Using Constraints on Beta Partners to Reconstruct Mainly Beta Proteins

Alessio Ceroni(1) , Paolo Frasconi (2)

Dipartimento Sistemi e Informatica, Università degli Studi di Firenze, via S. Marta, 3, 50139 Firenze -Italy
1)aceroni@dsi.unifi.it, 2)paolo@dsi.unifi.it

Keywords: Structural Genomics, Protein Structure Reconstruction, Beta Partners

Introduction

The knowledge of the spatial conformation of a protein can help the study of its function, but the number of resolved structures is still limited by the low throughput of the methods used. Structure prediction could bridge the sequence-structure gap, but no reliable and general methods have yet been proposed. An attempt to simplify the problem has been made by trying to predict the contact map of a protein instead of its atoms positions [1]. It has been demonstrated the protein structure can be reconstructed with sufficient precision even if the contact map contains error [2]. Unfortunately, the prediction of contact maps is still very unreliable and it is not clear whether the type of errors made by the predictor can be corrected by the reconstruction method.
A low-detail representation of the protein conformation could extract the relevant information to train more efficient predictors. The coarse-grain contact map is defined using contacts between secondary structure segments. The prediction of this type of contacts has been tried [3], but no results exists about the feasibility of a reliable method that uses only this type of information to reconstruct the protein structure.
In this work we concentrate on contacts defined by beta partners. The geometry and connectivity of beta strands imposes strong constraints on the overall structure of the protein, especially for those chains thar are formed mainly by residues in beta conformation. The reconstruction of the structure of this kind of proteins would be enhanced by the knowledge of the secondary structure and the indication of which strands are partners. We propose here an efficient procedure to find a structure that matches the aforementioned characteristics of a given protein in its native conformation.

Methods

The reconstruction procedure performs a minimization of the energy of a protein model. The knowledge about secondary structure and beta partners in the native conformation is enforced as a set of constraints on the solution. We used a model comprising all the heavy backbone atoms plus a single atom for the C_{β} to represent the side chain occupation. The free parameters of this model are the dihedral φ and φ angles. The ω angle is set fixed to 180°, bond lengths and angles are set to their average values calculated on the whole PDB dataset, the same for the coordinates of the C_{β} atom in the reference system defined by the N and Cα atoms.
Constraints on secondary structure are imposed on the values of the dihedral angles. Alpha helices (H) and beta strands (E) corresponds to two compact regions in the √□√ plot. For every residue in the H and E classes, the distance between its coordinates in the √□√ space and the center of the corresponding region is forced to be lower than a specified threshold.
For each pair of beta strands we know if they are partners and in this case if they are parallel or anti-parallel. The geometry of two beta partners forces the hydrogen bonded residues to stay at a specific distance. Unfortunately, two partner strands can be of different dimensions and we did not want to specify the partnership in terms of connectivity between residues. Therefore, the procedure tests all the possible alignments for each pair of strands:

- for partner strands, given a particular alignment, the distance between every pair of (supposedely) bonded atoms must be in a strict range of values; the alignment that violates less constraints contributes to the solution: this force the existence of at least one good alignment between partners;
- for non-partner strands both orientations are tested; given a particular alignment, the distances between paired

atoms must be greater than a specified value; the alignment that violates more constraints contributes to the solution: we demand that no good alignments exist between non-partners.

Atomic forces impose a lower bound on the distance between two atoms, thus defining an excluded volume for each atom that prevents the protein to collapse in a single point. We introduced these constraints in our procedure by forcing the distance between all pairs of atoms to be higher than a specified threshold.

To simplify the optimization task we decided to transform all the constraints in quadratic penalty terms. Unfortunately, this translate in an highly non linear function of the model free parameters. Globally optimizing a non-linear cost function is generally a difficult task. However, we decided to implement a simple approach consisting of a quasi-newton local optimization procedure (LBFGS [4]) coupled with a multistart strategy.

Results

The experiments have been performed using a representative set of non homologous chains from the Protein Data Bank (PDBSelect, december 2002 release). From this set we retained only high quality proteins without any physical chain breaks. We determined the secondary structure class for the remaining chains using the procedure described in [5], the same used to build the CATH database. The final dataset contained only "mainly beta" proteins, a total of 154 chains with lenghts between 30 and 300 residues.

To demonstrate that our reconstruction procedure can satisfy all the given constraints, we measured its accuracy as the proportion of pairs of beta strands correctly assigned as partners or non-partners. The average value obtained is 98.5%, with a 74% of proteins with all the beta partners correctly assigned. We verified the quality of the reconstructed protein structures using the measure adopted in the CASP contest [6]. We obtained an average GDT_TS of 29.7, a value that is comparable to the performances of good predictors in CASP5 but far from the optimum value of 100 of a perfect reconstruction. We think that this gap will be partially filled by the introduction of the remaining types of segment contacts in our reconstruction procedure.

Acknowledgements

We thanks Bernardetta Addis, Dipartimento di Sistemi e Informatica of the Universitàà degli Studi di Firenze, for the implementation of the LBFGS algorithm.

References

[1] G. Pollastri and P. Baldi, "Prediction of Contact Maps by Recurrent Neural Network Architectures and Hidden Context Propagation From All Four Cardinal Corners", Bioinformatics, vol. 1, pp 1—9, 2002.

[2] M. Vendruscolo, E. Kussel and E. Domany, "Recovery of Protein Structure from Contact Maps". Folding and Design, vol. 2, pp 295—306, 1997.

[3] A. Vullo and P. Frasconi, "Prediction of Protein Coarse Contact Maps". Journal of Bioinformatics and Computational Biology, vol. 1, pp. 411—431, 2003.

[4] D. C. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale". Mathematical Programming, vol. 45, pp. 503—528, 1989.

[5] A. D. Michie, C. A. Orengo and J. M. Thornton, "Analysis of Domain Structural Class Using an Automated Class Assignment Protocol". Journal of Molecular Biology, vol, 262, pp 178—185, 1996.

[6] A. Zemla, C. Venclovas, J. Moult and K. Fidelis, "Processing and evaluation of predictons in CASP4". Proteins, suppl. 5, pp 13—21, 2001.