Using n-dimensional Mutual Information to measure protein multiple residues coevolution

Hongyun Gao[1, a, *], Xiaoqing Yu[2, b], Shuang Chen[1, c], Dan Li[1, d] and Xiaolei Zhu[3, e]

[1] College of Information and Engineering, Dalian University, Dalian 116622, China

[2] Department of Applied Mathematics, Shanghai Institute of Technology, Shanghai 201418, China

[3] School of Life Sciences, Anhui University, Hefei, Anhui 230601, China

[a] gaohongyun@dlu.edu.cn, [b] xqyu@sit.edu.cn, [c] chenshuang0707@163.com,

[d] 279675738@qq.com, [e] xlzhu_mdl@hotmail.com

*corresponding author

Keywords: Mutual Information, Co-evolution, n-dimensional Mutual Information, G Proteins.

Abstract. Many computational tools have been developed to study amino acid co-evolution in proteins. Most of them, such as McLachlan-based substitution correlation and Mutual Information (MI), only focus on co-evolution of two amino acid residues. On the other hand, co-evolution information of multiple residues is useful to study protein structures and functions. Therefore, there is an urgent requirement for methods to identify multiple co-evolved residues. The n-dimensional Mutual Information (n DMI) method is introduced to identify co-evolution of multiple residues. This algorithm gains the co-evolution information of multiple residues from analysis of a matrix of MI values. The method was applied to G proteins. Results show that n DMI method is effective in identification of multiple co-evolved amino acids. Moreover, n DMI performs better in discovering functional important residues than methods focusing on pairwise co-evolved residues.

1. Introduction

In protein evolutionary process, if one residue in a protein sequence has a mutation, relevant residues also mutate to compensate it for maintaining structure and function [1]. These residues undergo adaptive or constructive change without disruption of organism integrity [2, 3]. The change of a biological object triggered by the change of a related object is called “co-evolution” [4]. Many biological processes are related to co-evolution, such as metabolic pathways, signaling cascades, and transcription regulation. The knowledge of co-evolution can be used to predict the functional or structural important residues and guide experimental designs [5, 6]. To quantify the co-evolution of residues, many computational methods have been developed. Typically, they used the multiple sequence alignment (MSA) of a protein chain and its homologous sequences to search for correlated mutations.

Previous works suggested that a network of co-evolved amino acids comprise the essence of three-dimensional structure and function of a protein [7, 8], and hence the number of co-evolved residues are not limited to two. Since all existing methods focus on the co-evolution of two amino acid residues, the emerging problem is how to quantify the co-evolution of multiple residues. To solve these problems, a method is called n-dimensional Mutual Information (n DMI) was developed and applied to image registration [9]. The method of n DMI needs to calculate the eigenvalues of an n-dimensional Mutual Information matrix (MIM). In the study, n DMI was further improved and extended to the application of detecting co-evolution of multiple residues. This method was applied to G proteins and G Protein-Coupled Receptors (GPCRs). The results show that n DMI is effective in quantifying co-evolution degree for multiple residues. Moreover, it performs better in identification of functional important residues than the co-evolution method for two residues.

2. Method

2.1. N-dimensional Mutual Information method

Given a protein chain with H residues, a MSA M was built using the protein sequence and its homologous sequences. For a residue set with n residues $(r_1, r_2, \cdots r_n)$ in the chain, MIM for the residue set is defined as:

$$\mathrm{MIM}(r_1, r_2, \cdots r_n) = \begin{pmatrix} I_{11} & I_{12} & \cdots & I_{1n} \\ I_{21} & I_{22} & \cdots & I_{2n} \\ \vdots & & \ddots & \vdots \\ I_{n1} & I_{n2} & \cdots & I_{nn} \end{pmatrix}, \quad (1)$$

Where $I_{ij}(i, j = 1,2, \cdots, n)$ denotes the MI value between residues r_i and r_j calculated from the MSA *M*, which will be described in the following section. Let λ_i be the ith eigenvalue of above matrix$\mathrm{MIM}(r_1, r_2, \cdots r_n)$, then the nDMI method is defined as:

$$\mathrm{nDMI}(r_1, r_2, \cdots r_n) = 1 + \frac{\sum_{i=1}^{n} \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \ln\left(\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}\right)}{\ln n}. \quad (2)$$

This definition makes the *n*DMI positive, and $0 \leq \mathrm{nDMI} \leq 1$.

2.2. Mutual Information

Given an MSA *M* with *N* sequences, the amino acid position specific frequency of a column *K* for the *i*th ($i = 1,2, \cdots, 20$) amino acid is calculated as:

$$p(K^i) = \frac{count(K^i)}{N}. \quad (3)$$

Where the $\boldsymbol{count(K^i)}$ is the number of the *i*th amino acid in the column *K*. The joint probability distribution of columns K and L is defined as:

$$p(K^i, L^j) = \frac{count(K^i, L^j)}{N}. \quad (4)$$

Where the $count(K^i, L^j)$ is the residue pairs of the *i*th and *j*th amino acids. Based on the above definitions, the MI of column *K* and *L* is calculated as:

$$I_{ij} = \sum_{i=1}^{20} \sum_{j=1}^{20} p(K^i, L^j) \cdot \ln \frac{p(K^i, L^j)}{p(K^i) \cdot p(L^j)} \quad (5)$$

The MI model is used to measure the co-evolution of pairwise residues in the study.

3. Results and Discussion

3.1. Comparisons on G proteins

G proteins are involved in transmitting signals from a stimulus outside a cell into the inside of the cell. G proteins are so-called because they bind the guanine nucleotides GDP and GTP. The catalytic domains, about 160 amino acids, in all G proteins that are associated with the function of nucleotide binding are conserved. Three motifs are identified in this domain: NKXD motif defines the guanine specificity, DXXG motif is important for the ion and γ-phosphate group binding, and GXXXXGK motif is for *h*-phosphate binding [10].

To study co-evolution of amino acids in G proteins, the MSA for G protein family were collected from the SCA toolkit, and this MSA contains 678 G protein sequences. MI, 3DMI, and 4DMI were applied to the MSA, respectively. To illustrate co-evolution information of G proteins conveniently, the G protein raps (PDB ID: 1Q21) was chosen as a representative because its structure has been solved. In this protein, three G protein motifs are NKCD (116-119) (M1), DTAG (57-60) (M2), and GAGGVGK (10-16) (M3), respectively [11]. The top 20 ranked residue sets by 3DMI and 4DMI are shown in Table 1. According to the results, we can find that n DMI identifies co-evolved residue sets within one motif, such as (15G(M3), 16K(M3)) and (57D(M2) ,60G(M2)); or between two motifs, such as (57D(M2), 119D(M1)) and (16K(M3), 119D(M1)). For this kind of cases, MI also

has chance to identify some of them. However, 3DMI has ability to discover co-evolved amino acid residues from three different motifs. In Table 2, three sets, (15G(M3), 57D(M2), 119D(M1)), (16K(M3), 57D(M2), 119D(M1)), and 15G(M3), 60G(M2), 119D(M1), highly scored by 3DMI have residues from three different motifs. For 4DMI, 19 out of 20 top scored sets have residues from three different motifs. These results show that n DMI is effective in identification of multiple residue co-evolution.

The n DMI method was also tested for identification of functional important residues. For the purpose, the function of conn(k): a kind of summarization of multiple co-evolution scores was applied. Because previous works suggested that conn(k) function performs better than individual co-evolution scores. The function of conn(k) is defined as the number of high-scoring pairs a residue k is an element of these high-scoring pairs. To calculate conn(k), top 75 scored residue sets were selected for each method in the study and values of conn(k) with cutoff conn(k) = 4 are shown in Table 2. As shown in Table 2, MI only finds 4 residues 15G(M3), 57D(M2), 58T (M2), and 59A(M2) in functional motifs, and three of them 15G(M3), 57D(M2) and 58T (M2) are ranked in low positions. For 3DMI, 6 amino acid residues 119D(M1), 57D(M2), 59A(M2), 60G(M2), 15G(M3) and 16K(M3) are identified. Although only 5 resides are discovered by 4DMI, 119D(M1), 57D(M2), 60G(M2), 15G(M3), and 16K(M3), all of them are in those three motifs. Moreover, the residue 119D is a functional important residue and its mutation can produce constitutively activated or dominant-negative effects. Based on values of conn(k), functional important residues are ranked at top positions by 3DMI or 4DMI, but MI method fails to find it.

Table 1 Top 20 ranked residue sets by 3DMI and 4DMI on G proteins.

	3DMI			4DMI			
1	15^{G}(M3)	57^{D}(M2)	119^{D}(M1)	15^{G}(M3)	16^{K}(M3)	57^{D}(M2)	119^{D}(M1)
2	15^{G}(M3)	16^{K}(M3)	57^{D}(M2)	15^{G}(M3)	57^{D}(M2)	60^{G}(M2)	119^{D}(M1)
3	15^{G}(M3)	16^{K}(M3)	119^{D}(M1)	15^{G}(M3)	16^{K}(M3)	57^{D}(M2)	60^{G}(M2)
4	16^{K}(M3)	57^{D}(M2)	119^{D}(M1)	15^{G}(M3)	16^{K}(M3)	60^{G}(M2)	119^{D}(M1)
5	15^{G}(M3)	16^{K}(M3)	60^{G}(M2)	16^{K}(M3)	57^{D}(M2)	60^{G}(M2)	119^{D}(M1)
6	57^{D}(M2)	60^{G}(M2)	119^{D}(M1)	15^{G}(M3)	57^{D}(M2)	116^{N}(M1)	119^{D}(M1)
7	15^{G}(M3)	57^{D}(M2)	60^{G}(M2)	10^{G}(M3)	15^{G}(M3)	57^{D}(M2)	119^{D}(M1)
8	39^{S}	67^{M}	96^{Y}	15^{G}(M3)	57^{D}(M2)	119^{D}(M1)	146^{A}
9	67^{M}	68^{R}	96^{Y}	15^{G}(M3)	35^{T}	57^{D}(M2)	119^{D}(M1)
10	59^{A}(M2)	68^{R}	96^{Y}	15^{G}(M3)	57^{D}(M2)	116^{N}(M1)	119^{D}(M1)
11	15^{G}(M3)	60^{G}(M2)	119^{D}(M1)	15^{G}(M3)	57^{D}(M2)	61^{Q}	119^{D}(M1)
12	58^{T} (M2)	59^{A}(M2)	68^{R}	15^{G}(M3)	57^{D}(M2)	83^{A}	119^{D}(M1)
13	67^{M}	72^{M}	96^{Y}	15^{G}(M3)	20^{T}	57^{D}(M2)	119^{D}(M1)
14	39^{S}	72^{M}	96^{Y}	15^{G}(M3)	17^{S}	57^{D}(M2)	119^{D}(M1)
15	67^{M}	71^{Y}	96^{Y}	15^{G}(M3)	55^{I}	57^{D}(M2)	119^{D}(M1)
16	11^{A}(M3)	67^{M}	96^{Y}	15^{G}(M3)	57^{D}(M2)	77^{G}	119^{D}(M1)
17	57^{D}(M2)	116^{N}(M1)	119^{D}(M1)	15^{G}(M3)	57^{D}(M2)	58^{T} (M2)	119^{D}(M1)
18	59^{A}(M2)	67^{M}	68^{R}	9^{V}	15^{G}(M3)	57^{D}(M2)	119^{D}(M1)
19	59^{A}(M2)	67^{M}	96^{Y}	15^{G}(M3)	23^{L}	57^{D}(M2)	119^{D}(M1)
20	11^{A}(M3)	59^{A}(M2)	96^{Y}	15^{G}(M3)	57^{D}(M2)	112^{V}	119^{D}(M1)

Table 2 Values of conn(k) for amino acid in G proteins from co-evolution scores generated by MI, 3DMI and 4DMI methods.

MI		3DMI		4DMI	
k	conn(k)	k	conn(k)	k	conn(k)
39^{S}	20	119^{D}(M1)	54	119^{D}(M1)	74
59^{A}(M2)	12	57^{D}(M2)	54	57^{D}(M2)	74
67^{M}	11	96^{Y}	14	15^{G}(M3)	74
68^{R}	11	67^{M}	9	16^{K}(M3)	4
96^{Y}	11	59^{A}(M2)	9	60^{G}(M2)	4
71^{Y}	6	68^{R}	7		
72^{M}	6	15^{G}(M3)	6		
58^{T} (M2)	6	39^{S}	5		
57^{D}(M2)	4	16^{K}(M3)	4		
15^{G}(M3)	4	60^{G}(M2)	4		

4. Conclusion

In this study, nDMI method was developed and applied to measure coevolution of multiple residues for a given protein chain. Different to previous methods, nDMI can quantify co-evolution degree for more than two amino acid residues. G proteins were used to test this method. Results show that n DMI successfully discovers multiple co-evolved residues. This method can be used to a complex system that has more co-evolved residues, and help in predicting protein structural and functional important residues.

Acknowledgments

This paper is partially supported by the Natural Science Foundation of China (11626051, 11626052, 11501074, 21403002), and the Doctoral Fund of Liaoning Province (201601296).

References

[1]. F. Pazos, M. Helmer-Citterich, G. Ausiello, A. Valencia, Correlated mutations contain information about protein-protein interaction, J Mol Biol 271(4) (1997) 511–523.

[2]. F. Pazos, A. Valencia, Similarity of phylogenetic tree as indicator of protein protein interaction, Protein Engineering 14(9) (2001) 609–614.

[3]. H. B. Fraser, A. E. Hirsh, D. P. Wall, M. B. Eisen, Coevolution of gene expression among interacting proteins, PNAS 101(24) (2004) 9033–9038.

[4]. K. Y. Yip, P. Patel, P. M. Kim, D. M. Engelman, D. McDermott, M. Gerstein, an integrated system for studying residue coevolution in proteins, Bioinformatics 24(2) (2008) 290–292.

[5]. D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein coevolution. Nature Reviews Genetics 14 (2013) 249–261.

[6]. I. Sandler, M. Abu-Qarn, A. Aharoni, Protein co-evolution: how do we combine bioinformatics and experimental approaches? Mol. BioSyst. 9 (2013) 175–181.

[7]. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure, Cell 138 (2009) 774-786.

[8]. S. W. Lockless, R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286(5438) (1999) 295-299.

[9]. B.Wang, Y. Shen, A method on calculating high-dimensional mutual information and its application to resgistration of multiple ultrasound images, Ultrasonics 44 (2006) e79-e83.

[10]. W. Moller, R. Amons, Phosphate-binding sequences in nucleotide-binding proteins, FEBS Letters 186 (1985) 1-7.

[11]. G. M. Suel, S. W. Lockless, M. A. Wall, R. Ranganathan, evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol. 10 (1).