Evolutionarily conserved networks of residues mediate allosteric communication in proteins

Gürol M. Süel[1,2], Steve W. Lockless[1,2], Mark A. Wall[2] and Rama Ranganathan[2]

Published online 16 December 2002; doi:10.1038/nsb881

A fundamental goal in cellular signaling is to understand allosteric communication, the process by which signals originating at one site in a protein propagate reliably to affect distant functional sites. The general principles of protein structure that underlie this process remain unknown. Here, we describe a sequence-based statistical method for quantitatively mapping the global network of amino acid interactions in a protein. Application of this method for three structurally and functionally distinct protein families (G protein–coupled receptors, the chymotrypsin class of serine proteases and hemoglobins) reveals a surprisingly simple architecture for amino acid interactions in each protein family: a small subset of residues forms physically connected networks that link distant functional sites in the tertiary structure. Although small in number, residues comprising the network show excellent correlation with the large body of mechanistic data available for each family. The data suggest that evolutionarily conserved sparse networks of amino acid interactions represent structural motifs for allosteric communication in proteins.

Communication between distant sites in proteins is fundamental to their function and often defines the biological role of a protein family. In signaling proteins, it represents information transfer — the transmission of signals initiated at one functional surface to a distinct surface mediating downstream signaling. For example, ligand binding at an externally accessible site in G protein–coupled receptors (GPCRs) reliably triggers structural changes at distant cytoplasmic domains that mediate interaction with heterotrimeric G proteins[1,2]. Studies in many other protein systems indicate that long-range interactions of amino acids also are important in binding (and catalytic) specificity. Substrate recognition in the chymotrypsin family of serine proteases[3,4], the tuning of antibody specificity through B-cell maturation[5] and the cooperativity of oxygen binding in hemoglobin[6–9] all depend not only on residues directly contacting substrate, but also on distant residues located in supporting loops and other secondary structural elements. Crystallographic studies in all of these systems[5,9–11] indicate that the distant residues participating in substrate recognition do so by acting through intervening positions to control the structure of the substrate-binding site. These long-range interactions are remarkable because many other sites, even if closer to active site residues, show little contribution to function. Taken together, these studies indicate that proteins are complex materials in which perturbations at sites — for example, substrate binding, covalent modification or mutation — may cause conformational change to happen in a fracture-like manner that is not obvious in atomic structures. From a biological point of view, these fractures represent the energy transduction mechanisms that mediate signal flow, allosteric regulation and specificity in molecular recognition.

How can we globally map energetic interactions between amino acid residues in protein structures? Although methods such as the thermodynamic double mutant cycle[12–14] provide excellent tools for estimating such interactions, practical limitations restrict these techniques to small studies in specific model systems. An alternative approach is suggested by a new sequence-based statistical method for estimating thermodynamic coupling between residues in proteins[15]. The basis of this method is that the coupling of two sites in a protein, whether for structural or functional reasons, should cause those two positions to co-evolve[16–18]. In principle, this might be revealed in an analysis of a large and diverse multiple sequence alignment (MSA) of a protein family. Application of this method for one active site residue in a small protein interaction domain (the PDZ domain) family predicted energetic coupling to a small set of other residues that were organized into a chain-like network through the protein core, linking the active site residue with distant sites[15]. These predictions were verified through mutagenesis, suggesting that the statistical measurement of coupling through sequence analysis is a good reporter of thermodynamic coupling.

These results suggest the possibility that we can visualize the global network of energetic interactions between pairs of amino acids and explain long-range energetic interactions in proteins. Here, we describe this mapping for three protein families that represent completely distinct folds and biological activities: (i) a transmembrane signaling receptor family (GPCRs), (ii) an enzyme family that has served as a model system for catalytic specificity (the chymotrypsin class of serine proteases) and (iii) a multi-subunit protein family that is the classic model system for allosteric regulation (hemoglobin).

A statistical mapping of interactions in proteins

To illustrate the sequence analysis, we consider four positions of a hypothetical protein (*i*, *j*, *k* and *l*) and a corresponding MSA of the protein family (Fig. 1*a*,*b*). If the MSA is sufficiently large and diverse that it describes the evolutionary constraints on the family, we can make the following two postulates about the amino acid frequencies observed at specific sites. First, if site *l* contributes nothing to either the folding or function of the protein,

[1]These authors contributed equally to this work. [2]Howard Hughes Medical Institute and Department of Pharmacology, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA.

Correspondence should be addressed to R.R. *e-mail: rama@chop.swmed.edu*

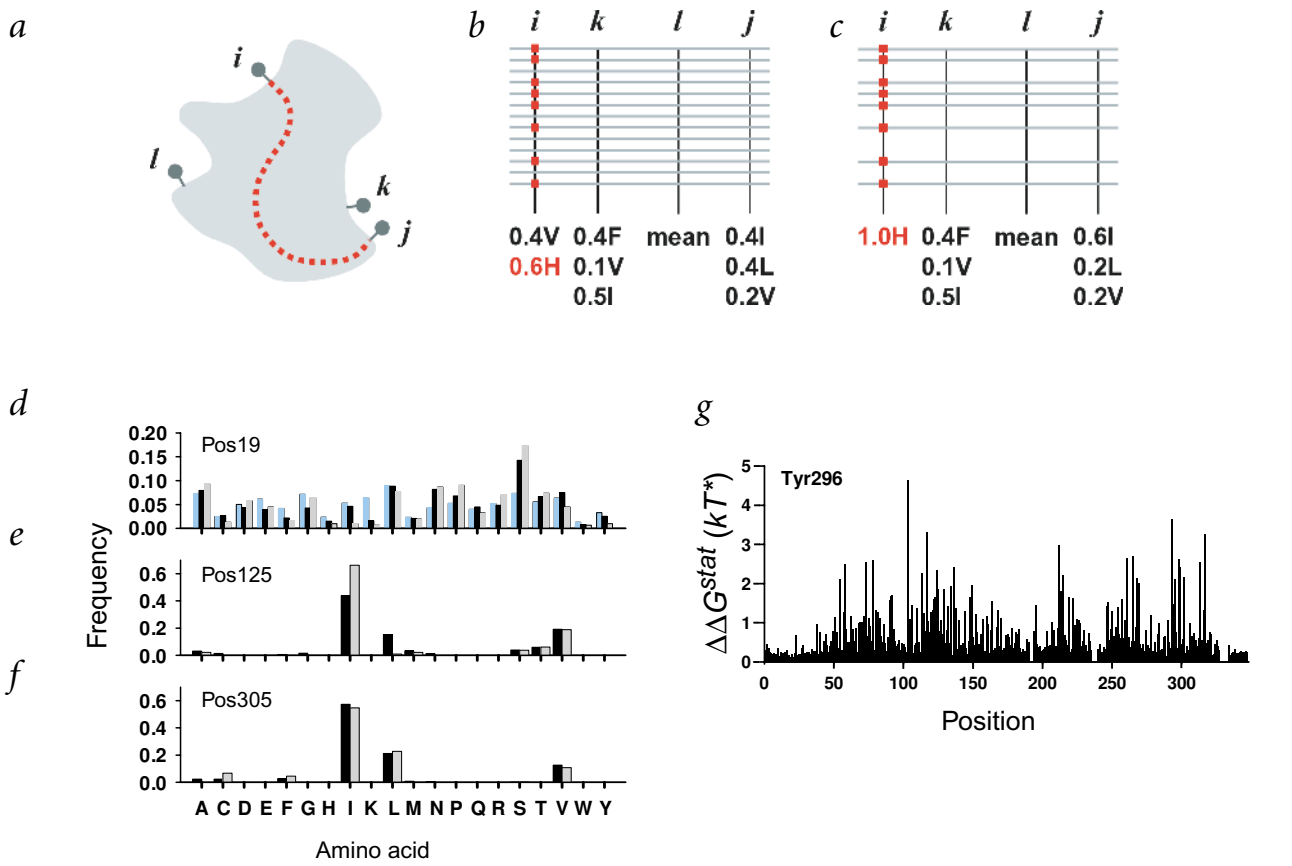

Fig. 1 A statistical perturbation method for measuring interactions between residues in proteins. ***a***, A hypothetical protein, showing four sites *i*, *j*, *k* and *l* with the following energetic properties: (i) *l* makes no contribution to structure or function of the protein, but *i*, *j* and *k* contribute in some way, and (ii) *l* and *k* act independently of one another, and *i* and *j* act cooperatively. Thus, *l* is energetically valueless, *i* and *k* are energetically additive, and *i* and *j* are energetically coupled. ***b***, Schematic representation of a large and diverse MSA of the protein family, where horizontal lines represent individual protein sequences. Site *l* (which is unconstrained) should show the mean distribution of amino acids found randomly in all natural proteins, whereas *i*, *j* and *k* should show some level of conservation (deviance from the mean distribution). ***c***, A subalignment resulting from a perturbation experiment on the MSA, where site *i* is fixed to histidine. If the subalignment is also large and diverse, site *l* should remain at the mean distribution, site *k* should display its independence from site *i* in the invariance of its distribution, and site *j* should reveal its coupling to site *i* in the change of its distribution. The change in an amino acid distribution at a given site upon perturbation at another can be calculated as a statistical coupling energy $\Delta\Delta G^{stat}$ between the two sites[15] (see Methods). ***d***, Amino acid distributions at positions ***d***, 19; ***e***, 125; and ***f***, 305 before (in black) and after (in gray) a perturbation at position 296 (fixed to tyrosine) in an alignment of 940 class A GPCRs. The graph for position 19 also shows the mean frequency of amino acids in all natural proteins (blue). Position 19 is nearly unconstrained in the GPCRs, shows little change upon perturbation Tyr296 and a low coupling energy ($\Delta\Delta G^{stat}_{19,296Y} = 0.12kT^*$). Position 125 shows changes to its distribution, and a large coupling energy ($\Delta\Delta G^{stat}_{125,296Y} = 1.86kT^*$). Position 305 displays conservation similar to 125, but shows little change upon the Tyr296 perturbation, and a low coupling energy ($\Delta\Delta G^{stat}_{305,296Y} = 0.3kT^*$). ***g***, The complete pattern of statistical coupling for the Tyr296 perturbation over all other GPCR positions *j* ($\Delta\Delta G^{stat}_{j,296Y}$). A relatively small set of positions, widely distributed in the primary structure, show large coupling to perturbation of 296.

the corresponding amino acid frequencies in the MSA should be unconstrained and, therefore, should approach their mean values in all proteins. However, if sites *i*, *j* and *k* make some contribution, the amino acid distributions at these sites should deviate from these mean values, and the extent of this deviation should provide a quantitative measure of the underlying evolutionary constraint (conservation). Second, the functional coupling of two sites *i* and *j* should exert a mutual evolutionary constraint between these sites, which should be encoded in the statistical coupling of the underlying distributions of amino acids. That is, the distribution of residues at site *j* should depend on those at site *i*. It then follows that a lack of functional interaction between two sites *i* and *k* should, regardless of conservation at both sites, result in independence of their amino acid distributions.

To measure the mutual dependence of two sites on a protein, we carry out a perturbation experiment on the MSA in which we introduce a change to the amino acid distribution at one position and examine the effect at another site (Fig. 1*b*,*c*). For example, extracting only the sequences that contain histidine at position *i* results in a subalignment (Fig. 1*c*) in which position *i* has experienced a substantial statistical perturbation (the fraction of histidine changes from 0.6 to 1.0). If the subalignment still retains sufficient size and diversity so that it remains a representative ensemble of the fold family, then the following properties should hold. First, site *l*, which was not conserved in the parent alignment, should still show an amino acid distribution near the mean in all proteins. Second, site *k*, which we defined as conserved but not coupled to site *i*, should remain unchanged in its amino acid distribution. Finally, the coupling of sites *i* and *j* should induce a change in the observed distribution at site *j* upon perturbation at *i*. As described[15], the magnitude of this change can be quantitatively measured as a statistical coupling energy between position *j* and the perturbation at *i* ($\Delta\Delta G^{stat}_{j,i}$; see Methods).

As an example of this sequence-based statistical experiment, consider an alignment of 940 members of the class A GPCR

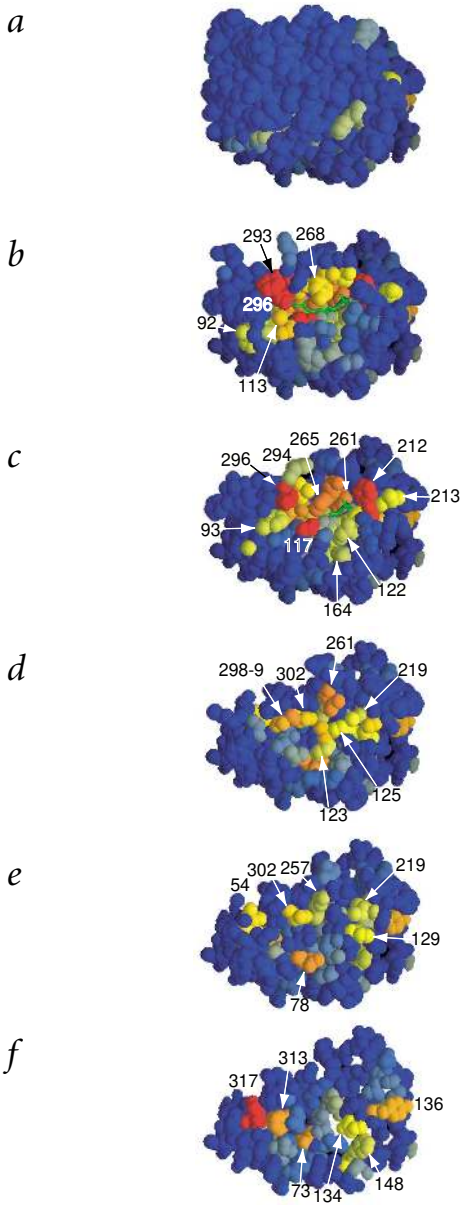

Fig. 2 Mapping statistical coupling for the Tyr296 perturbation in GPCRs (Fig. 1*g*) onto the structure of a representative member of the protein family, bovine rhodopsin. Shown are serial sections through the receptor ***a***, starting from the extracellular face and ***f***, ending at the cytoplasmic surface. The orientation is normal to the plasma membrane, looking from the extracellular side. The intermediate sections show the ***b***, top or ***c***, bottom of the ligand-binding pocket and ***d***,***e***, two successive sections below, essentially following the flow of information from the ligand interaction site to the G protein interaction site. The covalently bound ligand in rhodopsin (11-*cis*-retinal) is shown in green (*b*,*c*), and $\Delta\Delta G^{stat}$ values are mapped onto a linear color scale ranging from blue (0.6kT*) to red (3.5kT*). Coupled residues occur both in the neighborhood and far from position 296 and form a connected network through the structure. Residues scoring above 1.3kT* (yellow) comprise only 14% of all residues and 22% of core residues in the GPCR. The figure was prepared using GRASP[65].

family (see Methods). Position 296 (bovine rhodopsin numbering used here and below) is a moderately conserved site located at the middle of the seventh transmembrane helix that forms a key determinant of ligand interaction in GPCRs[1,19]. We made a perturbation at this site by extracting the subalignment containing only tyrosine at this site (Tyr296), a manipulation that retains 34.6% of sequences from the parent alignment. Both the full MSA and the Tyr296 subalignment are sufficiently diverse and over-represented so that unconserved sites, such as position 19, show amino acid frequencies near to their mean values found in all proteins (Fig. 1*d*). As expected, the perturbation of Tyr296 does little to alter the outcome at position 19 (Fig. 1*d*, compare black and dark gray bars) and shows a low statistical coupling energy to this site ($\Delta\Delta G^{stat}_{19,296Y} = 0.12kT^*$; kT^* is an arbitrary energy unit[15]). In contrast, position 125, a moderately conserved site in the third transmembrane helix of GPCRs, shows several changes in its distribution in response to the Tyr296 perturbation (Fig. 1*e*) corresponding to a larger coupling energy ($\Delta\Delta G^{stat}_{125,296Y} = 1.86kT^*$). This coupling is highly significant ($P <$ 0.001); an average of 100 trials of random selection of 34.6% of sequences from the full MSA results in little change to the distribution of residues at position 125, and low associated statistical coupling energies ($\Delta\Delta G^{stat}_{125,random} = 0.17 \pm 0.1kT^*$). However, not all conserved sites show such coupling to the Tyr296 perturbation; position 305 shows conservation similar to position 125 but shows virtually no changes to its amino acid distribution (Fig. 1*f*) and, consequently, only weak coupling ($\Delta\Delta G^{stat}_{305,296Y} = 0.3kT^*$). This coupling energy is similar to that observed at this site for random exclusion of sequences ($\Delta\Delta G^{stat}_{305,random} = 0.18 \pm 0.11kT^*$, $P < 0.29$) and, therefore, is statistically insignificant.

This site-specific coupling between residues in GPCRs is not particularly surprising. Studies in many proteins demonstrate the principle of compensatory mutagenesis in which the phenotype resulting from a perturbation introduced at one site can be rescued (or enhanced) by a second-site perturbation[20–22]. In some cases, this effect can be rationalized as local volume compensation of residues[23,24] or as local charge compensation[25]. However, in the general case, second-site interactions may also arise from distantly positioned residues through propagated interactions. Regardless of mechanism, the statistical coupling analysis provides a quantitative parameter ($\Delta\Delta G^{stat}$) to measure these inter-residue interactions in the evolutionary record of a protein family.

A sparse network of coupling for GPCR position 296

The complete calculation of statistical coupling energies for all sites j ($\Delta\Delta G^{stat}_{j,296Y}$) describes the effect of perturbation of position 296 over all GPCR positions (Fig 1*g*). These data were mapped on serial sections through the atomic stucture of a prototypcial GPCR family member, bovine rhodopsin (Fig. 2). Position 296 makes a direct interaction with ligands in several GPCRs[19]; in rhodopsin, a lysine at this position serves as the Schiff base–attachment site for the covalently bound chromophore, 11-*cis*-retinal (green, Fig. 2*b*,*c*). Similar to position 305, most positions show only weak coupling to the Tyr296 perturbation (blue, Fig. 2), demonstrating that co-evolution with position 296 is a special property of only a few positions. We identify three classes of residues that show highly coupled evolution with position 296. The first comprises a set of residues in the immediate environment of 296 (Fig. 2*b*,*c*)[2,26]. For example, Phe293, Leu294, Ala295, Ala299 and Phe91 in rhodopsin make intra- and inter-helical packing interactions with position 296, and Glu113 makes a salt-bridge interaction with the protonated form of the Schiff base[27], which is critical in maintaining the inactive configuration[28]. In adrenergic receptors, residues at both 296 and 113 contribute to ligand interactions, and mutations at either position in several GPCRs leads to loss of allosteric control and

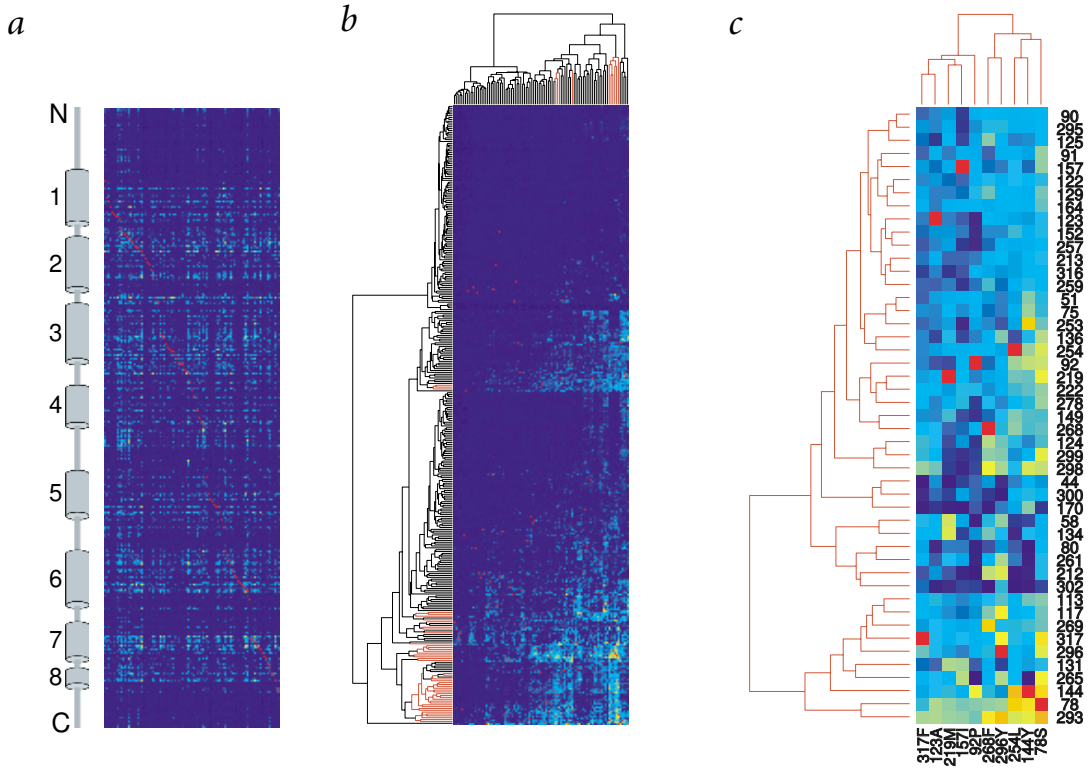

Fig. 3 Cluster analysis of statistical coupling in the GPCR family. ***a***, A matrix representation of all $\Delta\Delta G^{stat}$ values for 106 perturbation analyses on the GPCR MSA, where rows represent positions (from N to C terminus, top to bottom) and columns represent perturbations (N to C terminus, left to right). Thus, each column is one bar graph as in Fig. 1*g*, with $\Delta\Delta G^{stat}$ values shown colorimetrically on a linear scale covering the full range of the data (0kT*, blue, to 4.5kT*, red). In each column, sites of perturbation are shown in bright red. ***b***, Two dimensional clustering of the matrix in (*a*) reveals relationships between positions and perturbations through their pattern of coupling. Thus, two positions would cluster closely if they show a similar pattern of coupling to all perturbations, and two perturbations would cluster closely if they show a similar pattern of coupling to all positions. The data show that most positions are not coupled to any perturbation and act as if they are evolutionarily independent, and a few positions cluster together that show high amplitude $\Delta\Delta G^{stat}$ values for some perturbations. ***c***, Iterative focusing of the clustering around the regions in (*b*) of high $\Delta\Delta G^{stat}$ values produces a final cluster of 47 positions and 10 perturbations. The cluster shows self-consistency: the included positions are related by perturbations at sites within the cluster. Positions and perturbations in the initial clustering dendrogram that correspond to the final cluster are colored brown.

constitutive receptor activity[29]. These short-range statistical interactions revealed by perturbation of 296 intuitively make sense; we expect the effect of a perturbation in a protein to have at least some local effect in the structure.

The second class of residues coupled to position 296 comprises a linked network extending parallel to the plasma membrane from 296 to form the bottom of the ligand-binding pocket (Fig. 2*c*). These include a cluster of aromatic residues that, in rhodopsin, bind to the cyclohexenyl ring of retinal (Phe261, Trp265, Tyr268 and Phe212)[2,19,27]. These sites are also known determinants of ligand specificity or receptor activation in several GPCR subfamilies[1,19]. In summary, these two classes of residues describe much of the GPCR ligand-binding pocket that is thought to mediate transduction of ligand binding (or isomerization) to initial conformational change in the receptor molecule.

The third class of residues statistically coupled to perturbation of 296 forms a structure that is not obvious from inspection of the atomic structure: a sparse but contiguous network of interhelical interactions linking the ligand-binding pocket with the cytoplasmic surface (Fig. 2*d–f*). At the section immediately beneath the ligand-binding pocket (Fig. 2*d*), a small group of coupled residues form packing interactions between helix 3 (Gly121, Ile123 and Leu125) and helices 5 (Ile219), 6 (Phe261) and 7 (Ala299, Ser298 and Asn302)[27]. Many of these residues are known to be important in the allosteric activation of GPCRs[1,19]. For example, mutations at positions 125 and 261 cause constitutive receptor activity[30–32], and position Asn302 forms part of a well-known motif in GPCRs (NPXXY)[1,2,33], which has been linked to both stabilizing the inactive conformation[34] and mediating receptor desensitization[35]. In the next section below (Fig. 2*e*), coupled residues break up into discrete clusters of interhelical packing interactions: Asn302 and Met257 interact between helices 6 and 7, Val129 and Ile219 interact between helices 3 and 5, and Ile54 on helix 1 vertically contacts Ala299 at the next level up[27]. Met257 helps maintain the inactive conformation of the receptor and is thermodynamically coupled to both Glu113 and Glu134 (see below) in stimulus-dependent receptor activation[34]. Finally, these interactions then connect to two terminal sites at the cytoplasmic surface of the receptor: (i) the bottom of helix 3 and cytoplasmic loop 2 (Glu134, Tyr136 and Phe148) and (ii) residues in cytoplasmic helix 8 (Phe313 and Met317), helix 1 (Thr58) and the beginning of helix 2 (Asn73) (Fig. 2*f*). These two regions undergo a structural change upon light activation in rhodopsin[2,36–40], contain sites that likely contact the G protein α subunit and display loss of allosteric control upon mutation[2,41–46]. Thus, the pattern of coupling for the Tyr296 perturbation comprises a sparse but connected network of core residues that largely describes signal flow through the GPCR from initiation at the ligand-binding pocket to the final conformational change at the G protein–binding sites.

Cluster analysis of perturbations in the GPCR MSA

If the network of coupled residues described above constitutes a general functional unit for allosteric signaling in the receptor family rather than an isolated property of the Tyr296 perturbation, a similar pattern of coupling should be revealed through perturbation experiments at many sites in the GPCRs. To address this, we carried out a complete perturbation scan of the GPCR MSA involving 106 sites in the receptor that satisfy the statistical criteria established above for the Tyr296 perturbation. Subalignments in each case contain sufficient size and diversity such that unconserved sites show little change in amino acid distribution relative to the full MSA ($\overline{\Delta\Delta G^{stat}} = 0.13 \pm .03kT^{\star}$, see Methods).

The complete perturbation analysis for the GPCRs is displayed as a matrix of $\Delta\Delta G^{stat}$ values, with positions in the MSA as rows and perturbation experiments as columns (Fig. 3*a*). Thus, each column represents the set of coupling energies for one perturbation over all GPCR positions (Fig. 1*g*), and each row represents the coupling for one position over all perturbation experiments. In each column, the position that is the site of perturbation is indicated in bright red. This matrix is a global representation of this protein family in which relationships between amino acid positions are described not in terms of real distances or thermodynamic energies, but as statistical energies reporting the co-evolution of many pairs of sites in the protein family.

To understand the information contained in the matrix, we carried out a two-dimensional cluster analysis to identify groups of amino acid positions that show a profile of strong mutual statistical coupling[15]. The logic of this cluster analysis is based on a simple proposition: if evolutionarily coupled networks of residues occur and are robust and conserved features of a protein family, then perturbations at network positions should redundantly identify each other. In principle, this should be revealed as a cluster of both positions and perturbations comprising the network in the matrix of coupling energies. This analysis has strong similarities with transcriptional profiling through DNA microarrays; consequently, we used an iterative clustering algorithm originally developed for identification of co-expressed gene clusters to identify clusters of statistically coupled residues[47]. Clustering of the GPCR matrix (Fig. 3*b*) demonstrates surprising simplicity in the global pattern of statistical coupling. The majority of positions (rows) show no coupling to any perturbation, a finding that suggests evolutionary independence of many positions. Two iterations of the clustering, each time focusing the matrix around regions of large $\Delta\Delta G^{stat}$ values, identifies a group of 47 positions and 10 perturbations (Fig. 3*c*) that form a self-consistent cluster; these positions show similar patterns of coupling and all the perturbations used to identify them come from within this set of positions. Interestingly, these positions form a network of van der Waals interactions that link the ligand-binding pocket with regions of the cytoplasmic face that correspond to G protein–interaction sites (Fig. 4). Residues mapped by this global analysis of statistical coupling are nearly the same as those derived from the single site analysis of the Tyr296 perturbation (compare Fig.4 and 2*a–f*). This finding is consistent with the mutual reciprocity in coupling expected for a cluster of co-evolving residues. The physical linkage of the network is particularly striking given that it comprises only 14% of total residues and 22% of residues buried in the core of the molecule and that no structural information was used in the prediction of these sites. Thus, the sparse mapping of interactions between residues found in the statistical coupling matrix corresponds to a sparse network of physical associations in the molecule.

The strong correlation of the residues we have identified with the large body of mutagenic, structural and dynamic analyses of GPCRs supports the model that this network represents a canonical structural basis for signal transduction in these receptors. However, we note that this conclusion does not exclude the possibility that they may also participate in determining the cooperative stability of the protein fold. Indeed, recent work has suggested that determinants of fold stability and allosteric communication may be largely overlapping aspects of protein structures, such that the unequal distribution of cooperative stabilizing interactions also provides the framework for functional cooperativity[48]. The ability to rapidly identify these coop-

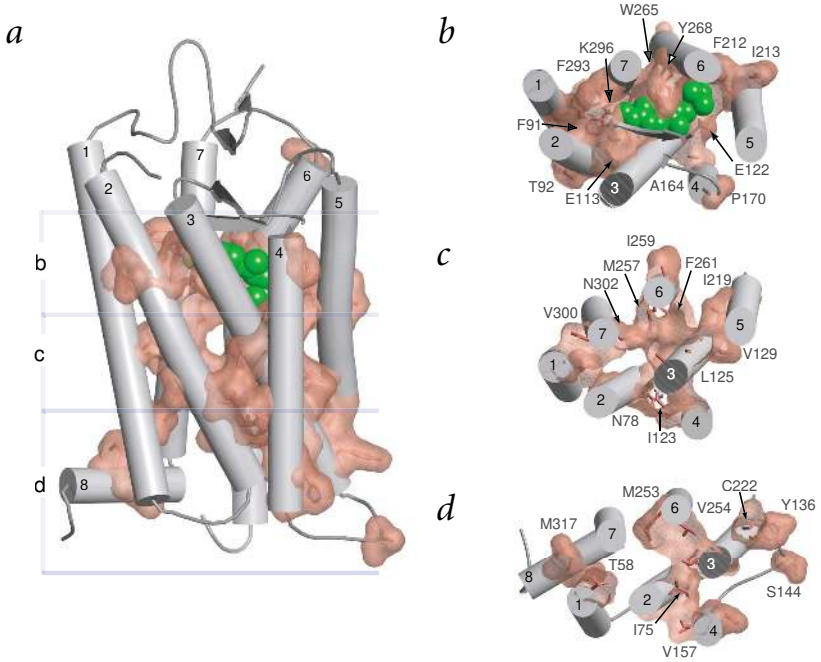

Fig. 4 A physically connected network of coupling between residues in the GPCR family. ***a***, The cluster of 47 positions identified in Fig. 3*c* is mapped onto the structure of bovine rhodopsin with the van der Waals surface associated with these residues in brown. ***b–d***, Serial sections through the receptor looking down from the external side (as in Fig. 2) at the levels indicated in (*a*). The data show that the full-scale analysis of coupling in the GPCR family describes a sparse but connected network of co-evolving residues within the core of the protein. This network connects the ligand-binding pocket with known G protein–binding regions through a few residues mediating packing interactions between helices. This figure and Figs. 6 and 8 were prepared with MolScript[66], GLRender (L. Esser, unpub.) and PovRay[67].

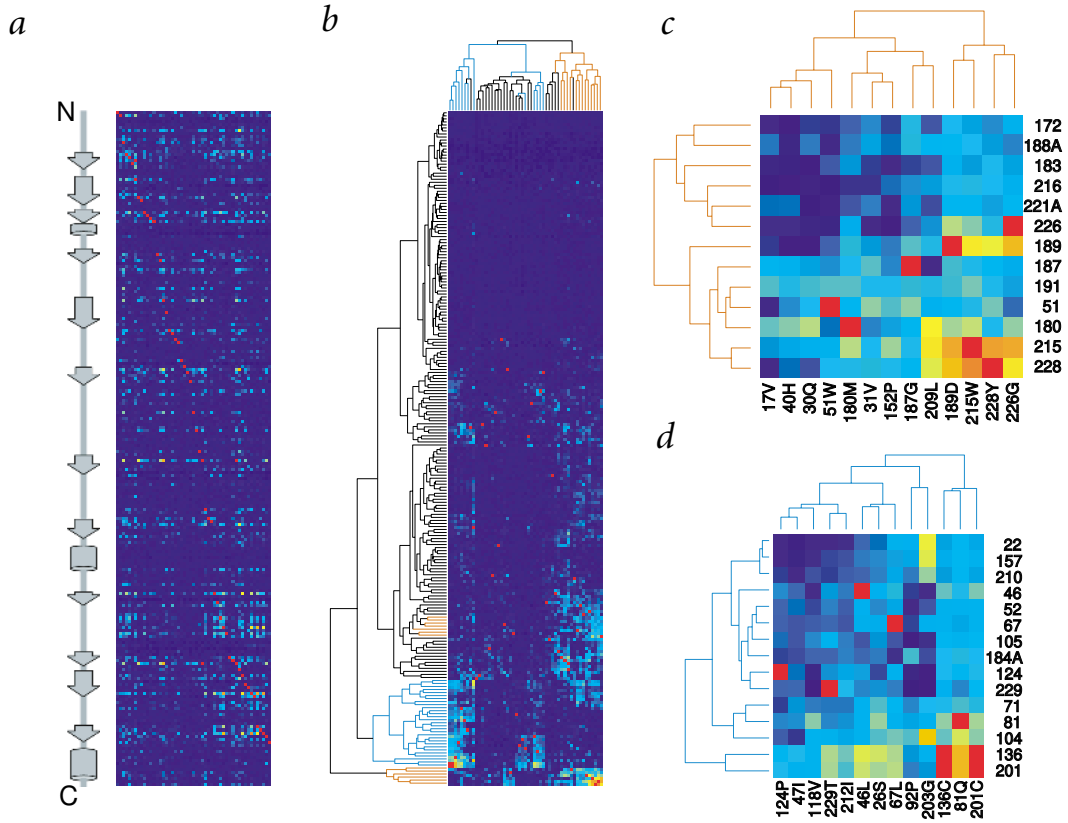

Fig. 5 A global analysis of statistical coupling in the chymotrypsin-family of serine proteases. ***a***, The statistical coupling matrix for this family containing 69 perturbations and organized as described for GPCRs (Fig. 3*a*). The color scale linearly maps the data from 0kT^* (blue) to 4.5kT^* (red). ***b***, Two dimensional clustering shows two separate co-evolving networks in this protein family (brown and blue in dendrograms). Iterative focusing around these two clusters leads to two final groups: ***c***, 1 and ***d***, 2.

erative interactions between residues may help in the further experimental testing of this idea.

Sparse but connected networks of coupling in other protein families

Do similar sparse networks of evolutionary coupling occur in other proteins, and do they correlate with functional mechanism? To study this, we carried out a global statistical coupling analysis of the chymotrypsin-like serine proteases and the hemoglobins, both well-studied cases of long-range energetic coupling.

Serine proteases. The major determinant of specificity in the chymotrypsin family of serine proteases is a deep pocket called the S1 site, which interacts with the so-called P1 residue on the substrate[49] (green, Fig. 6). In terms of catalytic power, a lysine at the P1 position is preferred in trypsin by >10^4-fold relative to phenylalanine, the preferred residue for chymotrypsin[4]. The bottom of the S1 pocket in trypsin contains a negatively charged residue (Asp189), suggesting a simple local electrostatic mechanism for selecting positively charged side chains at the P1 site. However, appreciable exchange of trypsin specificity to that of chymotrypsin requires not only mutation of position 189 and the entire S1 pocket, but also substitution of an unexpectedly distributed group of residues comprising three surface loops (L1, L2 and L3)[3,50,51]. These substitutions act cooperatively (rather than independently) in determining specificity at the S1 site, indicating some mechanism for thermodynamic coupling at a distance. Crystallographic studies of the native and chimeric proteins suggest a structural rationale: position 172 on loop L3, although distant from the S1 pocket, seems to influence specificity *via* stabilization of the L1 loop, which in turn supports the S1 pocket[10,52]. In support of this model, both computational and experimental studies show that the S1 pocket and its environment seem to display correlated dynamical motions upon substrate interaction[53,54]. Thus, specificity in substrate recognition at the S1 pocket depends on a set of distributed residues that act as a cooperative mechanical unit; however, the basic rules that define the spatial pattern of this unit are unknown.

To map the pattern of statistical coupling for this family, we constructed an alignment of 616 members of the chymotrypsin class of serine proteases (see Methods) and carried out a complete scanning perturbation analysis on the MSA comprising 69 site-specific perturbations (Fig. 5*a*). Two-dimensional clustering of the $\Delta\Delta G^{stat}$ data (Fig. 5*b*) shows that, similar to GPCRs, most positions show no coupling to all perturbations and a few positions, which show similar patterns of coupling, cluster together. Interestingly, we identify two independent clusters of co-evolving residues in this protein family (Fig. 5*b*), where the perturbations marked in blue or brown each identifies a distinct set of positions marked in the corresponding colors. Iterative focusing of the clustering around each of these two groups (Fig. 5*c*,*d*) leads to two self-consistent clusters. That is, each contains a small set of positions that show similar patterns of coupling, and most perturbations defining each cluster are at positions within the cluster.

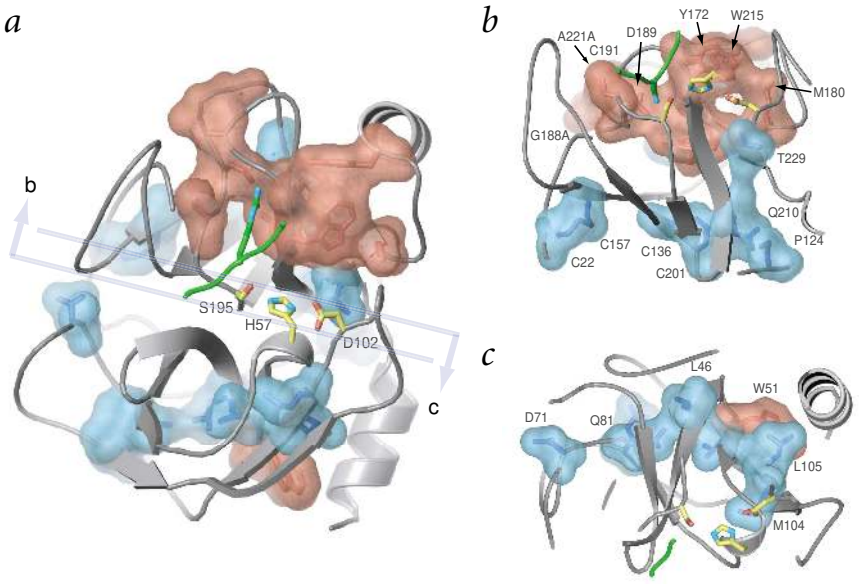

Fig. 6 Structural mapping of the two co-evolving clusters of residues in the chymotrypsin family of serine proteases. ***a***, Group 1 and group 2 residues from Fig. 5 mapped onto the structure of trypsin as brown and blue van der Waals surfaces, respectively. A substrate analog is shown (green) with the P1 residue bound within the S1 site on the protease, and the catalytic triad residues shown as stick bonds. Views of these data from slicing the trypsin molecule along the planes shown in (*a*) and rotating the ***b***, top and ***c***, bottom halves apart. Group 1 residues (with the exception of residue 51) form a connected unit around the S1 pocket that involves surface loops L1, L2 and L3 (brown). Group 2 residues form two pseudo-symmetric units within the structure, each in one of the β-barrel domains that comprise the chymotrypsin fold.

We mapped both groups of co-evolving residues as molecular surfaces on the atomic structure of bovine trypsin (Fig. 6). Group 1 (brown) comprises a largely contiguous network of residues related by van der Waals interactions. This network is built around the S1 site of the protease and propagates outward to include portions of the supporting surface loops L1, L2 and L3. Composed of a small fraction of either total residues (7%) or core residues (10.4%), the network is nearly exactly correlated with the experimental data mapping the propagated determinants of S1 pocket specificity. Indeed, the network includes both walls of the S1 pocket, supporting loops L1 and L2, and the distantly positioned Tyr172 on loop L3. We conclude that the group 1 network represents the mechanical coupling of residues that are the primary specificity determinant in the chymotrypsin class of serine proteases.

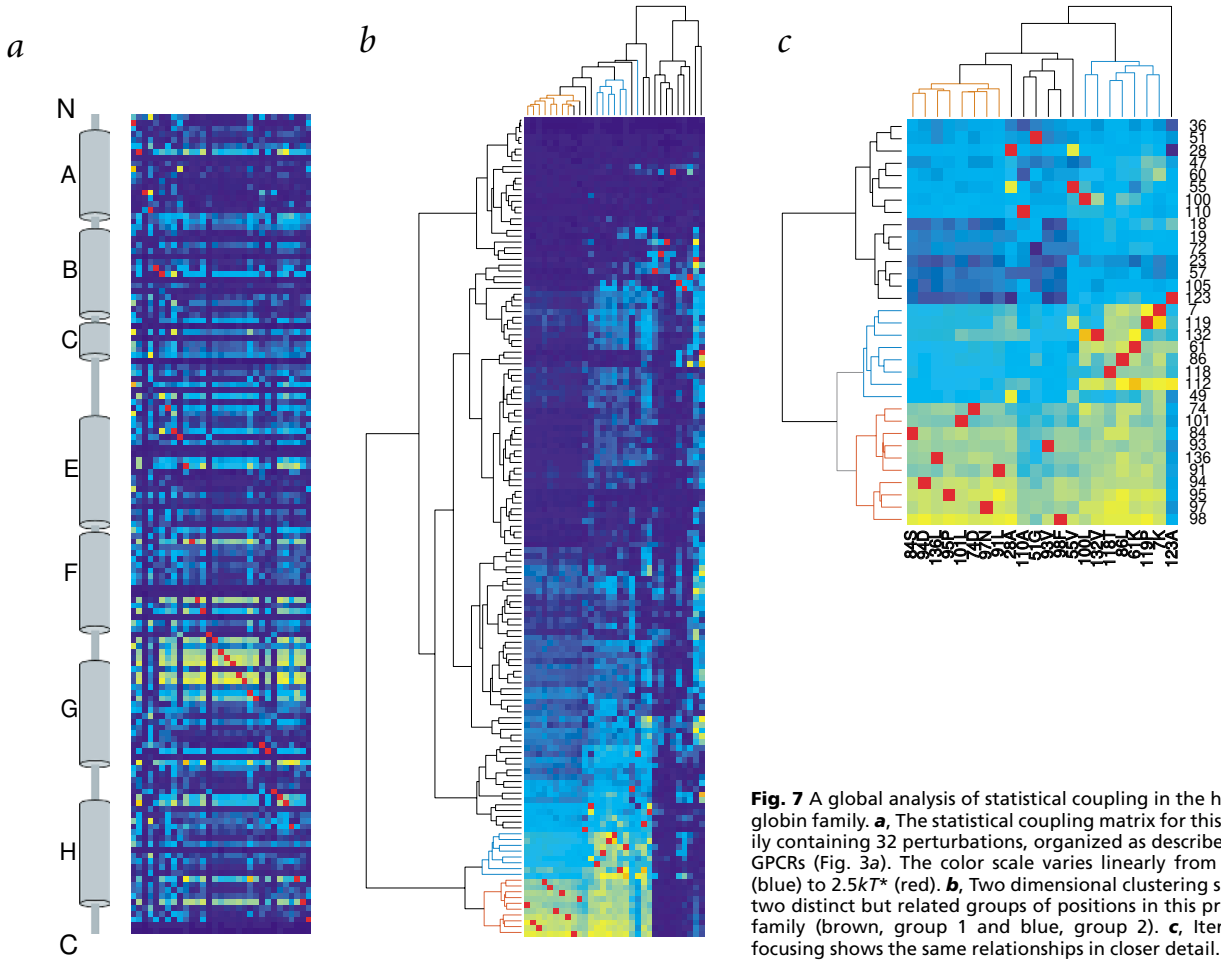

Fig. 7 A global analysis of statistical coupling in the hemoglobin family. ***a***, The statistical coupling matrix for this family containing 32 perturbations, organized as described for GPCRs (Fig. 3*a*). The color scale varies linearly from 0*kT** (blue) to 2.5*kT** (red). ***b***, Two dimensional clustering shows two distinct but related groups of positions in this protein family (brown, group 1 and blue, group 2). ***c***, Iterative focusing shows the same relationships in closer detail.

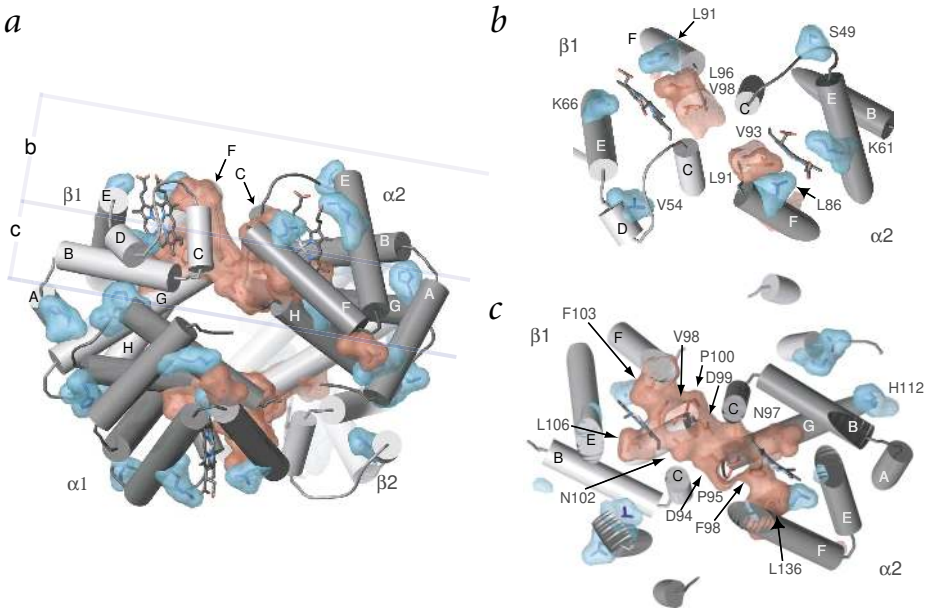

Fig. 8 Structural mapping of the two clusters of coupled residues in the hemoglobin family. ***a***, Group 1 and group 2 residues from Fig. 7 mapped onto the structure of the hemoglobin tetramer as brown and blue van der Waals surfaces, respectively. The hemes in each subunit are shown in stick bond representation. ***b***,***c***, Two serial sections through one pair of αβ subunits (α2β1) showing the tetramerization interface. Group 1 residues form a connected network relating the heme-binding sites on the two subunits across the tetramerization interface. Group 2 residues are generally peripheral to group 1 sites, and form either packing interactions with these sites, with heme, or form a few interhelical interactions.

The second group of statistically coupled residues (blue surface, Fig. 6) also constitutes a small set of positions (5.6% of total and 9.7% of the core residues) that comprise four contiguous units in the structure. The group 2 residues seem to form two pseudo symmetric substructures within the overall protease fold, each in one of the two β-barrels that comprise the structure (Fig. 6*b*,*c*). The interpretation of these observations is still unclear, but could represent other evolutionary constraints on the protease family, such as fold stability, zymogen activation or substrate recognition outside of the P1 site. However, the coevolution of this network around the catalytic triad suggests the alternate possibility that these residues may mediate the correlated motions of active site residues, a feature that may make precise positioning of catalytic residues possible. Further experiments will be necessary to address these possibilities.

The hemoglobin family. In the simplified but illustrative model for the mechanism of cooperative oxygen binding in hemoglobin, the tetramer of two α and two β subunits exists in rapid equilibrium between two stable conformations: a low-affinity oxygen-binding T-state and a high-affinity R-state[6]. The T-state maintains low-affinity oxygen binding through non-optimal positioning of residues of the heme-binding site, an arrangement that is stabilized through a distributed network of interactions with residues at the α1β2 and α2β1 tetramerization interfaces[6,7,55]. Crystallographic studies indicate that oxygen ligation to one subunit in the T-state initiates a local structural change in the heme-binding sites that propagates to these interfaces to eliminate key energetic interactions, allowing relaxation of the structure to the R form[6,9]. In effect, the network of interactions in each subunit connecting the heme-binding site to the α1β2 (and α2β1) protein interface represents a communication mechanism between the allosteric heme pairs in the T-state tetramer.

The resulting matrix of $\Delta\Delta G^{stat}$ values from a complete statistical perturbation scan of an alignment of 880 members (see Methods) of the globin family (Fig. 7) contains 32 site-specific perturbations (columns). Two-dimensional clustering shows that only a few positions show large amplitude coupling to perturbations and that these positions cluster together into well-defined groups. Two clusters of positions that show self-consistent patterns of coupling emerge, where again each cluster contains a set of positions related by perturbations within this set. Unlike the serine proteases, the clustering in this case also indicates that there is considerable overlap in the statistical coupling between the two clusters. For example, the perturbations representing group 2 (Fig. 7*c*, blue columns) also couple to group 1 positions, and perturbations in group 1 (brown columns) show significant coupling to group 2 positions.

Mapping of these data onto the tertiary structure of human hemoglobin shows that group 1 residues form a physically connected pathway of packing interactions that link the hemes across the α1β2 (and α2β1) tetramerization interfaces (brown surface, Fig. 8). Specifically, residues Leu136, Leu106, Phe98 and Val93 (α subunit numbering) make direct interactions with the heme (Fig. 8*b*,*c*) and, in turn, are linked to Pro95 and then to Asn97 and Asp94, which are located at the tetramer interface. In the hemoglobin nomenclature, the heme-contacting residues connect to the end of the F-helix, the FG corner and finally to the residues at the interface. A symmetric network then relates these residues on the other side of the interface to the β subunit heme. The group 1 network shows excellent correlation with the experimental data; the F helix and FG corner show structural changes upon oxygen ligation to the T-state of hemoglobin and the coupled residues at the tetramerization interface (α94, β102, α97 and β99; Fig. 8*c*) define the fulcrum around which the quaternary structure changes in making the T-to-R transition[9]. We conclude that this cluster of evolutionarily coupled residues is consistent with the known structural mechanism for allosteric communication across the tetramerization interfaces in hemoglobin.

The second group of statistically coupled residues (blue surface, Fig. 8) displays a pattern that is essentially peripheral to the group 1 network. Many of these residues directly contact the heme, pack against group 1 residues or comprise a few interhelical contacts between helices containing group 1 residues. These findings are consistent with the mutual coupling of these two groups as revealed in the clustering of the statistical coupling matrix; however, further experimental study of group 2 positions is necessary to understand their contribution to function.

Conclusions

Long-range communication is central to protein function, and it is not surprising that proteins have evolved specific mechanisms to address this constraint. Here we show that information about these mechanisms are embedded in the evolutionary record of a protein family and can be extracted through a systematic perturbation strategy measuring correlated evolution between residues. In three completely different protein families that represent signal propagation, catalytic specificity and allosteric binding, we find sparse networks of amino acid interactions. These networks connect known functional surfaces and show strong correlation with the large body of experimental data mapping the core functionality of each family. Thus, the statistical energy function representing co-evolution in a protein family provides a good mapping of the known physical interactions mediating protein function.

The statistical mapping described here is thermodynamic in nature and, therefore, provides no intrinsic information about the underlying mechanism of the interactions between residues. However, the finding of physical connectivity in co-evolving networks suggests that these may be mechanically coupled elements in the atomic structure that permit efficient propagation of local perturbations at a distance. Experimental data in all three protein families described here supports this hypothesis. In addition, recent work has provided important clues that site-specific correlated motions in proteins are key determinants of function[56–59]. For example, a small set of residues in the active site of the peptidyl-prolyl *cis-trans* isomerase cyclophilin show correlated motion in the time scale of substrate turnover, and these motions have been proposed to be fundamental for catalysis[59]. Thus, allostery within proteins may mechanistically amount to the coupled motions of a sparse and unequally distributed set of residues. These coupled motions may be the physical basis for the mutual evolutionary constraint that results in the statistical co-evolution of positions in protein families.

A central result from this study is the simplicity in the pattern of coupling between amino acids in proteins. Although, in principle, the pattern of all inter-residue interactions could be complex, reality seems to be much simpler. Most sites seem to act in an evolutionarily independent manner, uninfluenced by perturbations at many other sites, and a few positions form co-evolving linked networks through the structure. Such an architecture for mediating long-range communication in proteins is consistent with two seemingly incongruous properties of proteins: the extraordinary tolerance to mutagenesis at many positions but extreme sensitivity to perturbation at some sites, so that even subtle amino acid substitutions result in severe phenotypes. We suggest that these sparse networks are the result of a fundamental optimization process that guides sequence evolution: on one hand, the need for complexity (coupling) in proteins to make concerted activities possible but, on the other hand, the need for simplicity (independence) to make proteins robust to random mutagenesis.

Methods

Sequence alignments. The chymotrypsin-class serine protease and globin sequences were collected from the nonredundant database using PSI-BLAST[60] (*e*-score <0.001) and initially aligned using ClustalW[61]. Class A GPCR sequences were collected as an alignment from the GPCRDB and TinyGRAP database[62,63]. Alignments were then manually adjusted using standard structure-based sequence alignment techniques as described[64]. All alignments are available for download through our laboratory web site (*http://www.hhmi.swmed.edu/Labs/rr/SCA.html*).

Calculation of statistical parameters. The calculation of statistical coupling was carried out as described[15]. This analysis quantitatively measures the change in the amino acid distribution at one position *j* in an MSA given a perturbation at another position *i* as a statistical coupling energy between the two ($\Delta\Delta G_{j,i}^{stat}$). Briefly, each position *j* in the MSA is described as a 20-element vector of individual amino acid frequencies (for example, $\overline{f_j} = [f_j^{ala}, f_j^{cys}, f_j^{asp}, \ldots, f_j^{tyr}]$). The frequency vector is then converted to a vector of statistical energies ($\overline{\Delta G_j^{stat}} = [\Delta G_j^{ala}, \Delta G_j^{cys}, \Delta G_j^{asp}, \ldots, \Delta G_j^{tyr}]$), where each term is the value for amino acid *x* at site *j* and is given by $\Delta G_j^x = kT^* \ln (P_j^x / P_{MSA}^x)$, where kT^* is an arbitrary energy unit as described[15]. P_j^x is the binomial probability of observing amino acid *x* at site *j* given its mean frequency in all natural proteins.This calculation provides a logical basis for dealing with cases in which the frequency of an amino acid at a site is zero and accounts for the intuitive expectation that changes in the frequency of an amino acid when highly conserved should be scored higher than an equivalent frequency change when weakly conserved. P_{MSA}^x is the probability of observing amino acid *x* overall in the MSA and serves as a common reference state for all sites.

To measure coupling between a perturbation at *i* and any site *j*, we calculate the difference energy vector, $\overline{\Delta\Delta G_{j,i}^{stat}} = \overline{\Delta G_j^{stat}} - \overline{\Delta G_{j|i}^{stat}}$, where $\overline{\Delta G_j^{stat}}$ is the statistical energy vector of site *j* in the parent alignment and $\overline{\Delta G_{j|i}^{stat}}$ is that of site *j* in the subalignment derived from the perturbation at *i*. The scalar coupling energy ($\Delta\Delta G_{j,i}^{stat}$) is the magnitude of this difference vector and reports the combined effect of perturbation at on all amino acids at position *j*. If sites *i* and *j* are evolutionarily independent, the coupling energy is zero and is consistent with lack of interaction, but if the coupling energy is non-zero, the two sites interact to the extent measured by $\Delta\Delta G_{j,i}^{stat}$. Calculation of $\Delta\Delta G_{j,i}^{stat}$ for all sites *j* given a perturbation at *i* is a mapping of how all sites in the protein experience the effect of perturbing *i* (for example, Fig. 1e). The code implementing this algorithm and sample datasets are available to the scientific community by request, and further details are provided on our lab web site (*http://www.hhmi.swmed.edu/Labs/rr/SCA.html*).

Acceptance criteria for alignments and perturbations. Because the goal of this analysis is to expose functional (rather than historical) relationships between positions in the evolutionary record of a protein family, we apply two statistical criteria to validate the MSA for this analysis and one statistical criterion for validating sites on the MSA for perturbation. First, the MSA should be so diverse that several sites display amino acid distributions near the mean in all natural proteins (for example, Fig. 1*d*). If so, we conclude that the MSA has experienced substantial evolution and that the amino acid distributions at all sites are indeed reflective of the functional constraints on the protein family. Second, the MSA should be large enough that random elimination of sequences from the alignment does not considerably change the amino acid frequencies at the sites. If so, we can say that the MSA has reached a state of statistical equilibrium in sequence space, a necessary condition for applying Boltzmann statistics in the analysis. Finally, perturbations at sites in the MSA should produce subalignments that are also large and diverse such that they remain a representative subset of the parent MSA and do not substantially alter the state of statistical equilibrium. If so, unconserved sites (which by definition are not evolutionarily constrained) should remain unconserved and show coupling energies close to zero. A simple method that implements this rule is to apply a subalignment size cut-off for each protein family such that perturbations must produce subalignments with more than the cutoff number of sequences. The cutoff values were determined by graphing the average statistical coupling energy for five unconserved sites in each MSA upon random exclusion of varying numbers of sequences. In fractions of the total alignment size, the perturbations chosen for study in each family produce subalignments that are >0.32 (GPCRs), >0.49 (serine proteases) and >0.68 (globins). Information describing these criteria is provided on our lab web page (*http://www.hhmi.swmed.edu/Labs/rr/SCA.html*).

Matrix assembly and cluster analysis. The statistical coupling matrix for each protein family represents all MSA perturbations that meet the criteria described above and contains positions from N to C terminus as rows and perturbations (N to C terminus) as columns. The clustering algorithm to determine co-evolving net-

works of positions is based on iterative clustering methods developed for DNA microarray analysis[47]. The overall idea is to carry out sequential rounds of two-dimensional clustering, each time extracting the submatrix that contains positions and perturbations that cluster together in the previous iteration and that contain large signals. Thus, each iterative step is an attempt to refine the assignment of clusters by focusing the clustering algorithm around regions of positions *i* and perturbations *j* that show significant $\Delta\Delta G^{stat}_{j,i}$ values. If, as in the serine proteases, two independent clusters are found at a given iteration, then each submatrix is extracted and subjected to independent two-dimensional clustering at the next round. The city-block metric was used for calculating distances, and clustering was carried out using software written in MATLAB 6.1 (The Mathworks). Supplementary material describing this approach is provided on our lab web site (*http://www.hhmi.swmed.edu/Labs/rr/SCA.html*). Iterations are continued until no further refinement to clusters is found; protein families studied required three iterations at most to converge to stable clusters of positions.

Acknowledgments

We thank J. Albanesi, M. Brown, A. Gilman, and members of the Ranganathan lab for critical reading of the manuscript. This work was partially supported by a grant from the Robert A. Welch Foundation to R.R., who is also a recipient of the Burroughs-Wellcome Fund New Investigator Award in the Basic Pharmacological Sciences and the Mallinckrodt Scholar Award. M.A.W. is a Research Associate and R.R. is an Associate Investigator of the Howard Hughes Medical Institute.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 1 November, 2002; accepted 19 November, 2002.

1. Gether, U. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr. Rev.* **21**, 90–113 (2000).
2. Menon, S.T., Han, M. & Sakmar, T.P. Rhodopsin: structural basis of molecular physiology. *Physiol. Rev.* **81**, 1659–1688 (2001).
3. Hedstrom, L., Szilagyi, L. & Rutter, W.J. Converting trypsin to chymotrypsin: the role of surface loops. *Science* **255**, 1249–1253 (1992).
4. Hedstrom, L. Trypsin: a case study in the structural determinants of enzyme specificity. *Biol. Chem.* **377**, 465–470 (1996).
5. Patten, P.A. *et al*. The immunological evolution of catalysis. *Science* **271**, 1086–1091 (1996).
6. Perutz, M.F., Wilkinson, A.J., Paoli, M. & Dodson, G.G. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 1–34 (1998).
7. Perutz, M.F., Fermi, G., Luisi, B., Shaanan, B. & Liddington, R.C. Stereochemistry of cooperative mechanisms in hemoglobin. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 555–565 (1987).
8. Perutz, M.F. Stereochemistry of cooperative effects in haemoglobin. *Nature* **228**, 726–739 (1970).
9. Paoli, M., Liddington, R., Tame, J., Wilkinson, A. & Dodson, G. Crystal structure of T state haemoglobin with oxygen bound at all four haems. *J. Mol. Biol.* **256**, 775–792 (1996).
10. Perona, J.J., Hedstrom, L., Rutter, W.J. & Fletterick, R.J. Structural origins of substrate discrimination in trypsin and chymotrypsin. *Biochemistry* **34**, 1489–1499 (1995).
11. Williams, D.C. Jr., Benjamin, D.C., Poljak, R.J. & Rule, G.S. Global changes in amide hydrogen exchange rates for a protein antigen in complex with three different antibodies. *J. Mol. Biol.* **257**, 866–876 (1996).
12. Schreiber, G. & Fersht, A.R. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478–486 (1995).
13. Hidalgo, P. & MacKinnon, R. Revealing the architecture of a K^+ channel pore through mutant cycles with a peptide inhibitor. *Science* **268**, 307–310 (1995).
14. Carter, P.J., Winter, G., Wilkinson, A.J. & Fersht, A.R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus*). *Cell* **38**, 835–840 (1984).
15. Lockless, S.W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
16. Lichtarge, O., Bourne, H.R. & Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
17. Marcotte, E.M. *et al*. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
18. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
19. Ballesteros, J.A., Shi, L. & Javitch, J.A. Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors. *Mol. Pharmacol.* **60**, 1–19 (2001).
20. Saraceni-Richards, C.A. & Levy, S.B. Second-site suppressor mutations of inactivating substitutions at Gly247 of the tetracycline efflux protein, Tet(B). *J. Bacteriol.* **182**, 6514–6516 (2000).
21. Minor, D.L. Jr., Masseling, S.J., Jan, Y.N. & Jan, L.Y. Transmembrane structure of an inwardly rectifying potassium channel. *Cell* **96**, 879–891 (1999).
22. Cain, S.M., Matzke, E.A. & Brooker, R.J. The conserved motif in hydrophilic loop 2/3 and loop 8/9 of the lactose permease of *Escherichia coli*. Analysis of suppressor mutations. *J. Membr. Biol.* **176**, 159–168 (2000).
23. Zhang, H., Skinner, M.M., Sandberg, W.S., Wang, A.H. & Terwilliger, T.C. Context dependence of mutational effects in a protein: the crystal structures of the V35I, I47V and V35I/I47V gene V protein core mutants. *J. Mol. Biol.* **259**, 148–159 (1996).
24. Baldwin, E., Xu, J., Hajiseyedjavadi, O., Baase, W.A. & Matthews, B.W. Thermodynamic and structural compensation in 'size-switch' core repacking variants of bacteriophage T4 lysozyme. *J. Mol. Biol.* **259**, 542–559 (1996).
25. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U S A* **91**, 98–102 (1994).
26. Nakayama, T.A. & Khorana, H.G. Orientation of retinal in bovine rhodopsin determined by cross-linking using a photoactivatable analog of 11-*cis*-retinal. *J. Biol. Chem.* **265**, 15762–15769 (1990).
27. Palczewski, K. *et al*. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745 (2000).
28. Robinson, P.R., Cohen, G.B., Zhukovsky, E.A. & Oprian, D.D. Constitutively active mutants of rhodopsin. *Neuron* **9**, 719–725 (1992).
29. Porter, J.E., Hwa, J. & Perez, D.M. Activation of the α1b-adrenergic receptor is initiated by disruption of an interhelical salt bridge constraint. *J. Biol. Chem.* **271**, 28318–28323 (1996).
30. Yano, K., Kohn, L.D., Saji, M., Okuno, A. & Cutler, G.B. Jr. Phe576 plays an important role in the secondary structure and intracellular signaling of the human luteinizing hormone/chorionic gonadotropin receptor. *J. Clin. Endocrinol. Metab.* **82**, 2586–2591 (1997).
31. Andres, A., Kosoy, A., Garriga, P. & Manyosa, J. Mutations at position 125 in transmembrane helix III of rhodopsin affect the structure and signalling of the receptor. *Eur. J. Biochem.* **268**, 5696–5704 (2001).
32. Garriga, P., Liu, X. & Khorana, H.G. Structure and function in rhodopsin: correct folding and misfolding in point mutants at and in proximity to the site of the retinitis pigmentosa mutation Leu-125→Arg in the transmembrane helix C. *Proc. Natl. Acad. Sci. USA* **93**, 4560–4564 (1996).
33. Okada, T., Ernst, O.P., Palczewski, K. & Hofmann, K.P. Activation of rhodopsin: new insights from structural and biochemical studies. *Trends Biochem. Sci.* **26**, 318–324 (2001).
34. Han, M., Smith, S.O. & Sakmar, T.P. Constitutive activation of opsin by mutation of methionine 257 on transmembrane helix 6. *Biochemistry* **37**, 8253–8261 (1998).
35. Gripentrog, J.M., Jesaitis, A.J. & Miettinen, H.M. A single amino acid substitution (N297A) in the conserved NPXXY sequence of the human N-formyl peptide receptor results in inhibition of desensitization and endocytosis, and a dose-dependent shift in p42/44 mitogen-activated protein kinase activation and chemotaxis. *Biochem. J.* **352**, 399–407 (2000).
36. Meng, E.C. & Bourne, H.R. Receptor activation: what does the rhodopsin structure tell us? *Trends Pharmacol. Sci.* **22**, 587–593 (2001).
37. Farrens, D.L., Altenbach, C., Yang, K., Hubbell, W.L. & Khorana, H.G. Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* **274**, 768–770 (1996).
38. Altenbach, C., Klein-Seetharaman, J., Cai, K., Khorana, H.G. & Hubbell, W.L. Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 316 in helix 8 and residues in the sequence 60–75, covering the cytoplasmic end of helices TM1 and TM2 and their connection loop CL1. *Biochemistry* **40**, 15493–15500 (2001).
39. Altenbach, C., Cai, K., Klein-Seetharaman, J., Khorana, H.G. & Hubbell, W.L. Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 65 in helix TM1 and residues in the sequence 306–319 at the cytoplasmic end of helix TM7 and in helix H8. *Biochemistry* **40**, 15483–15492 (2001).
40. Dunham, T.D. & Farrens, D.L. Conformational changes in rhodopsin. Movement of helix f detected by site-specific chemical labeling and fluorescence spectroscopy. *J. Biol. Chem.* **274**, 1683–1690 (1999).
41. Cai, K. *et al*. Single-cysteine substitution mutants at amino acid positions 306-321 in rhodopsin, the sequence between the cytoplasmic end of helix VII and the palmitoylation sites: sulfhydryl reactivity and transducin activation reveal a tertiary structure. *Biochemistry* **38**, 7925–7930 (1999).
42. Altenbach, C., Cai, K., Khorana, H.G. & Hubbell, W.L. Structural features and light-dependent changes in the sequence 306–322 extending from helix VII to the palmitoylation sites in rhodopsin: a site-directed spin-labeling study. *Biochemistry* **38**, 7931–7937 (1999).
43. Fahmy, K. & Sakmar, T.P. Regulation of the rhodopsin-transducin interaction by a highly conserved carboxylic acid group. *Biochemistry* **32**, 7229–7236 (1993).
44. Franke, R.R., Konig, B., Sakmar, T.P., Khorana, H.G. & Hofmann, K.P. Rhodopsin mutants that bind but fail to activate transducin. *Science* **250**, 123–125 (1990).
45. Franke, R.R., Sakmar, T.P., Graham, R.M. & Khorana, H.G. Structure and function in rhodopsin. Studies of the interaction between the rhodopsin cytoplasmic domain and transducin. *J. Biol. Chem.* **267**, 14767–14774 (1992).
46. Kim, J.M., Altenbach, C., Thurmond, R.L., Khorana, H.G. & Hubbell, W.L. Structure and function in rhodopsin: rhodopsin mutants with a neutral amino acid at E134 have a partially activated conformation in the dark state. *Proc. Natl. Acad. Sci. USA* **94**, 14273–14278 (1997).
47. Getz, G., Levine, E. & Domany, E. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA* **97**, 12079–12084 (2000).
48. Luque, I., Leavitt, S.A. & Freire, E. The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annu. Rev. Biophys. Biomol. Struct.* **31**, 235–256 (2002).
49. Scheidig, A.J., Hynes, T.R., Pelletier, L.A., Wells, J.A. & Kossiakoff, A.A. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of Alzheimer's amyloid β-protein precursos (APPI) and basic pancreatic trypsin inhibitor (BPTI): engineering of inhibitors with altered specificities. *Protein Sci.* **6**, 1806 (1997).
50. Graf, L. *et al*. Electrostatic complementarity within the substrate-binding pocket of trypsin. *Proc. Natl. Acad. Sc.i USA* **85**, 4961–4965 (1988).
51. Hedstrom, L., Perona, J.J. & Rutter, W.J. Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry* **33**, 8757–8763 (1994).
52. Szabo, E., Bocskei, Z., Naray-Szabo, G. & Graf, L. The three-dimensional structure of Asp189Ser trypsin provides evidence for an inherent structural plasticity of the protease. *Eur. J. Biochem.* **263**, 20–26 (1999).
53. Mace, J.E. & Agard, D.A. Kinetic and structural characterization of mutations of glycine 216 in α-lytic protease: a new target for engineering substrate specificity. *J. Mol. Biol.* **254**, 720–736 (1995).
54. Davis, J.H. & Agard, D.A. Relationship between enzyme specificity and the backbone dynamics of free and inhibited α-lytic protease. *Biochemistry* **37**, 7696–7707 (1998).
55. Liddington, R., Derewenda, Z., Dodson, E., Hubbard, R. & Dodson, G. High resolution crystal structures and comparisons of T-state deoxyhaemoglobin and two liganded T-state haemoglobins: T(α-oxy)haemoglobin and T(met)haemoglobin. *J. Mol. Biol.* **228**, 551–579 (1992).
56. Stevens, S.Y., Sanker, S., Kent, C. & Zuiderweg, E.R. Delineation of the allosteric mechanism of a cytidylyltransferase exhibiting negative cooperativity. *Nat. Struct. Biol.* **8**, 947–952 (2001).
57. Nicholson, L.K. *et al*. Flexibility and function in HIV-1 protease. *Nat. Struct. Biol.* **2**, 274–280 (1995).
58. Osborne, M.J., Schnell, J., Benkovic, S.J., Dyson, H.J. & Wright, P.E. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry* **40**, 9846–9859 (2001).
59. Eisenmesser, E.Z., Bosco, D.A., Akke, M. & Kern, D. Enzyme dynamics during catalysis. *Science* **295**, 1520–1523 (2002).
60. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
61. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTALW: 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 (1994).
62. Beukers, M.W., Kristiansen, I., IJzerman, A.P. & Edvardsen, I. TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol. Sci.* **20**, 475–477 (1999).
63. Horn, F. *et al*. GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **26**, 275–279 (1998).
64. Doolittle, R., Abelson, J.N. & Simon, M.I. *Computer Methods for Macromolecular Sequence Analysis* (Academic Press, San Diego; 1996).
65. Nicholls, A., Sharp, K. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
66. Kraulis, P.J. MOLSCRIPT: a program to produce both detailed andschematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
67. Bacon, D. & Anderson, W.F. A fast algorithm for rendering space-filling molecule pictures. *J. Mol. Graph.* **6**, 219–220 (1988).

corrigendum

Evolutionarily conserved networks of residues mediate allosteric communication in proteins

Gürol M. Süel, Steve W. Lockless, Mark A. Wall and Rama Ranganathan

Nat. Struct. Biol. **10**, 59–69 (2003).

Because of resolution mismatch, Figs. 3*c*, 5*c*,*d*, 7*b*,*c* of this paper appear blurry in print. The pixels in these matrices have sharp edges, and the correct panels are now reprinted. We apologize for any inconvenience this may have caused.

Fig. 3c

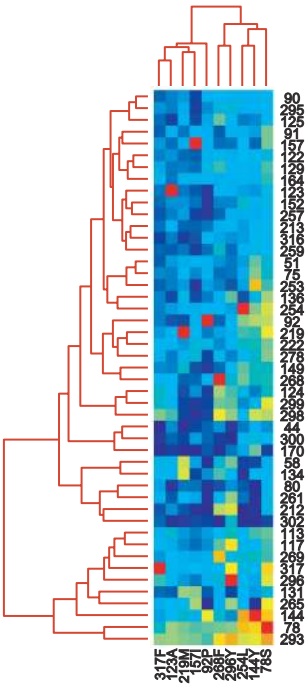

Fig. 5c

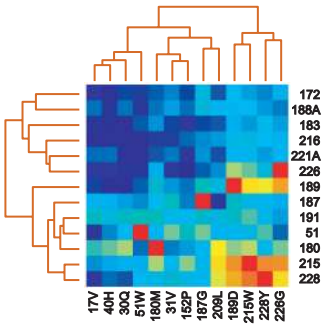

Fig. 5d

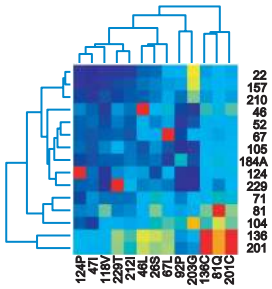

Fig. 7b

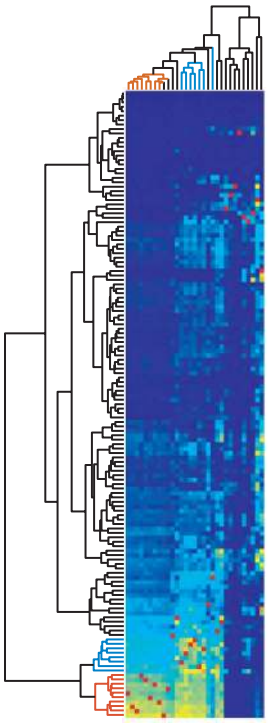

Fig. 7c