Protein Engineering vol. 16 no. 6 pp. 397–406, 2003
DOI: 10.1093/protein/gzg053

Using multiple sequence correlation analysis to characterize functionally important protein regions

Manish C.Saraf, Gregory L.Moore and Costas D.Maranas[1]

Department of Chemical Engineering, The Pennsylvania State University, 112 Fenske Laboratory, University Park, PA 16802, USA

[1]To whom correspondence should be addressed.
E-mail: costas@psu.edu

Protein co-evolution under structural and functional constraints necessitates the preservation of important interactions. Identifying functionally important regions poses many obstacles in protein engineering efforts. In this paper, we present a bioinformatics-inspired approach (residue correlation analysis, RCA) for predicting functionally important domains from protein family sequence data. RCA is comprised of two major steps: (i) identifying pairs of residue positions that mutate in a coordinated manner, and (ii) using these results to identify protein regions that interact with an uncommonly high number of other residues. We hypothesize that strongly correlated pairs result not only from contacting pairs, but also from residues that participate in conformational changes involved during catalysis or important interactions necessary for retaining functionality. The results show that highly mobile loops that assist in ligand association/dissociation tend to exhibit high correlation. RCA results exhibit good agreement with the findings of experimental and molecular dynamics studies for the three protein families that are analyzed: (i) DHFR (dihydrofolate reductase), (ii) cyclophilin, and (iii) formyl-transferase. Specifically, the specificity (percentage of correct predictions) in all three cases is substantially higher than those obtained by entropic measures or contacting residue pairs. In addition, we use our approach in a predictive fashion to identify important regions of a transmembrane amino acid transporter protein for which there is limited structural and functional information available.

Keywords: bioinformatics/directed evolution/functional domain/protein engineering/residue correlation analysis

Background and motivation

The contribution of different protein regions to function is determined by the interactions formed with substrates, cofactors and other residues. Considerable effort has been devoted to identifying functional protein regions from known amino acid sequences (Aloy *et al.*, 2001; Armon *et al.*, 2001; Voigt *et al.*, 2002). When families of sequences with similar structure and function are aligned, it is possible to glean conserved patterns that encapsulate important functional domains (Zvelebil *et al.*, 1987). However, in many cases residues are not conserved but rather co-evolve with their interacting partners to retain important interactions and hence function. Through evolution, many families have diverged substantially, so that it is typically difficult to observe directly any distinct patterns. The inability to identify these functionally important regions (especially those which co-evolve and hence vary in a coordinated manner) poses a long-standing challenge in protein engineering efforts, particularly in the context of directed evolution experiments.

Directed evolution experiments aim at creating diverse sequences with novel properties (e.g. enhanced catalytic activity, stereoselectivity, thermostability) in the form of combinatorial libraries generated by: (i) creating sequence permutations of parent sequences through recombination (Stemmer, 1994a,b; Zhao and Arnold, 1997; Zhao *et al.*, 1998; Ostermeier *et al.*, 1999); and/or (ii) introducing point mutations at either specific positions (Martin *et al.*, 2001) or randomly (Sakamoto *et al.*, 2001). Moore and co-workers proposed modeling frameworks (Moore and Maranas, 2000; Moore *et al.*, 2001) for quantifying the statistics of crossover allocations and mutations in the recombinant sequences. The key problem here is that usually such sequence modifications are not coordinated and, therefore, disrupt functional domains or introduce incompatible interactions (see Figure 1) that frequently results in the loss of their function (Moore and Maranas, 2003). It is therefore desirable to be able to predict these functional sites that are less likely to tolerate uncompensated mutation/crossover, so as to guide these combinatorial libraries towards retaining a higher level of functionality. This objective defines the scope of this study.

Multiple sequence alignment (MSA) and entropy calculations provide some insight into identifying protein building blocks preserved through evolution. However, there are currently no reliable techniques for identifying regions that may subtly affect functionality by taking part in large numbers of interactions that co-evolve under functional or structural constraints. There have been a few studies towards developing methods to identify these functional sites. Studies conducted by Voigt and co-workers (Voigt *et al.*, 2001, 2002) propose that contact between residue pairs can be considered to represent interaction between them. Therefore, residues that are in contact with many other residues participate in numerous interactions and hence are unlikely to tolerate uncompensated mutations. While this is an interesting hypothesis, many studies suggest that even distant residues may interact strongly through a network of related interactions and/or through electrostatic forces (Cannon *et al.*, 1997; Fisher *et al.*, 1998; Gong *et al.*, 2000; Radkiewicz and Brooks, 2000). Lichtarge and Sowa proposed an alternative approach in which they identify functional sites by mapping tertiary structures of sequences that form the nodes near the root of the evolutionary tree (Lichtarge and Sowa, 2002). Spatially clustered residues are assumed to be functionally important since changes in the amino acid composition of these regions are linked with evolutionary divergences and, hence, functional specificity. Similarly, work by Landgraph *et al.* identified residues with

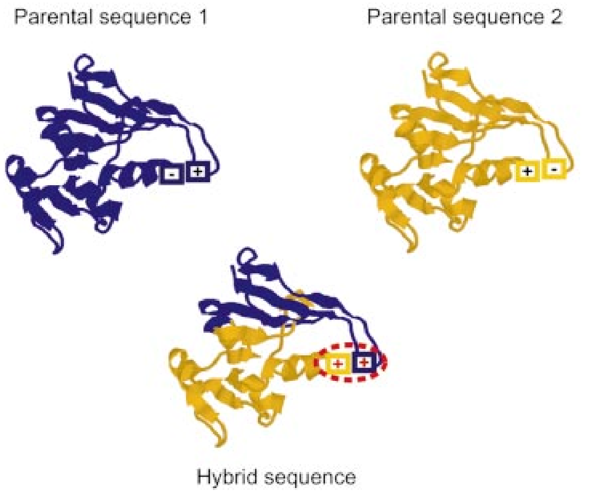


Fig. 1. Formation of a repulsive ion pair in a recombinant hybrid that may disrupt contacting pairs in addition to essential motions.

conserved structural neighbors and residue clusters that have high sequence similarity (Landgraf *et al*., 2001). Both studies, however, provide little insight into which residue pairs are interacting. Moreover, many of the variable, functional loop regions may not map on to the corresponding loop of other members of the family on the static molecule, but may have related motion during the catalytic cycle, thereby playing an important role in ligand binding and dissociation.

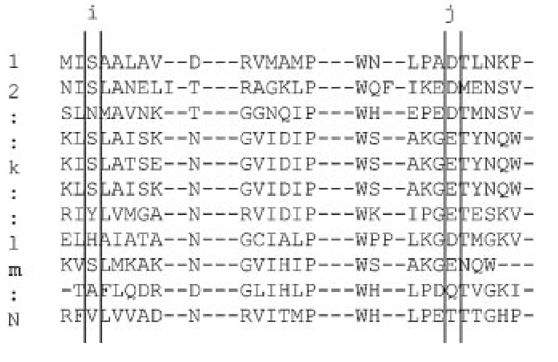
Studies have shown (Baldwin *et al*., 1993; Shindyalov *et al*., 1994) that changes in protein properties are brought about by cumulative effects of many small adjustments, many of which are propagated over significant distances in the three-dimensional structure. Trace evidence of such coordinated mutations brought about by evolution are present in the protein sequence data of the members of a protein family. It has been postulated that a substitution at one position is compensated by a substitution elsewhere in the sequence to ensure that structural features essential for the functioning of the protein are conserved (Lim and Ptitsyn, 1970; Altschuh *et al*., 1987; Gobel *et al*., 1994; Shindyalov *et al*., 1994). In fact, studies have revealed that residues distant in sequence but near in three-dimensional space undergo simultaneous compensatory variation to conserve their overall physiocochemical properties (Lesk and Chothia, 1980; Oosawa and Simon, 1986; Altschuh *et al*., 1987; Lim and Sauer, 1989; Bordo and Argos, 1990; Baldwin *et al*., 1993). This hypothesis has been used to predict residue contact maps by identifying correlated mutations (Gobel *et al*., 1994; Neher, 1994; Taylor and Hatrick, 1994) whose signals can be strengthened by comparing neighboring nodes in the phylogenetic tree (Fukami-Kobayashi *et al*., 2002). Despite the compelling logic behind this hypothesis, these studies have met with only limited success. The estimated accuracy of statistical contact prediction has at best been only 15–20% (Olmea *et al*., 1999).

We hypothesize that strongly correlated residue pairs do not necessarily have to be in contact; rather, they may affect the dynamics of protein function by participating in a network of distal motions involved in catalysis or by participating in important interactions. Generally, the motions of protein regions are associated with ligand binding/dissociation involved in catalysis. Our hypothesis is based on the assumption that uncompensated mutations in these regions will disrupt their motion/interaction and therefore attenuate important chemical steps in catalysis, affecting reaction rates by several orders of magnitude (Miller and Benkovic, 1998; Osborne *et al*., 2001). This leads us to postulate that regions that are involved in the dynamics of the reaction or in an uncommonly high number of interactions with other residues are likely to tolerate only coordinated mutations. For example, if a lysine in a loop region is replaced by a glutamine, it may be necessary to substitute a glutamine by a lysine at a position elsewhere so that the net charge remains the same, ensuring that no essential motions are disrupted due to charge repulsion (Figure 1). To detect these functionally important regions, we introduce the use of the *correlation tendency* metric to quantify the average number of other residues to which a particular residue/region is correlated. In this paper we demonstrate that highly mobile regions of the protein exhibit high correlation tendency values. This occurs mainly due to the many additional physical and functional contacts (i.e. through hydrogen bond, van der Waals interactions and long-range electrostatic interactions) that these regions make during their motion associated with catalysis (Miller *et al*., 2001).

To test our hypothesis, we performed residue correlation analyses (RCA) on three protein families: (i) dihydrofolate reductase (DHFR), (ii) cyclophilin and (iii) formyl-transferase. These families were chosen based on the differences in the degree of sequence alignment and conservation and availability of structural and functional data. In addition, we used our approach in a predictive fashion to identify important regions of a transmembrane amino acid transporter protein for which there is little structural and functional information available. The Pfam database (Bateman *et al*., 2002) was accessed to download protein family sequence data. RCA is comprised of two major steps: (i) identifying pairs of positions whose mutations occur in a coordinated manner and (ii) using these results to identify protein regions that interact with an uncommonly high number of other residues.

Residue correlation analysis (RCA)

Protein chemists discovered early on that certain residue substitutions commonly occur in homologous proteins from different species (Dayhoff *et al*., 1978). Because the protein retains its functionality after these substitutions, the substituted residues are either compatible with the protein structure and function or else the effects of these substitutions are compensated by some other changes (Lesk and Chothia, 1980; Oosawa and Simon, 1986; Altschuh *et al*., 1987; Lim and Sauer, 1989; Bordo and Argos, 1990; Baldwin *et al*., 1993). Since these substitutions are coordinated, there exists a measurable correlation between these mutation patterns (Gobel *et al*., 1994; Neher, 1994; Taylor and Hatrick, 1994). However, measuring amino acid variability requires the use of a metric of similarity that will reflect how likely one residue is to be substituted by another. Numerous methods have been suggested such as utilizing physicochemical vectors describing residue physical properties [e.g. side chain volume (Grantham, 1974), charge (Taylor and Hatrick, 1994), hydrophobicity (Levitt, 1976)] and similarity matrices that codify empirical information from phylogenetic trees [such as BLOSUM (Henikoff and Henikoff, 1992) and PAM (Dayhoff *et al*., 1978)]. PAM250, BLOSUM62 and McLachlan (McLachlan, 1971) scoring matrices were used in our study to compute the



```
        i                                   j
1   MISAALAV--D---RVMAMP---WN--LPADTLNKP-
2   NISLANELI-T---RAGKLP---WQF-IKEDMENSV-
:   SLNMAVNK--T---GGNQIP---WH--EPEDTMNSV-
:   KLSLAISK--N---GVIDIP---WS--AKGETYNQW-
k   KISLATSE--N---GVIDIP---WS--AKGETYNQW-
:   KLSLAISK--N---GVIDIP---WS--AKGETYNQW-
:   RIYLVMGA--N---RVIDIP---WK--IPGETESKV-
l   ELHAIATA--N---GCIALP---WPP-LKGDTMGKV-
m   KVSLMKAK--N---GVIHIP---WS--AKGENQW---
:   -TAFLQDR--D---GLIHLP---WH--LPDQTVGKI-
N   RFVLVVAD--N---RVITMP---WH--LPETTTGHP-
```

Fig. 2. The scores X_{ikl} and X_{jkl} are obtained from one of the similarity matrices [PAM250, BLOSUM62 or McLachlan (McLachlan, 1971)] for positions i, j corresponding to the residues at these positions in the sequences k and l. These residues (i, j) are reported to be correlated if correlation coefficient value (r_{ij}) is above a threshold value ($r_c = 0.4$).

correlation coefficient. Results obtained for the three scoring systems were very similar (data not shown) and the one selected here is the McLachlan scoring matrix. Alternatively, for identifying the functional coupling of two positions of the MSA, Lockless and Ranganathan used vectors of 20 binomial probabilities of individual amino acid frequencies (Lockless and Ranganathan, 1999). These probabilities are determined by the distribution of residues in each column of the alignment. Clearly, the sequence alignment obtained by randomly shuffling the residues in each column of the original alignment would yield identical results even though the resulting sequences may be significantly different from the parental sequences. Furthermore, since no metric of similarity between residues has been utilized in the above method, less frequent but conservative substitution patterns will not be recognized. In another recent study by Larson *et al.*, correlation signals are identified based on the probability of occurrences of residue pairs rather than by use of scoring matrices (Larson *et al.*, 2000). Clearly, some conservative substitutions (i.e. substitution of a residue with another that is very similar in physical and chemical properties) cannot be detected by this method. Hence, the use of similarity matrices has the advantage that it can detect conservative substitution patterns more accurately than other methods.

Highly correlated pairs may arise due to: (i) physical contact between them, (ii) distal interaction, (iii) interaction through conformational changes, or (iv) occurrence by chance. In our analysis, predictions made based on RCA are assumed to be correct if the residues of the correlated pair are interacting through conformational changes or through distal interactions. Inter-residue distances are calculated for identifying contacting residues. Various cutoffs have been proposed in the literature to define contacting pairs. Two residues are said to be in contact if the distance between them is below a given arbitrary threshold. These distances could be the distance between the two β-carbons (Cβ) (Thomas *et al.*, 1996; Lund *et al.*, 1997; Olmea and Valencia, 1997; Fariselli *et al.*, 2001) or the two α-carbons (Cα) (Vendruscolo *et al.*, 1997) of the corresponding residues. The average of the distances between all the atoms of the two residues or the distance between the nearest atoms belonging to the side chain or the backbone of the two residues (Fariselli and Casadio, 1999) have also been used in these definitions. Here, we consider a pair of residues to be contacting if the Cβ–Cβ (or Cα–Cα in the absence of Cβ) distance is <8 Å.

Residue correlation coefficients

The family of aligned sequences obtained from the Pfam database is assumed to be a randomly chosen sample of a population of all functional protein sequences. Correlation coefficients between any two positions (Figure 2) are calculated similarly to the method proposed by Gobel *et al.* (Gobel *et al.*, 1994). For a given pair of sequences (k, l), each substitution at a position (i or j) is associated with a similarity score (X_{ikl} and X_{jkl}, respectively) obtained from the McLachlan scoring matrix. The expression used for computing the correlation coefficient (r_{ij}) between two sequence positions (i, j) in the alignment is

$$r_{ij} = \frac{2}{N(N-1)} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left(\frac{X_{ikl} - <X_i>}{\sigma_i} \right) \left(\frac{X_{jkl} - <X_j>}{\sigma_j} \right) \quad (1)$$

where σ_i and σ_j are the standard deviations of the scores X_{ikl} and X_{jkl} at positions i and j about their means $<X_i>$ and $<X_j>$, respectively. The use of weights in computing the correlation coefficient has been avoided since they not only penalize genuinely correlated signals in a group of similar sequences but often cannot be quantified in a universal fashion. Studies also indicate that sequence weighing is not an important factor in achieving high accuracy in the covariation signal (Larson *et al.*, 2000). To prevent the correlation coefficient from being biased by over-represented groups of similar sequences, we eliminate combinations of pairs of sequences that have repeated patterns in the corresponding columns. For example, in computing the correlation coefficient between two positions i and j of the alignment shown in Figure 2, the k–m sequence combination is not considered. Note that the denominator of Equation 1 [i.e. $N(N - 1)/2$ = the number of combinations of sequences] is adjusted accordingly. Positions that have a high percentage of gaps (>70%) are omitted to avoid misleading results due to the small amount of data available at these alignment positions. Furthermore, the results depend on how well the sequences align and hence a few lengthy (or too short) sequences are deleted from the alignment to avoid introduction of excessive gaps. The remaining sequences are then realigned with CLUSTALW (Higgins and Sharp, 1988; Higgins *et al.*, 1996) using BLOSUM/PAM matrices.

Two positions are considered to be correlated if the absolute value of the correlation coefficient between the two is above a threshold value ($r_c = 0.4$) (Gobel *et al.*, 1994). The significance of the cutoff value is tested by performing a correlation analysis on a sequence alignment obtained by: (i) randomly shuffling the residues of each row of the alignment (i.e. each sequence still retains the same residues), thereby destroying the existing secondary structure elements, and (ii) randomly shuffling the residues in each column of the alignment. In either case it was observed that no residue pair yielded a correlation coefficient even close to, let alone above, the threshold value. Residue pairs with correlation coefficients above the threshold value are identified that are then used in computing the correlation tendencies of protein regions as outlined below. These pairs are also investigated for contacts to estimate the percentage of the correlated pairs that are contacting. Inter-residue distances <8 Å are considered to be contacting. For this purpose, the Euclidean distance between

Table I. Summary of statistical analyses for the three protein families (all values are shown as percentages)

Protein family	Specificity				Sensitivity		
	RCA[a]	Entropy[b]	Contacting pairs[c]	Random[d]	RCA	Entropy	Contacting pairs
DHFR	90	71	50	55	82	45	36
Cyclophilin	33	31	25	17	50	100	50
Formyl-transferase	56	50	38	24	83	100	50

Results are shown based on: [a]RCA: $r_{ij} > r_c$, $t_m > 1$; [b]domain entropy; [c]contacting pairs, average number of contacts per residue for a domain > overall residue average; and [d]random choice of domains.

two residues ($|r_i - r_j|$) is calculated from the coordinates obtained from the Protein Databank (PDB) database.

Correlation tendencies

The residue correlation coefficient (r_{ij}) is a measure of the relationship between the scores of two sequence alignment positions. However, it is more informative to identify protein regions (i.e. a contiguous string of residues) that show strong correlation with a relatively large number of other residues. Residues adjacent in the sequence are contacting and, therefore, identifying these signals provides no additional information. Thus, the correlation of the position under consideration with the adjacent three residues in the sequence is not taken into account. In general, the correlation signals are fairly noisy and therefore it is difficult to glean useful information from them. The noise level in the correlation data is reduced by averaging out these effects over secondary structure elements to calculate the correlation tendencies. The correlation tendency (t_m) of a segment m is defined as the ratio (x_m) of the number of correlated pairs with at least one of the residues in region m to the total number of correlated pairs, that is scaled by the ratio of its length l_m to the total sequence length L:

$$t_m = \frac{x_m}{l_m / L} \qquad (2)$$

Because each residue position of the segment is weighted by its frequency of occurrence in the correlated set (a set including all correlated residue pairs), the correlation tendency reflects the frequency of interactions in which these residues are engaged.

The protein segments, for the purpose of calculating correlation tendency values, are determined based on the secondary structure. Sequence alignment with no existing secondary structure elements (i.e. alignment obtained by randomly shuffling the columns of the original MSA without disrupting the relative order of the residues in the column) yielded t_m values close to 1. Hence, a correlation tendency value >1 is considered to be significant. Regions with t_m values >1 are identified and located on the three-dimensional structure of the corresponding protein molecule. These are then investigated for functional roles known from experimental and molecular dynamics studies.

Site entropy

Highly conserved positions do not carry correlation information and are not considered in the correlation analysis, but they do contain useful information with respect to functionality. Hence, in addition to identifying strongly correlated pairs, it is important to measure variability at residue sites to identify conserved regions. A widely used measure of site variability is the site entropy (S_i) that is calculated using the expression

$$S_i = -\sum_{a=1}^{20} p_a \log_2 p_a \qquad (3)$$

where p_a is the probability of occurrence of an amino acid a in the column i of the aligned sequences. The domain entropy, S_m, is derived by averaging positional entropy of residues in the domain m.

Entropy and correlation capture different statistical properties of family sequence data; therefore, we investigated whether correlation analysis captures information not accessible by simple residue variability measures such as entropy and whether the two can be used in conjunction to predict functional domains better. Prediction of functional sites are made based on RCA ($t_m > 1$) and entropic measures ($S_m <$ average sequence entropy). Functional sites have also been identified using contacting residues as a representation of all interacting residues (Voigt *et al.*, 2002). For this purpose, a similar analysis was performed using contacting residues as was done for correlated residues in calculating correlation tendency. The accuracy of prediction of functional sites by these methods is expressed in terms of *sensitivity* (the fraction of true functional sites identified by prediction) and *specificity* (the fraction of the predicted domains that form the functional sites) that are presented in Table I. These are compared with the specificity of identifying a functional site by random choice of a protein domain (the ratio of the number of functional domains (included as secondary structure elements) to the total number of secondary structure elements present in the protein sequence) that are also included in the same table. A detailed description of the results for the three protein families is presented next.

Computational results and comparisons

Correlation analyses are performed on the protein families of dihydrofolate reductase (DHFR), cyclophilin and formyl-transferase. These families include protein members that have different levels of alignment and conservation. Cyclophilin and DHFR families include sequences that are closely related (maximum average tree distance of 16.54 and 25.8, respectively) and therefore align fairly well, whereas the formyl-transferase family (maximum average tree distance of 30.89) contains distant sequences that do not align well. In addition, unlike the DHFR (average entropy = 2.300) and formyl-transferase (average entropy = 2.563) families, the cyclophilin family is much more conserved (average entropy = 1.908). In comparison with the cyclophilin and formyl-transferase families, much more is known about DHFR and its functional domains.

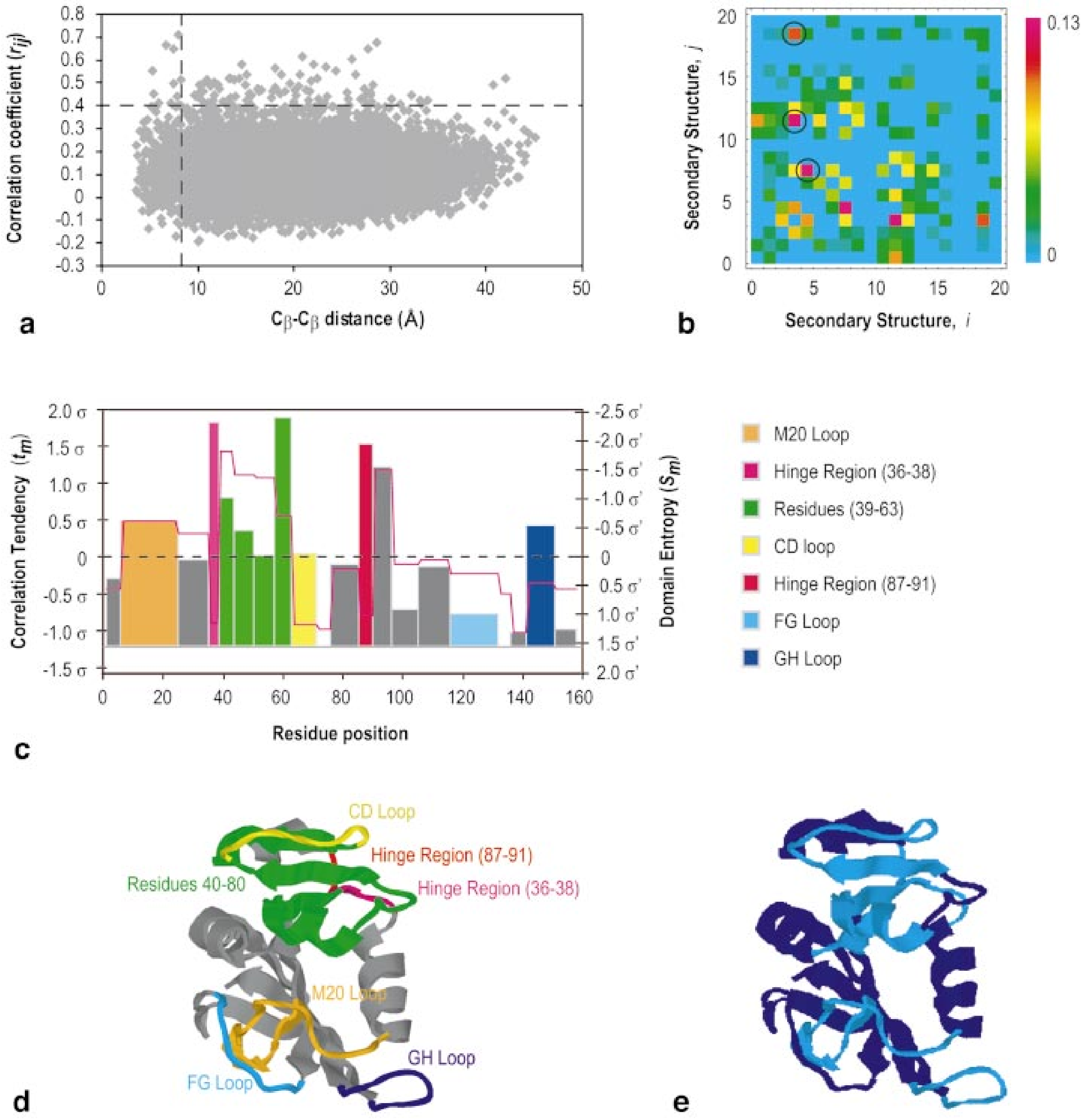


Fig. 3. (**a**) Plot of residue correlation coefficients (r_{ij}) versus Cβ–Cβ distances (calculated from the crystal structure of 1DDR) for pairs of residues in the DHFR enzyme. The vertical broken line partitions pairs that are in physical contact (i.e. inter-residue distance <8 Å), while the horizontal broken line indicates the cutoff value ($r_c = 0.4$) above which the pairs are considered to be strongly correlated. Note that many of the correlated pairs are not contacting and many of the contacting pairs are not correlated. (**b**) Plot of the average correlation coefficient between residues of various secondary structure elements of DHFR. Strongest correlation signals are detected between the two hinge regions, the CD loop and the GH loop (shown as circled regions). (**c**) The correlation tendency (at $r_c = 0.4$) for different segments of the DHFR enzyme based on its secondary structure is plotted against residue position. The zero line in the figure indicates the average correlation tendency (1) and the entropy (2.3). The colored bars represent correlation tendency values for different regions that include residues involved in motions during catalysis or important interactions as found through experimental and molecular dynamics simulation studies. The graph also shows the average entropy (red line) for these domains on the secondary axis. Note that the values are reversed on the secondary axis (i.e. peaks are conserved regions) for easy comparison with results obtained through RCA. The two axes are scaled based on their standard deviations (σ = 0.8, σ′ = 0.62) about their means (1 and 2.3, respectively). (**d**) Regions in the DHFR enzyme that are functionally important (as known from experimental and simulation studies) are highlighted in color. (**e**) Light-blue regions correspond to regions having t_m values >1, whereas dark-blue regions imply lower t_m values indicating less than average participation in the correlated set (PDB ID: 1DDR).

Dihydrofolate reductase

DHFR is an enzyme that is necessary for maintaining intracellular levels of tetrahydrofolate, an active form of the vitamin folic acid and an essential cofactor in the synthetic pathway of purines, pyrimidines and several amino acids. It catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) using nicotinamide adenine dinucleotide phosphate (NADPH) as a cofactor. X-ray crystallographic studies indicate that the members of the DHFR family contain an eight-stranded β-sheet and four α-helices interspersed with loop regions that connect these secondary structures (Figure 3d). Analysis of the DHFR complex with folate has revealed that isolated residues exhibit diverse backbone fluctuations on the nanosecond to picosecond time-scale. The most significant motions are observed in the M20 loop (residues 7–24), the neighboring FG loop (residues 116–132), the GH loop (residues 142–150) (Miller and Benkovic, 1998;

Table II. Residue pairs, their role in catalysis and correlation coefficients

Residue pair	Role	Correlation coefficient
13–121	The residues in this pair belong to the M20 and the FG loops which are hydrogen bonded to each other. Mutation in these loops affects the catalytic rate by 400-fold	0.710
53–104	Mutation of this pair diminishes the rate by a factor of $\geqslant$6. It lines the active site, implying that mutation alters the binding site geometry	0.649
60–42	These amino acids are involved in strong hydrogen bonding with each other	0.646
42–113	Residues 113 and 27 are hydrogen bonded to DHF and residues 42 and 60 are hydrogen bonded to NADPH.	0.616
59–113		0.605
60–113		0.535
28–42		0.494
21–122	A hydrogen bond between them stabilizes the closed conformation of the M20 loop. Conformation changes of the M20 loop regulate ligand binding	0.518

Radkiewicz and Brooks, 2000; Miller *et al.*, 2001; Agarwal *et al.*, 2002), the distant CD loop (residues 64–71) and the hinge region connecting the two subdomains (residues 87–91) (Falzone *et al.*, 1990; Lau and Gerig, 1997; Radkiewicz and Brooks, 2000) (see Figure 3c). Motions detected in the region between residues 40 and 80 are strongly anti-correlated to the fluctuations in the M20 and FG loops (Radkiewicz and Brooks, 2000). Fluctuations in these loops play crucial roles in the catalytic pathway; for example, conformational changes in the M20 loop may limit the rate of THF dissociation. Mutational studies reveal that only specific residue substitutions are permitted in these loops. The replacement of four M20 loop residues with a glycine results in a 500-fold decrease in the rate of hydride transfer and similar effects are observed for mutations in the FG loop ~17 Å from the active site (Li *et al.*, 1992; Miller and Benkovic, 1998). In addition, mutations in the NADPH (residues 42, 60) and DHF/THF (27, 113) binding regions have drastic effects.

The RCA analysis of the DHFR family includes 122 sequences accessed from the Pfam database. The scatter plot (Figure 3a) between r_{ij} values and Cβ–Cβ distances in Angstroms outlines the proximity of the correlated pairs in three-dimensional space. *Clearly most of the correlated pairs are not contacting and many of the contacting pairs are not correlated*. Of the 105 correlated pairs identified (r_c = 0.4), only 9.52% of them are contacting whereas contact by random choice of pairs has a likelihood of 3.4% alone.

Figure 3b shows the average correlation coefficient between the residues of various secondary structure elements. It has been observed that a strong correlation exists between functionally important regions. Particularly the strongest correlation signals (shown in circles in Figure 3b) are detected between the two hinge regions, the CD loop and the GH loop. Correlation tendency values for different segments (based on secondary structure) were calculated as described earlier. A cutoff value (r_c) of 0.4 resulted in a high specificity of 90% whereas the corresponding sensitivity is 81.8%. Of the 20 segments into which the DHFR enzyme is divided, 11 are functionally important, resulting in a likelihood of only 55% that a randomly chosen region (secondary structure) is functionally important. Our study shows that t_m values are >1 for almost all of the mobile loop regions (except for the FG loop) whereas the converse holds for regions outside these loops (Figure 3c and e). High positional entropy is observed for two important hinge regions (36–38 and 87–91) and for the CD and GH loops, indicating that these regions are not conserved. Correlation tendency values, however, show that residue changes at these positions are highly coordinated. The entropy calculations (Figure 3a) result in a specificity of 71.4% and a sensitivity of only 45%. Evidently, entropy alone does not capture functionality information clearly discernible with the correlation analysis. Furthermore, most of the t_m values in the region between residues 40 and 80 are >1 with an overall average of 1.22. Agreement with these results suggests that the proposed correlation analysis is indeed capturing information related to distal motions during catalysis. Low entropy and high correlation tendency values are observed for residues 91–96 and 40–60, indicating that these regions are fairly conserved and limited changes at these positions are coordinated (Figure 3c). Interestingly, even though functional roles of the residues 91–96 are unknown, residues 40–60 (a subset of the region 40–80) have been observed to fluctuate during catalysis (Radkiewicz and Brooks, 2000). A comparison of prediction results obtained by RCA, entropy measures, contacting pairs and random choice is summarized in Table I.

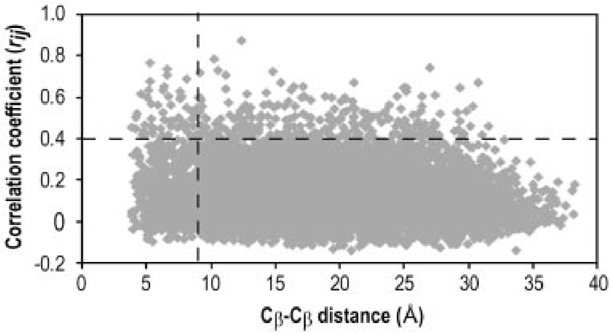


Fig. 4. Plot of residue correlation coefficients (r_{ij}) versus Cβ–Cβ distances (calculated from the crystal structure of 1RMH) for pairs of residues in the cyclophilin protein. The vertical broken line partition pairs that are in physical contact (i.e. Cβ–Cβ distances <8 Å), while the horizontal broken line indicates the cutoff value (r_c = 0.4) above which the pairs are considered to be strongly correlated.

A number of DHFR studies have delineated the role that specific residue pairs play during catalysis (Miller and Benkovic, 1998; Radkiewicz and Brooks, 2000; Osborne *et al.*, 2001; Agarwal *et al.*, 2002). Table II contrasts these results with a few of the pairs that are identified with the correlation analysis. Remarkably, for most of these pairs we find exceptionally high correlation coefficients, further

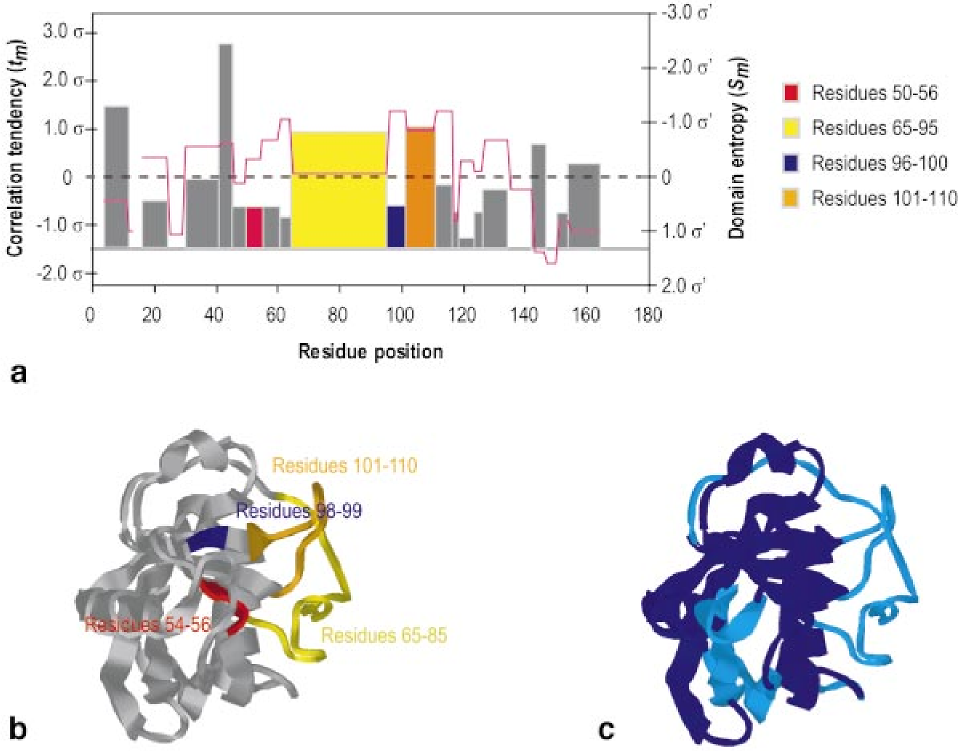


Fig. 5. (**a**) The correlation tendency plot (at $r_c = 0.4$) for the cyclophilin enzyme based on its secondary structure. The zero line in the figure indicates the average value for both the correlation tendency (1) and the domain entropy (1.908). The correlation tendencies of functionally important regions are represented by colored bars. The red line corresponds to the domain entropy shown on the secondary axis. As in the case of DHFR, the secondary axis is reversed to represent highly conserved regions as peaks. The two axes are scaled based on their standard deviations (σ = 0.68, σ′ = 0.73). (**b**) Loop regions in cyclophilin protein that are in motion during catalysis (as known from experimental studies) are highlighted in color. (**c**) Light-blue regions identify domains with t_m values >1, whereas dark-blue regions imply lower t_m values (PDB ID: 1RMH).

strengthening the hypothesis that important information regarding function can be recovered from protein family sequence data through residue correlation analysis.

Cyclophilin

Cyclophilin is a binding protein for the immunosuppressive drug cyclosporin and also an enzyme with *cis–trans* isomerase activity. It catalyzes the interconversion between *cis* and *trans* conformations of X–Pro peptide bonds, where X could be any amino acid. Studies have indicated that internal protein dynamics are intimately connected with enzyme catalysis that influences the substrate turnover (Eisenmesser *et al*., 2002). As in the case of DHFR, rapid fluctuations are observed in the loop regions of cyclophilin during catalysis. Significant conformational exchange dynamics were observed in the residue regions 54–56 and the loops 65–80 and 101–110 (Eisenmesser *et al*., 2002), as shown in Figure 5b. Furthermore, a narrow pass separates the two loops that provide a possible location of the extended substrate binding (Kallen and Walkinshaw, 1992). Studies indicate that residues L98 and S99, during the catalytic cycle, interact with the *trans* peptide while the *cis* isomer binds near residues 55, 82, 101–103 and 109 (Zhao and Ke, 1996; Eisenmesser *et al*., 2002).

For the RCA analysis of cyclophilin, 304 sequences were downloaded from the Pfam database. Correlation coefficients (r_{ij}) for all pairs are calculated as described earlier and are plotted against Cβ–Cβ distances as shown in Figure 4. A cutoff of 0.4 is chosen to identify strongly correlated residues, resulting in the selection of 310 pairs as members of the correlated set of which only 12.16% are contacting. The correlation tendency plot (Figure 5a) indicates that of the three mobile regions mentioned above, two have t_m values significantly >1. This results in a high specificity of 33% and a sensitivity of 50%, whereas the likelihood of identifying important regions based on random choice is only 17.4%. These predictions are in good agreement (see Figure 5b and c) with the results obtained through NMR relaxation experiments

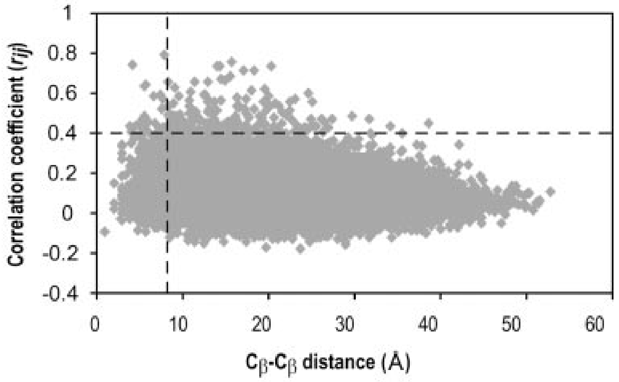


Fig. 6. Plot of correlation coefficient versus Cβ–Cβ distance (calculated from the crystal structure of 1GRC) for pairs of residues in the formyl-transferase protein family. The vertical broken line partitions pairs that are in physical contact and the horizontal broken line indicates the cutoff value (r_c = 0.4).

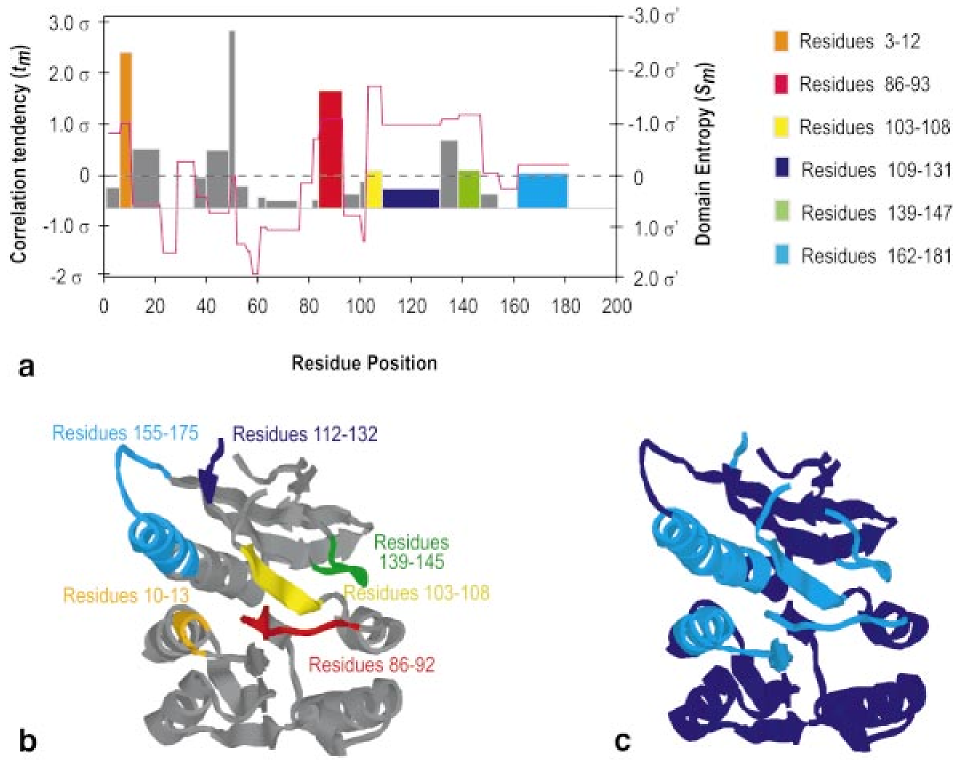


Fig. 7. (**a**) The correlation tendency plot (at $r_c = 0.4$) of the formyl transferase family based on its secondary structure. The zero line in the figure represents both the average correlation tendency (1) and the average entropy (2.563). The correlation tendencies of functionally important regions (identified through experimental and simulation data) are shown as colored bars, while the red line corresponds to the domain entropy shown on the secondary axis. Values on the secondary axis are in reverse order for easy comparison between the results obtained from RCA and entropy measurements. The two axes are scaled based on their standard deviations (σ = 1.42, σ′ = 0.54). (**b**) Functionally important regions of formyl-transferase (known from experimental and simulation studies) are highlighted in color. (**c**) Light-blue regions correspond to regions with t_m values > 1, whereas dark-blue regions have lower t_m values (PDB ID: 1GRC).

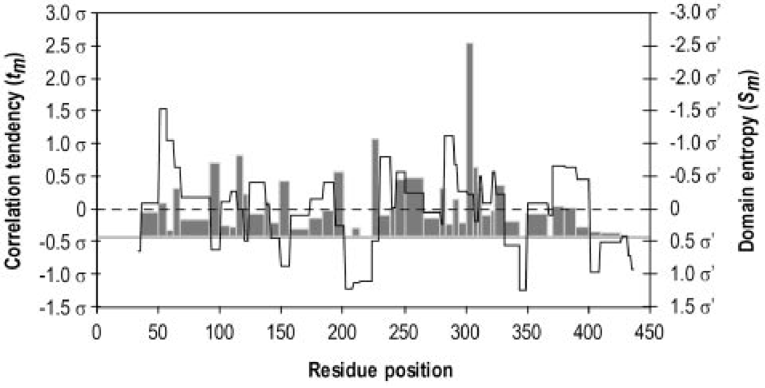


Fig. 8. The correlation tendency plot (at $r_c = 0.4$) for different segments of the transmembrane amino acid transporter protein based on the predicted secondary structure. The zero line in the figure corresponds to the average value for correlation tendency (1) and also for the domain entropy (2.843). The dark line corresponds to the domain entropy shown on the secondary axis. As in the other cases, values on the secondary axis are in reverse order. The two axes are scaled based on their standard deviations (σ = 1.26, σ′ = 0.61).

conducted by Eisenmesser *et al.* (Eisenmesser *et al.*, 2002). The motions in the loops are associated with *cis* isomer binding and therefore also include residues necessary for interacting with the *cis* isomer. The average entropy of the loop 65–95, where the most prominent motion is observed, is very close to the overall average, clearly indicating that the loop region is not highly conserved. Residues 98 and 99, even though they are functionally important, did not show high t_m values. However, these positions are well conserved as low entropy values are observed at these positions (Figure 5a). In addition, regions 4–11, 42–45 and 143–146 show coordinated mutations resulting in correlation tendency values >1, suggesting that they may be functionally important and hence require further investigation. Table I summarizes the statistical analyses carried out for the cyclophilin family.

Formyl-transferase

Glycinamide ribonucleotide transformylase (GART) catalyzes the transfer of a formyl group from 10-formyltetrahydrofolate to glycinamide ribonucleotide (GAR), a reaction in the purine biosynthetic pathway. The GART structure can be subdivided into two sub-domains with the N-terminal domain consisting of a central core of smoothly twisted, parallel β-sheet of four strands surrounded on both sides by two pairs of α-helices (Figure 7b). The C-terminus consists of the remaining six β-sheets with a long α-helix (Almassy *et al.*, 1992; Chen *et al.*, 1992). X-ray structure analysis of a ternary complex with the substrate GAR has confirmed that the phosphate group of GAR is tightly bound by the loop consisting of residues 10–13 (Almassy *et al.*, 1992; Klein *et al.*, 1995). The terminal three oxygen atoms of the phosphate interact primarily with the main-chain NH groups of residues 11, 12 and 13, while the fourth phosphate oxygen is within hydrogen-binding distance of the NH of the Gln170 side chain (Almassy *et al.*, 1992; Chen *et al.*, 1992). The ribose hydroxyl group interacts with residues in the α-helix (162–185), while the rest of the sugar lies in a hydrophobic pocket formed by residues 86, 88, 107, 121, 166 and 171. The key residues in the active site are 103, 106, 108 and 144 and these are highly conserved as reflected by their average entropy shown in Figure 7a. Study of a high-resolution structure of GART with a multisubstrate adduct has revealed that residues 112–132, 140–145 and 155–175 are highly mobile (Klein *et al.*, 1995) (see Figure 7b). Mutations at position 144, a part of the loop with highest mobility and also a part of the active site, resulted in an enzyme that was 10^4 times less active than the wild-type (Inglese *et al.*, 1990). The folate derivative binds in the hydrophobic pocket formed by the residues 85, 88, 92, 197, 104 and 139 where it forms six hydrogen bonds interacting with residues 90, 92, 140, 141 and 144 (Klein *et al.*, 1995; Morikis *et al.*, 2001).

The formyl-transferase family includes the following members: (i) GART, (ii) formyltetrahydrofolate deformylase, and (iii) methionyl-tRNA formyl-transferase. In total, 169 sequences were downloaded from the Pfam database, of which seven had large insertions and, hence, were removed. The MSA includes the first 181 residues of the reference sequence (glycinamide ribonucleotide transformylase from *Escherichia coli*; PDB ID: 1GRC). As observed in the previous two cases, most of the correlated pairs in formyl-transferase are not contacting (Figure 6). A cutoff value of 0.4 captures 207 pairs as members of the correlated set that results in a high specificity of 55.6%. Note that the corresponding specificity for entropy measures and contacting pairs are only 50 and 37.5%, respectively. Figure 7a clearly shows that all the regions mentioned above, including residues 41–51 and 132–138, exhibit high correlation tendency values, with the only exception being residues 112–132. However, entropy values show that the loop 112–132 includes residues that are highly conserved, but the low correlation tendency indicates that most mutations observed in this region are not coordinated. Residues 132–138, similar to the loop 112–132, are fairly conserved but are also involved in numerous interactions, as suggested by the observed high correlation tendency. Comparisons between prediction results obtained by various methods are shown in Table I. The sensitivity value obtained by domain entropy is 100% compared with 83.3% of that of RCA, while contacting pairs perform poorly with a value of only 50%. Although most of the functionally important regions (especially the residues 8–13, 86–93, 103–108, 139–147 and 162–181) show low entropy, they are also discernible by RCA, further strengthening the point that correlation analysis does identify regions with mutations that are highly coordinated.

Transmembrane amino acid transporter protein

The common element of the three protein families for which we carried out the RCA analyses is that they all have substantial structural and functional information available. Here we consider a protein family for which there are very limited structural and functional data. The transmembrane amino acid transporter protein family (further referred to as the permease family) mainly includes proline and amino acid permeases that are integral membrane proteins involved in the transport of amino acids into the cell. The multiple sequence alignment obtained from the Pfam database consisted of 167 sequences with the reference sequence (in our study) as the amino acid permease [PID: g152(7493)]. The total length of the sequence is 446 residues; however, the sequence alignment part includes only residues 34–437.

Correlation tendency values are calculated for segments based on predicted secondary structure. The widely used methods for protein structure prediction based on neural network methods [HNN and PROF (Ouali and King, 2000), accessible at http://www.expasy.ch/] are utilized. The correlation tendency values along with the average entropy for different regions are shown in Figure 8. A high degree of conservation is observed at positions 51–65, 233–239, 283–291 and 370–381. Residue 64, which is in the conserved region and adjacent to the completely conserved glycine (residue 65), shows a relatively high entropy. Extremely high correlation tendency and correlation coefficient values corresponding to residue 64 suggests that mutation at this position is largely coordinated with other mutations. Similarly to residue 64, positions 96–98, 114, 154, 194, 195, 225, 226, 247–267 and 304–307 include regions that are strongly correlated to a fairly large number of other residues resulting in high correlation tendency values. Nevertheless, positional entropy values show that these regions are not highly conserved, implying that possible mutations need to be coordinated, thus requiring other adjustments. It has been observed that almost all of the regions with high correlation tendency, as in the other three cases, lie in the predicted loop regions that are likely to be crucial to protein dynamics during catalysis.

Summary and discussion

In this paper, a computational framework for identifying protein regions that are correlated to a large number of other residues is proposed. Residue correlation analyses were performed on three protein families: (i) DHFR, (ii) cyclophilin, and (iii) formyl-transferase. It was shown that residues/regions that have a high correlation tendency are either involved in important interactions or participate in conformational changes necessary for retaining function. RCA and entropy calculations were used to identify the protein building blocks that co-evolve under structural and functional constraints. Predictions of these methods were tested against experimental and molecular dynamics data available in the literature. Table I summarizes prediction results obtained by RCA, entropic measures, contacting pairs and random choice. The DHFR case study demonstrated that entropy alone cannot capture functionality information clearly discernible with RCA. However, in the other two cases, sensitivity values achieved by the entropy measures are 100%. A close examination of these two cases reveals that functionally important regions not found by RCA are those which have relatively low entropy, leading to a weakened correlation signal not detected by RCA. No clear relationship between entropy and correlation tendency was detected. It was observed that many of the conserved functional sites have high correlation tendency values reflecting the conservative nature of mutation patterns in these regions. However, we also found a number of regions, particularly the functional loops that are highly variable, with high correlation tendency values. Contacting pairs captured less information about important interactions than the RCA approach and entropic measures for all three protein families. This confirms that strongly correlated pairs and *not* contacting pairs are better descriptors for identifying interacting residues. Combined results of the RCA and entropy measures identified all of the functionally important regions (except for the FG loop in the DHFR family) in the three families. Hence, RCA and entropic measures collectively form a better technique to identify functionally important regions.

In the current implementation, the correlation analysis examines the relationship among the leaves of the phylogenetic tree. However, one would expect that correlation signals would be stronger between the nodes of the tree at shorter evolutionary distances and a number of groups have observed this while attempting to detect contacting residues (Shindyalov *et al.*, 1994; Fukami-Kobayashi *et al.*, 2002). We are currently developing methods for further refining the correlation calculations presented here by measuring signals at the tree nodes.

Acknowledgements

We thank Professor Stephen Benkovic and Professor Sharon Hammes-Schiffer for helpful discussions. Financial support from National Science Foundation Award BES0120277 and hardware support from the IBM-SUR program are gratefully acknowledged.

References

Agarwal,P., Billeter,S., Rajagopalan,P., Benkovic,S. and Hammes-Schiffer,S. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 2794–2799.

Almassy,R., Janson,C., Kan,C. and Hostomska,Z. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 6114–6118.

Aloy,P., Querol,E., Aviles,F. and Sternberg,M. (2001) *J. Mol. Biol.*, **311**, 395–408.

Altschuh,D., Lesk,A., Bloomer,A. and Klug,A. (1987) *J. Mol. Biol.*, **193**, 693–707.

Armon,A., Graur,D. and Ben-Tal,N. (2001) *J. Mol. Biol.*, **307**, 447–463.

Baldwin,E., Hajiseyedjavadi,W. and Matthews,B. (1993) *Science*, **262**, 1715–1718.

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S., Griffiths-Jones,S., Howe,K., Marshall,M. and Sonnhammer,E. (2002) *Nucleic Acids Res.*, **30**, 276–280.

Bordo,D. and Argos,P. (1990) *J. Mol. Biol.*, **211**, 975–988.

Cannon,W., Garrison,B. and Benkovic,S. (1997) *J. Am. Chem. Soc.*, **119**, 2386–2395.

Chen,P., Schulze-Gahmen,U., Stura,E., Inglese,J., Johnson,D., Marolewski,A., Benkovic,S. and Wilson,I. (1992) *J. Mol. Biol.*, **227**, 283–292.

Dayhoff,M., Schwartz,R. and Orcutt,B. (1978) *Atlas Protein Sequence Struct.*, **5**, 345–352.

Eisenmesser,E., Bosco,D., Akke,M. and Kern,D. (2002) *Science*, **295**, 1520–1523.

Falzone,C., Benkovic,S. and Wright,P. (1990) *Biochemistry*, **29**, 9667–9677.

Fariselli,P. and Casadio,R. (1999) *Protein Eng.*, **12**, 15–21.

Fariselli,P., Olmea,O., Valencia,A. and Casadio,R. (2001) *Proteins*, **5**, 157–162.

Fisher,B., Schultz,L. and Raines,R. (1998) *Biochemistry*, **37**, 17386–17401.

Fukami-Kobayashi,K., Schreiber,D. and Benner,S. (2002) *J. Mol. Biol.*, **319**, 729–743.

Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins*, **18**, 309–317.

Gong,X., Wen,J., Fisher,N., Young,S., Howe,C., Bendall,D. and Gray,J. (2000) *Eur. J. Biochem.*, **267**, 3461–3468.

Grantham,R. (1974) *Science*, **185**, 862–864.

Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Higgins,D. and Sharp,P. (1988) *Gene*, **73**, 237–244.

Higgins,D., Thompson,J. and Gibson,T. (1996) *Methods Enzymol.*, **266**, 383–402.

Inglese,J., Smith,J. and Benkovic,S. (1990) *Biochemistry*, **29**, 6678–6687.

Kallen,J. and Walkinshaw,M. (1992) *FEBS Lett.*, **300**, 286–290.

Klein,C., Chen,P., Arevalo,J., Stura,E., Marolewski,A., Warren,M., Benkovic,S. and Wilson,I. (1995) *J. Mol. Biol.*, **249**, 153–175.

Landgraf,R., Xenarios,I. and Eisenberg,D. (2001) *J. Mol. Biol.*, **307**, 1487–1502.

Larson,S., Di Nardo,A. and Davidson,A. (2000) *J. Mol. Biol.*, **303**, 433–446.

Lau,E. and Gerig,J. (1997) *Biophys. J.*, **73**, 1579–1592.

Lesk,A. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.

Levitt,M. (1976) *J. Mol. Biol.*, **104**, 59–107.

Li,L., Falzone,C., Wright,P. and Benkovic,S. (1992) *Biochemistry*, **32**, 7826–7833.

Lichtarge,O. and Sowa,M. (2002) *Curr. Opin. Struct. Biol.*, **12**, 21–27.

Lim,V. and Ptitsyn,O. (1970) *Mol. Biol. (USSR)*, **4**, 372–382.

Lim,W. and Sauer,R. (1989) *Nature*, **399**, 31–36.

Lockless,S. and Ranganathan,R. (1999) *Science*, **286**, 295–299.

Lund,O., Frimand,K., Gorodkin,J., Bohr,H., Bohr,J., Hasen,J. and Brunak,S. (1997) *Protein Eng.*, **10**, 1241–1248.

Martin,A., Sieber,V. and Schmid,F. (2001) *J. Mol. Biol.*, **309**, 717–726.

McLachlan,A. (1971) *J. Mol. Biol.*, **61**, 409–424.

Miller,G. and Benkovic,S. (1998) *Biochemistry*, **37**, 6327–6335.

Miller,G., Wahnon,D. and Benkovic,S. (2001) *Biochemistry*, **40**, 867–875.

Moore,G. and Maranas,C.D. (2000) *J. Theor. Biol.*, **205**, 483–503.

Moore,G. and Maranas,C.D. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 5091–5096.

Moore,G., Maranas,C., Lutz,S. and Benkovic,S. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 3226–3231.

Morikis,D., Elcock,A., Jennings,P. and McCommons,J. (2001) *Protein Sci.*, **10**, 2379–2392.

Neher,E. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

Olmea,O. and Valencia,A. (1997) *Fold. Des.*, **2**, S25–S32.

Olmea,O., Rost,B. and Valencia,A. (1999) *J. Mol. Biol.*, **295**, 1221–1239.

Oosawa,K. and Simon,M. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 6930–6934.

Osborne,M., Schnell,J., Benkovic,S., Dyson,H. and Wright,P. (2001) *Biochemistry*, **40**, 9846–9859.

Ostermeier,M., Nixon,A., Shim,J.-H. and Benkovic,S. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 3562–3567.

Ouali,M. and King,R. (2000) *Protein Sci.*, **9**, 1162–1176.

Radkiewicz,J. and Brooks,C.L. (2000) *J. Am. Chem. Soc.*, **122**, 225–231.

Sakamoto,T., Joern,J., Arisawa,A. and Arnold,F. (2001) *Appl. Environ. Microbiol.*, **67**, 3882–3887.

Shindyalov,I., Kolchanov,N. and Sander,C. (1994) *Protein Eng.*, **7**, 349–358.

Stemmer,W. (1994a) *Nature*, **370**, 389–391.

Stemmer,W. (1994b) *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.

Stemmer,W. and Soong,N. (2000) *Nat. Biotechnol.* **18**, 1279–1282.

Taylor,W. and Hatrick,K. (1994) *Protein Eng.*, **7**, 341–348.

Thomas,D., Casari,G. and Sander,C. (1996) *Protein Eng.*, **9**, 941–948.

Vendruscolo,M., Kussell,E. and Domany,E. (1997) *Fold. Des.*, **2**, 295–306.

Voigt,C., Mayo,S., Arnold,F. and Wang,Z.-G. (2001) *J. Cell. Biochem. Suppl.*, **37**, 58–63.

Voigt,C., Martinez,C., Wang,Z., Mayo,S. and Arnold,F. (2002) *Nat. Struct. Biol.*, **9**, 553–558.

Zhao,H. and Arnold,F. (1997) *Nucleic Acids Res.*, **25**, 1307–1308.

Zhao,H., Giver,L., Shao,Z., Affholter,J. and Arnold,F. (1998) *Nat. Biotechnol.*, **16**, 258–261.

Zhao,Y. and Ke,H. (1996) *Biochemistry*, **35**, 7356–7361.

Zvelebil,M., Barton,G., Taylor,W. and Sternberg,M. (1987) *J. Mol. Biol.*, **195**, 957–961.

Received December 1, 2002; revised February 7, 2003; accepted April 15, 2003