions (Vassura et al. 2007 a, b).

2 FT-COMAR

FT-COMAR implements a heuristic procedure for recovering a set of 3D coordinates from a possibly erroneous and incomplete contact map (Vassura et al. 2007 a, b; Vassura et al. 2008). The algorithm is in two phases. The first phase relies on the metric matrix embedding algorithm (Havel, 1998) to retrieve an initial set of 3D coordinates from partially randomized statistical values of inter-residues distances. The second phase of the algorithm iteratively applies a correction/perturbation procedure to the randomly generated set of coordinates. Correction applies to each residue a pseudo-force derived from the input contact map, while perturbation slightly moves each residue coordinate in order to relax the current computed structure, allowing a better correction. This is performed in order to obtain a new set of coordinates as consistent as possible with the given contact map (Vassura et al. 2007 a, b; Vassura et al., 2008). The refinement applies until the set of coordinates is consistent with the given contact map or until some control parameter reaches the stop condition. The current version of FT-COMAR applies a correction-perturbation procedure to ensure that the distances of each Cα atoms preserve the protein constraints. Since a set of 3D coordinates consistent with an erroneous contact map may not exist, the quality of the solution found by FT-COMAR is highly influenced by the degree of correctness of the input contact map. To improve the reconstruction it can be useful to mark as unknown some (possibly) highly unsafe areas of the input contact map. FT-COMAR does not consider the coordinate points related to unknown entries. Ignoring faulty regions during the refinement phase it eventually avoids the propagation of errors, increasing reconstruction quality.

The FT-COMAR implementation takes as input an ASCII file containing a symmetric contact map matrix whose possible entries are 0 (non contact), 1 (contact) and -1 (unknown entry) or a list of contacts in the widely used EVAc on format (Graña et al., 2005). Such contact map may be computed within any threshold value, even if previous results show that higher thresholds allow the contact map to carry more information (Vassura et al., 2007; Vassura et al., 2008).

Marking unsafe areas as unknown can greatly improve reconstruction quality when the contact map is noisy. For such reason, a simple filtering procedure is provided: it can be used to detect possibly unsafe entries of the input contact map. This filtering procedure (Vassura et al. 2007 b) is based on the common neighbors property of contact maps at threshold 12 Å (so it has better performances when the input contact map threshold is 12 Å). FT-COMAR has been tested on physical contact maps, obtaining structures with RMSD from the native ones on average less than 4 Å, under the following conditions:

- contact thresholds ranging from 7 to 18 Å;
- protein chains ranging from 55 to 786 residue long;

*To whom correspondence should be addressed.
3 RESULTS

The current version of FT-COMAR has been tested on a set of 100 nonredundant mono domain protein chains evenly distributed among the 4 main structural classes: all Alpha, all Beta, Alpha+Beta and Alpha/Beta. The reconstruction performances have been compared in terms of RMSD by introducing on the native contact maps three different types of random errors: generic errors, missing contacts and false contacts. Our tests show that in presence of errors the reconstruction quality decreases with the length of the protein and that FT-COMAR largely tolerates missing contacts. In particular, the experimental results show that the reconstruction quality of contact maps with 50% missing contacts is comparable to the reconstruction quality of contact maps with 15% false contacts. FT-COMAR is much more tolerant to under prediction than to over prediction of contacts. Moreover, FT-COMAR can ignore up to 75% of the contact map and still compute a protein 3D structure whose native RMSD from the native one is less than 4 Å (assuming that the remaining 25% contains no errors). Furthermore, in this last case, the reconstruction quality is independent from the protein length (Vassura et al. 2007 b). This suggests that, to improve protein reconstruction from contact maps, contact map prediction should put much more emphasis on the quantity more than on the quality of the prediction. In all our tests, FT-COMAR runs within minutes, allowing its adoption for a large-scale number of predictions. Figures showing results for all experiments updated to the current version of FT-COMAR can be found as supplementary material at http://bioinformatics.cs.unibo.it/FT-COMAR/index.html/exp. A direct comparison with other results has been previously described (Vassura et al., 2007 a, b; Vassura et al., 2008), from which it appears that FT-COMAR outperforms the other methods.

As a blind test for FT-COMAR we evaluate its reconstruction performances on predicted contact maps for the test set of 100 proteins described above. The performances are predicted with CORNET (Fariselli et al., 2001). When the number of predicted contacts is set equal to the protein length/5, on the set of 100 proteins, CORNET average accuracy (sequence separation >24 residues) is 16.9%. This performance is comparable with those assessed in CASP7, where the mean across all groups was 13% (Izarzugaza et al. 2007). We tested FT-COMAR on these predicted contact maps performing 10 reconstructions per map without using the filtering procedure. The mean RMSD values obtained are shown in Fig. 1 as a function of the whole accuracy of the map (sequence separation >6 residues). Not surprisingly, the quality of the reconstructed structure increases at increasing accuracy values. Although we never reached RMSD < 5 Å (the lowest average RMSD equals 7.4 Å), we note that, even in presence of extremely noisy and erroneous contact maps, many reconstructed structures have average RMSD lower than 10 Å.

ACKNOWLEDGEMENTS

We thank MIUR for the following grants: PNR 2001-2003 (FIRB art. and PNR 2003 projects (FIRB art. on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica, both delivered to RC. This work was also supported by the Biosapiens Network of Excellence project no LSHG-CT-2003-503265 (a grant of the European Unions VI Framework Programme).

REFERENCES


Fig. 1. Average RMSD (Å) between reconstructed and native protein 3D structures at increasing percentages of accuracy. The line represents the linear least square fitting.