MULTICOM-NOVEL (Tpscx)

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

Badri Adhikari, Jianlin Cheng*

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

chengji@missouri.edu

Our server MULTICOM-NOVEL participated in four contact-based prediction categories in CASP11 experiment, including Tp (predicted contacts) category, Tc (correct predicted contacts) category, Ts (sparse experimental data) category and Tx (cross-link assisted contacts) category. For each category, it used a different strategy to pre-process and select contacts, and then build 3D models using contacts and secondary structure restraints as input for a distance geometry simulated annealing protocol.

Methods

For targets in all categories, MULTICOM-NOVEL used PSpro[1] and PSIPRED[2] to predict 3-class secondary structure and built models using distance geometry simulated annealing protocol implemented in Crystallography & NMR System[3,4]. For each model building task, MULTICOM-CLUSTER generated a fully extended structure from the input sequence, prepared contact restraints, secondary structure dihedral and distance restraints, and then used the distance geometry simulated annealing protocol implemented in the “dg_sa.inp” script to generate 20 structure models, as shown in Fig 1B. To obtain secondary structure restraints, as shown in Fig 1A, it translated helix and strand predictions to ideal dihedral angle restraints and added atomic distance restraints as discussed in[5].

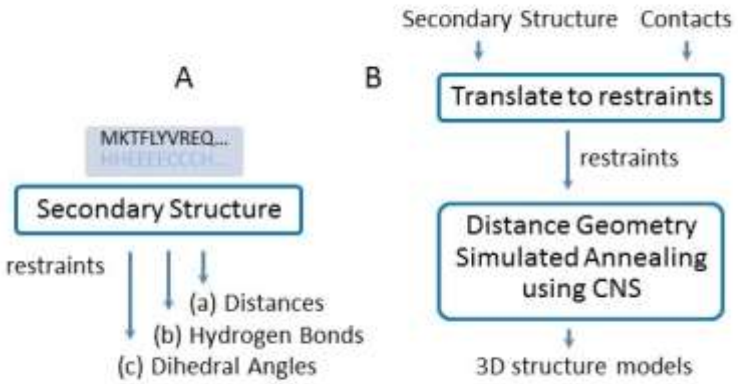

Figure 1. (**A**) Translation of secondary structure prediction to distance restraints, hydrogen bonding, and ideal dihedral angles. (**B**) The overall structure modelling pipeline.

Target sequences in Tp category were provided with predicted L/5 (1/5 times sequence length) contacts from 10 different contact prediction servers. MULTICOM-NOVEL 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, it combined all contacts such that their original ranking by prediction confidence was preserved. From the top L contacts in this list, it built 30 sets of contacts (considering the availability of processors) each consisting of randomly chosen 75% contacts. It built models for each contact set and best models built from each contact set were selected for ranking based on contact energy (how well the input contacts were satisfied).

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. In order to avoid ambiguous restraints MULTICOM-NOVEL ignored all NOESY peaks having more than 1 restraint. With the stringent selection strategy of selecting much fewer contacts it obtained a small but correct contacts set. To build models it added dihedral angle restraints translated from secondary structure predictions (with a lower weight compared to the weight for true contacts) along with the true contacts so that true contacts were satisfied even when the secondary structure predictions were not fully correct. MULTICOM-NOVEL followed the same method for targets in Tc 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, and vice versa for targets in Ts category.

Realizing that the contacts provided for targets in Tx category contained some false positives, MULTICOM-NOVEL built 20 sets of contacts by selecting L contacts randomly for each set. It built models using the same technique as that used for Ts and Tc targets, and ranked models by their contact energy for further selection.

Since the contact assisted categories were first of its kinds, our method for selecting contacts and building models evolved as the CASP11 experiment proceeded. The methods described above were the final methods applied to later targets. With the targets released earlier, we tested many techniques for selecting and combining contacts, building models, adding secondary structure restraints, and selecting predicted models. In Tp category, we tested aiding the top selected 0.5L long-range targets with 0.5L short- and medium-range contacts predicted by DNcon[6]. We built models using top 0.25L, 0.5L and L contacts, and observed many conflicting contacts. For targets released later, we used only top .25L contacts. In the Ts category, for the first two targets released we experimented by selecting all distance restraints with confidence greater than 0.9 and 1.0. For targets in Tc and Ts categories, we experimented building models with subsets of regular secondary structure elements to identify and use only the secondary structure elements that did not conflict with the true contacts. Models for the earlier targets in Tx category were built using all the contacts supplied until we realized that there were some false positives as well. For selecting models we experimented by ranking models based on total energy of the models, by only contact energy, and also by counting the secondary structure elements that matched the input. Ranking based on contact energy was the final technique we chose for most of the targets released later.

1. Cheng, J., Li, J., Wang, Z., Eickholt, J., Deng, X. (2012). The MULTICOM toolbox for protein structure prediction. BMC Bioinformatics, 13, 65.
2. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 292(2), 195-202.
3. Brunger, Axel T., et al. "Crystallography & NMR system: a new software suite for macromolecular structure determination." Acta Crystallographica Section D: Biological Crystallography 54.5 (1998): 905-921.
4. Adhikari, A., Bhattacharya, D., Deng, X., Li, J., Cheng, J. (2013). A contact assisted aporach to protein structure prediction and its assessments in CASP10. The workshop on artificial

intelligence and robotics methods in computatinal biology of 27th AAAI Conference, Bellevue, WA, USA.

5. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PloS one, 6(12), e28766.
6. Eickholt, J., & Cheng, J. (2012). Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics, 28(23), 3066-3072.