Report 1: 10.02.1015

**Purpose**: assessment of Direct Coupling Analysis (*DCA*) and Normalized Mutual Information (*NMI*) for detection of residue contact intra-protein.

**Data set**: 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**

**1) Create MSA**

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

**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 <ci, cj, sij>, where 1<=i<j<=L, ci, cj are two residues in protein chain and sij 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 A.

**2.1) DCA**

Employ Direct Coupling Technique described in [Morcos, F. (2011)] with the default parameter (pseudocount\_weight = 0.5; theta = 0.2).

**2.2) NMI**

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

Result

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Top Score | DCA | NMI1 (w.oDSM) | NMI2 (w. DSM) | DCA - NMI1 | DCA-NMI2 | 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:

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-NMI1 is the number which shows the overlap between the best 50 site pairs of DCA and NMI1. 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 Ni for chain i) and choose the top-Ni residue pair to evaluate.

**Remarks**

The performance of DCA is definitely outperform NMI. Moreover, the overlap between DCA and NMI is quite small and approximately encompasses the true contacts predicted by NMI.

Date: 13.02.15

Purpose: measure the correlation between DCA, NMI score and distance in 3D.

Before eliminating neighborhood pair sites

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

After eliminating neighborhood pair sites

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