PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Coevolution in defining the functional specificity

Saikat Chakrabarti* and Anna R. Panchenko*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

ABSTRACT

Covariation between sites can arise due to a common evolutionary history. At the same time, structure and function of proteins play significant role in evolvability of different sites that are not directly connected with the common ancestry. The nature of forces which cause residues to coevolve is still not thoroughly understood, it is especially not clear how coevolutionary processes are related to functional diversification within protein families. We analyzed both functional and structural factors that might cause covariation of specificity determinants and showed that they more often participate in coevolutionary relationships with each other and other sites compared with functional sites and those sites that are not under strong functional constraints. We also found that protein sites with higher number of coevolutionary connections with other sites have a tendency to evolve slower. Our results indicate that in some cases coevolutionary connections exist between specificity sites that are located far away in space but are under similar functional constraints. Such correlated changes and compensations can be realized through the stepwise coevolutionary processes which in turn can shed light on the mechanisms of functional diversification.

Proteins 2009; 75:231–240.

Key words: coevolution; correlated mutation; covariation; functional diversification; specificity determinants; subfamily specificity; mutual information; protein evolution.

INTRODUCTION

Coevolution between residues in proteins is a very important factor in the molecular evolution which has not been studied extensively and taken into account in the modeling of evolutionary processes. Coevolution occurs when residues in one site change depending on the residues at another site.[1–3] Many studies aimed at the detection of correlations between protein residues at different sites.[2,4–15] Despite the comparative success in prediction of protein secondary, tertiary structures,[4,6,16,17] and protein interaction partners[18–20] coevolution has been found to be rather weak in many cases, with the strongest signal coming from the sites in alpha helical stacking[2] and from charge compensating covariations.[8] Indeed, coevolution is difficult to detect due to the variable nature of compensatory mutations, the strong dependence of covariations on evolutionary distances, number of sequences in the alignment and residue environment. Moreover, the coevolution signal must be separated from strong background signals resulting from various correlations between the noncoevolving residues.

The apparent covariation between different sites can arise due to several factors. Associations between sites can persist over time as a natural result of a common evolutionary history (phylogenetic linkage). At the same time, structure, folding, and function put certain constraints on the evolvability of the sites that are not directly connected with the common ancestry. For example, sites that are within the proximity from each other in a native structure or come into contact upon folding do not evolve independently from each other as they need to maintain amino acid interactions important for protein stability and foldability. Function is another source of covariability and functionally important sites (FIS) which are not necessarily in direct contact, might covary, because they bind to the same ligand or participate in the allosteric regulation of functional activity. It is very difficult to separate structural and functional signals from the phylogenetic linkage.[21–24] It has been shown that the removal of the background phylogenetic signal improves the contact prediction in some cases although does not have much effect in the others.[21]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS

*Correspondence to: Saikat Chakrabarti, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894. E-mail: chakraba@ncbi.nlm.nih.gov or Anna R. Panchenko, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894. E-mail: panch@ncbi.nlm.nih.gov

Received 7 May 2008; Revised 27 June 2008; Accepted 27 July 2008
Published online 2 September 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22239

The nature of forces which cause residues to coevolve is still not thoroughly understood; especially it is unclear how coevolutionary process is related to functional changes. Some protein functional sites are under stringent evolutionary constraints to preserve their amino acid identity while other functional sites are under less evolutionary pressure. They can mutate in unison with certain sites from the same or a different (interacting) protein to maintain the overall protein functionality. It has been argued that the change in the conservation or evolutionary rate at a particular site is connected with the functional divergence after the gene duplication. After duplication of a gene, its one copy evolves under relaxed evolutionary constraints which allow it to accumulate changes and develop new functions and specificities.[25,26] Two types of functional specificity have been distinguished (Type I and Type II).[27] In the first case, a group of proteins (protein subfamily) develops recognition specificity towards certain types of ligands while maintaining an overall function of a protein family. It results in the conservation of this binding site in one subfamily and its variability in the rest of the family. The second type of specificity occurs when protein subfamilies in parallel evolve different recognition specificities. In this case purifying selection leads to similar levels of site conservation in the subfamilies but different amino acid types. It is very difficult to experimentally detect these *specificity determining sites* and recently new computational algorithms[28–35] have been developed which predict specificity determining sites and provide the foundation for detailed analysis of evolutionary mechanisms of protein specificity.

In this article, we examined the coevolution of specificity determining sites (called "*subsites*" hereafter) that have been annotated in the literature or predicted by algorithms reported earlier.[30,32,35] These sites generally determine the protein specificity either by binding to specific substrate or through interaction with specific protein partner. Our ultimate goal is to understand what changes cause the evolution of new specificity which would require a redesign of the finely tuned interaction network within the protein molecule to maintain structure–function relationship. To achieve this goal, we analyzed how such successive changes and compensations can be realized through the stepwise coevolutionary processes by looking at functional and structural factors that cause covariation of subsites in a protein family.

We showed that specificity determining sites covary more often with each other and other sites compared to sites that are not found under strong functional constraints. We found that individual sites that establish larger number of coevolutionary connections with other sites have a lower variability pointing to slower evolutionary rate. Our findings also suggest possible long range evolutionary coupling mechanisms in some families which might explain the coevolution between subsites that are located far away in space but are under similar functional constraints.

METHODS

Dataset of family alignments and subsites

We have collected reliable alignments of twelve protein families, for majority of which experimental evidence was available on the locations of subsites. The alignments were constructed by existing alignment methods and were subjected to the additional round of careful manual curation. The family alignments were grouped into subfamilies based on different criteria including sequence and structural properties, kinetic properties, substrate specificity, taxonomy, and function (see Supporting Information for details). This dataset with the exception of one protein family was used in our previous study for prediction of subsites.[35] Five of the test families were used for validation of previously published prediction methods and seven other families were taken from the version 2.10 of the Conserved Domain Database (CDD).[36]

Subsites were assigned based on extensive literature search for experimental evidences (some subsites were selected based on consistent predictions made by several methods;[29,31] *actual subsites*, see Supporting Information for details on data collection) and predicted by the SPEER algorithm which was devised in our previous work[35] (*predicted subsites*). SPEER algorithm makes predictions of subsites' locations based on amino acids' physico-chemical properties, evolutionary rate, and combined relative entropy. Altogether our test set contained 97 actual subsites and 180 predicted subsites (only top 15 predicted sites were taken). A complete list of the dataset families together with the number and locations of actual subsites is provided in Table SM1 Supporting Information data. The family alignments and subsite information can be can be also downloaded *via* ftp site ftp://ftp.ncbi. nih.gov/pub/SPEER.

All specificity determining sites were categorized into three groups, Type I, Type II, and marginally conserved (MC) sites.[27,35] Type I sites were defined as those conserved for one subfamily and variable in another while Type II sites were defined as those where different types of amino acids were conserved across different subfamilies. Here, we considered a site to be conserved for one subfamily if any amino acid type is represented more than 75% of the time. The sites that failed to satisfy the above criteria are marked as MC (none of the subfamilies are conserved in this site). For families with more than two subfamilies, sites were categorized into different types based on the category assigned to the majority of subfamily pairs. To distinguish subsites from globally conserved sites we excluded from the subsite set those highly conserved positions within the overall alignment where any amino acid type was represented more than 80% of the time (only one highly conserved subsite was present).

Sites which are globally important for the general function of a family (called "*functionally important sites*"

or FIS thereafter) were extracted from six CDD families. One hundred and eighty-one FIS (Table SM2 Supporting Information) were manually annotated based on literature and experimental data as being catalytic, metal-binding, nucleic acid, or protein binding sites.[36] Representative 3D structures were collected for each family from the PDB database.[37] Spatial distances between two residues (minimum distances between the nearest atoms coordinates) were calculated by in-house scripts utilizing the 3D coordinates supplied in individual PDB files.

Identification of coevolving residue pairs

Mutual information (MI) is a measure of reduction of uncertainty.[38] MI of pairs of positions in the multiple sequence alignments (MSA) has been used successfully to analyze coevolved protein positions in previous studies.[39,40] For this study the Homolmapper program[41] was applied to calculate the normalized MI *Z*-score for each pair of ungapped positions in the multiple sequence alignment. The entropy of a site "*c*" in the alignment was determined as:

$$H_c = -\sum_{i=1}^{20} p(x_i)\log_{20}p(x_i)$$

Here, $p(x_i)$ is the observed frequency of amino acid x_i. The joint entropy H_{cd} (c and d are two columns in the alignment) was calculated by the same method using the frequencies of occurrence of each combination of residues in positions c and d. Joint entropy scores range from the maximum of H_c or H_d to the sum, $H_c + H_d$. MI was calculated as: $MI = H_c + H_d - H_{cd}$, where MI scores range from 0 in the case if amino acid patterns in two columns do not correlate with each other to the minimum of H_c or H_d. Since the greater entropy leads to greater MI values, the raw MI values were normalized by dividing by the joint entropy, H_{cd}, to reduce the influence of entropy on MI.[40] At the same time the MI scores were converted to the *Z*-scores. It has been shown that such normalizations increase the ability of MI to detect coevolved sites and has been implemented in the Homolmapper[41] program. In the current study, two individual sites were considered to be coevolved if the MI-Z score between them was greater than or equal to 3. Selection of a more stringent MI-Z score threshold might result in a poor sensitivity and/or coverage as was ascertained earlier.[39] Additionally, our aim is to emphasize the relative differences in coevolutionary behavior between subsites and other sites and recognize even subtle tendencies toward coevolution.

To quantify the relative tendency of sites to coevolve, *the fraction of coevolution* (FC) for a given site *i* was calculated as: $FC_i = n_i^{coev}/(N-1)$, where n_i^{coev} is the number of sites coevolved with site *i*, *N* is the number of sites in a protein family/alignment, $(N-1)$ is the number of all possible site pair combinations of any given site with all other sites. FC has been calculated between subsites (normalized by possible combinations between all subsites), and between subsites and all other sites.

We have also used alternative algorithms to predict coevolution between protein sites, namely, a continuous-time Markov process (CTMP) algorithm[42] that incorporates phylogenetic information to predict coevolution between two sites. Unfortunately, the current version of the CTMP program was unable to produce results for some of the families in our dataset and the obtained coverage of coevolved sites was too low for any meaningful statistical characterization. Nevertheless, we also applied other available algorithms such as OMES[43,44] that detects differences between observed versus expected frequencies of residue pairs, the McLachlan Based Substitution correlation method[4,45,46] and another MI based method (MIp)[47] that removes background phylogenetic signal which is caused by the phylogenetic relationships rather than by true coevolutionary relationships. Accounting for phylogenetic linkage is very important and it has been shown that MIp, correctly identifies more coevolving positions in protein families than any other methods.

Using different models of simulated protein evolution it has been observed on the whole that in order to identify the majority of true correlation events, the family should be diverse enough and contain more than 16 sequences.[13] It should be mentioned that except for one family (Phosphofructokinase, cd00363) which has just 11 sequence members, all these requirements are satisfied by our dataset (the average sequence identity of families in our dataset ranges from 8 to 60%). Moreover, the overall family diversity is taken into account upon the MI normalization and *Z*-score calculation.[41]

Calculation of evolutionary conservation

Evolutionary conservation at each site in multiple sequence alignment was calculated by weighted sum-of-pairs measure implemented in AL2CO program[48] and by maximum likelihood approach implemented in Rate4-Site[49] program. A maximum likelihood approach (ML) allows one to estimate evolutionary conservation taking into account the topology and branch lengths of the phylogenetic tree and the rate heterogeneity over different sites in a protein family.

RESULTS

Coevolution of specificity determining and other sites

As mentioned in the Introduction, coevolution is one of the various mechanisms which allow proteins to develop new functions and functional specificities. Since

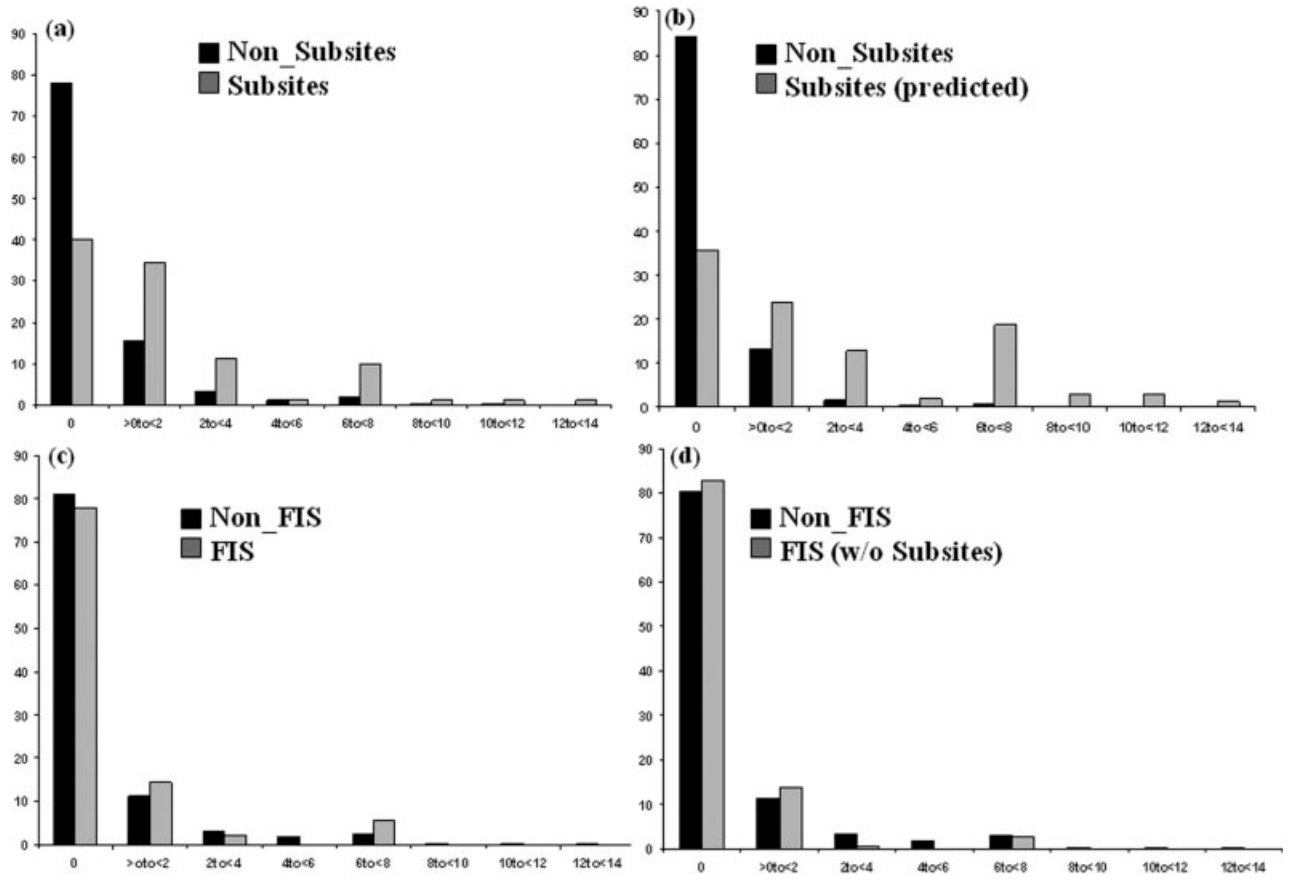

Figure 1

Fraction of coevolution (FC) for subsites and functionally important sites (FIS). Fraction of coevolution is shown on *X*-axes in bins of FC (in % scale). The relative frequency (*Y*-axes, in % scale) for sites with a given FC is shown for different site categories: (**a**) the fraction of coevolution of actual subsites ("Subsite") and non-subsites ("Non_Subsites"); (**b**) FC for top 15 predicted subsites ("Subsites (predicted)") and FC calculated for all other sites that were not predicted as subsites ("Non_Subsites"); (**c**) FC for FIS extracted from seven CDD families and FC for all other sites ("Non_FIS"); (**d**) similar data as panel c after exclusion of actual subsites from the list of FIS ("FIS (w/o Subsites)."

coevolved sites are clearly under selection pressure which is more pronounced for functional sites and specificity determining sites, we analyzed the coevolution between these sites.[39–41] To analyze the coevolution of subsites we calculated the fraction of coevolution (FC; see Methods for definition) for each experimentally annotated and predicted subsite. Subsites were found to coevolve with up to 13% of other sites and the average FC for them is 1.05%. Although this is a relatively low number, the FC of subsites is statistically significantly higher than coevolution of nonspecific sites as indicated by the two-sample *t*-test with *P*-value <0.02 [Fig. 1(a)]. A similar trend of higher FC for actual and predicted subsites compared with other sites is also observed when other coevolution prediction methods, like OMES,[43,44] McBASC,[4,45,46] and MIp[47] were applied to our dataset (Fig. SM1 in Supporting Information).

Subsites predicted by SPEER[35] (we analyzed the top 15 high scoring sites for each family) have an average 3.0% FC [Fig. 1(b)], which points out the ability of SPEER to identify coevolving subsites. The elevated FC for sites predicted by SPEER is consistent with our observation of a higher correlation (Pearson correlation coefficient: 0.54, *P*-value <0.0001) between its SPEER score and FC for any given site (Fig. SM2 in Supporting Information).

We analyzed the coevolution of sites that have been annotated as being functionally important (FIS, Table SM2 in Supporting Information) for the overall protein family [Fig. 1(c,d)] and compared fraction coevolution of those FIS which do not overlap with the subsites with the fractions coevolution of the subsites. FIS show much lower FC (with an average FC of 0.52%) compared with subsites, (the difference is statistically significant with *P*-value <0.001). This can be explained by their invariability and low capacity to change due to the strong functional constraints, although we observed that 10% of FIS actually do exhibit some coevolution of about 2% FC.

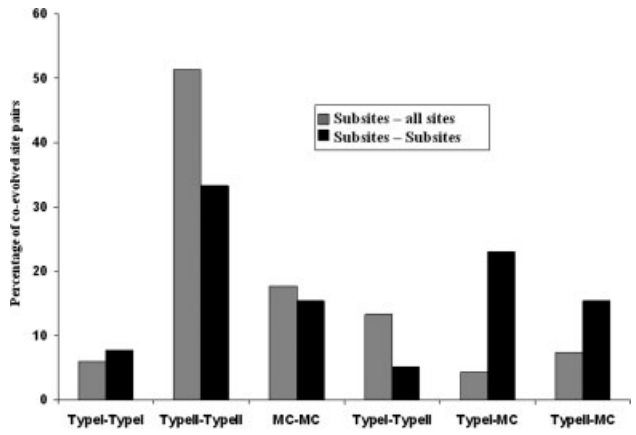

Figure 2

Fraction of coevolution between different types of subsites. All coevolved subsites were categorized into Type I, Type II and MC sites based on their sequence conservation patterns (see Methods) and percentage of each coevolved site pairs from each type category (e.g. Type I to Type I, Type I to Type II) is plotted. Percentage of coevolved pairs is calculated by dividing the number of coevolved subsite–subsite pairs of a given type by the overall number of coevolved subsite–subsite pairs. Black bars show the percent of coevolved site pairs for the actual subsites while grey bars show the percent of coevolved site pairs between the actual subsites and all other sites.

Dissecting subsites into different types shows that different types of subsites (Type I, Type II, and MC) coevolve with each other to different degree (see Fig. 2). Indeed, different subsites can be the result of different mechanisms of functional diversification and might exhibit various degrees of conservation and other properties. For example, Type II sites coevolve the most with other Type II sites although they are also largely connected to many other non-subsites. MC subsites show considerable coevolution with other types of subsites probably due to their frequent occurrence together with the other site types within one domain family.

Relation between sites' covariation and evolutionary conservation

We further examined the FC for the actual and predicted subsites separately for each test family. Among families with the highest FC are the families of cd00985, cd00423, G-protein, cd00363, and GST. At the same time some families (cd00333, cd00120, cd00264, and Ricin) do not exhibit any coevolution for subsites (see Fig. 3). This encouraged us to investigate the correlation of number of coevolutionary connections with respect to evolutionary conservation for all coevolved sites. Evolutionary conser-

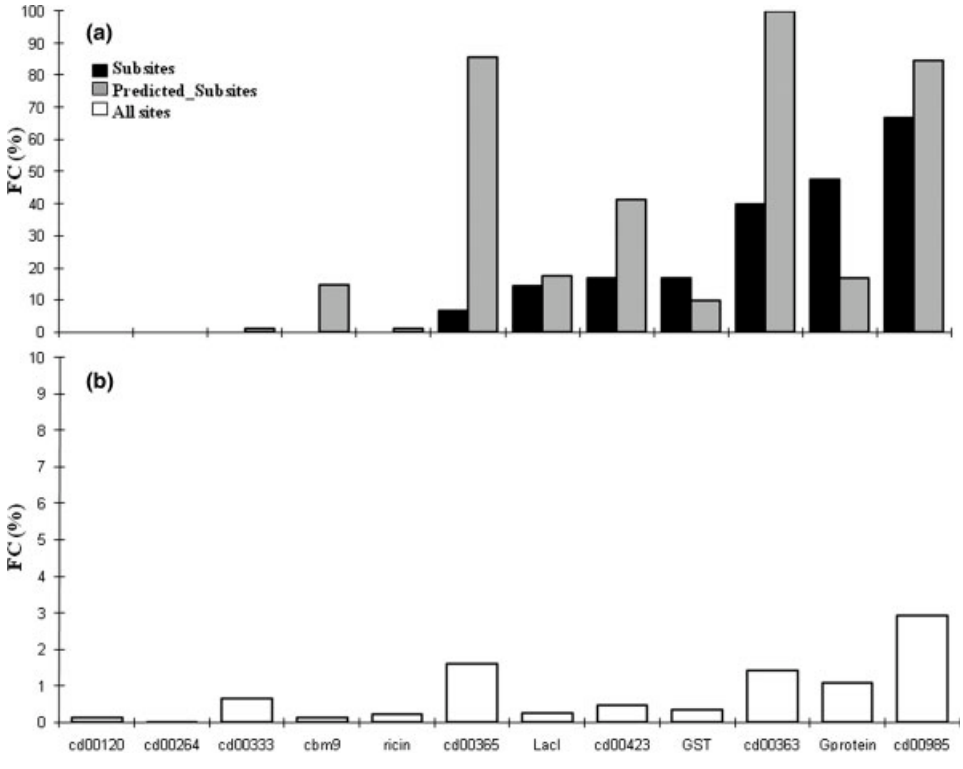

Figure 3

Fraction of coevolution in different families. Average fraction of coevolution is plotted between (**a**) actual subsites (black bar) and between predicted subsites (grey bar; top 15 predicted sites by SPEER[35]) and (**b**) all coevolved sites (open bars).

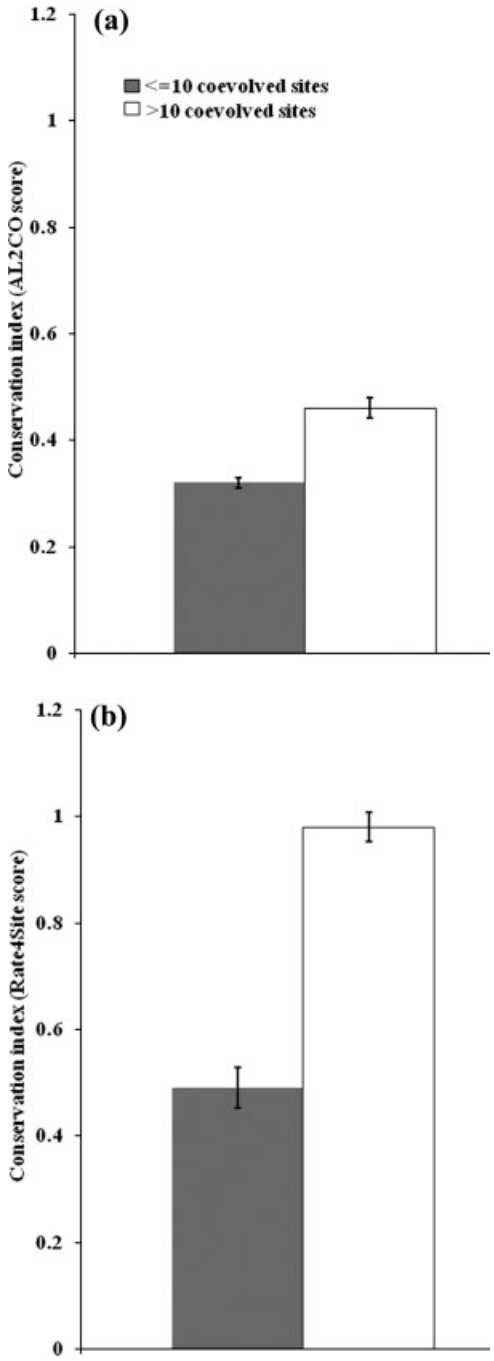

Figure 4

Correlation of evolutionary conservation with the number of coevolutionary connections for a given site. Evolutionary conservation scores were calculated by the (**a**) AL2CO[48] and (**b**) Rate4Site[49] programs where higher values indicate higher conservation (for comparison purpose, Rate4Site scores are projected on a reverse scale). Each box shows the mean value and standard error of evolutionary conservation score for sites with less that 10 (grey box) and more than 10 coevolutionary connections (open box).

vation of each coevolved site was calculated using weighted sum-of-pairs measures (AL2CO[48] scores) and a probabilistic evolutionary model that takes into account the phylogenetic tree and the rate heterogeneity over different sites (Rate4Site[49] scores). Figure 4 shows the mean values of AL2CO[48] and Rate4Site[49] scores together with the standard error for sites that have relatively low ($<=10$) and high (>10) coevolutionary connections. We found that the sites with higher number of coevolutionary connections have a statistically significant tendency to be more conserved (t-test P-value $<10^{-8}$) compared with the sites with the smaller number of connections. These sites might act as 'hub points' and therefore changes in these sites would affect many other connected sites. The same trend is also observed between the number of coevolutionary links and evolutionary conservation for specificity determining sites (Fig. SM3 in Supporting Information).

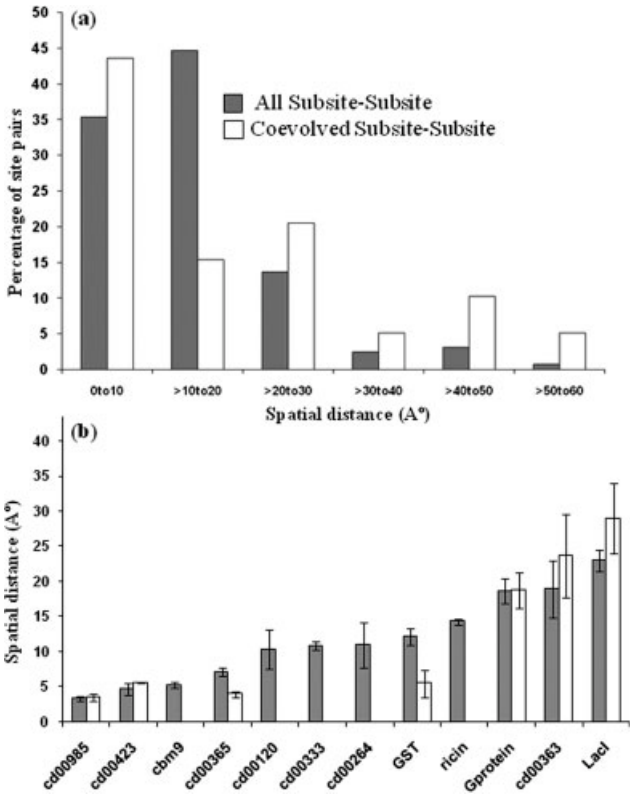

Figure 5

Spatial distance distribution for all subsites and coevolved subsites. Panel (**a**) shows spatial distances between all actual subsites (grey bars) and between coevolved actual subsites (open bars). Panel (**b**) shows average spatial distances (along with standard errors) between all actual subsites (grey bars) and between coevolved actual subsites (open bars) for each family in our dataset.

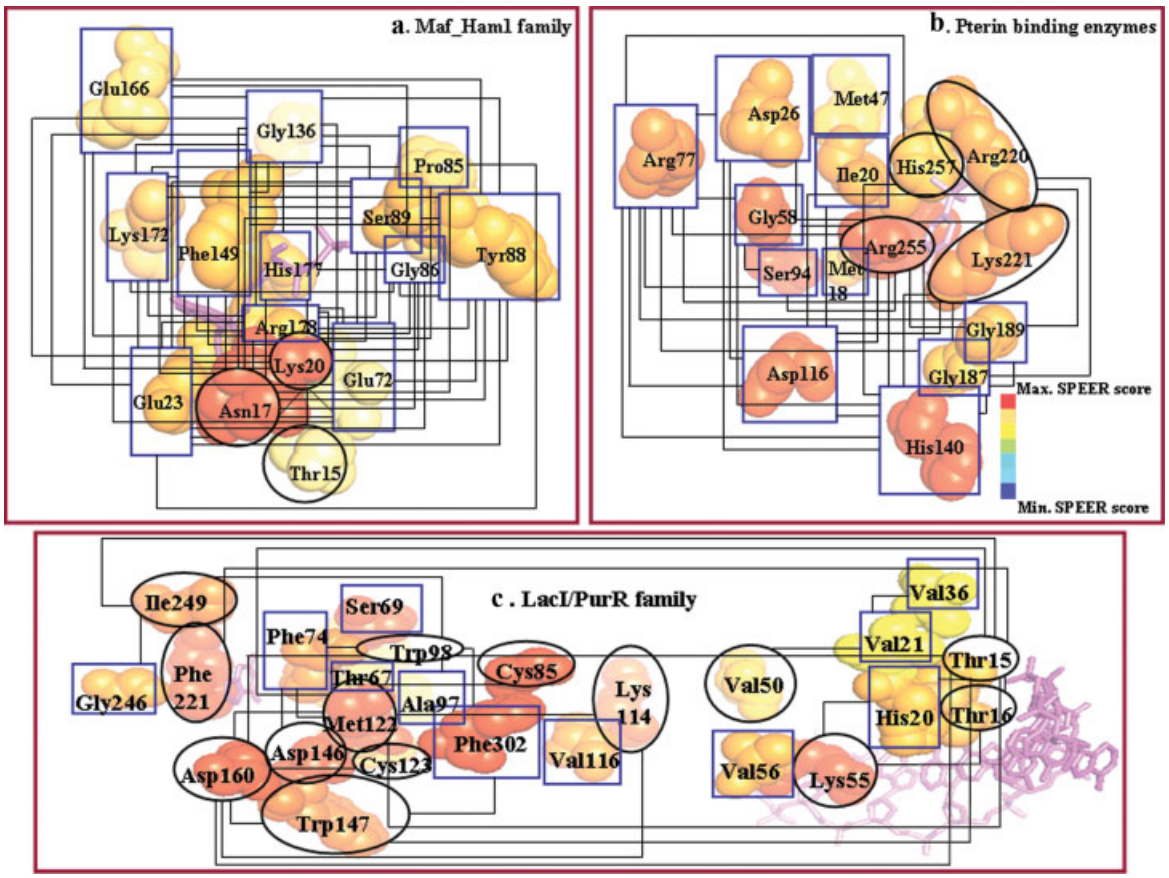

Figure 6

Examples of families with different number of coevolutionary connections among the actual and predicted subsites. Coevolutionary connections between two sites are shown as line while the actual and predicted subsites are outlined as black circles and blue boxes, respectively. Panels a and b show top 15 predicted subsites (including actual subsites), panel c shows top 25 predicted subsites (including actual subsites) which are projected on the representative structures of (**a**) Maf_Ham1 family (cd00985; 2MJP, chain a), (**b**) Pterin binding enzyme family (cd00423; 1AJ0), and (**c**) LacI/PurR family (1WET). Color coding on 3D structures (residues shown as space fill models) indicates the range of SPEER score. Ligands such as ATP analog (panel a), *p*-aminobenzoic acid (pABA) analog (panel b) and guanine and DNA (panel c) are shown in purple (stick representation).

Covariation of subsites with respect to spatial and sequence distance

Intuitively, one would expect that coevolving residues should interact with each other in a protein molecule. A few studies addressed this question and showed a certain tendency for residues exhibiting higher covariation to be structurally coupled.[39,42,65] We have attempted to go further and evaluate the degree of coevolution of specificity determining sites with respect to their spatial distance between each other. There is a very broad distance distribution for all coevolved pairs with the average spatial distance between coevolved sites in proteins from our dataset being 19.6 Å with an average distance along the sequence of 96 residues (see Figs. SM3 and SM4 in Supporting Information for details). It should be noted that structural variability between different structures of the same family is very limited with the average root mean square deviation (RMSD) of less than 2.5 Å and loop (dis)similarity metric[50] of less than 5 Å (with one exception), the average values of structural similarity measures are listed in Table SM1 in Supporting Information. The distance distributions shown in Figure 5(a) for coevolved subsites and overall subsite–subsite distances are statistically different (P-value <0.003). Coevolved subsites show two preferable distance regions: at small distances (less than 10 Å) there are 44% of coevolved subsite pairs compared to 35% of all subsite pairs, whereas at larger distances (more than 20 Å) there are 41% of coevolved subsite pairs compared to only 20% of all subsite pairs. The majority of these distantly located coevolved subsites belong to three families, Gprotein, Phosphofructokinase, and LacI, each of which interacts with multiple ligands and/or proteins at different binding sites. Figure 5(b) also demonstrates a great variability in distance distributions of coevolved subsites for different families.

Three dimensional graphical representations of the spatial locations of actual and predicted subsites of three families are provided in Figure 6. It shows a varying degree of coevolutionary connections within actual and top predicted subsites: forming a very dense network of coevolutionary connections (Maf_HamI) in one case and a sparse network in another (LacI/PurR). However, this graphical representation of coevolutionary connections for actual and predicted subsites emphasizes the observation that some sites can covary despite their large spatial distances if they are selected under similar functional constraints. In the LacI/PurR family the specificity of the transcription factors is regulated by binding to small molecules, such as nucleotides. For example, sites such as 221Phe and/or 160Asp in PurR protein (PDB code: 1WET), involved in nucleotide binding seem to coevolve with the DNA binding sites (15Thr and/or 16Thr) that are located on the opposite sides of the protein molecule.

DISCUSSION

As a result of our analysis, we found that subsites more often participate in coevolutionary connections with each other and other sites compared to FIS or protein sites that are not under strong functional constraints. It was also reported earlier that in about half of the studied proteins (13 out of 25) FIS were near coevolving positions.[42] Unlike FIS which are mostly highly conserved and important for an entire family, specificity determining sites are not globally conserved and subsequent correlated changes at these sites can lead to functional diversification. Interestingly, we found that the networks of coevolutionary connections between the subsites can be quite dense [for example, 87 coevolutionary connections between 15 sites for the Maf_Ham1 family; Fig. 6(a)] and subsites with many coevolutionary connections have a tendency to change slower in evolution compared to subsites with a few connections (this trend is also observed for all other sites). These sites might act as 'hub points' and therefore, could be responsible for establishing new coevolutionary connections or rewiring the existing finely tuned coevolutionary network.

Similar to previous studies performed on all protein sites,[39,42] we observed that almost half of all coevolved subsites are located at distances less than 10 Å which might be the result of coevolution through compensatory mutations, an important mechanism which shapes the finer details of the specificity in evolution. In addition, we found that forty percent of coevolved subsites are located at distances of more than 20 Å, the effect which is most pronounced for three families of Gprotein, Phosphofructokinase, and LacI/PurR. This observation points to the possibility of allosteric mechanisms other than preservation of the stereochemistry of residue contacts. Indeed, several other studies also showed that the residues which are important for the specificity of binding are not necessarily located within a distance of direct interaction with the ligand or other protein,[51–63] although statistical coupling inferred from MSA is not necessarily a true reporter of allosteric changes.[64] Another explanation of long-range coevolutionary coupling between subsites comes from the consideration that many proteins function as homooligomers and sites, which are located far from each other in a monomer, can form contacts between the monomers at nonsymmetrical homodimer interface.

ACKNOWLEDGMENTS

We thank one of the anonymous reviewers for the interesting comment about the possible effect of homodimerization on long-range coupling. The work was supported by Intramural Research program of the National Library of Medicine at NIH.

REFERENCES

1. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng 1997;10:647–657.
2. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 1999;287:187–198.
3. Yanofsky C, Horn V, Thorpe D. Protein structure relationships revealed by mutational analysis. Science 1964;146:1593–1594.
4. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
5. Neher E. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91:98–102.
6. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
7. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Eng 1994;7:341–348.
8. Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. An analysis of simultaneous variation in protein structures. Protein Eng 1997;10:307–316.
9. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
10. Fukami-Kobayashi K, Schreiber DR, Benner SA. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 2002;319:729–743.
11. Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C. Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 2003;328:1061–1069.
12. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180.
13. Pritchard L, Bladon P, Mitchell JMO, Dufton MJ. Evaluation of a novel method for the identification of coevolving protein residues. Protein Eng 2001;14:549–555.

14. Tuff P, Darlu P. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol 2000;17:1753–1759.
15. Dutheil J, Pupko T, Jean-Marie A, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 2005;22:1919–1928.
16. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.
17. Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503.
18. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol 2000;299: 283–293.
19. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 2001;14:609–614.
20. Jothi R, Cherukuri PH, Tasneem A, Przytycka TM. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein–protein interactions. J Mol Biol 2006;362:861–875.
21. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17:164–178.
22. Oliveira L, Paiva AC, Vriend G. Correlated mutation analyses on very large sequence families. Chembiochem 2002;3:1010–1017.
23. Tillier ER, Lui TW. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 2003;19:750–755.
24. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA 2000;97:3288–3291.
25. Ohno S. Evolution by gene duplications. Berlin: Springer-Verlag; 1970.
26. Doolittle RF. Similar amino acid sequences: chance or common ancestry? Science 1981;214:149–159.
27. Gu X. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 1999;16:1664–1674.
28. Hannenhalli SS, Russell RB. Analysis and prediction of functional subtypes from protein sequence alignments. J Mol Biol 2000;303: 61–76.
29. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity- determining residues in bacterial transcription factors. J Mol Biol 2002;321:7–20.
30. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res 2004;32:W424–W428.
31. Pirovano W, Feenstra KA, Heringa J. Sequence comparison by sequence harmony identifies subtype-specific functional sites. Nucleic Acids Res 2006;34:6540–6548.
32. Pei J, Cai W, Kinch LN, Grishin NV. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics 2006;22:164–171.
33. Capra J, Singh M. Characterization and prediction of residues determining protein functional specificity. Bioinformatics 2008; 24:1473–1480.
34. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 2007;8:R232.
35. Chakrabarti S, Bryant SH, Panchenko AR. Functional specificity lies within the properties and evolutionary changes of amino acids. J Mol Biol 2007;373:801–810.
36. Marchler-Bauer A, Anderson JB, Derbyshire MK, Deweese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 2007;35:D237–D240.
37. Henrick K, Feng Z, Bluhm WF, Dimitropoulos D, Doreleijers JF, Dutta S, Flippen-Anderson JL, Ionides J, Kamada C, Krissinel E, Lawson CL, Markley JL, Nakamura H, Newman R, Shimizu Y, Swaminathan J, Velankar S, Ory J, Ulrich EL, Vranken W, Westbrook J, Yamashita R, Yang H, Young J, Yousufuddin M, Berman HM. Remediation of the protein data bank archive. Nucleic Acids Res 2008;36:D426–D433.
38. Cover TA, Thomas JA. Elements of information theory. New York: Wiley; 1991.
39. Gloor GB, Martin LC, Wahl LM. Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44:7156–7165.
40. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics 2005;21:4116–4124.
41. Rockwell NC, Lagarias JC. Flexible mapping of homology onto structure with Homolmapper. BMC Bioinformatics 2007;8:123–137.
42. Yeang CH, Haussler D. Detecting coevolution in and among protein domains. PLoS Comput Biol 2007;3:e211.
43. Larson SM, Di Nardo AA, Davidson AR. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions J Mol Biol 2000;303:433–446.
44. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations Proteins 2002;48:611–617.
45. Olmea O, Rost B Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
46. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments Proteins 2004;56:211–221.
47. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 2008;24:333–340.
48. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;8:700–712.
49. Pupko T, Bell RE, 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.
50. Panchenko AR, Madej T. Analysis of protein homology by assessing the (dis)similarity in protein loop regions. Proteins 2004;57:539–547.
51. 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.
52. Yano T, Oue S, Kagamiyama H. Directed evolution of an aspartate aminotransferase with new substrate specificities. Proc Natl Acad Sci USA 1998;95:5511–5515.
53. Nettles KW, Sun J, Radek JT, Sheng S, Rodriguez AL, Katzenellenbogen JA, Katzenellenbogen BS, Greene GL. Allosteric control of ligand selectivity between estrogen receptors alpha and beta: implications for other nuclear receptors. Mol Cell 2004;13:317–327.
54. Chen Z, Zhao H. Rapid creation of a novel protein functions by in vitro coevolution. J Mol Biol 2005;348:1273–1282.
55. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;63:832–845.
56. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 2004;116:417–429.
57. Gouldson PR, Dean MK, Snell CR, Bywater RP, Gkoutos G, Reynolds CA. Lipid-facing correlated mutations and dimerization in G-protein coupled receptors. Protein Eng 2001;14:759–767.
58. 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.
59. Buck MJ Atchley WR. Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol 2005;22:1627–1634.

60. Daily MD, Upadhyaya TJ, Gray JJ. Contact rearrangements form coupled networks from local motions in allosteric proteins. Proteins 2008;71:455–466.
61. Flynn TC, Swint-Kruse L, Kong Y, Booth C, Matthews KS, Ma J. Allosteric transition pathways in the lactose repressor protein core domains: asymmetric motions in a homodimer. Protein Sci 2003; 12:2523–2541.
62. Taraban M, Zhan H, Whitten AE, Langley DB, Matthews KS, Swint-Kruse L, Trewhella J. Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. J Mol Biol 2008;376:466–481.
63. Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. Nature 2007;450:983–990.
64. Chi CN, Elfström L, Shi Y, Snäll T, Engström A, Jemth P. Reassessing a sparse energetic network within a single protein domain. Proc Natl Acad Sci USA 2008;105:4679–4684.
65. Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. An analysis of simultaneous variation in protein structures. Protein Eng 1997;10:307–316.