BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 14 2005, pages 3179–3180
doi:10.1093/bioinformatics/bti450

Structural bioinformatics

HCPM—program for hierarchical clustering of protein models

Dominik Gront* and Andrzej Kolinski

Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

Received on February 4, 2005; revised on April 6, 2005; accepted on April 12, 2005
Advance Access publication April 19, 2005

ABSTRACT

Summary: HCPM is a tool for clustering protein structures from comparative modeling, *ab initio* structure prediction, etc. A hierarchical clustering algorithm is designed and tested, and a heuristic is provided for an optimal cluster selection. The method has been successfully tested during the CASP6 experiment.

Availability: HCPM program can be downloaded from http://www.biocomp.chem.uw.edu.pl/HCPM/

Contact: dgront@chem.uw.edu.pl

INTRODUCTION

Recently several successful approaches to protein structure prediction have been proposed. Typically, a large number of decoys is generated and scored according to an energy function or a MQAP. CABS (Kolinski, 2004) is a high-resolution lattice model employing Monte Carlo dynamics with a large set of conformational micro-modifications. In Rosetta approach (Simons *et al*., 1997, 1999), protein conformations are built using short fragments extracted from Protein Data Bank (PDB). Correct identification of native-like structure still remains a challenging task. Force fields used in protein simulations usually reflect only energy terms. The idea behind clustering is to take also the entropic effects into account (Shortle *et al*., 1998).

Clustering procedures combine elements into groups according to a defined distance measure. Clustering approaches fall into two main categories (Jain *et al*., 1999): hierarchical and partitional.

In the hierarchical clustering a binary tree is created (Fig. 1A). At the beginning each structure forms a separate cluster. In a single step of an iterative procedure, a pair of closest clusters is identified, using distance (drmsd) or coordinate root-mean square deviation (crmsd) as a measure, and merged into a new cluster represented as a vertex of a binary tree. Distance between two merged clusters, the 'merging distance' is denoted as **r**. With the progress of the clustering, the average size of clusters increases and the number of clusters decreases. At the end, only a single cluster remains—the topmost vertex (root) of the tree. Typically, that final cluster, containing all the structures, is not the best answer. The procedure must be stopped at a threshold value of **r**, say $\mathbf{r_C}$, provided by the user. Each vertex located below the $\mathbf{r_C}$ value, with its parent located above the $\mathbf{r_C}$, is a root of a sub-tree corresponding to a cluster. Proper selection of the $\mathbf{r_C}$ value is crucial. Too high $\mathbf{r_C}$ produces big clusters, containing noise structures. With too small $\mathbf{r_C}$, not all the neighbors are combined. This is not so dangerous since the putative near-native structures are clustered at early stages of the process, although very small clusters are prone to statistical errors. A set of PDB files (for instance, decoys generated by Rosetta) for CA traces (or CABS trajectory) is the input for the HCPM program.

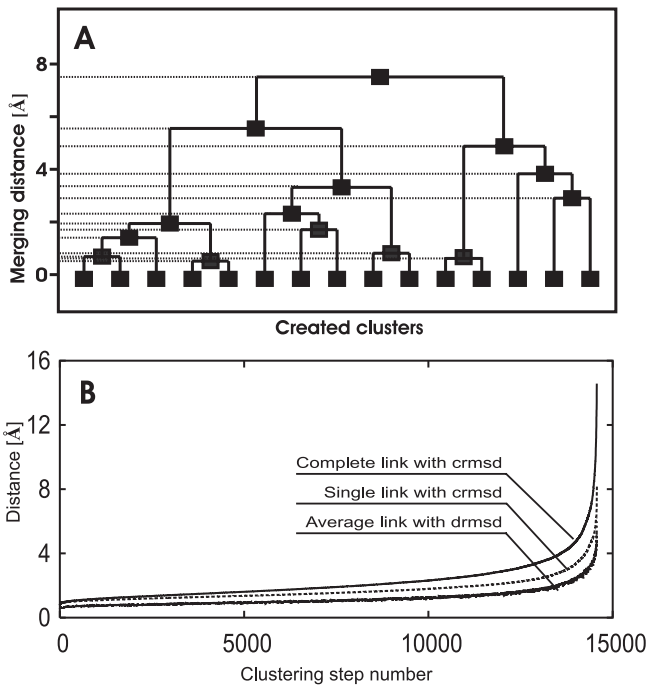

Fig. 1. An example of the clustering tree (**A**), and the plot of merging distance against the number of cycle of the clustering (**B**).

In the partitional algorithms each element has to be assigned to a certain, predefined, cluster. K-means, is the algorithm belonging to this group (Betancourt and Skolnick, 2001). K-means is faster than hierarchical clustering [$O(N^2)$ versus $O(N^3)$] however, it is less accurate. Moreover, the user must predefine the number of clusters. The cost of HCPM is 7 h of a PC unit for 20 500 decoys of a protein composed of 56 amino acids.

DETAILS OF IMPLEMENTATION

There are various ways to measure the distance between clusters. HCPM uses the average link measure. The distance between two clusters is defined as the drmsd (or crmsd in cases of very diverse structures) between maps of the average distances between atoms. The map is calculated by averaging corresponding interatomic distances in a cluster. A structure, which is the closest to the average distance map is treated as the cluster representative.

*To whom correspondence should be addressed.

Proper selection of the $\mathbf{r_C}$ depends on the origin of decoys. In comparative modeling conformations are very similar to each other, thus **r** values are small and the number of clusters is small. In *ab initio* simulations many different topologies have to be sampled in order to find the native fold. Thus, the **r** values have to be much higher. The clustering stops at a specified cut-off value of the merging distance. Alternatively, HCPM examines several probe values of **r**, specified by the user. Cluster's parameters for various probes can be used as criteria for an optimal selection of the cut-off.

The clustering should finish at the center of the plateau region of the sigmoid plot of the merging distance against the number of the clustering cycles (Fig. 1B). This point remains unknown until the whole clustering tree is created. The final clusters, for the selected merging distance, are recalculated afterwards. Clustering tree is stored in a file. Thus, the user can recalculate clusters for an arbitrary value of $\mathbf{r_C}$.

Each cluster emerges from a sub-tree, with a main branch and shorter side branches. The main branch is the branch with the shortest average length of the side branches. This value provides an additional criterion for the selection of the optimal cluster.

ACKNOWLEDGEMENT

This work was partially supported by the Grant no. PBZ-KBN-088/P04/2003.

REFERENCES

Betancourt,M.R. and Skolnick,J. (2001) Finding the needle in a haystack: educing native folds from ambiguous *ab initio* protein structure predictions. *J. Comp. Chem.*, **22**, 339–353.

Jain,A.K. *et al*. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.

Shortle,D. *et al*. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.

Simons,K.T. *et al*. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.

Simons,K.T. *et al*. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence- independent features of proteins. *Proteins*, **34**, 82–95.