Contact-Based Protein Structure Prediction in Tp, Tc, Ts and Tx Categories by MULTICOM-CLUSTER Server

Badri Adhikari, Jianlin Cheng*

Department of Computer Science, University of Missouri, Columbia, MO 65211 USA

chengji@missouri.edu

Four contact-based prediction categories were introduced in CASP11 experiment: Tp (predicted contacts), Tc (correct predicted contacts), Ts (sparse experimental data) and Tx (cross-link assisted contacts). For each category, MULTICOM-CLUSTER used a different strategy to pre-process and select contacts, and use used contacts with Rosetta[1] to build 3D models.

Methods

Target sequences in Tp category were provided with predicted L/5 (1/5 times sequence length) contacts from 10 different contact prediction servers. MULTICOM-CLUSTER considered a contact predicted by a server as an outlier if it did not have a similar contact (contacts within a residue shift of ±2 residues) in any other server predictions. After removing outlier contacts from all the contact predictions, MULTICOM-CLUSTER combined all contacts such that their original ranking by prediction confidence was preserved. For firstly released 2 targets (Tp761 and Tp763) MULTICOM-CLUSTER experimented by building models with top 2L, L, 3L/4, L/2, and L/4 contacts and for others it always selected top L/2 contacts. MULTICOM-CLUSTER translated each contact as a bounded restraint between 3.5Å to 8.0Å and built 50 models using Rosetta[1] (configured to use PSIPRED[2] for secondary structure prediction) with contact restraints weight as 1. It selected top 5 models after ranking them using their Rosetta generated energy score.

In the sparse experimental data files released for targets in Ts category, for each NOESY peak one or more distance restraints were provided. Many of the contacts were at less than 6 residue separation. For the first two targets released (Ts761 and Ts763) MULTICOM-CLUSTER experimented by selecting distance restraints with confidence greater than 0.9 and 1.0, and also tried building models with all the restraints given. For all other targets, MULTICOM-CLUSTER ignored all NOESY peaks that have more than 1 restraint, in order to avoid ambiguous restraints. With the stringent selection strategy of selecting much fewer contacts MULTICOM-CLUSTER obtained a small but likely correct contacts set. For restraints that involved hydrogen atoms in the side chains, it replaced the hydrogen atoms with CB atoms and increased the distance threshold by 1.5A. Assuming all contacts to be true MULTICOM-CLUSTER built models using Rosetta with contact restraint weight as 50. We observed that the pool of just 50 had many models that satisfied all the restraints supplied. It also used the contacts released in Tc category, if the target was already released in the Tc category as well.

For the targets in Tc and Tx categories, MULTICOM-CLUSTER simply used all the contacts as restraint with contact restraints weight as 50. Similar to targets in Ts category, for targets in Tc category it used the contacts from the Ts category as well, if the target was already released in the Ts category.

1. Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. Methods in enzymology, 383, 66-93.
2. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 292(2), 195-202.