Improving the selection of close-native protein structures in decoy sets using a graph theory-based approach

Casadio R (1), Fariselli P (1), Margara L (2), Filippo M (2), Vassura M (2)

(1) Biocomputing Group Department of Biology , University of Bologna.
(2) Computer Science Department, University of Bologna.

Motivation

One of the still unsolved problems in the ab initio protein structure prediction is the ability of distinguishing from near-native and distant-native protein structures. Indeed ab initio methods generate several structures, often spanning all the known structural types and in the absence of the known real solution (the real protein structure) it is very difficult to select the most likely fold/s of a given chain. Due to the intrinsic errors of the force fields presently available, several filtering procedure have been therefore developed and implemented, including combinations of different energy functions (1). The problem at hand can be addressed only when for a given protein, a good decoy set is also available (1). This is necessary in order to test the discriminative ability of the different methods

Methods

Here we take a new approach introducing a graph representation of each protein and associated decoys. The data set of selected proteins and decoys is somewhat modified from that computed in a previous published work (1). In (1) good decoys were computed and made available following a very stringent definition of "good" decoy set. Decoys were produced with the Rosetta method for a large set of proteins, following four criteria: 1) contain conformations for a wide variety of different proteins; 2) contain conformations close (<0.4 nm) to the native structures; 3) consist of conformations that are at least near local minima of a reasonable scoring function; 4) be produced by a relatively unbiased procedure. We consider a set of 41 proteins with 1800 decoys per protein, for a total of about 76000 protein structures. Given the data set of protein and decoy structures, we consider fifteen properties taken from graph theory and we test their ability to distinguish correct from incorrect folds. Our method stems from the notion that according to each property decoys can be ranked with respect to the corresponding protein structure as a function of structural similarity. In our approach each protein and related decoys are represented with a graph adjacent matrix or "contact map". For each decoy set (or set of decoy contact maps per each protein) we computed the number of edges (number of contacts), average degree (average number of contacts per residue), contact order, diameter, complexity, flow, connectivity, and also the variances and weighted versions of each property, respectively. For each property and for a given decoy set, we evaluated the Enrichment measure introduced by Tsai et al. (1) and the Z score. The ability of a given graph property to act as a scoring function is then evaluated by computing the Enrichment measure and Z score (1). More specifically, to compute the Enrichment due to a given property, we count how many of the best decoys are among the best decoys according to Ca-RMSD (Root Mean Square Deviation of the C alpha backbone of the decoy to that of the corresponding protein) and compare this number with what would be expected for a uniform distribution. Due to redundancy of decoy structures, it may occur that very similar structures are present in different decoy sets. To avoid this overlapping in the final evaluation, we keep into consideration only those decoys found in the intersection set between the top high scoring 15% decoys, as obtained according to the property at hand, and the top 15% decoys with the lowest Ca-RMSD for a given protein. This figure is then divided by the number of random assignment (15%*15%. number in the set) to highlight the performance of the method. Values greater than one indicate an Enrichment over a uniform distribution.

Results

We evaluate the Enrichment and Z-score due to each selected graph functions and we obtain that according to some of these properties (such as connectivity, network flow, and particularly complexity) the selection of optimal decoys outperforms previously described methods based on a combination of all-atom energy functions (1). With complexity the score is 1.4 times better than with other methods. We suggest that our approach can be applied when in "ab initio" predictions a set of "good structures" need to be selected among a population of predicted structures. This method can therefore complement the classical approach of energy-based scoring functions and help in solving the protein folding problem.

Contact email: casadio@alma.unibo.it

References

1. J Tsai, R Bonneau, AV Morozov, B Kuhlman, CA Rohl, and D Baker, "An improved protein decoy set for testing energy functions for protein structure prediction." Prot Struct Func Genetics, 2003