Residue-Residue Contact Prediction by MULTICOM Servers

Badri Adhikari, Jianlin Cheng*

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

chengji@missouri.edu

To predict residue-residue contacts, we used our sequence-based machine learning methods NNcon[1], SVMcon[2], and DNcon[3] participating as MULTICOM-NOVEL server group, MULTICOM-CONSTRUCT server group and MULTICOM-CLUSTER server group respectively.

Methods

DNcon is a sequence-based residue-residue contact prediction tool built using deep networks and boosting techniques empowered by GPUs and CUDA parallel computing technology. It was trained separately for predicting medium/long range contacts and short-range contacts. The data set used for training and testing consisted of 1426 proteins of which 1230 were used for training and 196 for testing. For predicting medium/long range contacts, multiple ensembles of deep networks were trained using several pairwise potentials, global features and values characterizing the sequence between contact pairs. For making short range predictions one ensemble of deep networks were trained on fixed window size of 12 residues. DNcon was tested rigorously by comparing its performance with the best sequence based contact predictors in CASP9 experiment including SVMcon.

NNcon, an ab initio contact prediction server and tool, predicts general residue contacts and beta-residue contacts in beta sheets using 2D-Recursive Neural Network models. For general contact prediction at thresholds of 8A and 12A, an ensemble of 10 models were trained using a data set of 482 proteins and validated on a data set of 48 proteins. For predicting inter-strand residue contacts, an ensemble of 10 models were trained and validated using 10-fold cross validation on a data set consisting of 916 chains and 2533 beta sheets. NNcon was evaluated by comparing it with SVMcon and other contact predictors in the CASP8 experiment using all 116 targets and also using the 11 ab initio targets.

SVMcon is sequence based medium- and long-range contact map predictor built using support vector machines. For training and testing it uses 5 categories of features: local window features, pairwise information features, residue type features, central segment window features, and protein information features. This includes features like sequence profiles, secondary structure, relative solvent accessibility, mutual information of sequence profiles, polar/non-polar/acidic/basic residue types, sequence separation length, etc. The models were trained using radial basis function (RBF) kernel. The data set used to train the system had 220,994 negative (residues not in contact) examples and 94,110 positive (residues in contact) examples. SVMcon was benchmarked by comparing it with other predictors participating in the CASP7 experiment.

Availability

NNcon, SVMcon, and DNcon are available as web service and/or software at http://sysbio.rnet.missouri.edu/multicom_toolbox/.

1. Tegge, A. N., Wang, Z., Eickholt, J., & Cheng, J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic acids research, 37(suppl 2), W515-W518.
2. Cheng, J., & Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. BMC bioinformatics, 8(1), 113.
3. Eickholt, J., & Cheng, J. (2012). Predicting protein residue‐residue contacts using deep networks and boosting. Bioinformatics, 28(23), 3066-3072.