Report

Linh Dang

15.02.15

Purpose

Assessment of *Direct Coupling Analysis* (*DCA*) and *Normalized Mutual Information* (*NMI*) for detection of residues contacts intra-protein.

Data Set

The data set consists of 17 protein chains: 1A70_A, 1A71_A, 1BQU_A, 1ELV_A, 1G2E_A, 1GDC_A, 1O0W_A, 1O47_A, 1WVN_A, 2BOL_A, 2HDA_A, 2O72_A, 2VI6_A, 3BFR_A, 3FHI_A, 5PTI_A, 6GSU_A.

Method

0.1 Create MSA

MSA is retrieved from BLAST, get as much sequence as possible, but use maximum 20,000 sequences to create MSA.

0.2 Algorithms

I test with two algorithms. The first one is *DCA*, and the second one is *NMI* from Compensatory Mutation Finder. For both algorithms:

Input	MSA of a protein chain (#row should be larger than 1000) with length of L
Output	a list of tube (c_i, c_j, s_{ij}) where $1 \leq i < j \leq L$, c_i, c_j are two residues in protein chain and s_{ij} is the corresponding score.

Note: only long-range residues contacts are taken into account, any pair of residues with 4 units is not considered as contact. The cut-off distance of site contact is 8.5 Å.

0.2.1 DCA

Employ Direct Coupling Technique described in [1] with the default parameter (pseudocount_weight = 0.5; $\theta = 0.2$).

0.2.2 NMI

Utilize normalized mutual information with and without transformed through DSM to calculate the connection score between two residues in chain.

Result

Residue Contacts Scores

Top Score	DCA	NMI_1	NMI_2	DCA-NMI_1	DCA-NMI_2	total
50	**317**	57	32	47	20	7953
60	**370**	69	40	61	28	7953
80	**455**	92	56	90	41	7953
100	**539**	112	73	132	60	7953
(*)	**1730**	380	277	789	358	7953

Explanation

NMI_1: normalized Mutual Information without transformed via DSM

NMI_2: normalized Mutual Information with transformed via DSM

The first column (Top Score) is the number of best residue pair, based on the value from DCA or NMI. For example, in the the second row of above table, we choose the best 50 residue pairs (based on its score, after ignoring the neighborhood) of each protein chain and assume them as the contacts. Based on DCA method, there are 317 true positive, while NMI has only 57 ones. Besides, the index, for example, DCA-NMI_1 is the number which shows the overlap between the best 50 site pairs of DCA and NMI_1. And the final column is the amount of real pair contacts.

(*) Choosing a fixed number such as 50, 100 could be problematic because it is regardless to the protein chain sequence length. In this evaluation, we calculate the true number of residue contacts (called N_i for chain i) and choose the top-N_i residue pair to evaluate.

Correlations

Measure the correlation between DCA, NMI score and distance in 3D among amino acids pairs.

	3D	DCA	NMI
3D	1	**-0.29**	0.04
DCA	**-0.29**	1	0.08
NMI	0.04	0.08	1

Table 1: Correlation of scores among DCA, MNI and 3D distance before eliminating neighbor pair sites

	3D	DCA	NMI
3D	1	**-0.21**	0.05
DCA	**-0.21**	1	0.17
NMI	0.05	0.17	1

Table 2: Correlation of scores among DCA, MNI and 3D distance after eliminating neighbor pair sites

Remarks

1. DCA definitely outperforms NMI because it could capture the real direct coupling and minimize the effect of indirect coupling.

2. Statistical model in physics could be promising to solve many problem in Bioinformatics.

3. DCA is indeed the inverse Ising model problem. In original Ising model, the lattice & bonds parameters (analogous to vertexes and edges) are given, and people are interested in finding the configuration of the model (σ_i is +1 or -1) such as $< \sigma >$ (the magnetization). In contrast, we assume that the configuration is given (20 amino acids and the gap), and we are interested in estimating the bonds between each pair sites.

References

[1] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108, E1293-E1301 (2011).