

MULTICOM-NOVEL, MULTICOM-CONSTRUCT, MULTICOM-CLUSTER (RR)

Machine Learning, Coevolution-Based and Hybrid Methods for Contact Prediction

B. Adhikari¹, J. Cheng¹

¹ - Department of Computer Science, University of Missouri, Columbia
chengji@missouri.edu

We participated in the contact prediction (RR) category using three methods. Contact predictions by our sequence-based deep learning contact predictor DNcon¹ were submitted as MULTICOM-NOVEL. For predicting contacts using coevolution-based approach, we implemented an in-house method for generating alignments followed by running a coevolution based method to predict contacts and submitted as MULTICOM-CONSTRUCT. Finally, we use a novel contact combination approach to combine the two predictions and submitted as MULTICOM-CLUSTER.

Methods

Our MULTICOM-CONSTRUCT contact predictor relies on our new alignment generation algorithm to predict contacts with MetaPSICOV². For coevolution-based contact prediction tools like MetaPSICOV, coming up with the right size of alignment file is crucial for efficiency. We develop an alignment generation method that produces at least some sequences whenever possible even if the quality is not high, and, on the other hand, not have too many sequences for faster execution even when there are many homologous sequences available. For alignment generation, we run JackHMMER when the number of alignments produced by HHblits³ is less than $2.5L$, L being the length of the input sequence. We observed that a range of e-values are required for running JackHMMER⁴ because for some input protein sequences, stringent e-value criteria like e-40 produces too few sequences (just a hundred or so) while much lesser stringent criteria of e-4 produces too many sequences (25K, 50K, etc.). Our algorithm for generating alignments is summarized below:

```
T = 2.5
alsize() computes the ratio of the number of sequences in alignment and L
coverage = 75, 68, 60
for each C in coverage:
    run HHblits (with -n 3 -maxfilt 500000 -diff inf -e 0.001 -id 99 -cov C) and get
    alignment hhbc
    e-value = 40, 30, 20, 10, 4, 0
    for each e in e-value:
        run JackHMMER (with -N 5 and -E e) and get alignment jhe
    # select the right alignment
    alignment = hhbcov75, hhbcov60, jhe40, jhe30, jhe20, jhe10, jhe4, jhe0
    for each aln in alignment (in that order):
        if alnsize(aln) > T:
            accept aln and quit
accept jhe0
```

Our MULTICOM-CLUSTER method is a meta-predictor that combines contacts predicted by the two other methods – CONSTRUCT and NOVEL, whenever the selected alignment for CONSTRUCT method is less than 50 in size. Otherwise, the predicted contact confidence values of the MULTICOM-CONSTRUCT predictions are updated and submitted as MULTICOM-CLUSTER predictions. As the first step for contact combination, we select top 5L contacts from each of the two methods, and replace the confidence values with integer numbers, starting from 5L and ending at 1 for most confident contact prediction to the least confident one. Then, the confidence scores for the MULTICOM-CLUSTER predicted top 5L contacts is calculated as the normalized mean of the confidence values from each method, i.e. $MCL_{ij} = (MCO_{ij} + MNO_{ij})/5L$, where MCO_{ij} is the confidence of contact pair (i,j) predicted by MULTICOM-CONSTRUCT method and MNO_{ij} is the confidence of contact pair (i,j) predicted by MULTICOM-NOVEL. The final confidence values are normalized such that top L predicted contacts have confidence values more than 0.5. **Table 1** presents the summary of our preliminary evaluation of our predictions on targets for which the best predicted model in the CASP released models has less than 0.5 TM-score, assuming the targets to be potentially free-modeling targets. MULTICOM-CONSTRUCT and MULTICOM-CLUSTER methods have similar performance, which is higher than MULTICOM-NOVEL.

Table 1. Precision of contacts predicted by MULTICOM-NOVEL (MNO), MULTICOM-CONSTRUCT (MCO), and MULTICOM-CLUSTER (MCL) methods based on our preliminary evaluations. N is the number of sequences in the multiple sequence alignment and N_{eff} is the effective number of sequences for the same alignment.

Target	L	N	N_{eff}	Top L/10 Contacts			Top L/5 Contacts		
				MCL	MCO	MNO	MCL	MCO	MNO
T0859	133	2	1	0.0	0.0	0.0	0.0	0.0	0.0
T0862	239	163	23	33.3	33.3	33.3	27.8	27.8	27.8
T0863	670	453	44	1.7	1.7	5.2	1.7	1.7	4.3
T0864	246	526	134	68.2	68.2	40.9	65.9	65.9	36.4
T0869	120	17	12	60.0	60.0	60.0	52.4	42.9	47.6
T0870	138	137	69	25.0	25.0	50.0	16.7	16.7	37.5
T0904	341	23741	147	75.9	75.9	37.9	69.0	69.0	27.6
Avg	270	3577	61	37.7	37.7	32.5	33.3	32.0	25.9

Availability

DNcon is available for download at http://sysbio.rnet.missouri.edu/multicom_toolbox/tools.html.

1. Eickholt, J. and J. Cheng, Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics*, 2012. 28(23): p. 3066-3072.
2. Jones, D.T., et al., MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 2014: p. btu791.
3. Eddy, S.R., Profile hidden Markov models. *Bioinformatics*, 1998. 14(9): p. 755-63.
4. Johnson, L.S., S.R. Eddy, and E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 2010. 11: p. 431.