

Determination of network of residues that  
regulate allostery in protein families using  
sequence analysis

Ruxandra I. Dima<sup>a</sup> D. Thirumalai<sup>a,b,c</sup>

<sup>a</sup>*Biophysics Program*

*Institute for Physical Science and Technology*

<sup>b</sup>*Department of Chemistry and Biochemistry*

*University of Maryland, College Park, MD 20742, USA*

<sup>c</sup>*thirum@glue.umd.edu; Phone: 301-405-4803; Fax: 301-314-9404*

---

**Abstract****Abstract:**

Allosteric interactions between residues that are spatially apart and well separated in sequence are important in the function of multimeric proteins as well as single domain proteins. This observation suggests that, among the residues that are involved in long-range communications, mutation at one site should affect interactions at a distant site. By adopting a sequence-based approach, we present an automated approach, which uses a generalization of the familiar sequence entropy in conjunction with coupled two way clustering algorithm, to predict the network of interactions that trigger allosteric interactions in proteins. We use the method to identify the subset of dynamically important residues in three families, namely, the small PDZ family, G protein-coupled receptors (GPCR), and the Lectins which are cell-adhesion receptors that mediate the tethering and rolling of leukocytes on inflamed endothelium. For the PDZ and GPCR families our procedure predicts, in agreement with previous studies, a network containing a small number of residues that are involved in their function. Application to the Lectin family reveals a network of residues interspersed throughout the C-terminal end of the structure that are responsible for binding to ligands. Based on our results and previous studies we propose that functional robustness requires that only a small subset of distantly connected residues be involved in transmitting allosteric signals in proteins.

---

**Introduction**

Long range communications among a network of residues, that are far apart in sequence and in structure, are crucial for biological functions. The classic example is the allosteric communication in which binding of a ligand to

a specific region of a protein often triggers large conformational changes in a distant part (Monod et al., 1965). Starting from the well studied case of oxygen binding to hemoglobin (Perutz et al., 1998), large scale domain movements, in response to ligand binding, have been noted in other systems. For example, binding of ATP and the co-chaperonin GroES to the oligomeric *E. coli* chaperonin triggers dramatic rigid body motions in different subdomains of GroEL (Xu and Sigler, 1998; Horovitz et al., 2001). Similarly, examination of the structures of DNA polymerases and the dNTP-polymerase-DNA complexes shows evidence of such large movements (Steitz, 1999). In the case of polymerases, whose structure is analyzed using the right hand metaphor, the initial step involving binding of the unliganded polymerase to DNA triggers the thumb to close around the DNA. Subsequent binding of dNTP to the binary complex results in the rotation of the fingers from the open conformation to a closed catalysis-ready state. These and other examples show that the set of residues that induce the long range allosteric communications in oligomeric biological nanomachines may well be encoded in the structures and hence, in the sequence itself.

More recently, it has been argued that allosteric communication may well be a part of the observed dynamical fluctuations in small single domain proteins (Kern and Zuiderweg, 2003). Many NMR experiments have shown that even after reaching the native state proteins undergo conformational fluctuations with time scales from several nanoseconds to milliseconds. Surprisingly, it has been suggested that such functionally important fluctuations are triggered by long-range interactions among a network of residues. Just as with multimeric proteins these functionally important sets of residues are also encoded in the structures.

The above mentioned, rather ubiquitous examples, naturally raise an important question, namely, how can we identify the network of residues that mediate allosteric communication? There have been several distinct approaches to answer this question each with varying degree of success. Although they differ significantly in detail the methods are either based on probing changes in the thermodynamics and dynamics of proteins with known structures or using probabilistic methods to analysis of evolutionary related sequences. The double mutant cycles, which probe the responses of specific sequence pairs to mutations (Schreiber and Fersht, 1995), can be used to obtain the thermodynamics of interactions between the mutated residues. It is also possible to probe interactions among residues by examining the dynamics of specific residues using NMR (Kern and Zuiderweg, 2003). Recently, we introduced a method that monitors the response of a region of a folded protein, represented using the Elastic Network Model, in response to a perturbation at another site (Zheng et al., 2005). The application of the method to the polymerase family successfully identified a network of dynamically relevant residues that are involved in the open/closed transition.

In contrast to the relatively few structure-based methods, numerous techniques that exploit the properties of families of sequences have been developed to infer correlations between amino acids in protein families. A sequence-based method (Lesk and Chothia, 1980; Altschuh et al., 1987; Neher, 1994; Taylor and Hatrick, 1994; Pollock and Taylor, 1997; Pazos et al., 1997; Olmea et al., 1999; Valencia and Pazos, 2002; Lichtarge et al., 1996; Sowa et al., 2001; Kass and Horovitz, 2002) is desirable because a larger database of evolved sequences with related functions can be studied. It is logical to postulate that the distribution of amino acid types at a given position in a multiple sequence alignment

(MSA) is the manifestation of evolutionary changes under constraints imposed by function. In addition, it is likely that for functional reasons co-evolution of a network of residues in a sequence also occurs. If so, such correlations should appear as statistically significant signals when analyzing a MSA. Using sequence correlation entropy (SCE) (Dima and Thirumalai, 2004), which does not involve the standard preaveraging of site-dependent probabilities in the MSA, we showed that statistically significant correlations between charged residues exist in a number of protein families. Interestingly, most of these proteins are associated with various diseases. Thus, functionally important signals can be obtained using a large dataset of sequences alone.

In a series of insightful papers Ranganathan and coworkers (Lockless and Ranganathan, 1999; Suel et al., 2003; Hatley et al., 2003; Shulman et al., 2004) have introduced a method based on statistical coupling analysis (SCA) to identify the relevant energetically-coupled residues. The basic premise of the SCA method, is that, from the sequence family, the co-evolution of positions, either for structural or functional reasons, can be captured by comparing the statistical properties of amino acids in the full MSA and its statistically-significant subsets. In the applications so far, the SCA method has revealed co-evolution between distant sites with functional role. The good agreement between the SCA predictions and limited experimental data (Lockless and Ranganathan, 1999; Suel et al., 2003; Hatley et al., 2003; Shulman et al., 2004) lends credence to this approach.

In this paper we introduce a variant of the SCA method which builds on the hypothesis that signals for co-evolving residues are encrypted in the database of sequences provided the number of sequences in the MSA is large enough. By insisting that the central limit theorem be obeyed, namely, the statistical

properties of a large enough subset of the MSA be the same as in the MSA, we present an automated method for identifying allosteric sites using a family of sequences. After establishing that our method provides consistent results for the PDZ and GPCR families, we describe the mapping of interacting residues for the selectin family which are cell adhesion proteins. Our results for all three families show that the predictions of the automated sequence based approach can be used to target the functionally or dynamically relevant residues in double mutant cycle experiments (Schreiber and Fersht, 1995) or NMR.

## Results

### Size of subalignments

The choice of appropriate size for subalignments is critical in obtaining statistically significant correlations between residues in a protein family. The smallest value of  $f$  should be chosen to satisfy the central limit theorem. We applied our criteria to the families included in the present study (GPCR and Lectin C) and also to the globin family analyzed by LR in (Suel et al., 2003).

#### *GPCR family:*

The full MSA for the GPCR family, reported in the supplementary material of (Suel et al., 2003), contains 940 sequences. From the MSA, we build subalignments with different  $f$  ( $< 1$ ) values by randomly choosing  $fN_{MSA}$  sequences. To perform the averages in Eqs.(6) and (7), we generated 1000 subalignments for a given  $f$  and we computed  $\overline{\Delta G_\lambda}$  using Eqn.(6). The dis-

tribution of these values for increasing  $f$  from Fig.(1(a)) shows that for  $\frac{N_S}{940} > 0.35$  the subalignments satisfy the conditions from Eqn.(6) and (7) and are therefore statistically significant. This value of  $f$  is virtually identical to  $f = 0.33$  value chosen by LR (Suel et al., 2003) for the minimal size of a subalignment.

*Globin family:*

The MSA for the globin family contains 880 sequences (Suel et al., 2003). The distribution of  $\overline{\Delta G}_\lambda$  for increasing fraction of sequences in a subalignment shows (Fig.(1(b))) that subalignments containing more than 55% of the original MSA are statistically significant. LR chose, based on their adhoc criterion,  $f = 0.68$ . In globins  $\overline{\Delta G}_{MSA} \sim 0.65$  in globins, while in GPCR  $\overline{\Delta G}_{MSA} \sim 0.17$ . There is also a greater variation in GPCR between distributions corresponding to different sizes. These results reflect the degree of variation among the sequences in a family, with the sequence similarity in GPCR being smaller than in globins.

*Lectin C:*

The Lectin C family, obtained starting from the Pfam (Bateman et al., 2002) entry and the Clustalw software (Higgins et al., 1994) for realignment of the sequences that remain after weeding out all sequence fragments, contains 1126 sequences. The distribution of  $\overline{\Delta G}_\lambda$  for increasing  $f$  shows (Fig.(1(c))) that subalignments containing more than 15% of the MSA are statistically signif-

icant. In our application, we chose a cut-off of 20%. This example illustrates that for  $f > 15\%$  the average of the distribution no longer changes, while the variance decreases as  $\frac{1}{N_S}$  and the distribution becomes more Gaussian-like. We also note that the average sequence similarity in the Lectin C family is smaller than in globins, but still bigger than in GPCR. These examples clearly show that the choice of  $f$ , based on Eqs.(6) and (7), depends on the family.

### **Correlation between residues in PDZ and GPCR families**

To test the efficacy of our procedure we applied our version of the SCA to identify a set of correlated residues in the PDZ and GPCR families. These two cases allow us to compare our results with those of LR. After demonstrating the equivalence between the two procedures, we apply our formulation to identify the network of correlated residues in the family of cell adhesion molecules.

We calculated  $\Delta\Delta G_{ij}$  for all positions in the MSA using Eqs.(4) and (5) for the PDZ domains which represent protein binding motifs. Following LR, we chose the subset of sequences in the MSA in which His at position 76 is perfectly conserved and calculated the response to this perturbation at all other positions. Comparison of the LR results and our calculations for  $\Delta\Delta G_{i,76}$  shows excellent agreement with a better than 0.95 correlation (Fig.(2A)). The set of identified residues that are coupled, a few long range pairs, may be relevant in the dynamics of the PDZ domains.

The application of the present simplified procedure to the transmembrane G-protein-coupled receptor (GPCR) family also yields results in near quantitative agreement with the LR algorithm. Using the moderately conserved

position Tyr296, which is involved in the ligand interactions in GPCR family, as a perturbation we calculated  $\Delta\Delta G_{i,296}$ . There is a small subset of residues that are uniformly spread throughout the sequence and are coupled to Tyr 296. The magnitudes of  $\Delta\Delta G_{i,296}$  for all  $i$  calculated using the two procedures are nearly identical (Fig.(2B)). We conclude that the present version of the statistical coupling analysis can identify the network of interacting residues provided the dataset of sequences is large enough that meaningful statistical mechanics can be used.

## Correlation between residues in the GPCR family

In order to identify the network of residues that are correlated we performed the CTWC analysis using the  $\Delta\Delta G_{ij}$  values. We used the Euclidean distance (see Eqn.(10)) as similarity metric for comparison of the  $\Delta\Delta G_{ij}$  values at two sites  $k$  and  $j$  (Eq.(9)). We used  $K = 20$  order nearest neighbors and  $q = 20$  in clustering the positions and  $K = 10$  in clustering of perturbations. We performed two rounds of coupled SPC clustering: (i) the clustering of positions in the presence of all the perturbations and the clustering of perturbations in the presence of all the positions. The size of the clusters is determined as described in the **Methods** at each temperature. (ii) In the second round we cluster the positions using the already clustered perturbations in the previous step. In addition we cluster the perturbations in the presence of the positions clustered at step (i). At each step we selected the cluster corresponding to the largest  $\Delta\Delta G_{ij}$  values. The results of clustering the  $\mathcal{G}$  matrix led to 55 clustered positions and 18 clustered perturbations (see Tables (1) and (2)). Out of the 18 clustered perturbations, 17 correspond to clustered positions which shows that the statistical procedure leads to self-consistent results. Moreover,

all 10 perturbations that were reported to be clustered in (Suel et al., 2003) are among the 18 clustered perturbations and 41 (including two positions found at (-1) or (+1) from an actually clustered position) of the 47 clustered positions in (Suel et al., 2003) are among our 55 clustered positions. Therefore, we can recognize 100% of the perturbations and 87% of the positions identified by (Suel et al., 2003). In addition, we also identify three (121, 294 and 313) new functionally important for GPCRs. These were not found using the LR procedure (Suel et al., 2003).

### **Network of functionally important residues in the Lectin C family**

*Background:* The selectin family contains proteins that are involved in the cell adhesion process. These proteins and their glycoconjugate ligands are implicated in the tethering and rolling of circulating leukocytes on blood vessels endothelial cells and platelets. The first step in a multistage dynamics process involves binding of proteins in the selectin family, which are expressed in leukocytes, to ligands in the endothelial cells. The recognition of the ligands involve a coordinated interaction between L-selectin and the various glycoproteins. The crystal structures of the complex between selectin constructs and different ligands have provided insights into the set of residues in selectins that mediate the initial steps in the tethering and rolling process. The sequence-based approach allows us to map the network of residues that signal the tethering process. The success of our predictions is validated by explicit comparison of the predicted binding sites to those identified in the crystal structures.

The structures of the complexes between P-selectin and E-selectin and a weakly bound ligand and the stereo specifically bound P-selectin glycoprotein ligand (PSGL)-1 (Somers et al., 2000) identify a number of sites that not only bind to the ligand, but also respond dynamically to the ligand binding process. Comparison of the liganded and non-liganded structures also reveal large scale movements in the loops connecting Asn83-Asn89 (we use the numbering in PDB entry 1g1s for LEM3-HUMAN positions 42-188). The crystal structures clearly identify specific discrete binding sites. There are three classes of sites that are coordinated to different ligands: (1) The metal ( $\text{Ca}^{2+}$ ) ion dependent weak binding of certain ligands occurs by coordination of  $\text{Ca}^{2+}$  to side chains of Gln80, Asn82, Asn105, and Asp106. (2) The interaction of glycans with P or E-selectin occurs by coordination to residues Tyr48, Gln92, Tyr94, Ser97, Pro98, and Ser99. (3) It is well known that selectins bind strongly to glycoproteins like PSGL-1. A number of residues have been identified in the crystal structure of PSGL-1 in complex with P- and E-selectins. These include Ser47, Arg85, His108, Lys111, Lys112, and Lys113. Several of these residues form a network of hydrogen bonds upon ligand binding.

Besides the three classes of residues, regions of the P- and E-selectins apparently undergo large conformational changes upon complexation (Somers et al., 2000). Comparison of the structures of unliganded and liganded complexes shows that upon binding the loop Asn83 - Asn89 moves from the periphery to the sugar binding sites. In addition, the group of residues Arg54 - Glu74 also undergoes large scale displacement into the region occupied by the Asn83 - Asn89 loop in the unliganded state. From the perspective of allosteric signalling it is unclear which of the 21 residues in the Arg54 - Glu74 loop are directly linked in the signalling pathway. It is logical to suggest that only a

subset of these residues are coupled directly to other parts of the structure. The movement of the Arg54 - Gln74 as a whole is likely to be a consequence of chain connectivity and stereochemistry.

*Sequence entropy is not an indicator of sequence correlation:* Sequence entropy for the Lectin C family shows that not many residues are strongly conserved ( $S(i) < 0.5$ ) (Fig.(3)). With this as the cutoff we find that only residues 19Cys, 26Leu, 50Trp, 52Gly, 81Pro, 90Cys, 109Cys, 117Cys exhibit strong conservation. A key characteristic of cell-adhesion proteins is the preponderance of a large number of Cys residues that form disulfide bonds. From this perspective, strong conservation of the four Cys residues is not a surprise. The other six strongly conserved residues do not seem to be associated with identified functions. In contrast, the crystal structures identify many non-conserved long-range interacting residues to be relevant for some aspects of function. It is obvious from  $S(i)$  alone that it is impossible to decipher the set of coevolving or interacting residues.

*Signalling involves a sparse network:* We have obtained the network of coupled residues which may be involved in the allosteric signalling in the selectin family upon binding to glycoproteins and sugars. We used the Lectin\_C Prosite (PS50041) (Hulo et al., 2004) entry which contains 287 sequences. The sequences are aligned against the sequence of LEM3\_HUMAN from positions 39-159 using the Clustalw package (Higgins et al., 1994). The total length of the alignment (including gaps) is 214 residues. With  $f = 0.2$ , 98 perturbations are allowed at the various positions in the Lectin\_C family. We applied successive rounds of SPC clustering for the 214 positions and the 98 perturbations. For all steps we used  $K = 10$ ,  $q = 20$  and the standard re-scaling of input matrix values (described in the **Methods** section). The measure of

similarity used to cluster both the positions and the perturbations is  $SE_{ik}$  from Eqn.(10). After 3 rounds of SPC, we obtained a cluster of 28 positions (see Table (3)) and a cluster of 24 perturbations (see Table (4)). Nineteen of these 24 perturbations occur at positions that cluster as well.

It is logical to suggest that the list of residues whose interactions are coupled and form a network for functional reasons are the union of positions and perturbations that are coupled. Based on this supposition, we find that all of the residues involving binding to  $\text{Ca}^{2+}$  are identified (Table (3)). Among the six residues that bind directly to sugar our method is able to predict only 3. We do find that residues that are close to the crystallographically identified binding sites are found in the largest cluster. Similarly, nearly all the residues in the neighborhood of these that interact directly with PSGL-1 are correctly identified by our method (Table (3)). In addition to these discrete binding sites, we successfully identify the beginning and end of the Asn83 - Asn89 loop. Several of the residues in the Arg54 - Glu74 loop are also predicted to be significant in the response to ligand binding. Taken together the comparison between the prediction of the sequence-based approach and the crystallographic method that provide a direct glimpse of the network of residues that play a key role in response to ligand binding is good.

*Allosteric network involves predominantly long range contacts:* The mapping of the 28 clustered positions on the structure of the complex between P-selectin and PSGL-1 (1g1s) from Fig.(4) reveals the extent to which they form a spatially correlated network. We find, from the contact map of the complex, that the identified positions are connected either by covalent bonds or by non-bonded contacts. The various views of the cluster of residues and the set of experimentally proposed positions with functional roles (Fig.(4)) show a large

degree of spatial overlap between the two sets. The nature of interactions among the residues in the network may be classified in terms of the contact map of 1g1s. If the distance between a pair of heavy atoms in two residues is within 5.2 Å we assume that they are in contact. There are 244 non-bonded contacts among the 117 residues of chain A in 1g1s.

We denote all contacts between amino acids separated by 11 or more positions as long-range and all other contacts as short-ranged. In 1g1s there are 146 long-range contacts and 98 short-ranged contacts. In the contact map of 1g1s we find 64 long-range and 18 short-range contacts between the 33 clustered positions, while the crystallographically identified 43 functional positions, which include all the 21 residues in the Arg54-Glu74 loop, have 53 long-range and 63 short-range contacts. Thus, the set of 33 clustered positions is composed of very well inter-connected residues that are located far away along the sequence. Therefore, in view of their large structural overlap, we propose that the correlated residues can act as connectors between various functional regions. This finding resembles the idea that allosteric communications are transmitted throughout a structure by means of a sparse but well connected network of interacting residues. In the lectin family the signalling occurs predominantly through long-range contacts.

## Discussion

Discovering residues involved in long-range communications is important for understanding the molecular basis of allostery (Perutz et al., 1998; Kern and Zuiderweg, 2003; Dima and Thirumalai, 2004; Lockless and Ranganathan, 1999), and in a number of contexts including proteins that are implicated in

diseases (Dima and Thirumalai, 2004). It is logical to suggest, as proposed in several previous studies (Lesk and Chothia, 1980; Altschuh et al., 1987; Neher, 1994; Taylor and Hatrick, 1994; Pollock and Taylor, 1997; Pazos et al., 1997; Olmea et al., 1999; Valencia and Pazos, 2002; Dima and Thirumalai, 2004), that the interaction between residues must be encrypted as a statistically significant signature in the evolutionary catalogue of sequences. Here we have proposed a variant of the sequence entropy, that embodies the principles of the SCA, to infer the network of interacting residues in three protein families. For the PDZ and the GPCR families our predictions coincide with the ones reported elsewhere. Moreover, for the GPCR family we not only reproduce known functionally relevant residues that are involved in signalling and binding, but also predict previously unidentified residues that could play a relevant role in the interhelical packing of the rhodopsin family. The application of our procedure to the lectin family leads to the correct prediction of all the important sets of residues that could play a key role in the tethering and rolling processes. Just as in the previous applications (Suel et al., 2003) our results also show that the network of correlated residues is sparse. This finding may be fairly general and is consistent with the notion that proteins can tolerate substantial number of mutations at many positions without sacrificing functional efficiency.

In the current formulation of the SCA only pairwise interactions between residues are probed. It is likely that variations among more than two residues are possible due to simultaneous interactions among three or more residues. In order to decipher correlations between three residues it is necessary to perturb sites  $j$  and  $k$  and probe the response at site  $i$ . Because this will require obtaining subsets of the MSA in which sites  $j$  and  $k$  are conserved, the statistics

might not be as good as for probing pairwise interactions. Nevertheless, by using our procedure the coupling  $\Delta\Delta G_{i,\{jk\}}$  can be computed to test which of the perturbations are pairwise additive. This valuable information is extremely difficult to obtain from experiments.

All sequence based approaches are “thermodynamic” in nature and only consider evolutionary sequence changes. From the perspective of function it is necessary to consider dynamic changes to perturbations which can be either induced by mutations or by changes in external conditions (pH, temperature, denaturant, mechanical force, etc.). Such changes require structural probes. Our previous work using elastic network model (Zheng et al., 2005) was an attempt to integrate sequence and structure based methods to identify the sparse network of correlated residues that dynamically trigger allosteric transitions in polymerases. It is desirable to develop a theoretically based method, along the lines developed here, that focuses on residue-dependent structural perturbations for probing dynamical responses.

## Methods

### Response of a position in the MSA to perturbations:

Following Lockless and Ranganathan (LR) (Lockless and Ranganathan, 1999), we defined, for each position  $i$  in the MSA, a statistical “free energy”

$$\frac{\Delta G_i}{kT^*} = \sqrt{\frac{1}{C_i} \sum_{x=1}^{20} [p_i^x \ln(\frac{p_i^x}{p_x})]^2} \quad (1)$$

where  $kT^*$  is an arbitrary energy scale (which we set to 10 in our calculations),  $L_{MSA}$  is the length of the sequences in the MSA including the gaps, and  $i =$

1,2,3,...,L<sub>MSA</sub>. The number of types of amino acids that appear at least once at  $i$  is  $C_i$  (i.e.,  $C_i \leq 20$ ),  $p_x$  is the mean frequency of amino acid type  $x$  in the MSA. The statistical free energy is the excess value of  $\Delta G_i$  when  $p_i^x$  deviates from  $p_x$ . We computed  $p_i^x$  using

$$p_i^x = \frac{n_i^x}{N_i} \quad (2)$$

where  $n_i^x$  is the number of sequences in the MSA with amino acid  $x$  at position  $i$ , and  $N_i$  is the total number of sequences in the MSA that have one of the twenty types of amino acids at position  $i$ , so that  $N_i = \sum_{x=1}^{20} n_i^x$ .

Our procedure differs from the one used by LR in two important aspects: (i) Instead of the binomial density function (Lockless and Ranganathan, 1999) for the distribution of amino acid type  $x$  at a random position in the MSA we use the typical frequency of  $x$  in the MSA; (ii) We use the frequency of each of the 20 types of amino acids in the given MSA rather than in the whole Swiss-Prot database (Appel et al., 1994). With our procedure, the free energy  $\Delta G_i$  in Eq.(1) is a straightforward extension of the familiar sequence entropy

$$S(i) = - \sum_{x=1}^{20} p_i^x \ln(p_i^x) \quad (3)$$

at  $i$  in the MSA.

It is preferable to use, for each protein family, its typical amino acid distribution, rather than a general and maybe sometimes even incorrect set of frequencies. For example, the native state of BPTI has three disulfide bonds formed between six cysteine residues (out of the total of 56 positions in the chain). Therefore, for the BPTI family the frequency of finding CYS is  $\sim 12\%$  which is considerably bigger than the 1% frequency (Creighton, 1993) found

in a random protein sequence. Thus, context dependence is automatically accounted for in our procedure.

For a given MSA we build subsets of the whole alignment in which we retain only sequences that have only one type of amino acid at a position  $j$ , i.e., in the subset  $p_j^x = 1$  for  $x$ . Let us assume that for functional and structural reasons positions  $i$  and  $j$  are coupled, i.e., substitutions in  $i$  would affect  $j$ . If this is the case, we expect that evolutionary pressure, under functional constraints, would lead to a correlated change in the amino acid type at  $i$  in the restricted (the subset of the original MSA) alignment. In other words, the residues at  $i$  and  $j$  might “communicate” in the course of function or upon binding to ligands. A measure of the correlation between two positions in the MSA is the statistical free energy change at  $i$  as a result of a perturbation at  $j$

$$\frac{\Delta\Delta G_{ij}}{kT^*} = \sqrt{\frac{1}{C_i} \sum_{x=1}^{20} [p_{i,R}^x \ln(\frac{p_{i,R}^x}{p_x}) - p_i^x \ln(\frac{p_i^x}{p_x})]^2} \quad (4)$$

with

$$p_{i,R}^x = \frac{n_{i,R}^x}{N_{i,R}} \quad (5)$$

where  $n_{i,R}^x$  is the number of sequences in the restricted MSA that have amino acid  $x$  at  $i$  and  $N_{i,R}$  is the total number of sequences in the restricted MSA that have a valid type of amino acid at position  $i$ , and  $N_{i,R} = \sum_{x=1}^{20} n_{i,R}^x$ . It follows that  $\Delta\Delta G_{ij} = 0$  for both a perfectly conserved position ( $p_i^x = 1$ ) and for a position where all amino acids are found at their mean frequencies in the MSA ( $p_i^x = p_x$ ).

In what follows, we will give examples to show that there is an almost perfect correlation between the  $\Delta\Delta G_{ij}$  obtained with our method and the LR method.

As stated earlier, our method is an extension of the sequence entropy method which is used to infer conservation of amino acids in a MSA. Our formulation of SCA lends support to the finding of Fodor and Aldrich (Fodor and Aldrich, 2004) that the predictions of the SCA method are similar to those obtained from the sequence entropy alone. In contrast with the SCE method, the SCA approach involves averaging of the probabilities  $p_{i,R}^x$  and  $p_i^x$  over sequences. Such an averaging can sometimes obscure real correlations.

## Size of subalignments must obey central limit theorem

In the implementation of the procedure it is crucial to choose an optimal size of the subalignment which previously (Suel et al., 2003) the size of the subalignments was chosen arbitrarily based only on intuitive arguments. A number of subsets each containing a fraction of the total number of sequences in the MSA were chosen. For each set,  $\Delta\Delta G_{ij}$  values for a few (usually about 5) least conserved positions are computed. The size of the subalignment was chosen so that  $\langle \Delta\Delta G_i \rangle = \sum_{j=1}^{N_S} \Delta\Delta G_{ij} \sim 0$ , where  $\langle \rangle$  is an average over the  $N_S$  sequences in the subalignment.

We appeal to statistical mechanics in choosing the size of the subalignments that contain  $P = fN_{MSA}$  sequences. The number of alignments for a fixed  $f$  is  $k = \frac{N_{MSA}!}{P!(N_{MSA}-P)!}$ . To obtain statistically meaningful results, the general properties of the subalignments must be similar to the original MSA. In analogy with statistical mechanics we suggest that the smallest value of  $f$  be chosen so that the law of large numbers be obeyed. In particular, we choose  $f$  so that

the following criteria are satisfied

$$\langle \overline{\Delta G} \rangle_f = \sum_{\lambda=1}^k \overline{\Delta G}_\lambda \approx \overline{\Delta G}_{MSA} \quad (6)$$

$$\sigma_f^2 = \langle \overline{\Delta G}^2 \rangle_f - \langle \overline{\Delta G} \rangle_f^2 \sim \frac{1}{N_S} \quad (7)$$

where  $\overline{\Delta G}_\lambda = \frac{1}{L_{MSA}} \sum_{i=1}^{L_{MSA}} \Delta G_i^\lambda$ ,  $\overline{\Delta G}_{MSA} = \frac{1}{L_{MSA}} \sum_{i=1}^{L_{MSA}} \Delta G_i$ ,  $\sigma_f$  is the width of the distribution of  $\overline{\Delta G}_\lambda$ , and  $N_S$  is the number of sequences in the subalignment. Operationally, the second criterion (Eqn.(7)) is valid provided that the variance in the subalignments satisfies

$$\sigma_{f_2} \sqrt{N_{S_2}} = \sigma_{f_1} \sqrt{N_{S_1}} \quad (8)$$

where  $N_{S_1}$  and  $N_{S_2}$  are the number of sequences in subalignments with  $f_1$  and  $f_2$  respectively. The advantage of using our criteria (Eqs.(6) and (7)) is that  $f$  is *automatically chosen from the MSA alone without having to compute  $\Delta \Delta G_{ij}$* . Failure to satisfy these criteria can give spurious results in the application of SCA.

## Clustering procedure and similarity measures

The matrix  $\mathcal{G}$ , whose elements are the  $\Delta \Delta G_{ij}$  values for a protein family, represents the response of positions  $i$  in the MSA to all allowed perturbations at site  $j$  provided the perturbations satisfy the acceptance criteria stated above. The rows of the matrix correspond to positions in the MSA and the columns to perturbations. Our objective is to reliably determine the network(s) of positions that change in a correlated manner starting from this matrix. To this end, we used the coupled two-way clustering (CTWC) that was developed to

analyze DNA-microarray data (Getz et al., 2000). The basic idea is to carry out successive elementary rounds of Superparamagnetic clustering (SPC) (Blatt et al., 1996). At each step, the submatrix that contains positions and perturbations that cluster together in the previous iteration with large signals is extracted.

An important ingredient in the SPC technique is the choice of a similarity measure between a pair of entries that are to be clustered. In the context of clustering of positions in an MSA, there are at least two natural choices for similarity measures. (1) the Euclidean distance and (2) the Pearson correlation coefficient. In what follows, we give the rationale and the details for using these measures. The collection of the  $\Delta\Delta G_{ij}$  values (with  $j$  varying from 1 to  $L_{MSA}$  with  $L_{MSA}$  being the total number of positions (including gaps) of the alignment) for a given position  $i$  in the MSA can be thought of as a vector with  $L_{MSA}$  components,  $\vec{v}_i = \{\Delta\Delta G_{i1}, \dots, \Delta\Delta G_{iL_{MSA}}\}$ . Therefore, the degree of similarity between two positions  $i$  and  $k$  can be represented by the Euclidean distance between the two corresponding vectors, i.e.,

$$D_{ik} = \sqrt{\sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \Delta\Delta G_{kj})^2} \quad (9)$$

For each MSA there is a spread in the magnitudes of the  $\Delta\Delta G_{ij}$  values (e.g.: from  $\sim 0.01$  to  $\sim 10$ ). Thus, for a pair of small matrix elements  $D_{ik}$  will be small even if the two vectors are not similar. On the other hand, for two related positions  $i$  and  $k$  with large  $\Delta\Delta G_{ij}$  values a difference in any of their components could lead to a large  $D_{ik}$  value which would not reflect their true similarity. Positions with small  $\Delta\Delta G_{ij}$  values are of little interest because they show basically no response to changes in other positions in the MSA.

To correct for the potentially spurious results indicated above we use the following protocol: (i) We eliminate entries (positions) that show virtually no response to the overwhelming majority of perturbations, (ii) We scale the  $\Delta\Delta G_{ij}$  values so that only a few categories of the matrix elements are included in the analysis. To a large extent the results do not depend on the precise boundaries used in the classification of  $\Delta\Delta G_{ij}$ . (iii) The  $D_{ik}$  values are suitably normalized. If all or all but one of the corresponding  $\Delta\Delta G_{ij}$  values are less than 1.0, then the row corresponding to position  $i$  is deleted from the input data matrix. The scaling of the  $\Delta\Delta G_{ij}$  values is achieved by assigning them to two or three characteristic entries. For example, all the small  $\Delta\Delta G_{ij}$  values (i.e.,  $\Delta\Delta G_{ij} < 1.0$ ) are kept unchanged, while the intermediate  $\Delta\Delta G_{ij}$  values (i.e.,  $1.0 \leq \Delta\Delta G_{ij} < 2.0$ ) are assigned a value  $\alpha_1$  and the remaining (large)  $\Delta\Delta G_{ij}$  values are assigned a value  $\alpha_2$  such that  $\alpha_1 \sim 10 \times \max_{ij}\{\min(\Delta\Delta G_{ij})\}$  and  $\alpha_1 < \alpha_2$ . We normalize  $D_{ik}$  using

$$SE_{ik} = \frac{D_{ik}}{0.5 \times (||\vec{v}_i|| + ||\vec{v}_k||)} \quad (10)$$

where  $||\vec{v}_i||$  is the norm of the vector  $\vec{v}_i$ . The  $SE_{ik}$  is small for pairs of vectors that have components of similar values and it is independent of the actual magnitude of the individual components. In addition, because positions that show reduced or no response to the majority of the perturbations are eliminated, a small  $SE_{ik}$  value indicates that the two positions show large responses to the same set of perturbations.

A second similarity measure that can be used is the Pearson correlation coefficient

$$P_{ik} = \frac{\sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \langle \Delta\Delta G_i \rangle) (\Delta\Delta G_{kj} - \langle \Delta\Delta G_k \rangle)}{\sigma_i \sigma_k} \quad (11)$$

where  $\langle \Delta\Delta G_i \rangle = \frac{\sum_{j=1}^{L_{MSA}} \Delta\Delta G_{ij}}{L_{MSA}}$  is the average  $\Delta\Delta G_{ij}$  value for position  $i$  and  $\sigma_i^2 = \sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \langle \Delta\Delta G_i \rangle)^2$  is the variance. Just as with the Euclidean distance similarity measure  $P_{ik}$  is small for two positions with little or no responses to the majority of perturbations. Prior to calculating  $P_{ik}$  we eliminate all such positions. The procedure employed is the same as described above. For two perfectly correlated (anti-correlated) positions  $P_{ik} = 1$  (-1), while for uncorrelated positions  $P_{ik} = 0$ . Because the Euclidean distance measure ( $SE_{ik}$ ) has small values for two correlated positions and we want to be able to use the two similarity measures interchangeably, we replaced  $P_{ik}$  with

$$SP_{ik} = 1 - |P_{ik}| \tag{12}$$

where  $|P_{ik}|$  is the absolute value of  $P_{ik}$ . Therefore, *both  $SE_{ik}$  and  $SP_{ik}$  are zero when the two positions are perfectly correlated.*

The Euclidean similarity measure  $SE_{ik}$  is best suited when the individual  $\Delta\Delta G_{ij}$  values are not broadly distributed, i.e., when the largest  $\Delta\Delta G_{ij}$  value is  $\sim 3.0$ . In such a case the responses of each position in the alignment to the various perturbations are similar in magnitude. Therefore, the use of the Pearson correlation coefficient  $SP_{ik}$  would lead to the majority of positions being clustered. On the other hand, using the  $SE_{ik}$  and the associated re-scaling of entries allows us to distinguish between positions and therefore only a handful of positions turn out to be clustered after the application of the CTWC procedure. It follows then that the Pearson similarity measure is best suited for MSA for which there is a broad distribution in the magnitude of the responses of positions to perturbations.

## Clustering Algorithm:

The clustering algorithm consists of several steps. We first calculate the similarity measure between all the pairs of data points. Based on these values, we select, for each input data point, its  $K$  nearest neighbors (n.n.) ( $K$  varies between 10 and 20). The next step consists in retaining, for each input data point, only its  $K$ -order neighbors, i.e.,  $i$  is considered a  $K$ -order neighbor of  $j$  if and only if  $i$  is one of the  $K$  n.n. of  $j$  and  $j$  is one of the  $K$  n.n. of  $i$ . Using only the pairs of data points in this  $K$ -order neighbors list, we calculate the parameters in the  $q$  state Potts spin representation of the data points (Blatt et al., 1996). The Swendsen-Wang algorithm (Wang and Swendsen, 1990), which we use to determine the conformations of the spin system in the SPC step (Blatt et al., 1996), starts with all the data points being assigned to the same value of the spin (i.e., with the spin system in the ferromagnetic phase) and unconnected with each other and the temperature  $T \sim 0$ . To map out the conformational space of the spin system, we increase the temperature linearly in small steps (such that at step  $t$  the temperature is  $T_t = T_{t-1} + \delta T$  with  $\delta T \sim 0.001$ ) from  $T \sim 0$  to  $T = T_{max}$  (usually  $T_{max} \sim 1.0$ ). At each temperature we go through each site  $i$  that has at least one  $K$ -order neighbor and we randomly connect it with its n.n. with the same spin value using the probability  $P(ij)$  from (Wang and Swendsen, 1990). We pick a random number  $r_1$  between 0 and 1 and we connect  $i$  with  $j$  (with  $s_i = s_j$ ) if and only if  $r_1 \leq P(ij)$ . Then we identify the clusters in the system of spins as the collections of connected points with the same value of the spin. In the next step, we reassign the value of the spin in each of these clusters to a randomly picked value (picked with equal probability among the  $q$  available values). Finally, we loop through all the data points and their  $K$ -order n.n. and determine all the points that belong to the

same cluster. Two positions with the same spin value are considered part of the same cluster. Like in a typical Monte-Carlo simulation, this procedure is repeated for  $N_{steps}$  (usually 10000) number of steps at each temperature and, to allow for the equilibration of the system at the given temperature, in the calculation of the averages that enter in the average correlation between spins and the susceptibility we disregard the data from the first  $N_{eq}$  steps (usually 3000).

The spin conformations are used to calculate the average spin-spin correlation ( $\langle \delta_{s_i, s_j} \rangle$ ) for each data point and its K-order nearest neighbors at each temperature. From the distribution of spin-spin correlations we choose the clustering temperature ( $T_c$ ) as the temperature where this distribution has equal height peaks at values 1 (more precisely,  $1 - \frac{2}{q}$ ) and 0 (more precisely,  $\frac{1}{q}$ ) and is small in between. At  $T_c$  any two points for which the corresponding  $\langle \delta_{s_i, s_j} \rangle > \Theta$  are assigned to a cluster. This assigns the core region for each of the clusters of data points. Because at non-zero temperatures some data points might not have any pair that satisfies the threshold for  $\Theta$ , following Domany ((Domany, 1999)), we capture the points lying at the periphery of the clusters by linking each data point  $i$  with its K-order n.n.  $j$  of maximum correlation  $\langle \delta_{s_i, s_j} \rangle$ .

### **Acknowledgments:**

We acknowledge Jie Chen for useful discussions. This work was supported in part by grants from the National Institutes of Health and the National Science Foundation through grant number NSF CHE05-14056.

## Figure Captions

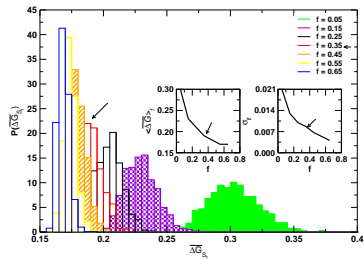
Fig.1. Distribution of  $\overline{\Delta G}_\lambda$  for various sizes ( $f = \frac{N_S}{N_{MSA}}$ ) of subalignments for GPCR, globin and Lectin families. The index  $\lambda$  refers to the one of the subsets with  $N_S$  sequences. Insets show the dependence of  $\langle \overline{\Delta G} \rangle_f$  and  $\sigma_f$  for various  $f$ . As expected from Eqn.(6),  $\langle \overline{\Delta G} \rangle_f$  shows a plateau starting at a given value of  $f$  (indicated by the arrow), while  $\sigma_f$  for  $f$  larger than this value decreases according to Eqn.(7). The minimum size of  $f$  should correspond to the value indicated by the arrow. (a) GPCR family (940 sequences in full MSA). (b) globins family (880 sequences in full MSA). (c) Lectin C family (1126 sequences in full MSA).

Fig.2. Comparison between  $\Delta\Delta G_{ij}$  values obtained as described in (Lockless and Ranganathan, 1999) and using our procedure. The panel on the left is for the GPCR family while the one on the right corresponds to the PDZ family. For each family we selected a perturbation at the most functionally important position. The data for the two families shows a very large degree of correlation ( $\geq 0.95$ ) indicating that our procedure captures all the crucial details of the previous implementation of the statistical coupling analysis (SCA) method.

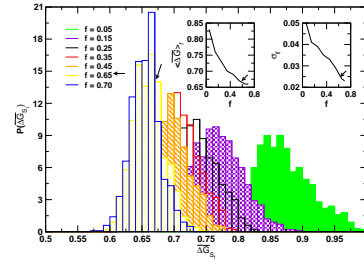
Fig.3. Sequence entropy for the 117 positions in 1g1s using the multiple sequence alignment from PROSITE (with 287 sequences). The black circles correspond to  $S(i)$  for the 20 types of amino acids obtained according to Eqn.(3), while the red diamonds represent the chemical sequence entropy obtained by classifying the amino acids into four classes, namely hydrophobic, polar, positively and negatively charged. Hydrophobic residues are: Ala, Leu, Ile, Val, Trp, Tyr, Cys, Met, Phe. Positively charged residues are: Arg and Lys. Negatively charged residues are: Glu and Asp. Polar residues are: Thr, Gly, Ser,

His, Gln, Asn, Pro. The reduction in the number of the classes of amino acids leads to conservation at a larger number of positions. By using only four types of amino acids and a cut-off for strong conservation at  $S(i) < 0.25$ , we find that positions 2Thr, 3Tyr, 5Tyr, 12Trp, 15Ser, 27Val, 29Ile, 37Tyr, 38Leu, 49Tyr, 51Ile, 52Gly, 53Ile, 60Trp, 62Trp, 76Trp, 91Val, 93Ile, 104Trp, 115Ala, 116Leu are also conserved. Among these 21 sites, only Trp60 and Trp62 are crystallographically identified to be functionally relevant (they belong to the Arg54-Glu74 loop). Therefore, sequence entropy alone cannot lead to the network of amino acids identified by our procedure.

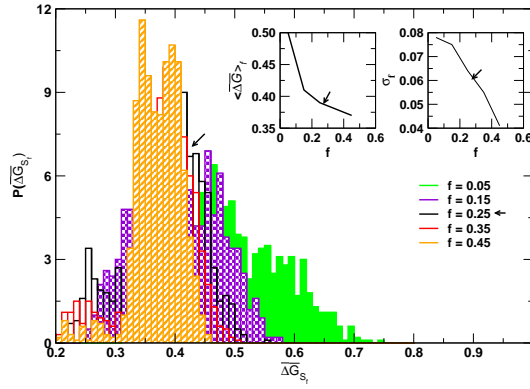
Fig.4. The network of 28 clustered residues in 1g1s (two different orientations: on the left there is the top view, while a side view is shown on the right). The yellow surface represents the 28 clustered positions while the orange surface represents the 42 positions that have been identified using crystal structures of liganded and unliganded P- and E-selectins to functionally important. Half of the 42 residues are in the Arg54-Glu74 loop. It is unlikely that all of these residues are equally important for function. Their coordinated motion upon glycoprotein binding is, in all likelihood, due to chain connectivity. There is a large degree of overlap between our predictions and experiments. The clustered positions also comprise a physically connected network either by bonded or non-bonded contacts. The mapping to the structure shows that, except for residues in the N-terminus end, they are interspersed throughout the structure. The figures were produced using the packages VMD (Humphrey et al., 1996) and Povray (<http://www.povray.org/>). Cylinders represent  $\alpha$ -helices and arrows represent  $\beta$ -strands.



(a)



(b)



(c)

Fig. 1.

$\Delta\Delta G_{GPCR}$  (position 296 = TYR)

$\Delta\Delta G_{PDZ}$  (position 76 = HIS)

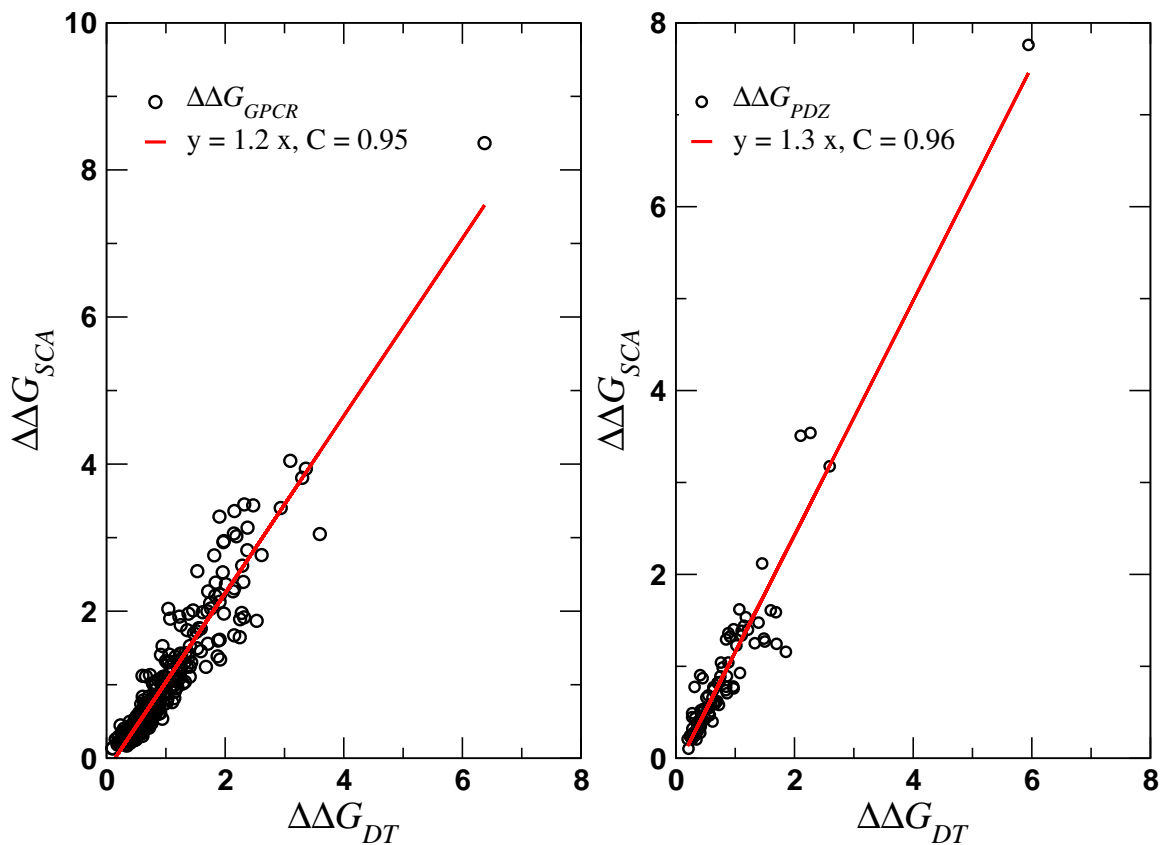


Fig. 2.

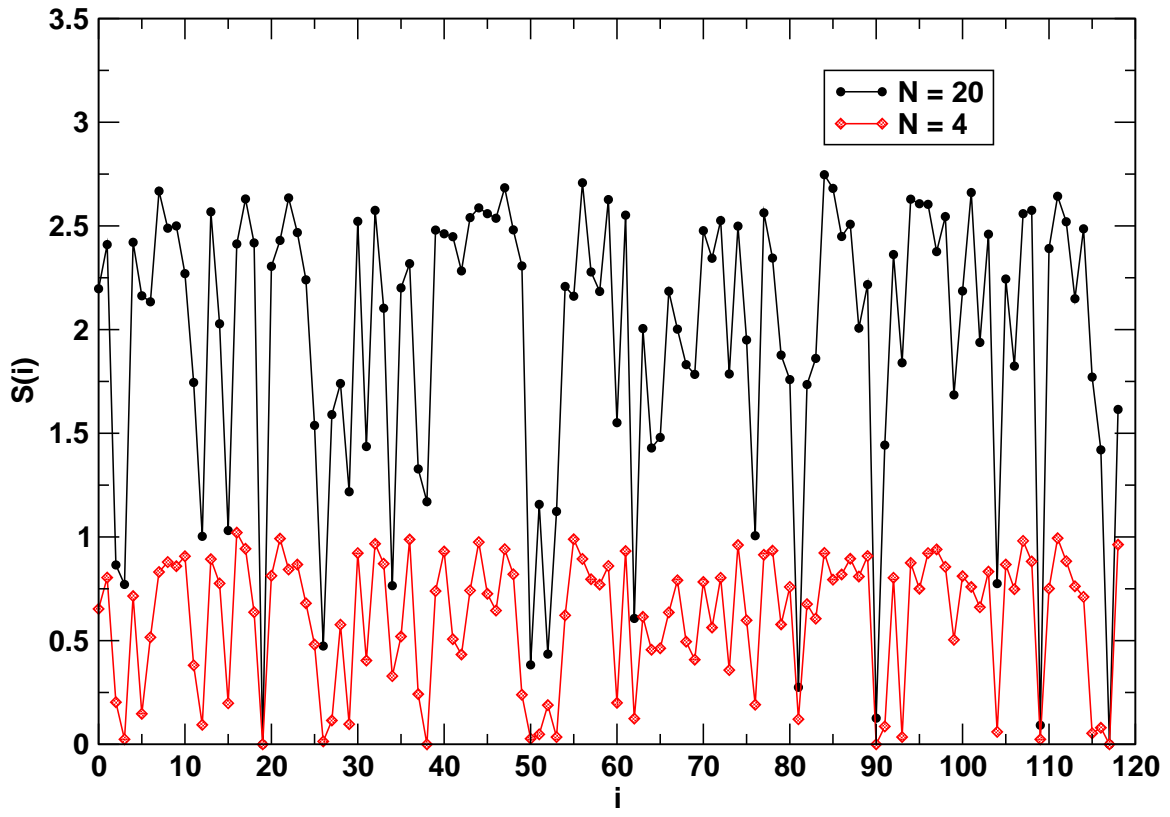
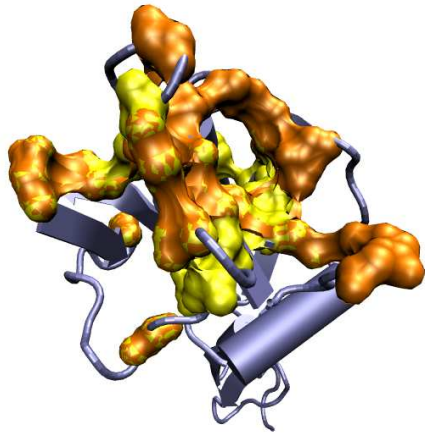
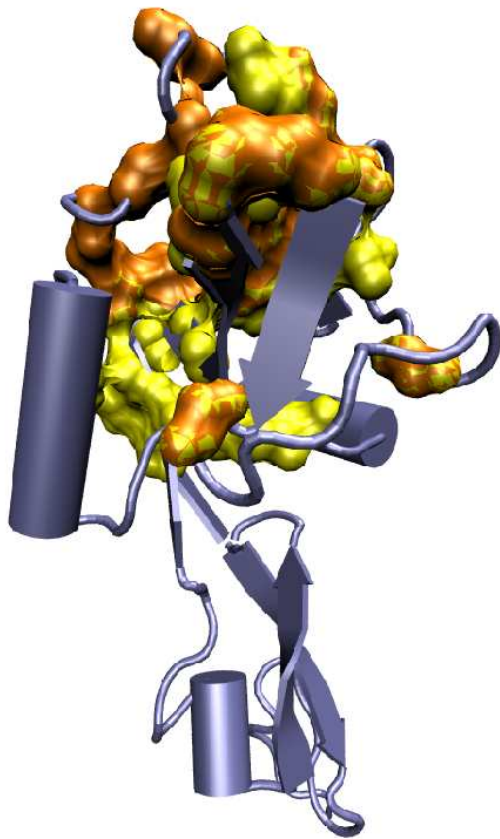


Fig. 3.



(a)



(b)

31

Fig. 4.

Table 1

List of 55 positions clustered (GPCR – notation from PDB file 1L9H)

---

<b>44</b>	<b>51</b>	<b>58</b>	59	61	66	67	<b>75</b>	<b>78</b>	<b>90</b>
<b>91</b>	<b>92</b>	<b>113</b>	115	<b>117</b>	120	<i>121</i>	<b>122</b>	<b>123</b>	
<b>124</b>	<b>125</b>	126	<b>129</b>	<b>131</b>	<b>134</b>	<b>136</b>	140	141	<b>149</b>
<b>152</b>	<b>157</b>	<b>164</b>	168	<b>213 (212)</b>	<b>219</b>	<b>222</b>	230	249	<b>253</b>
<b>254</b>	<b>257</b>	258	<b>259</b>	<b>265</b>	<b>268</b>	<b>269</b>	<b>293</b>	<i>294</i>	<b>295</b>
<b>296</b>	<b>298</b>	<b>299</b>	<b>300</b>	312 ( <i>313</i> )	<b>317 (316)</b>				

---

Table 2

List of 18 perturbations (position and amino acid type selected) clustered (GPCR – notation from PDB file 1L9H)

---

<i>51G</i>	<i>61V</i>	<i>66K</i>	<i>69R</i>	<b>78S</b>	<b>92P</b>	<b>123A</b>	<b>144Y</b>	<b>157I</b>	
<i>164A</i>	203Y	<b>219M</b>	<i>222C</i>	<b>254L</b>	<b>268F</b>	<i>268Y</i>	<b>296F</b>	<b>317Y</b>	

---

Table 3

List of 28 clustered positions in Lectin C (notation from PDB file 1g1s)

---

4	5	6	15	27	28	29	38	49	51
<b>55</b>	<b>57</b>	<b>58</b>	<b>63</b>	<b>73</b>	76	<b>83</b>	<b>89</b>	<b>90</b>	92
<b>93</b>	94	104	<b>105</b>	<b>106</b>	<b>113</b>	115	116		

---

Table 4

List of 24 perturbations (position and amino acid type selected) clustered in LectinC

(notation from PDB file 1g1s)

---

<i>5A</i>	<i>6H</i>	<i>28S</i>	35Q	<i>38L</i>	<i>38V</i>	<b>54S</b>	<i>55D</i>	<i>57E</i>	<i>58G</i>
<i>63V</i>	<i>63S</i>	<b>80E</b>	<b>82N</b>	<i>83N</i>	87G	<i>89E</i>	<i>90D</i>	<i>92A</i>	<i>93E</i>
<i>105N</i>	<i>106D</i>	<i>115F</i>	<i>116V</i>						

---

## References

- Altschuh, D., Lesk, A. M., Bloomer, A. C., Klug, A., 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* 193, 693–707.
- Appel, R. D., Bairoch, A., Hochstrasser, D. F., 1994. A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.* 19, 258–260.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Ewinger, L., Eddy, S. R., Griffith-Jones, S., Howe, K. L., Marshall, M., Sonnhammer, E. L., 2002. The Pfam protein families database. *Nucl. Ac. Res.* 30, 276–280.
- Blatt, M., Wiseman, S., Domany, E., 1996. Superparamagnetic clustering of data. *Phys. Rev. Lett.* 76, 3251–3254.
- Creighton, T. E., 1993. *Proteins: Structures and molecular properties*. W.H. Freeman and Company, New York.
- Dima, R. I., Thirumalai, D., 2004. Proteins associated with diseases show enhanced sequence correlation between charged residues. *Bioinformatics* 20, 2345–2354.
- Domany, E., 1999. Superparamagnetic clustering of data - the definitive solution of an ill-posed problem. *Physica A* 263, 158–169.
- Fodor, A., Aldrich, R. W., 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56, 211–221.
- Getz, G., Levine, E., Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA* 97, 12079–12084.

Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G., Ranganathan, R., 2003. Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci. USA* 100, 14445–14450.

Higgins, D., Thompson, J., Gibson, T., Thompson, J. D., Higgins, D. G., Gibson, T. J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Horovitz, A., Fridmann, Y., Kafri, G., Yifrach, O., 2001. Allostery in chaperonins. *J. Struct. Biol.* 135, 104–114.

Hulo, N., Sigrist, C. J. A., Saux, V. L., Langendijk-Genevaux, P., Bordoli, L., Gattiker, A., Castro, E. D., Bucher, P., Bairoch, A., 2004. Recent improvements to the PROSITE database. *Nucl. Ac. Res.* 32, D134–D137.

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* 14, 33–38.

Kass, I., Horovitz, A., 2002. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 48, 611–617.

Kern, D., Zuiderweg, E. R., 2003. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* 13, 748–757.

Lesk, A. M., Chothia, C., 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136, 225–270.

Lichtarge, O., Bourne, H. R., Cohen, F. E., 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257, 342–358.

Lockless, S. W., Ranganathan, R., 1999. Evolutionary conserved pathways of energetic connectivity in protein families. *Science* 286, 295–299.

Monod, J., Wyman, J., Changeux, J. P., 1965. On the nature of allosteric

transitions: a plausible model. *J. Mol. Biol.* 12, 88–118.

Neher, E., 1994. How frequent are correlated changes in families of protein sequences ? *Proc. Natl. Acad. Sci. USA* 91, 98–102.

Olmea, O., Rost, B., Valencia, A., 1999. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* 295, 1221–1239.

Pazos, F., Helmer-Citterich, M., Ausiello, G., Valencia, A., 1997. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* 271, 511–523.

Perutz, M. F., Wilkinson, A. J., Paoli, M., Dobson, G. G., 1998. Stereochemistry of cooperative mechanisms in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.* 27, 1–34.

Pollock, D. D., Taylor, W. R., 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.* 10, 647–657.

Schreiber, G., Fersht, A. R., 1995. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* 248, 478–486.

Shulman, A. J., Larson, C., Mangelsdorf, D. J., Ranganathan, R., 2004. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116, 417–429.

Somers, W. S., Tang, J., Shaw, G. D., Camphausen, R. T., 2000. Insights into the molecular basis of leukocyte tethering and rolling revealed by structures of P- and E-Selectin bound to SLex and PSGL-1. *Cell* 103, 467–479.

Sowa, M. E., Hen, W., Slep, K. C., Kercher, M. A., Lichtarge, O., Wensel, T. G., 2001. Prediction and confirmation of a site critical for effector

regulation of RGS domain activity. *Nat. Struct. Biol.* 8, 234–237.

Steitz, T. A., 1999. DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.* 274, 17395–17398.

Suel, G. M., Lockless, S. W., Wall, M. A., Ranganathan, R., 2003. Evolutionary conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10, 59–69.

Taylor, W. R., Hatrick, K., 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng.* 7, 341–348.

Valencia, A., Pazos, F., 2002. Computational methods for prediction of protein interactions. *Curr. Op. Struct. Bio.* 12, 368–373.

Wang, J.-S., Swendsen, R. H., 1990. Cluster Monte Carlo algorithms. *Physica A* 167, 565–579.

Xu, Z., Sigler, P. B., 1998. GroEL/GroES: structure and function of a two - stroke folding machine. *J. Struct. Biol.* 124, 129 – 141.

Zheng, W. J., Brooks, B. R., Doniach, S., Thirumalai, D., 2005. Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure* 13, 565–577.