PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Analysis of the residue–residue coevolution network and the functionally important residues in proteins

Byung-Chul Lee, Keunwan Park, and Dongsup Kim*

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

ABSTRACT

It is a common belief that some residues of a protein are more important than others. In some cases, point mutations of some residues make butterfly effect on the protein structure and function, but in other cases they do not. In addition, the residues important for the protein function tend to be not only conserved but also coevolved with other interacting residues in a protein. Motivated by these observations, the authors propose that there is a network composed of the residues, the residue–residue coevolution network (RRCN), where nodes are residues and links are set when the coevolutionary interaction strengths between residues are sufficiently large. The authors build the RRCN for the 44 diverse protein families. The interaction strengths are calculated by using McBASC algorithm. After constructing the RRCN, the authors identify residues that have high degree of connectivity (hub nodes), and residues that play a central role in network flow of information (C^I nodes). The authors show that these residues are likely to be functionally important residues. Moreover, the C^I nodes appear to be more relevant to the function than the hub nodes. Unlike other similar methods, the method described in this study is solely based on sequences. Therefore, the method can be applied to the function annotation of a wider range of proteins.

Proteins 2008; 72:863–872.

Key words: multiple sequence alignment; correlated mutations analysis; covariance algorithm; hub; information centrality; network analysis.

INTRODUCTION

Developing experimental and computational methods to identify the functionally important residues of proteins have long been investigated. Since functionally and structurally important sites of proteins are more conserved than other residues, many methods based on the conservation pattern have been developed.[1] A different kind of approach is the correlated mutation analysis (CMA).[2–7] In CMA, coevolutionary relationships between residues are inferred by analyzing the correlated mutational patterns between columns of a multiple sequence alignment (MSA) of a protein family. One typical application of this analysis is to predict the inter-residue contacts.[2,5,6] Another type of application is to find the functional sites such as allosterically modulating residue pairs in proteins.[4] There are several recent studies that apply CMA methods to specific proteins. Such examples are a study on clustering structural and ligand portal residues in the iLBP protein family,[8] and another study showing that the coevolutionary information is very helpful to specify the artificial WWW domains.[7,9]

Recently, viewing a protein as a network of interacting residues has been gaining much interest.[10–15] In such studies, efforts to find the functionally important sites from protein structures are made. They represent a protein as a network using interatomic physical interactions and calculate some network properties such as "residue centrality"[10] to examine whether these network properties are related to protein structure and function.[10,15] Motivation for this kind of approach is based on experimental findings that mutations of most of the residues have little effect on protein function, while perturbation of a few residues break down the function entirely.[16–18] This property is reminiscent of the network property of scale-free network where the network is robust against random attack, but highly vulnerable to pointed attack on so-called hub nodes. The importance of hub nodes arises from the fact that they play a critical role in information flow of the network because by definition they are connected to a large number of neighboring nodes. Removal of a hub node, therefore, is detrimental to the proper functioning of the network. From this observation, it is reasonable to conjecture that we may be able to identify structurally and/or functionally important residues in proteins by locating some sort of hub nodes in the network made of residues of a protein.[10]

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Dongsup Kim, Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, Korea. E-mail: kds@kaist.ac.kr

Received 16 May 2007; Revised 4 December 2007; Accepted 18 December 2007
Published online 14 February 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21972

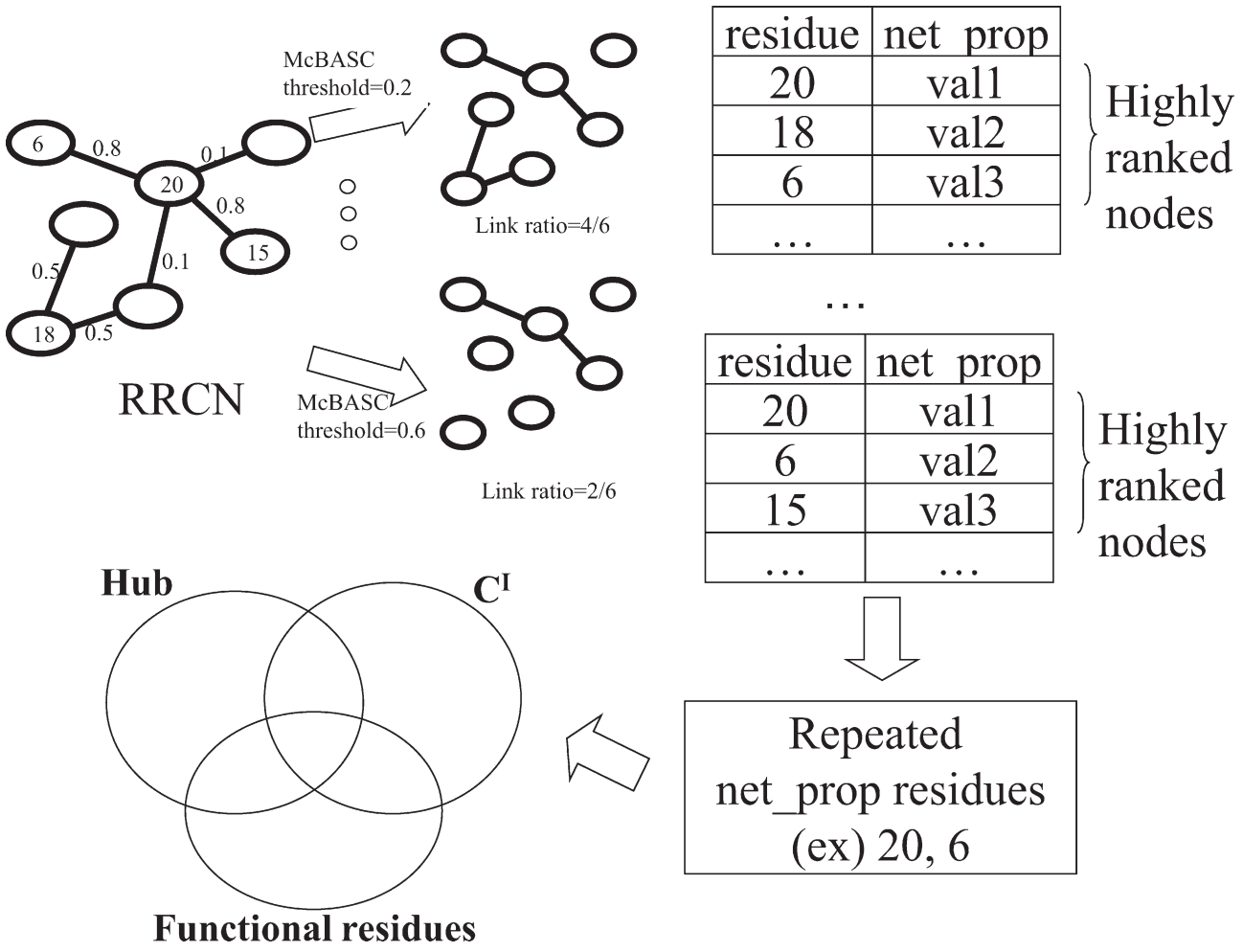

Figure 1

The overall workflow of our experiment. RRCN is the residue–residue coevolution network built in this research, and net_prop stands for the network properties: hubs and information centrality (C^I). The RRCN is a network whose nodes are residues and links are set when the coevolutionary interaction strengths between residues are sufficiently large. We use the McBASC scores for the interactions. Hub nodes have high degree of connectivity; C^I nodes play a central role in network flow of information. From the RRCN, the two network properties are calculated, and then the comparisons with the functional sites are made.

In this work, we propose a new type of protein network, and develop a new way to predict the functionally important residues of proteins by analyzing certain network properties. We assume that there is a network of evolutionarily related residues of a protein, which is named the residue–residue coevolution network (RRCN). This RRCN is an unweighted undirected graph where nodes are residues of a protein, and links are set depending on the coevolutionary interaction strengths between residues. In previous works,[10,11] links between residues are inferred from the direct physical interactions between residues. In this work, we calculate from the MSA the interaction strengths that are a measure of how closely the two residues are evolutionarily related. After constructing the RRCN, we identify the residues that have high degree of connectivity (hub nodes), and the residues that play a central role in network flow of information (C^I nodes). Next, we show that these residues are likely to be the functionally important residues. In addition, we provide evidence that the C^I nodes are more relevant to the protein functions than the hub nodes. Finally, we compare our results with those of simple conservation and previous related works.[10,19]

METHODS

General work flow of this study is shown in Figure 1. We first construct the RRCN by performing a series of calculations such as building MSA, calculating residue–

residue coevolutionary information by CMA, and making links between residues. After constructing RRCN, network properties are inferred, and a few top-ranked residues are selected. Finally, overlap between the selected residues and the functional residues are analyzed.

Building multiple sequence alignments

To retrieve the homologs of sequences that are used in this study, we use PSI-BLAST[20] against the protein database NR65, which is prepared from "nr" database by cd-hit.[21,22] The options for PSI-BLAST are −h 0.001 −e 0.001 −j 3. To build the MSAs, we use MUMMALS,[23] instead of more popular program CLUSTALW.[24] MUMMALS is based on the hidden Markov model (HMM), and its parameters are optimized by using the structural information. According to the study on MUMMALS, MUMMALS outperforms CLUSTALW. We use the model "HMM_1_3_1" of MUMMALS, HMM consisting of 1 solvent accessibility category, 3 secondary structure types, and 1 unmatched state.

Calculating residue–residue coevolutions

Two residues in a protein are said to be interacting if these two residues are structurally or functionally coupled. However, deciding whether two residues are structurally or functionally coupled is not a trivial task. A common experimental method to detect energetically coupled interactions is the double mutant cycle analysis.[25–28] On the other hand, if protein's structure is known, physical interactions between residues can be deduced from the structure by examining if two residues make a direct contact, as done in the previous work by del Sol *et al.*[10] As seen in many examples of allosterically interacting residues,[29] however, interacting residues are not necessarily spatially close to each other.

These direct and indirect interactions aforementioned can be inferred by CMA, that is, by analyzing the correlated evolutionary patterns between residues embedded in MSA of the protein. One of the most widely used and successful methods in CMA is McBASC algorithm.[5] Recently, Ranganathan and coworkers[4,30–32] have proposed the statistical coupling analysis (SCA) algorithm that is designed to detect the coevolution of amino acid residues in a protein and applied to detecting functionally coupled interactions. Dekker *et al.*[3] proposed a modified algorithm that shows a better performance when compared with SCA algorithm. We have tried both SCA algorithm and McBASC algorithm among a variety of coevolution analysis algorithms. It turns out that McBASC algorithm gives a somewhat better and more consistent result than SCA. Accordingly, only the results obtained by using McBASC algorithm are presented in this article. The programs for SCA and McBASC algorithms are downloaded from A. Fodor's homepage.[3]

Constructing and analyzing residue–residue coevolution network

We assume that there is a network of evolutionarily related residues of a protein, which is named as the RRCN. This RRCN is an unweighted undirected graph where nodes are residues of the protein, and links are set depending on the coevolutionary interaction strengths between residues. To set the link between residues, we need to set the threshold value of interaction strength, which inevitably affects the network structure, and therefore our inferred network properties. To alleviate negative effect of choosing arbitrary threshold value, we generate multiple networks by choosing multiple threshold values, and then for functional analysis we only chose those hub nodes and C^{I} nodes that repeatedly appear in all networks. We select the thresholds in such a way that the resulting 7 networks have 2, 1, 0.5, 0.20, 0.15, 0.10, and 0.05% of the maximum number of links, respectively. In addition, before making networks, we remove all pairs of residues with −2.0 McBASC score because the score represents too many gaps in MSA columns.

After setting up the RRCN of proteins, we analyze the networks using two network properties: highly linked nodes (hub nodes) and nodes with high information centrality scores (C^{I} nodes).[33] We chose the top high-ranking nodes the number of which ranges from 20 up to 70, and then finally select the nodes that appear in all 7 networks. Depending on the number of chosen nodes, the number of finally selected residues per protein varied from 8.86 to 32.05 for hub nodes, and from 3.70 to 27.82 for C^{I} nodes. Hub nodes are defined as the most highly linked nodes. In the scale free network, hub nodes are important in that they guarantee the stability of the network against random attack. C^{I} is a measure of how efficiently the information propagates through a network. C^{I} is defined by

$$C_i^{\mathrm{I}} = \frac{\Delta E}{E} = \frac{E[G] - E[G_i']}{E[G]}$$

where $E[G]$ is global efficiency of a network and $E[G_i']$ is also global efficiency but after removing the node i from the RRCN. Global efficiency is defined by

$$E[G] = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$$

where N is the total number of nodes and d_{ij} is the shortest path between ith and jth nodes in the graph G. Hubs and bridges in a network usually show high C^{I} score.

Table I
The Representative PDB Structures and Their SCOP Class Distribution of the Protein in the Data Set

SCOP class						
A (8)	B (11)	C (14)	C and D (2)	D (6)	E (2)	G (1)
1a59	1a2yA	1a4sA	1bwvA	1ayu	1bsg	1cdtA
1aeiA	1a30A	1a50B	1chrA	1b02A	1g0hA	
1aokA	1a78A	1a5cA		1b5eA		
1aru	1a8mA	1aoeA		1bmkA		
1auwA	1aac	1b00A		1dzaA		
1bbhA	1afcA	1bj4A		1ehwA		
1ch4A	1am5	1bwkA				
1dkfA	1ao5A	1bxkA				
	1b07A	1cl1A				
	1cbrA	1d2rA				
	1fljA	1d5cA				
		1dciA				
		1dx4A				
		1lqaA				

PDB id and chain are used for the representative ids. The numbers in parentheses are the number of the structures of each SCOP family.

Functional residues

Information on the functional residues is taken from the del Sol *et al.*'s article.[10] In their study, the functional residues such as catalytic, ligand and metal binding, and protein–protein interface residues of 46 diverse protein families are listed. In this study, we only use 44 proteins (Table I) because 2 proteins have too small number of PSI-BLAST hits and, therefore, they are excluded. The SCOP classes and the representative structures of each family are shown in Table I. The number of structures of each family is shown in parentheses. It is clear that the test set is well selected from diverse structural classes. We name the set of functionally important residues of these 44 proteins *Functional Site Set*, which has 1175 functional residues. To compare our work with del Sol *et al.*'s work, we also prepare *Expanded Functional Site Set* that includes residues in direct contact with the residues in functional site set, which has 4092 functional residues. Two residues are assumed to be in direct contact if at least one pair of heavy atoms is within 5 between the two residues.

Previous sequence-based method

We compare our result with ConSeq server.[19] ConSeq server consists of two independent modules: One is to calculate the evolving rate and the other is to predict solvent accessibility. The core program of ConSeq is Rate4-Site,[34] which takes a role of calculating the evolving rate from a given MSA. We have obtained this program from ConSeq server administrator and performed local experiments with default options (Bayesian method). The same MSAs of RRCN have been used for the Rate4Site. The outputs of Rate4Site have been converted to the scores from 1 to 9 using the script given by the administrator, and the score 9 is used as the threshold score for the most conserved residues.

Information contents

We also perform a simple experiment on the conservation scores of the 44 proteins. To calculate the conservation, we calculate the information content using the following formula,

$$\mathrm{IC}_j = \sum_{i=1,\ldots,20} P_{i,j} \log \frac{P_{i,j}}{Q_i}$$

where i is one of 20 amino acids, j is the column number of MSA. IC_j is the information content of jth column, P_{ij} is the observed frequency of amino acid i of jth column, and Q_j is the background probability. Background probability are from the Astral SCOP 1.67 with 40% homology cutoff.[35] While changing the threshold score from 1.5 to 2.5, the sensitivities and specificities are plotted in Figure 3.

P-value, sensitivity, and specificity

In our study, a selected number of top high-ranking nodes that have the most links and/or the highest C^1 scores are used to analyze their relationship with the functional residues of proteins. We have examined how many selected nodes are in fact the functional residues. To test if the overlap between the selected residues and the functional residues is statistically significant, we estimate the *P*-values using the hypergeometric distribution.[36] If N is the total number of residues, $n1$ the number of hub nodes in the RRCN (in our case, $n1$ is the number of repeatedly appearing nodes), $n2$ the number of the functional residues, and m the intersect of $n1$ and $n2$, the probability and *P*-value are given by

$$P(m) = \frac{\binom{n1}{m}\binom{N-n1}{n2-m}}{\binom{N}{n2}},$$

$$P\text{-}value = \sum_{k=m}^{\min(n1,n2)} P(k) = \sum_{k=0}^{\min(n1,n2)} P(k) - \sum_{k=0}^{m-1} P(k)$$

The sensitivity and specificity are defined by

$$\mathrm{S_n} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{S_p} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

where TP, FN, and FP are the number of true positives, false negatives, and false positives, respectively. In our experiment, TP = m, FP = $n1-m$, and FN = $n2-m$. For this calculation, we use an R[37] function: *P*-value = phyper (min($n1$,$n2$),$n1$,$n-n1$,$n2$)−phyper($m-1$,$n1$,$n-n1$,$n2$).

Table II
Comparison of the Results from RRCN and Previous Methods

Methods	S_n	S_p	*P-value*	TP	FN	FP	# Pred
Functional site set							
Top 20							
Hub	0.07	0.23	5.73 E−13	88	1087	302	8.86
C^I	0.05	0.33	6.77 E−15	53	1122	110	3.70
hub and C^I	0.04	0.34	1.31 E−14	49	1126	96	3.30
Top 70							
Hub	0.23	0.19	0.00 E+00	265	910	1145	32.05
C^I	0.21	0.20	0.00 E+00	247	928	977	27.82
hub and C^I	0.20	0.20	0.00 E+00	230	945	920	26.14
ConSeq	0.27	0.24	0.00 E+00	313	862	1006	29.98
res_cent	0.03	0.27	5.23 E−07	30	1145	81	2.52
info_cont (threshold 2.5)	0.11	0.17	2.24 E−09	127	1048	611	16.77
Expanded functional site set							
Top 20							
Hub	0.06	0.64	0.00 E+00	248	3844	142	8.86
C^I	0.03	0.65	2.88 E−14	106	3986	57	3.70
hub and C^I	0.02	0.66	3.21 E−13	95	3997	50	3.30
Top 70							
Hub	0.20	0.58	0.00 E+00	813	3279	597	32.05
C^I	0.18	0.59	0.00 E+00	724	3368	500	27.82
hub and C^I	0.17	0.59	0.00 E+00	681	3411	469	26.14
ConSeq	0.18	0.56	0.00 E+00	744	3348	575	29.98
res_cent	0.02	0.74	3.33 E−16	82	4010	29	2.52
info_cont (threshold 2.5)	0.09	0.48	2.8 E−12	354	3738	384	16.77

S_n, S_p, *P*-value, TP, FP, and FN stand for sensitivity, specificity, *P*-values, true positives, false positives, and false negatives of the entire dataset of 44 proteins, respectively. Functional site set is functionally important sites itself. Expanded functional site set includes the residues in direct contact with functional site set. Top 20 (Top 70) indicates that top 20 (70) high-ranking hub or C^I nodes were initially chosen from the network for analysis. ConSeq, res_cent, and info_cont stand for ConSeq program (Rate4Site), residue centrality method, and information contents, respectively. #Pred denotes the average number of predictions per protein. See Supplementary Data (STable V–VII) for the results with other numbers of chosen top high-ranking nodes. The detailed results are also depicted in Figure 3.

RESULTS

Relationship between residue–residue coevolution network analysis and functional sites

To investigate the relationship between the RRCN and the functional sites, we calculate the sensitivity (S_n), specificity (S_p), and *P*-value. We use the functionally important site information from the previous research[10] as a golden standard. As shown in Table II where the overall summary of the comparison results are listed, the majority of residues that are predicted to be the hub nodes and/or the C^I nodes are overlapped with either functional sites or in direct contact with them. We have changed the number of initially chosen top high-ranking nodes from 20 to 70. For hub nodes, the sensitivity on functional site set increases from 0.07 to 0.23. However, the specificity on the same set slightly decreases from 0.23 to 0.19. On expanded functional site set, the specificity changes from 0.64 to 0.58. With respect to C^I nodes, the sensitivity increases from 0.05 to 0.21, but the specificity decreases from 0.33 to 0.20. On expanded functional site set, the specificity changes from 0.65 to 0.59. More detailed results with comparison to other methods are displayed in Figure 3.

Table II also shows *P*-values. Overall, small or near-zero *P*-values indicate that the overlaps are statistically significant. However, if we calculate *P*-values for each individual protein, in majority of cases, statistical significance is not guaranteed (STable I–STable IV). This seemingly contradictory result is simply due to the fact that *P*-value is greatly affected by the number of samples, and obviously for individual protein, the number of samples is very small when compared with the number of samples of entire proteins in the set. Nonetheless, nearly half of the proteins show the statistically significant results (STable I–STable IV) if we use somewhat relaxed significance level, 0.1. When the number of initial high-ranking nodes is 70, with respect to the hub nodes, 30 proteins show the statistical significance for expanded functional site set, and with respect to the C^I nodes, 20 proteins are statistically significant. For functional site set, 15 (hub nodes and C^I nodes) proteins show statistically significant results. These results indicate that the RRCN contains important information on protein functional sites.

Since we exploit two different network properties, it is useful to investigate the relationship between the hub nodes and C^I nodes. As mentioned in the previous section, the hub nodes play an important role in maintaining the network stability, and C^I nodes do a crucial role in the information flow of a network. All C^I nodes are not necessarily the hub nodes. However, as shown in Figure 2, 89% of C^I nodes are also the hub nodes and 37% of hub nodes are also the C^I nodes, which indicate that the two network properties of RRCN are largely overlapped. In general, as shown in Figure 2 and Table II, S_p of C^I nodes is higher than that of hub nodes, while S_n of C^I nodes is lower than that of hub nodes. Figure 2 and Table II also show that the nodes that are both the hub node and the C^I node simultaneously are more likely to be the functional sites than the hub or C^I nodes are.

These results suggest that the special nodes of the RRCN (hub nodes and C^I nodes) are in fact important for the functionality of proteins. Moreover, the C^I nodes, an obscure term in the field of network theory,[33] which have not been gaining much attention when compared with the hub nodes, may be more significant than the hub nodes when we interpret the biological meaning of the network.

Comparison with other methods

To further investigate the performance of our method, the comparison with simple conservation method and the other previous structure or sequence-based stud-

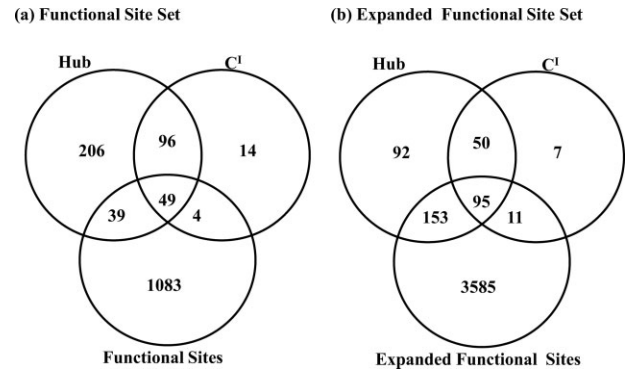

Figure 2

*Venn diagram showing overlap between the hub nodes or C^I nodes with functional sites when initial top 20 high-scoring nodes are initially chosen. Each number represents the number of predictions for each category for each network property. For example, in (**a**), there are 302 (= 206 + 96) hub nodes that are not a functional site, while 88 (= 39 + 49) hub nodes are a functional site. There are 1087 (= 1083 + 4) residues that are not predicted to be a hub node. Therefore, the specificity is 0.23 [= 88/(302 + 88)], and the sensitivity 0.07 [= 88/(88 + 1087)]. (**b**) The numbers are for expanded functional site set, and can be interpreted in the same way.*

ies[10,19] are made. In the previous structure-based study,[10] they define a new concept of residue centrality (similar to the C^I), and the functional residues are predicted if the positions of residues are in the alignments with statistically highly significant residue centrality values (z-score $\geq$ 2.0) in at least 70% of the structures of the family members. The definition of S_p and S_n is different from ours, so we recalculate the S_n, S_p, and P-values of the previous study. The average size of our predicted sites is 8.86 for hub nodes and 3.70 for C^I nodes, but that of the residue centrality nodes of the previous study is only 2.52. The comparison results are shown in Table II and Figure 3. S_p is almost the same, but S_n of our result is better than that of the previous result. It should be noted that the present method is more general than the previous method because it does not require structural information. On top of that, the comparison study indicates that the residues found by the present method have more overlap with the functional sites than those by the previous method.

The second comparison is made with ConSeq server.[19] The results are shown in Table II and Figure 3. For functional site set, it is evident that the performance of ConSeq is better than those of our result and the previous method based on residue centrality. However, for expanded functional site set, the performance of the present method is comparable to, or even better than, that of ConSeq. However, it should be noted that ConSeq method does not just rely on sequence conservation; it also uses phylogenetic relations between the sequences, and estimate the evolutionary rates at each site of the protein. Therefore, it is interesting to test how the performance of our method compares with that of a method based on sequence conservation information only. To do this, we have calculated the information contents (see Methods) for the same input MSA of RRCN. If the information content of one column is higher than a certain threshold value, we decide that the column is conserved. The results, shown in Table II and Figure 3, indicate that the method based on simple sequence conservation performs much worse than ConSeq and the present method, implying that RRCN contains more information that is relevant to the protein function than simple sequence conservation pattern.

In addition to the comparison of several methods, we examine the overlap between the present method, previous methods, and the functional sets. The results are

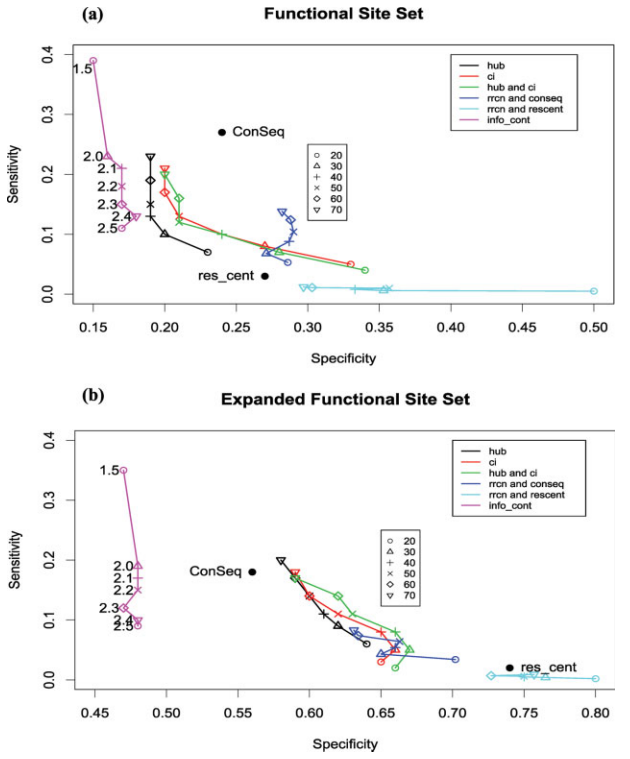

Figure 3

*The sensitivity and specificity plot of Table II. The meanings of colors are depicted at the upper-right corner. "hub" is hub nodes, "ci" is C^I nodes, "hub and ci" means the simultaneous predictions of hub and C^I nodes. "rrcn and conseq" and "rrcn and rescent" are the cases of using our method and ConSeq (sequence-based previous work) or residue centrality (structure-based previous work), respectively, and "info_cont" is information contents. The numbers next to the information contents are the threshold. The number of top high-ranking nodes changes from 20 to 70. (**a**) Functional site set results. As the number of selected top high-ranking nodes increases, S_p decreases and S_n increases. C^I nodes show better performance than hub nodes. When we combine our method and other previous methods, the specificities of previous methods increase. (**b**) Expanded functional site set results. C^I nodes maintain better performance than hub nodes. Residue centrality nodes show the best specificity but lowest sensitivity. ConSeq shows good sensitivity but low specificity. Combining effect of our method and previous methods are weaker than for the case of functional site set.*

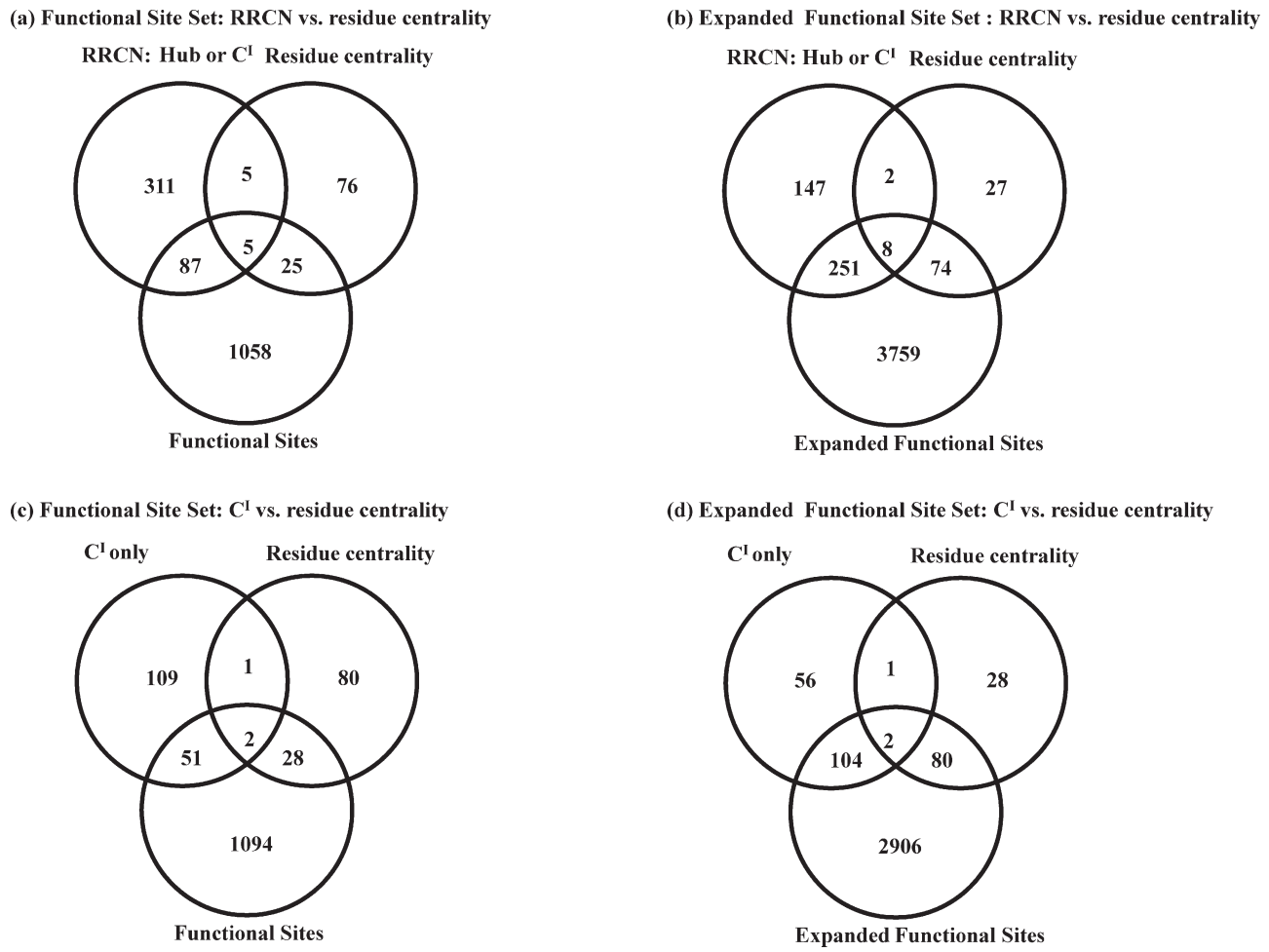

Figure 4

*Venn diagram of current work (RRCN, top 20 nodes) and the previous structure-based work (residue centrality). In this diagram, RRCN represents the union of hub and C^I nodes. (**a**),(**b**) S_p of RRCN is 0.023 and 0.63, but the intersection of RRCN and previous study is 0.5 and 0.8. (**c**),(**d**) few residues are overlapped between C^I and previous study.*

depicted in Figures 4 and 5 as a Venn diagram. The specificity–sensitivity plots for overlapped residues are shown in Figure 3. The overlaps between the RRCN method and the previous methods are quite low, indicating that they identify quite different set of residues as the functional sites. Meanwhile, it is interesting that the specificity is highest for the residues predicted by RRCN and residue centrality methods (S_p = 0.5 for functional site set, S_p = 0.8 for expanded functional site set, Figure 4(a,b). Although the concept of C^I and residue centrality is similar, few residues are overlapped [Fig. 4(c,d)]. From the observations, we can conclude that RRCN (especially hub nodes) and residue centrality methods are complementary to each other. Similar tendency is also observed for the residues predicted by current work and ConSeq: Specificity is improved. Similarly to the case of RRCN and residue centrality methods, we can also conclude that RRCN and ConSeq methods are complementary and

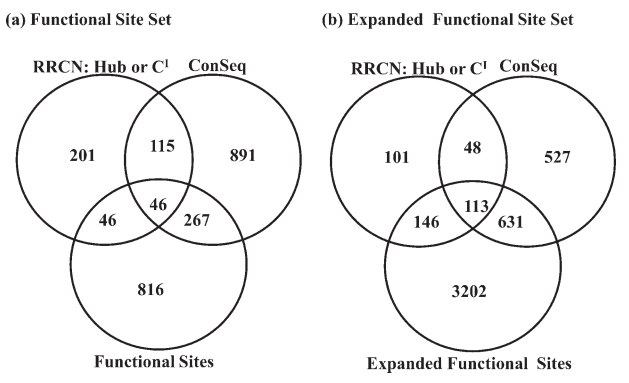

Figure 5

Venn diagram for current work (RRCN, top 20 nodes) and the previous sequence-based work (ConSeq). In this diagram, RRCN represents the union of hub and C^I nodes.

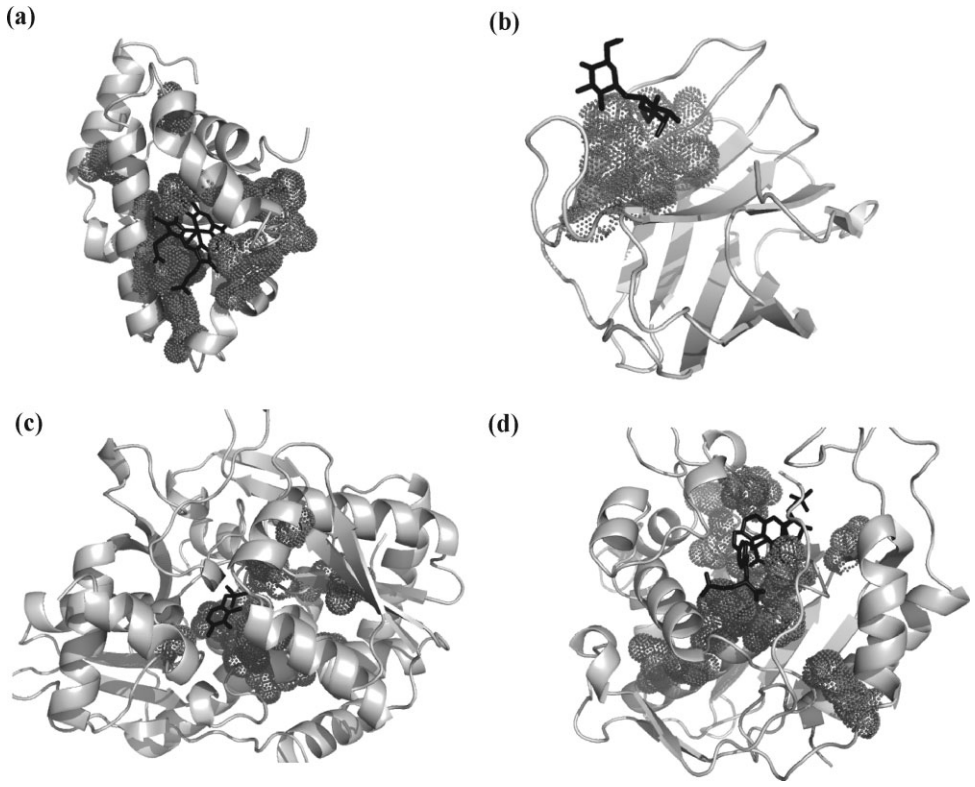

Figure 6

*Examples of functionally important sites and the predicted residues of top high-ranking 20 nodes. Hub and C^I residues are depicted in dark gray dot spheres. Black sticks are ligands. (**a**) SCOP A class example (pdbid 1ch4A) and its ligand heme. The predicted residues are around the ligand heme. (**b**) SCOP B class example (pdbid 1a78A) and its ligand thiogalactosamine. The predicted residues are clustered near the ligand. (**c**) SCOP C class example (pdbid 1a50B) and its ligand pyridoxal phosphate. The ligand is well located on the predicted residues (**d**) SCOP D class example (pdbid 1b02A) and its ligand FdUMP and cofactor CH2H4-folate. The ligands are packed in the most of predicted residues. All pictures are drawn using PyMol.*

cover different area of predictions. Therefore, we expect that by combining our method with the other previous methods we can make better prediction on the protein's functional sites.

Analysis of the worst examples

To analyze what affects the performance, we investigate the relation between MSA quality and the specificities of 44 proteins using linear regression. A few explanatory variables are examined: (i) the mean, (ii) the standard deviation of sequence similarities between query sequence and homologs, (iii) the standard deviation of sequence length differences between query sequence and homologs, (iv) the number of sequences in MSA, (v) the size of the functional set, and so forth. None of them show clear correlation with the S_p, except that the mean of sequence similarities seems to have weak inverse correlation. On the basis of this observation, we have checked how the performance varies when we throw out from the MSAs some highly similar sequences with the higher sequence similarity than a certain cutoff value. We perform this experiment with two worst examples, 1ayu and 1d5cA. It is observed that the specificities of hub and C^I for functional site set increase by 0.2, when the cutoff value is set below 50, suggesting that maintaining sequence diversity in MSAs is important.

Examples of the RRCN analysis

We have shown that the network properties of RRCN are closely related with the functional residues. In Figure 6, four examples of successful RRCN analysis are shown; (a) Class A: Hemoglobin 1ch4A and its ligand heme,[38] (b) Class B: Lectin 1a78A and its ligand thiogalactosamine, (c) Class C: Tryptophan synthase 1a50B and the ligand pyridoxal phosphate, and (d) Class D: Thymidylate synthase 1b02A and its ligand FdUMP and cofactor CH2H4-folate.[39–41] In our examples, the hub nodes and C^I nodes are depicted in dark gray dot spheres. In the case of 1ch4A, the predicted residues are around the ligand heme. With the respect to functional site set the specificities are 0.67 and 0.55, and the sensitivities are 0.15 and 0.23 for C^I and hub, respectively. For the

other SCOP classes, the predicted residues contact well their ligands. The specificities are 0.64 and 0.54 for 1a78A, 0.75 and 0.63 for 1a50B, and 1.0 and 0.33 for 1b02A for the functional sets. The sensitivities are 0.25 and 0.25 for 1a78A, 0.14 and 0.23 for 1a50B, and 0.04 and 0.17 for 1b02A for the sets (STable I). All pictures are made with PyMol.[42]

DISCUSSION

The results we present in this article show that the RRCN based on the coevolutionary relationship between residues is closely related to the functionality of proteins. After we identify some residues that have a special role in the network, we show that those residues are likely to be the functionally important sites. Those special nodes of the network include the nodes with high connectivity (hub nodes) and the nodes with information centrality (C^I) score. Although the importance of the hub nodes has been addressed in many areas of the biological systems, the importance of the C^I nodes has not been gaining much attention. We found only one related publication on the application to the human immune cell network.[43] In this work, we demonstrate that the C^I nodes are more likely to be the functional sites than the hub nodes are. We also show that the nodes being both the hub node and the C^I node simultaneously are most likely to be the functional residues. Comparison between the present method and the previous structure-based method reveals that the overlap between the two methods is low and the specificity is highest for the residues predicted by both methods, indicating that both methods are highly complementary to each other. We also show that our method and the previous sequence-based method, ConSeq, cover different area, and by combining them we can increase the prediction accuracy.

Our approach has several limitations. The critical step in building the RRCN is to determine the threshold score to set the links. Even though we employ a robust method that is not sensitive to the chosen threshold value, it nonetheless affects the final results. In this study, the interaction strengths are calculated by using McBASC algorithm, which requires high-quality MSAs of many diversified homologs to ensure the prediction accuracy, which is not always guaranteed in many cases. It should be noted that despite the main conclusion of this work that the RRCN is closely related with the functional residues, our method at the present form still cannot compete with traditional methods in finding the functional residues in proteins.[44] However, the results from a series of calculations suggest that an alternative view of proteins as a network is quite relevant, and it can give many important clues about the protein function to the researchers who are interested in the relationship between sequences and functions.

ACKNOWLEDGMENTS

The authors thank all Protein BioInformatics Laboratory members for their helpful advice. They also thank the McBASC program of Dr. Fodor and Rate4Site program of Dr. Nir Ben-Tal and Elana Erez. This work was supported in part by a grant from the Ministry of Science and Technology of Korea (Grant #: 2007-03994) and KAIST.

REFERENCES

1. Pazos F, Bang J-W. Computational prediction of functionally important regions in proteins. Curr Bioinformatics 2006;1:15–23.
2. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
3. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 2004;20: 1565–1572.
4. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
5. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
6. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
7. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437:512–518.
8. Marcelino AM, Smock RG, Gierasch LM. Evolutionary coupling of structural and functional sequence information in the intracellular lipid-binding protein family. Proteins 2006;63:373–384.
9. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature 2005;437:579–583.
10. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 2006;15:2120–2128.
11. Chennubhotla C, Bahar I. Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2006;2:36.
12. Thibert B, Bredesen DE, del Rio G. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics 2005;6:213.
13. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.
14. Brinda KV, Vishveshwara S. A network representation of protein structures: implications for protein stability. Biophys J 2005;89: 4159–4170.
15. Muppirala UK, Li Z. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. Protein Eng Des Sel 2006;19:265–275.
16. Terwilliger TC, Zabin HB, Horvath MP, Sandberg WS, Schlunk PM. In vivo characterization of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis. J Mol Biol 1994;236: 556–571.
17. Reddy BV, Datta S, Tiwari S. Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. Protein Eng 1998;11:1137–1145.
18. Taverna DM, Goldstein RA. Why are proteins so robust to site mutations? J Mol Biol 2002;315:479–484.

19. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 2004; 20:1322–1324.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
21. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
22. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 2002;18:77–82.
23. Pei J, Grishin NV. MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 2006;34:4364–4374.
24. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
25. Carter PJ, Winter G, Wilkinson AJ, Fersht AR. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus*). Cell 1984;38:835–840.
26. Horovitz A Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold Des 1996;1:R121–R126.
27. Horovitz A, Serrano L, Avron B, Bycroft M, Fersht AR. Strength and co-operativity of contributions of surface salt bridges to protein stability. J Mol Biol 1990;216:1031–1044.
28. Horovitz A, Fersht AR. Co-operative interactions during protein folding. J Mol Biol 1992;224:733–740.
29. Shi Z, Resing KA, Ahn NG. Networks for the allosteric control of protein kinases. Curr Opin Struct Biol 2006;16:686–692.
30. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
31. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 2004;116:417–429.
32. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R. Allosteric determinants in guanine nucleotide-binding proteins. Proc Natl Acad Sci USA 2003;100:14445–14450.
33. Latora V, Marchiori M. A measure of centrality based on the network efficiency. New Journal of Physics 2007(9):188.
34. Pupko T, Bell R, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18 (Suppl 1):S71–S77.
35. Chandonia J, Hon G, Walker N, Lo Conte L, Koehl P, Levitt M, Brenner S. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32:D189–D192.
36. Fury W, Batliwalla F, Gregersen PK, Li W. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4030165; 2006.
37. Team RDC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2006.
38. Shirai T, Fujikake M, Yamane T, Inaba K, Ishimori K, Morishima I. Crystal structure of a protein with an artificial exon-shuffling, module M4-substituted chimera hemoglobin β α, at 2.5 A resolution. J Mol Biol 1999;287:369–382.
39. Bianchet MA, Ahmed H, Vasta GR, Amzel LM. Soluble β-galactosyl-binding lectin (galectin) from toad ovary: crystallographic studies of two protein-sugar complexes. Proteins 2000;40:378–388.
40. Fox KM, Maley F, Garibian A, Changchien LM, Van Roey P. Crystal structure of thymidylate synthase A from *Bacillus subtilis*. Protein Sci 1999;8:538–544.
41. Schneider TR, Gerhardt E, Lee M, Liang PH, Anderson KS, Schlichting I. Loop closure and intersubunit communication in tryptophan synthase. Biochemistry 1998;37:5394–5406.
42. DeLano WL. The PyMOL molecular graphics system. Available at: http://www.pymol.org; 2002.
43. Tieri P, Valensin S, Latora V, Castellani GC, Marchiori M, Remondini D, Franceschi C. Quantifying the relevance of different mediators in the human immune cell network. Bioinformatics 2005;21: 1639–1643.
44. Chakrabarti S, Lanczycki CJ. Analysis and prediction of functionally important sites in proteins. Protein Sci 2007;16:4–13.