Protein Sectors: Evolutionary Units of Three-Dimensional Structure

Najeeb Halabi,[1,4] Olivier Rivoire,[2,4] Stanislas Leibler,[2,3] and Rama Ranganathan[1,*]
[1]The Green Center for Systems Biology, and Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX 75390-9050, USA
[2]The Center for Studies in Physics and Biology and Laboratory of Living Matter, Rockefeller University, New York, NY 10065, USA
[3]The Simons Center for Systems Biology and the School of Natural Sciences, The Institute for Advanced Study, Princeton, NJ 08540, USA
[4]These authors contributed equally to this work
*Correspondence: rama.ranganathan@utsouthwestern.edu

DOI 10.1016/j.cell.2009.07.038

SUMMARY

Proteins display a hierarchy of structural features at primary, secondary, tertiary, and higher-order levels, an organization that guides our current understanding of their biological properties and evolutionary origins. Here, we reveal a structural organization distinct from this traditional hierarchy by statistical analysis of correlated evolution between amino acids. Applied to the S1A serine proteases, the analysis indicates a decomposition of the protein into three quasi-independent groups of correlated amino acids that we term "protein sectors." Each sector is physically connected in the tertiary structure, has a distinct functional role, and constitutes an independent mode of sequence divergence in the protein family. Functionally relevant sectors are evident in other protein families as well, suggesting that they may be general features of proteins. We propose that sectors represent a structural organization of proteins that reflects their evolutionary histories.

INTRODUCTION

How does the amino acid sequence of a protein specify its biological properties? Here, we intend the term "biological properties" to broadly encompass chemical activity, structural stability, and other features that may be under selective pressure. A standard measure of the importance of protein residues is sequence conservation—the degree to which the frequency of amino acids at a given position deviates from random expectation in a well-sampled multiple sequence alignment of the protein family (Capra and Singh, 2007; Ng and Henikoff, 2006; Zvelebil et al., 1987). The more unexpected the amino acid distribution at a position, the stronger the inference of evolutionary constraint and therefore of biological importance. However, protein structure and function also depend on the cooperative action of amino acids, indicating that amino acid distributions at positions cannot be taken as independent of one another (Gobel et al., 1994; Lichtarge et al., 1996; Lockless and Ranganathan, 1999; Neher, 1994). A more informative formulation of sequence conservation should be to include pairwise or even higher-order correlations between sequence positions—the statistical signature of conserved interactions between residues. Indeed, analyses of correlations have contributed to the identification of allosteric mechanisms in proteins (Ferguson et al., 2007; Hatley et al., 2003; Kass and Horovitz, 2002; Lee et al., 2008, 2009; Peterson et al., 2004; Shulman et al., 2004; Skerker et al., 2008) and were found to be sufficient for recapitulating native folding and function in a small protein interaction module (Russ et al., 2005; Socolich et al., 2005).

These findings motivate a deeper theoretical and experimental analysis of correlations of sequence positions with the goal of understanding how protein sequences encode the basic conserved biological properties of a protein family. Here, we carry out this analysis using a classic model system for enzyme catalysis, the S1A family of serine proteases (Hedstrom, 2002; Rawlings and Barrett, 1994; Rawlings et al., 2008). We find that the nonrandom correlations between sequence positions indicate a decomposition of the protein into groups of coevolving amino acids that we term "sectors." In the S1A proteases, the sectors are nearly statistically independent, are physically connected in the tertiary structure, are associated with different biochemical properties, and have diverged independently in the evolution of the protein family. Functionally relevant and physically contiguous sectors are evident in other protein domains as well, providing a basis for directing further experimentation using the principles outlined in the serine protease family. Overall, our data support two main findings: (1) protein domains have a heterogeneous internal organization of amino acid interactions that can comprise multiple functionally distinct subdivisions (the sectors), and (2) these sectors define a decomposition of proteins that is distinct from the hierarchy of primary, secondary, tertiary, and quaternary structure. We propose that the sectors are features of protein structures than reflect the evolutionary histories of their conserved biological properties.

RESULTS

From Amino Acid Sequence to Sectors

The S1A family consists primarily of enzymes catalyzing peptide bond hydrolysis through a conserved chemical mechanism, but

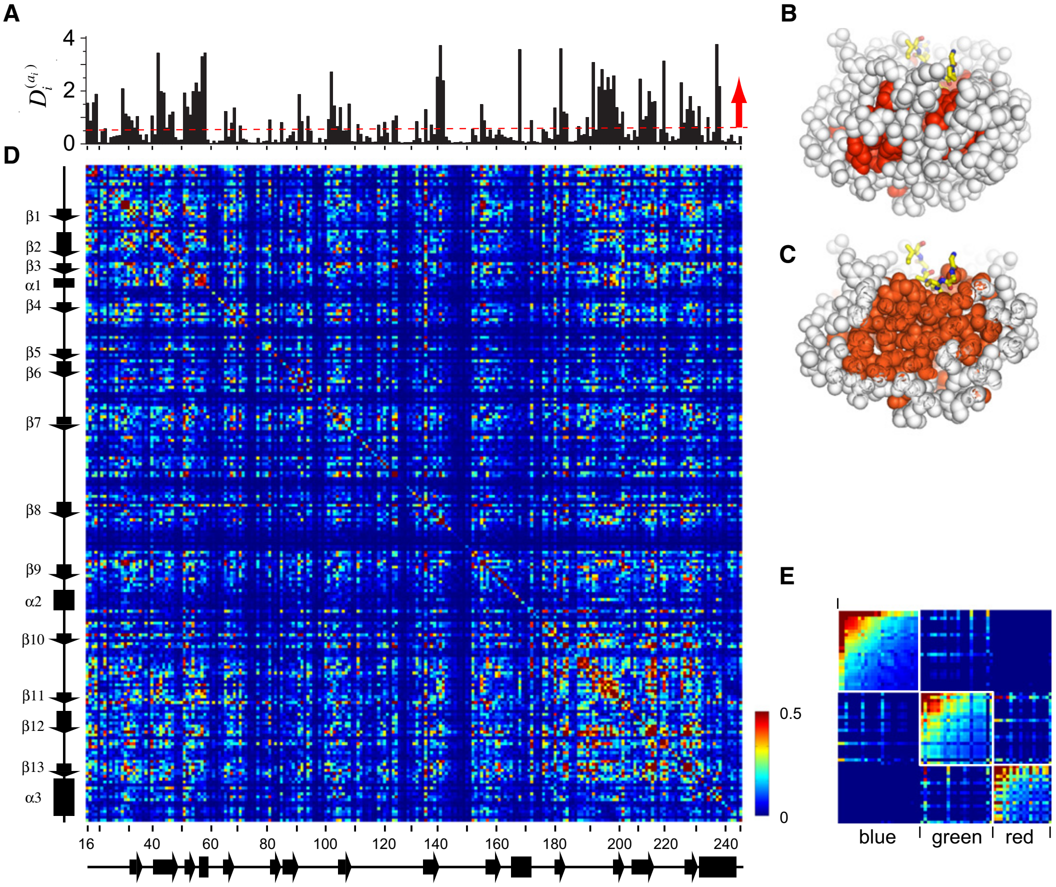

Figure 1. Position-Specific and Correlated Conservation in the S1A Protease Family

(A) The conservation of each position *i* in a multiple sequence alignment of 1470 members of the S1A family, computed by the relative entropy $D_i^{(a_i)}$(position numbering according to bovine chymotrypsin, and graph is aligned with the matrix below).

(B and C) Mapping of the moderate to strongly conserved positions in a surface view (B) and a slice through the core (C) of rat trypsin shows a simple and intuitive arrangement. Residues with $D_i^{(a_i)} > 0.5$ (in orange) occupy the protein core and regions contacting substrate, while less conserved positions are mostly located on the surface. The cutoff is chosen to color ~50% of residues to illustrate the pattern of conservation in the protein structure.

(D) SCA matrix $\tilde{C}_{ij}$ for a sequence alignment of 1470 members of the protease family, showing a pattern of correlated conservation that is distributed throughout the primary structure and across secondary structure elements.

(E) SCA matrix after reduction of statistical noise and of global coherent correlations (see the Supplemental Experimental Procedures and Supplemental Discussion). The 65 positions that remain fall into three groups of positions (red, blue, and green, termed "sectors"), each displaying strong intragroup correlations and weak intergroup correlations. In each sector, positions are ordered by descending magnitude of contribution (Figure S3), showing that sector positions comprise a hierarchy of correlation strengths.

its members show a broad range of substrate specificities and environments within which they operate. Analysis of positional conservation (see the Supplemental Experimental Procedures) in a multiple sequence alignment of 1470 members of the family reveals a pattern over sequence positions (Figure 1A) that has a simple and well-known structural interpretation: more conserved positions tend to be located in the core of the protein or at functional surfaces, and less conserved positions tend to occur on the remainder of the protein surface (Figures 1B and 1C) (Bowie et al., 1990; Chothia and Lesk, 1982; Lesk and Chothia, 1982).

To examine the contribution of correlations to conservation, we followed the statistical coupling analysis (SCA) approach (Lockless and Ranganathan, 1999) to compute a conservation-weighted covariance matrix between all sequence positions in the S1A family ($\tilde{C}_{ij}$, Figure 1D, see the Supplemental Experimental Procedures). Inspection of the matrix clearly indicates that correlations are not simply dominated by proximity in primary structure; many positions show only weak correlation to neighboring positions but significant correlation to positions that are distant along the sequence (Figure 1D).

What pattern of functional correlations within the serine protease does this matrix indicate? The essence of addressing this problem is two-fold: (1) to separate the functionally significant correlations in the $\tilde{C}_{ij}$ matrix (the "signal") from correlations that could arise due to limited sampling of sequences ("statistical noise") or phylogenetic relationships between sequences ("historical noise"), and then (2) to analyze the pattern of the remaining significant correlations.

Our approach for isolating signal from noise in the SCA correlation matrix derives from work more than 50 years ago on

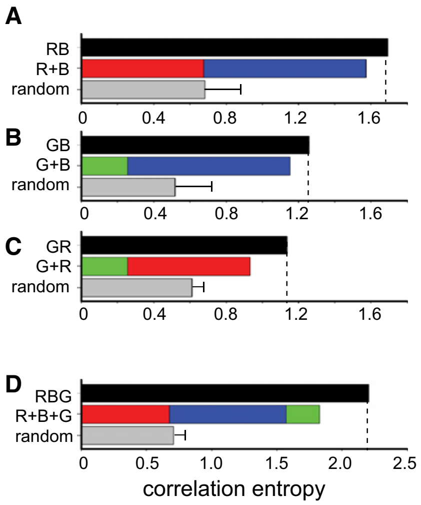

Figure 2. Statistical Independence of the Three Sectors

For each pairwise combination of sectors (A, red-blue [RB], B, green-blue [GB], and C, green-red [GR]) and the combination of all three sectors (D, red-blue-green [RBG]), the graph shows the total correlation entropy (black bar), summed sector correlation entropies (stacked colored bars), and the average summed correlation entropy for 100 random groupings (top five constituent residues; error bars represent the standard deviation). In each case, the summed entropies of the sectors are close to the total entropies and are far from that expected randomly.

random matrix theory (Wigner, 1967). The basic idea is to model the effect of statistical noise by examining correlation matrices for randomized versions of the data; significant patterns of correlations are then deduced by comparison. This approach was used in finance to extract nonrandom correlations of stock performance over a finite time window (Bouchaud and Potters, 2004; Plerou et al., 2002). This analysis showed that only a small fraction of observed correlations are relevant because most could arise simply by the limited period of time over which stock prices are sampled. The remaining significant correlations are organized in a few collective modes that decompose the economy into business sectors—groups of business entities whose performance fluctuates together over time. We applied these same methods to extract the nonrandom correlated modes of the SCA matrix, effectively “cleaning” the matrix of statistical noise (Figure S2A available online). As for the financial markets, we find that only a few top modes (five out of 223 total) contain correlations that are clearly distinct from random expectation.

The work in finance also provides a clue for reducing the effect of historical noise. Global, coherent correlations in stock performance occur due to fluctuations in the overall economy and are responsible for a dominant first mode of the correlation matrix. This mode is irrelevant for identifying the nonglobal, heterogeneous correlations between stocks that define the different business sectors and is therefore removed (Bouchaud and Potters, 2004; Plerou et al., 2002). Similarly, global, coherent correlations between positions should occur due to phylogenetic relationships between sequences and are expected to produce the dominant first mode observed for the SCA matrix (see the Supplemental Experimental Procedures and Supplemental Discussion). This mode is irrelevant for decomposing the protein sequence into functional units and is removed. Though the principles of computing correlations vary, similar approaches for partial elimination of purely phylogenetic correlations in protein sequence alignments have been previously described (Atchley et al., 2000; Buck and Atchley, 2005; Ortiz et al., 1999). The process is summarized in the Experimental Procedures, and a script for reproducing this analysis is provided in the Supplemental Data. The final result is shown in Figure 1E, a highly simplified representation of the SCA matrix that shows the statistically relevant pattern of correlation. For the S1A family, this analysis reveals two main findings: (1) the 223 sequence positions in the multiple sequence alignment are reduced to 65 positions that show significant patterns of correlations, and (2) these 65 positions can be separated into three seemingly distinct groups (labeled red, blue, and green). By analogy with the work in finance, and to distinguish from other terminologies used in describing protein structures, we refer to these groups of correlated positions as protein sectors—units of a protein that have coevolved within a protein family.

The concept of business sectors in the economy is clear, but what is the meaning of sectors in proteins? Using the S1A family as a model system, we identify four characteristics of sectors: (1) statistical independence, (2) physical connectivity in the tertiary structure, (3) biochemical independence in mediating protein function, and (4) independent phenotypic variation in the protein family.

Statistical Independence

Figure 1E provides a qualitative picture of independence between sectors in the S1A family, but quantitatively, how independent are they? To address this, we computed a measure called the correlation entropy—the degree to which a selected group of residues are statistically coupled to each other in the multiple sequence alignment (Figures 2 and S6, and see the Supplemental Discussion). If two sectors are independent, then the correlation entropy of two taken together must be the sum of their correlation entropies taken individually. Figure 2 shows that for all pairs of sectors (Figures 2A–2C) and for the three sectors combined (Figure 2D), this condition holds to a remarkable degree. For example, the correlation entropy of the red and blue sectors taken together (Figure 2A, black bar) is nearly that of the sum of the individual sector correlation entropies (stacked red and blue bars), and much different from random expectation (gray bar). Overall, the data in Figures 2A–2D show that the red, blue, and green sectors represent highly independent statistical units in the serine protease family. This analysis also permits a quantitative comparison of the degree of independence; the red/blue and green/blue sectors emerge as the most independent (Figures 2A and 2B), while the red and green sectors show less independence (Figure 2C). For example, G216 and V213 are jointly shared by both the red and green sectors (Figure S3F), suggesting that they represent sites of interaction between these two otherwise independent sectors.

Structural Connectivity

The identification of independent protein sectors in the serine protease is entirely based on statistical analysis of the sequence alignment without any consideration of the protein structure or its

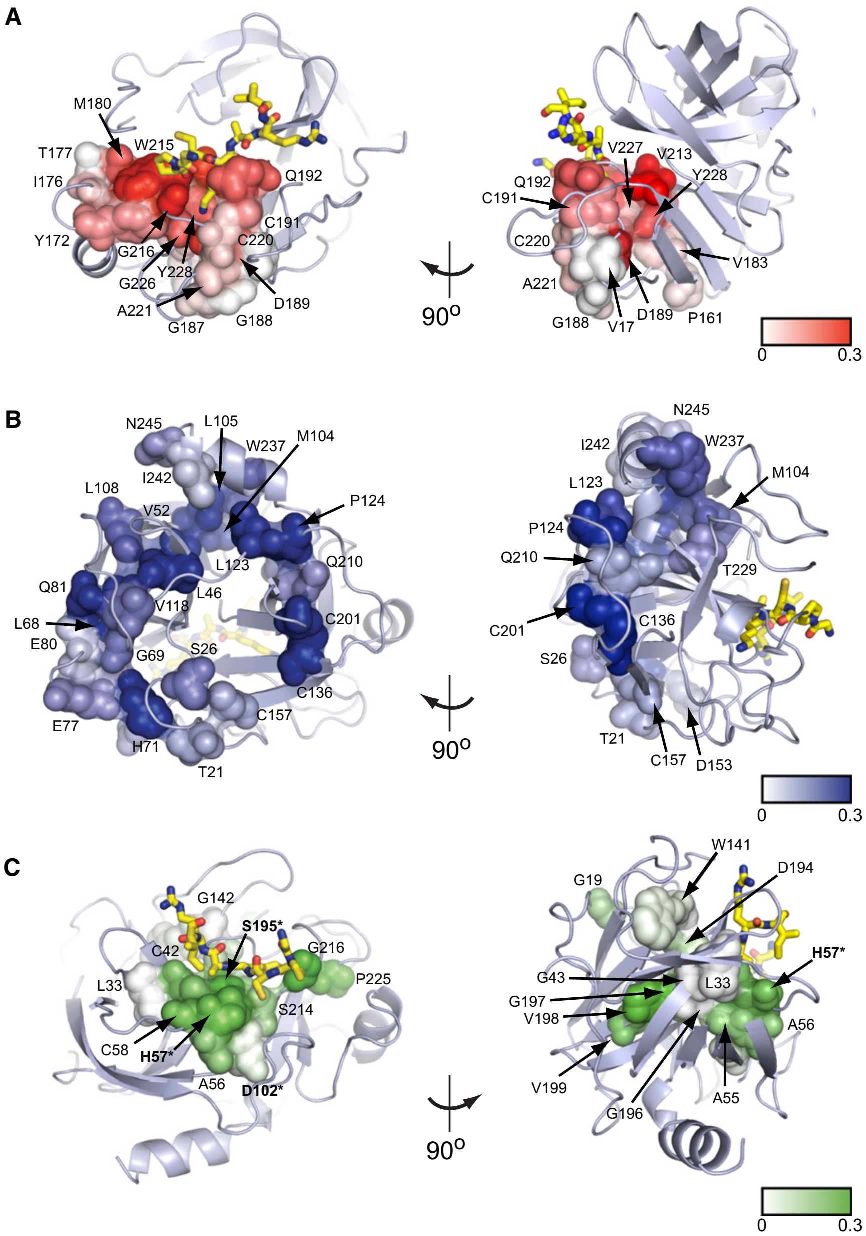

Figure 3. Structural Connectivity of the Three Sectors

Residues comprising each sector displayed in space filling representation with a van der Waals surface on the tertiary structure of rat trypsin (PDB 3TGI [Pasternak et al., 1999]). The blue sector comprises a ring of residues within the core of the two β barrels (B), the red sector comprises the S1 pocket and its environment (A), and the green sector comprises the catalytic mechanism of the protease located at the interface of the two β barrels (C). In each panel, residues are gradient colored by strength of contribution to the sector (Figures 1E and S3).

biochemical properties. Nevertheless, the sectors have clearly interpretable tertiary structural properties (Figure 3). The red sector comprises a contiguous network of amino acids built around the S1 pocket, the primary determinant of substrate specificity (Hedstrom, 2002) (Figure 3A). The color gradient in Figure 3A represents residue weights, revealing a tertiary structural organization in which the strongest contributors are centered around the S1 pocket and weaker positions comprise the surrounding. This sector includes residues in the environment of the S1 pocket that are known to contribute to its mechanical stability, providing a rationale for their cooperative action (Bush-Pelc et al., 2007; Perona et al., 1995). This sector is clearly involved in catalytic specificity; mutation of residues comprising this sector are known to influence specificity for substrates in several S1A family members (Craik et al., 1985; Hedstrom, 1996; McGrath et al., 1992; Perona et al., 1993; Wang et al., 1997), and this sector correlates well with positions mutated in transferring chymotryptic specificity into trypsin (Hedstrom et al., 1994).

The blue sector comprises another contiguous group of amino acids, but is structurally distinct from the red sector; the constituent residues run through the interior of both of the β barrels that comprise the core structure of the protease (Figure 3B), but also extend from both β barrels to directly contact the catalytic triad residues (Figure 4B). Mapping of residue weights in this sector indicates a few foci joined by intervening positions with lower weights, as if the activity of this sector is a more distributed rather than localized property of the protein structure. Unlike the red sector, prior work establishes no unified role for this sector, likely because blue sector residues are not obviously distinguishable from the general milieu of residues in the protein core that are similarly conserved and well-packed (Figures 4 and S7).

Finally, the green sector forms another contiguous group of amino acids, located at the interface between the two β barrels that make up the protease (Figure 3C). Residues within this sector include the catalytic triad (H57, D102, and S195), and surrounding residues known to be important for the basic chemical mechanism of this enzyme family (Baird et al., 2006; Hedstrom, 2002), and for some forms of allosteric control over this activity (Guinto et al., 1999; Huntington and Esmon, 2003). Like the red sector, residue weights are largest around a hotspot (the catalytic residues), and fall off in surrounding positions. This sector includes one disulfide bond pair (C42-C58), substitution of which has been shown to cooperatively interact with mutation of S195 (Baird et al., 2006). Indeed, triple mutation of C42A, C58A/V, and S195T is sufficient to convert trypsin from a serine protease to a threonine protease. We conclude that the green sector

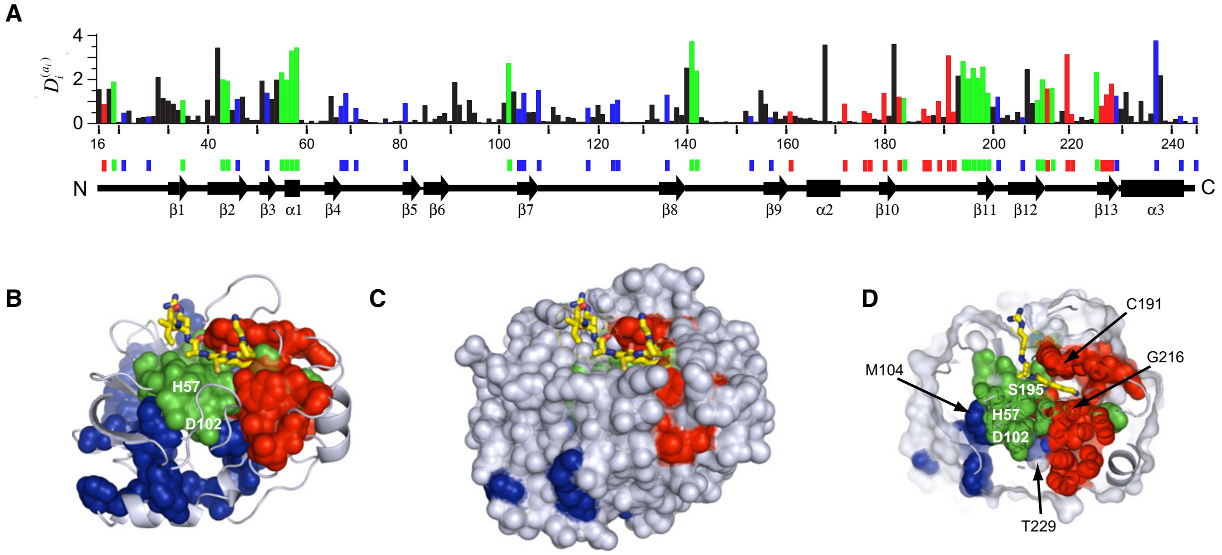

Figure 4. Relationship of Sectors to Primary, Secondary, and Tertiary Structure

(A) Positions colored by sector identity on the primary and secondary structure of a member of the S1A family (rat trypsin); the bar graph shows the global conservation of each position.

(B) The red, blue, and green sectors shown together on the three dimensional structure of rat trypsin (PDB 3TGI); sectors occupy different regions but make contacts with each other at a few positions.

(C) A space filling representation in the same view as (B), showing that all sectors are similarity buried in the protein core.

(D) A slice through the core of rat trypsin at the level of the catalytic triad residues (labeled in white), with sector positions in colored spheres and the molecular surface of the protein in gray. Two blue sector positions (M104 and T229) and two red sector positions (C191 and G216, which is shared with the green sector) that are similarly buried and proximal to catalytic triad residues are highlighted.

represents the catalytic core of the protease family. Consistent with joint contribution to both the red and green sectors, positions 213 and 216 are found to form a major part of the packing interface between these two sectors (data not shown).

More generally, the physical connectivity of each sector is striking, given that no information about tertiary structure was used in their identification and that only ~10% of total sequence positions contribute strongly to each sector. Shown together, the three sectors occupy largely distinct subdivisions within the core of the tertiary structure, making contacts only at a few positions (Figures 4B–4D). The considerable prior experimental work on the serine proteases permits the partial functional interpretation of sectors provided above, but it is important to note that the sectors are otherwise not obvious. No sector corresponds to any known subdivision of proteins by primary structure segments, secondary structure elements, or subdomain architecture (Figure 4A). In addition, the three sectors are not distinguishable by degree of solvent exposure, by the conservation of positions taken independently, or with the obvious exception of the green sector, by proximity to the active site (Figure 4).

Biochemical Independence

What is the functional meaning of independence between protein sectors? To address this question, we carried out alanine mutagenesis of residues spanning the range of correlation strengths in the red and blue sectors in rat trypsin and measured the effect on two basic properties of these enzymes: catalytic power and thermal stability. Catalytic power was measured using a standard chromophore-based assay on a model trypsin substrate peptide (NH_2-AAPK-pNA) (Hedstrom et al., 1994), and stability by following the denaturation temperature (T_m) using the fluorescence of buried tryptophan residues as a probe for the native state (Figures S8–S10 and Table S2).

Consistent with prior work (Hedstrom, 2002), we find that mutations in the red sector have significant effect on catalytic activity (red circles, Figure 5A). However, mutations in this sector have only minor effect on thermal stability. The same result holds for a multiple mutant in the red sector (Hswap, Figure 5B), in which a large number of red sector positions are exchanged for corresponding amino acids in chymotrypsin (Hedstrom et al., 1994). Strikingly, blue sector mutations have the opposite phenotype—a wide range of effects on thermal stability, but only marginal effect on catalytic activity (Figure 5A, blue circles). Double alanine mutants within the blue sector reinforce this result: these proteins show exclusive effects on thermal stability, with little or no effect on catalytic activity (Figure 5B). In addition, the data suggest that mutations within the sector act cooperatively. For example, L105A and T229A destabilize trypsin by 10.4 K and 8.0 K, respectively, but the effects of these mutations are reduced or abrogated in the background of M104A, an indication of epistatic interactions within this sector. These findings are structurally nontrivial; some blue sector positions (e.g., M104, T229) are as close to catalytic residues as some red sector positions (e.g., C191, G216) and are just as buried (Figures 4C and 4D), but nevertheless show distinct functional properties upon mutation.

One further experiment tests the independence of the red and blue sectors in affecting structural stability and catalytic activity.

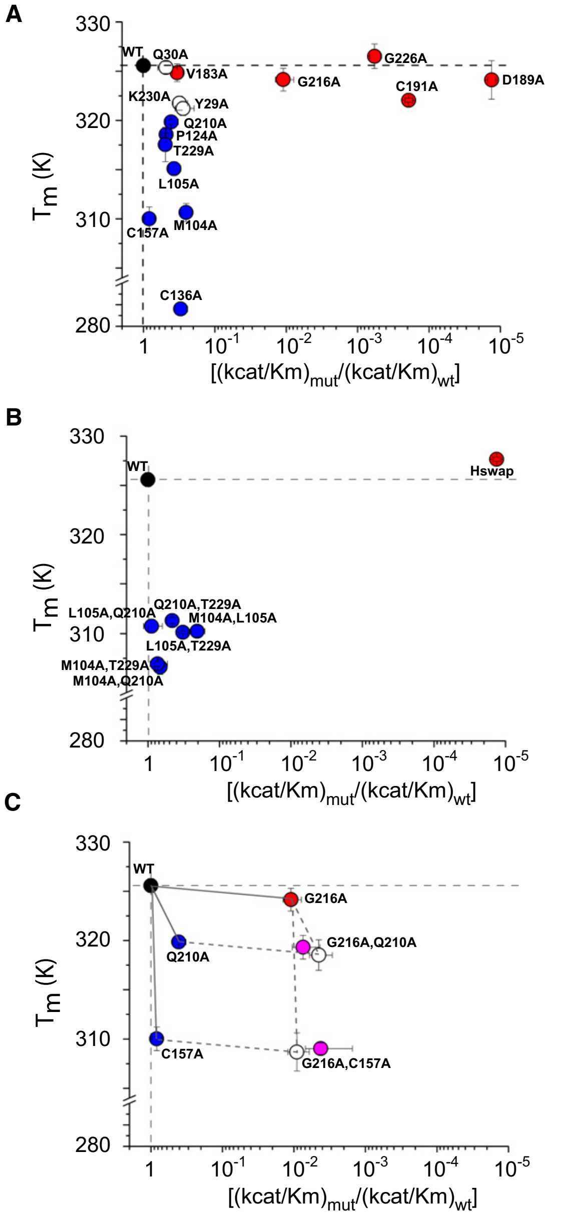

Figure 5. Mutational Analysis of the Red and Blue Sectors

(A) Single alanine mutations at a set of red and blue sectors positions in rat trypsin, evaluated for effects on catalytic power, and thermal stability (T_m). Residues selected for single mutation were chosen to sample the range of statistical contributions to the red and blue sectors. Wild-type rat trypsin is indicated in black, and mutations are colored according to sector identity (position 216 belongs to both red and green sectors). White circles represent nonsector mutants.

(B) Multiple mutants within each sector evaluated as in (A). In (B) and (C), residue pairs selected for double mutation analysis were chosen to have midrange single-mutation effects to permit assessment of additivity. Hswap indicates a multiple mutant largely within the red sector (Hedstrom et al., 1994). The multiple mutants show nonadditive but selective effects on either stability or catalytic power.

(C) Two double-mutant cycles between red and blue sectors, evaluated as described in (A). The white circle indicates the effect of the double mutant predicted from the independent action of the single mutants, and the magenta circle is the measured effect of the double mutant. All error bars indicate standard deviation from at least three independent experiments.

If these groups act independently, then combinations of mutations between these two groups should show additive effects on measured parameters. Figure 5C confirms this prediction for two pairs of intersector mutations; the measured effect of the double mutants (magenta circles) is nearly that predicted from the single mutant experiments (white circles). Thus, the red and blue sectors are associated with near-independent biochemical properties of the protease.

A small sampling of nonsector mutants in the core of trypsin shows little effects on either catalytic activity or thermal stability (white circles, Figure 5A). Further work will be required to more broadly test the role of conserved but nonsector positions in contributing to protease function. In addition, one aspect of the mutational effects in the blue sector is worth noting. Blue sector mutants affect thermal stability, but it is possible that this may actually reflect changes in the local stability of regions involved in functional processes such as control over protease lifetime through autocatalytic degradation. Consistent with this notion, the C136-C201 disulfide bond and other blue sector positions are located in regions that flank known autocatalytic sites (Bodi et al., 2001; Lee et al., 2004).

Independent Sequence Divergence

The finding of independent sectors in the serine protease has important implications for phylogenetic analysis of this protein family. Specifically, the data suggest that no single measure of the divergence of protein sequences can correctly represent their differences in functional properties. Instead, sequence divergence should be treated as a fundamentally multidimensional problem—using separate measures for each sector. To illustrate this, we calculated sequence similarities between sequences within the multiple sequence alignment using only the positions that contribute to the red, blue, or green sectors separately. As a control, we also calculated sequence similarity conventionally, using all positions in the sequence. Principal components analysis of the corresponding similarity matrices provides a simple representation of the relationships among the sequences as defined by each sector (Figures 6A–6C) or by all positions taken together (Figure 6D). Thus, sequences with a similar motif in the red sector are grouped in Figure 6A regardless of their divergence in other positions. Similarly, sequences with a similar motif in the blue sector are grouped in Figure 6B, and sequences with a similar motif in the green sector are grouped in Figure 6C, regardless of divergence elsewhere. Sequences are grouped in Figure 6D only if they are globally similar.

Consistent with the role of the red sector in substrate recognition, sequence divergence in this sector classifies the proteases effectively by primary catalytic specificity (Figure 6A, left panel). The trypsins (magenta) and chymotrypsins (blue) are separated, while the trypsins, tryptases (yellow), and kallikreins (orange), diverse proteases with similar specificity (Kam et al., 2000; Olsson et al., 2004), are found together. The granzymes (green) come in several specificity classes (A and K [tryptic], B [aspartic], and M [chymotryptic] [Bell et al., 2003; Kam et al., 2000;

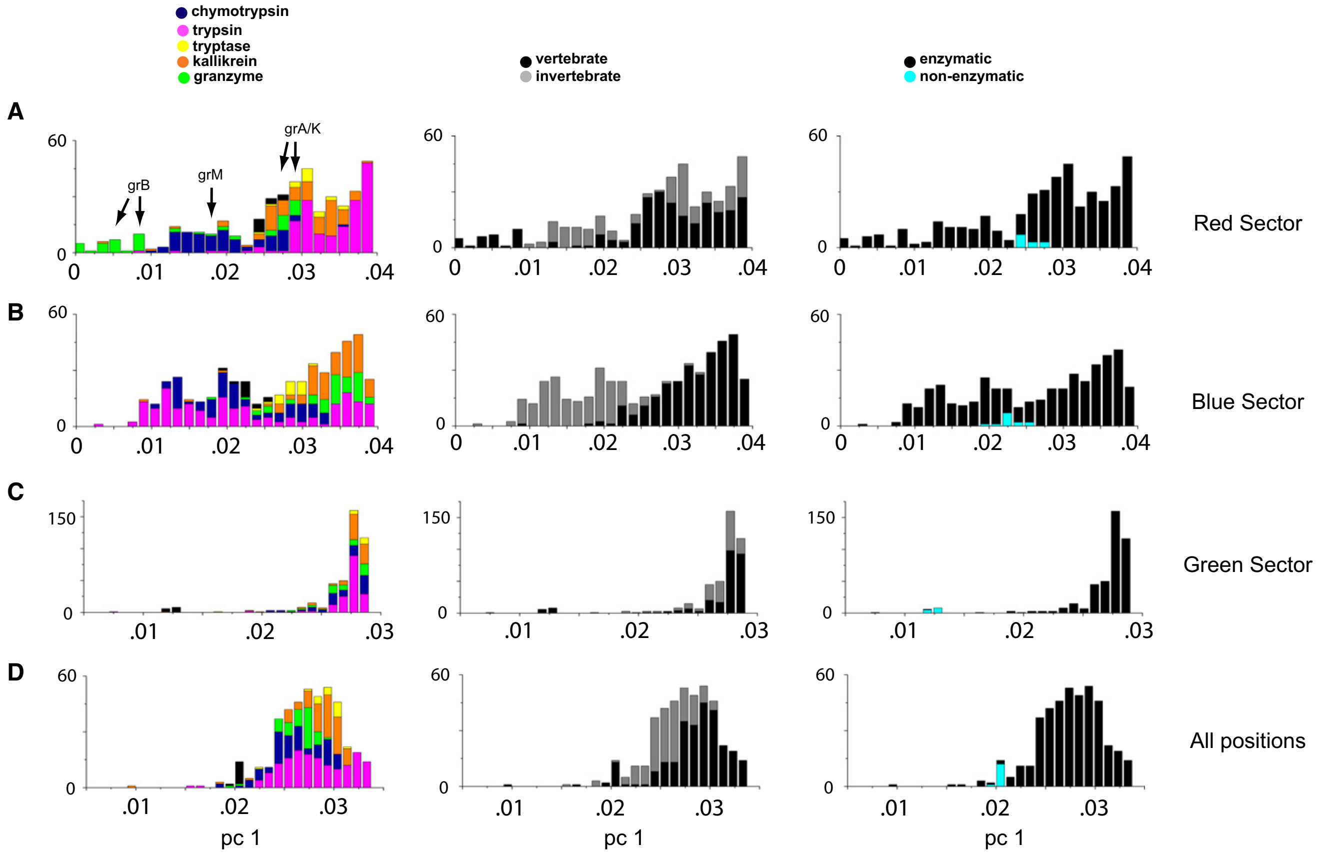

Figure 6. Multidimensional Sequence Divergence within the Serine Protease Family

Each stacked histogram shows the principal component of a sequence similarity matrix between the 442 members of the S1A family for which functional annotation is available. Similarity is calculated either for the red sector alone (18 positions) (A), the blue sector (23 positions) (B), the green sector (22 positions) (C), or for all 223 sequence positions (D). In each case, the left panel indicates the annotated primary catalytic specificity, the middle panel indicates organism type (invertebrate or vertebrate) from which the sequences originate, and the right panel indicates whether the protein has catalytic function.

Ruggles et al., 2004]) and occupy regions that correlate with their specificity class. However, this sector fails to separate the sequences according to the organism type in which they occur (Figure 6A, middle; vertebrate and invertebrate sequences are mixed) or by the existence of the catalytic mechanism (Figure 6A, right; nonenzymatic and enzymatic members of the S1A family are mixed). In contrast, the blue sector has a completely distinct effect in classifying protease sequences. This sector fails to group sequences by their catalytic specificity (Figure 6B, left) or by catalytic mechanism (Figure 6B, right), but does effectively classify sequences by organism type (Figure 6B, middle). Finally, the green sector displays a third classification; it fails to separate sequences by catalytic specificity (Figure 6C, left) or by organism type (Figure 6C, middle), but does separate the nonenzymatic and enzymatic members of the S1A family (Figure 6C, right). Similarity calculated over the entire protein sequence fails to effectively classify by catalytic specificity, organism type, or chemical mechanism (Figure 6D), indicating (1) that these phenotypic classifications are specific properties of the sectors and (2) that this result cannot be trivially explained by phylogenetic proximity of sequences. Thus, sectors represent independent modes of selection, a result that should provide important constraints in developing models for the evolutionary origins of the S1A family.

Sectors in Other Protein Families

To begin to examine the generality of the sector concept, we carried out spectral analysis of the SCA matrix from four other protein families for which substantial prior experimental data permit a meaningful interpretation (Figure 7). The results show that functionally relevant sectors are found in each case and provide a nontrivial basis for experiment design. For example, two sectors are evident in the PSD95/Dlg1/ZO1 (PDZ) domain family of protein interaction modules (blue and red, Figures 7A and S11), each of which comprises a small fraction of total residues. Interestingly, each sector is involved in a distinct regulatory mechanism in the PDZ family. The blue sector is connected through peptide ligand (Lockless and Ranganathan, 1999) and defines an allosteric mechanism for regulating binding affinity at the α2-β2 groove through molecular interactions at a distant surface site on the α1 helix (Peterson et al., 2004), and the red sector corresponds to a redox-based conformational switch that regulates the shape of the ligand-binding pocket (Mishra et al., 2007) (Figure 7A). These regulatory mechanisms have been experimentally demonstrated to date in only a few members of the PDZ family and could be seen as idiosyncratic features of specific PDZ domains. However, the sector hypothesis suggests that they are more general features of the protein family—variations within a sparse network of correlated positions that have the capacity to generate a diversity of regulatory phenotypes through stepwise modification of a few amino acid positions. The identification of PDZ sectors provides a basis for testing this hypothesis.

Two sectors are also evident in the Per/Arnt/Sim (PAS) domain family of allosteric signaling modules in which ligand binding (or chromophore isomerisation) at a surface pocket located on one side triggers conformational changes at N- and C-terminal structural motifs docked at the opposite surface (Halavaty and Moffat, 2007; Harper et al., 2003) (Figures 7B and S12). In the PAS family, one sector (blue) forms a network of amino acids within the core domain that links the ligand-binding pocket to the allosteric surface sites, and the other sector (red) comprises a cluster of amino acids at one surface site that connects the PAS core to a modular C-terminal ''output'' motif (Halavaty and Moffat, 2007) (Jα helix in Figure 7B). This mapping motivated the design of a synthetic two-domain allosteric protein by connecting sectors in two different proteins across their surface sites (Lee et al., 2008). The concept of allosteric coupling through sector linkage provides a starting point for testing a more general hypothesis that surface exposed regions of sectors represent ''hotspots'' for the establishment of cooperative functional interactions between protein domains.

Physically contiguous sectors are also evident in the SH2 and SH3 families of interaction modules (Figures 7C, 7D, S13, and S14). A full discussion of the extensive literature regarding these domains is a matter for future work, but an initial analysis reveals consistency with known functional mechanisms. In the phosphotyrosine-binding SH2 domains, the blue sector is largely buried within the core, while the red and green sectors make direct interactions with substrate peptide. The red sector surrounds the P-Tyr and the immediately N-terminal residue (positions 0 and −1, respectively, Figure 7C), and extends to a surface of the αA helix through network of intervening residues. Interestingly, αA is a major aspect of the interface between the SH2 domain and the catalytic domain of the Fes tyrosine kinase and experiments show that SH2 ligands allosterically influence kinase activity through this interdomain interaction (Filippakopoulos et al., 2008). The green sector includes residues interacting with the peptide ligand at positions that are C-terminal to the P-Tyr (+1 to +5, Figure 7C); these positions are known to contribute to determining the specificity of SH2 domains for target ligands (Kuriyan and Cowburn, 1997). In SH3, the blue sector identifies the residues that bind the canonical poly-proline motif that occurs in peptide ligands for this domain family (Yu et al., 1994; Zarrinpar et al., 2003). This finding suggests that the different subsites within the SH3 binding pocket (Figure 7D) should act cooperatively rather than separately in binding ligands. The SH3 red sector comprises a contiguous network of residues that link a region formed by a short 3_{10} helix and a portion of the n-Src loop to the so-called distal loop via residues within β strand c. Prior work indicates that these residues contribute to SH3 domain stability (Martinez and Serrano, 1999) and form part of a conserved ''folding nucleus'' that is partially ordered at the transition state for the folding reaction (Martinez and Serrano, 1999; Riddle et al., 1999). It will be interesting to test the sector-based prediction that determinants of substrate specificity and folding kinetics in the SH3 domain can be independently tuned through targeted variation of sector positions.

DISCUSSION

Classical analyses describe proteins as a hierarchy of primary, secondary, tertiary, and quaternary structures. This description derives from the basic chemical properties of polypeptide chains and empirical observation, and is the basis for current classifications and comparative analyses of protein families (Holm and Sander, 1996; Orengo and Thornton, 2005; Thornton et al., 1999). However, biological properties of proteins arise from the cooperative action of amino acid residues, and the pattern of residue cooperativity in the three-dimensional structure is generally unknown. Here, we show that generalizing the principle of conservation to account for correlations between positions reveals a novel structural organization for proteins that is distinct from traditional hierarchical descriptions. Statistically nonrandom correlations are arranged into physically connected groups of coevolving amino acids—the sectors—that involve amino acids spread out throughout primary structure, and across various secondary structure elements and tertiary structural subdomains. In the S1A family, the sectors manifest as strikingly independent features, controlling distinct biochemical properties and corresponding to orthogonal modes of sequence variation. The degree of independence of sectors in other domain families has yet to be investigated, and indeed, strict independence of sectors need not hold in every case (see below). Nevertheless, the fact that sectors correspond to important structural and functional properties in several protein families provides strong support for their biological relevance. Overall, we hypothesize that sectors represent the structural organization within proteins reflecting, at least in part, the functional interactions between amino acid residues that underlies conserved biological properties.

The finding of multiple independent sectors within a single protein domain has implications for physical properties of proteins. Atomic structures typically show a tightly packed and nearly homogeneous pattern of contacts between atoms, an observation that suggests a uniform importance of local interactions between amino acid residues. However, the finding of sparse, physically connected, and functionally quasi-independent sectors indicates that out of the uniform pattern of contacts between residues emerges a heterogeneous pattern of functional interactions. Indeed, a large body of experimental work now argues that amino acids contribute cooperatively but unequally in specifying protein structure and function (Agarwal et al., 2002; Benkovic and Hammes-Schiffer, 2003; Clackson and Wells, 1995; Datta et al., 2008; Eisenmesser et al., 2002; Fuentes et al., 2004; Ota and Agard, 2005; Sadovsky and Yifrach, 2007; Smock and Gierasch, 2009). The heterogeneity of correlations emerging from the statistics of conservation in protein families may be the representation of this feature. Support for this idea is provided by the ability of the pattern of correlations in the

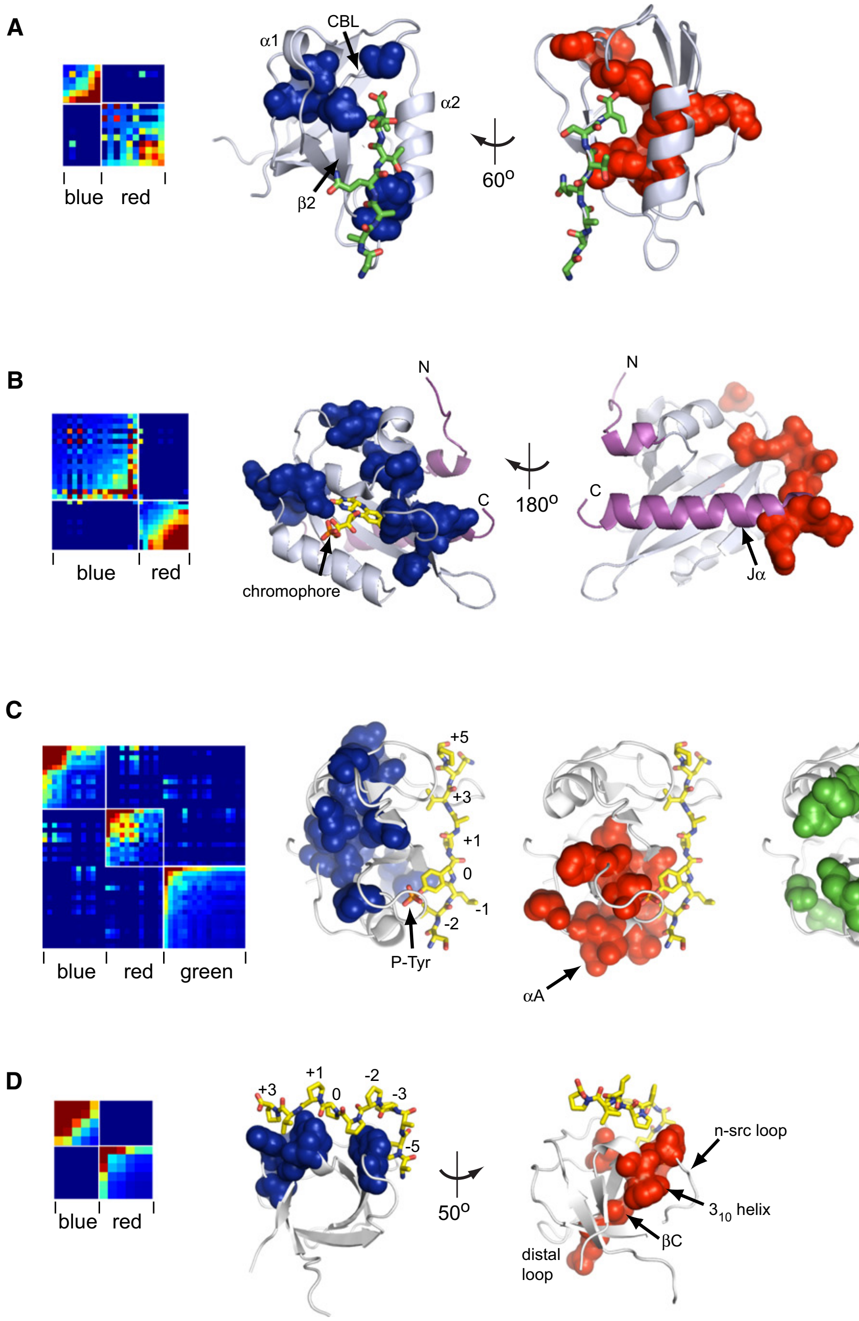

Figure 7. Functional Sectors in Other Protein Families

SCA correlation matrices for the PDZ (A), PAS (B), SH2 (C), and SH3 (D) domain families after reduction of statistical and historical noises ($\tilde{C}'_{ij}$, analogous to Figure 1E). In each case, the nonrandom correlations are described by sectors (labeled blue, red, and, if applicable, green), each comprising less than 20% of total positions.

(A) The blue and red sectors of the PDZ family, respectively, shown as spheres within a molecular surface on a member of the protein family (PDB 1BE9 [Doyle et al., 1996]; substrate peptide in green stick bonds). The peptide-binding pocket is bounded by the β2 strand, the α2 helix, and the "carboxylate binding loop" (CBL). Blue sector positions are either in direct contact with each other or are connected through interactions with substrate peptide and link a distant allosteric surface site on the α1 helix with the peptide-binding site (Lockless and Ranganathan, 1999; Peterson et al., 2004). Red sector positions comprise another contiguous group within the PDZ core, and correspond to a mechanism for regulating the conformation of the peptide-binding pocket (Mishra et al., 2007).

(B) The blue and red sectors of the PAS family, respectively, shown on a member of the protein family (PDB 2V0W [Halavaty and Moffat, 2007]; bound flavin mononucleotide [FMN] ligand shown as yellow stick bonds). The blue sector connects the environment of FMN to two "output" regions undergoing allosteric conformational change (in magenta): the N-terminal helix and the C-terminal region of the core domain that attaches to the Jα helix. Red sector positions comprise the linker connecting the PAS core to the Jα helix.

SCA matrix alone to enable the design of artificial proteins that recapitulate the atomic structure and biochemical activity of the WW domain family (Russ et al., 2005; Socolich et al., 2005).

These results underlie a conceptual departure between the SCA and some previous methods of analyzing residue covariation in protein families. Indeed, several reports have proposed approaches for the calculation of residue covariance, but often with the goal of identifying the pattern of contacts in the three dimensional structure (Gobel et al., 1994; Hamilton et al., 2004; Larson et al., 2000; Neher, 1994; Olmea and Valencia, 1997; Ortiz et al., 1999; Shindyalov et al., 1994; Thomas et al., 1996). The methodological details vary, but the conclusion of these studies is consistently clear: residue covariation is a poor indicator of the overall pattern of contacts in protein structures. One possible interpretation of this result is that covariation analyses fail to capture the essential design of proteins (Fodor and Aldrich, 2004), but, consistent with other studies (Kass and Horovitz, 2002; Lapedes et al., 1999; Lichtarge et al., 1996), the sector hypothesis suggests an alternate view: the pattern of constraints underlying the biological properties of proteins fundamentally differs from the pattern of observed contacts. More specifically, the hypothesis is that many contacts have weak or idiosyncratic roles, while a fraction of contacts are organized into collective systems—the sectors—that contribute most significantly to biological properties. In such a heterogeneous organization, different sectors could operate with near independence. The identification and validation of sectors in a few model proteins should help direct physical studies to experimentally test this proposal for the organization of amino acid interactions.

The results presented here imply the possibility of sector mapping for many protein families, but we caution that significant technical challenges remain in the development of general approaches for sector identification. The S1A, PDZ, PAS, SH2, and SH3 families represent cases in which both the extent and uniformity of sampling in the alignment permits straightforward application of the computational methods introduced in this work. In contrast, nonuniform sampling can lead to complications in sector analysis. An illustrative example of this problem is even evident in the S1A family; the presence of a small clade of snake venom proteases results in a weak “pseudo-sector” that emerges on one of the lower modes of the SCA matrix (Figure S4, and the Supplemental Data). This pseudo-sector is easily recognized and disregarded in this case, but it serves to highlight a potential challenge in the analysis of other protein families (Buck and Atchley, 2005). However, several strategies exist for correcting for biased sampling in alignments and for improving the recognition of statistical independent subgroups from correlation matrices that could be exploited in developing more powerful methods for sector identification. By taking the simplest approach in families suitable for forward and retrospective experimental analysis, this work provides a starting point for future studies.

Regardless of methodological issues, the validation of sectors in a few experimentally tractable model systems opens the possibility of addressing basic questions about the design of natural proteins. What is the origin of sectors in proteins and what controls their independence? Indeed, why should there be sectors at all? The answer to these questions ultimately involves the largely unknown evolutionary histories of protein families. In the case of serine proteases, it is interesting to note that enzymes with the same specificity are found in a variety of chemical environments and enzymes with different specificities are found in the same chemical environment. For example, tryptic specificity occurs in the gut, but also in the plasma and at sites of wound healing. At the same time, tryptic and chymotryptic specificities are often found together in the same environments. Thus, the capacity for independent control over enzyme activity, selectivity, and stability may provide an important adaptive advantage for the serine protease family. An implication of this line of thinking is that strict sector independence need not be guaranteed in every protein family. Instead, the emergence of independent functional sectors in proteins might be fundamentally tied to the independent variation of selective pressures acting on members of a protein family. More generally, we suggest that information about the statistics of the selective pressures is stored in the pattern of correlations in the protein sequence. The identification of sectors reported here provides a necessary first step in testing this hypothesis.

EXPERIMENTAL PROCEDURES

Sequence Alignment Construction and Annotation

Sequences comprising the S1A, PAS, SH2, and SH3 families were collected from the NCBI nonredundant database (release 2.2.14, May-07-2006) through iterative PSI-BLAST (Altschul et al., 1997) and aligned with Cn3D (Wang et al., 2000) and ClustalX (Thompson et al., 1997) followed by standard manual adjustment methods (Doolittle, 1996). The alignment of PDZ domains is from previous work (Lockless and Ranganathan, 1999). See the Supplemental Experimental Procedures for more information.

Sequence Analyses

The analysis of conservation and pairwise correlation in the multiple sequence alignment uses updated versions of the SCA method (Lockless and Ranganathan, 1999; Suel et al., 2003). Because of size considerations, methodological details for this analysis (including a MATLAB script for reproduction of all of the calculations) are provided in the Supplemental Data. A MATLAB (Mathworks) toolbox implementing the methods described here is available by request.

Minimum Discriminatory Information Method

The minimum discriminatory information (Kullback, 1997) (MDI) method generalizes the notion of positional conservation to include correlations between positions. In the binary approximation where only the most frequent amino acid a_i is considered at each position i; this is achieved by minimization of the relative entropy $D(P\|Q) = \sum_x P(x)\ln P(x)/Q(x)$ over the probability distributions $P(x)$, whose marginals reproduce the frequencies $f_{ij}^{(a_i a_j)}$. Here, x represents a sequence in the binary approximation, and $Q(x)$ denotes its background probability, $Q(x) = \prod_i (q^{(a_i)})^{x_i}(1-q^{(a_i)})^{1-x_i}$. We performed the

(C) Three sectors in the SH2 family of phosphotyrosine binding domains (blue, red, and green, shown on PDB 1AYA). The blue sector is nearly fully buried in the core, the red sector is built around the P-Tyr and −1 side chains and extends to the αA helix (an allosteric surface [Filippakopoulos et al., 2008]), while the green sector interacts with substrate positions 0 to +5.

(D) Two sectors in the SH3 family of polyproline binding domains (blue and red, shown on PDB 2ABL). The blue sector defines the polyproline binding site, while the red sector is nearly fully buried and connects the distal loop with a short 3_{10} helix through residues in β strand c.

minimization numerically for small subsets *S* of positions using the generalized iterative scaling algorithm (Darroch and Ratcliff, 1972). In Figure 2, the entropies represent the case when *S* was composed of the top five positions contributing to each sector. The statistical dependence between two sectors, S_1 and S_2, was measured by $D_{S_1 \cup S_2} - D_{S_1} - D_{S_2}$.

Protein Purification and Kinetic Assays

Purification of wild-type and mutant rat trypsins and measurement of kinetic parameters (V_{max} and K_m) were as previously described (Hedstrom et al., 1994) with minor modifications as detailed in the Supplemental Experimental Procedures. The substrate used was Suc-Ala-Ala-Pro-Lys-PNA (Bachem) dissolved in dimethylformamide (DMF) to 50 mM, and enzyme activity was measured at 23°C in 50 mM HEPES, 10 mM $CaCl_2$, and 100 mM NaCl, at a pH 8.0 by spectroscopic monitoring of p-nitroaniline release (extinction coefficient of 10204 M^{-1} cm^{-1} at 410 nm). So that k_{cat} (as V_{max}/active site concentration) could be obtained, active site concentration was measured by 4-methylumbelliferyl p-guanidobenzoate (MUGB, Sigma-Aldrich) titration (see the Supplemental Experimental Procedures). Kinetic assays were verified by comparison of data for WT rat trypsin and mutants with previously reported data (Craik et al., 1985; Hedstrom, 1996; McGrath et al., 1992; Wang et al., 1997).

Thermal Denaturation Assays

The fold stability of enzymes was measured using thermal denaturation and monitoring of the intrinsic tryptophan fluorescence of enzymes. Stability was assayed in 0.1 M formic acid to keep enzymes inactive (Bittar et al., 2003; Brumano et al., 2000). The fluorescence (excitation at 295 nm/emission at 340 nm) was measured in the range of 4°C to 85°C (at a rate of 4°C/min; sampling interval 0.1°C for most proteins) in a 3 ml quartz cuvette with stirring. The total volume was kept at 2.1 ml to ensure that the rate of temperature increase was the same across different assays. Pre- and posttransition baselines were fit by linear regression and subtracted from the raw data, and the T_m was calculated by the differential method (John and Weeks, 2000; Naganathan and Munoz, 2008). In brief, baseline subtracted data were smoothed by the robust Lowess method (MATLAB) and differentiated, and the T_m measured as the extremum of the differential melt. C136A showed no observable transition in the range of the experiment (Figure S10). All data were collected at least in triplicate; the data in Figure S9 show the mean and standard deviation of the individual trials.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, Supplemental Discussion, Supplemental MATLAB Script, two tables, and 14 figures and can be found with this article online at http://www.cell.com/supplemental/S0092-8674(09)00963-5.

ACKNOWLEDGMENTS

We thank members of the Ranganathan and Leibler laboratories for discussion and critical review of the manuscript, L. Hedstrom for materials and protocols, and A. Poole and W. Russ for the SH domain alignments. This study was supported by the Robert A. Welch foundation (R.R.) and the Green Center for Systems Biology at UT Southwestern Medical Center. O.R. is a fellow of the Human Frontier Science Program.

Received: March 9, 2009
Revised: July 3, 2009
Accepted: July 30, 2009
Published: August 20, 2009

REFERENCES

Agarwal, P.K., Billeter, S.R., Rajagopalan, P.T., Benkovic, S.J., and Hammes-Schiffer, S. (2002). Network of coupled promoting motions in enzyme catalysis. Proc. Natl. Acad. Sci. USA *99*, 2794–2799.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., and Dress, A.W. (2000). Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol. Biol. Evol. *17*, 164–178.

Baird, T.T., Jr., Wright, W.D., and Craik, C.S. (2006). Conversion of trypsin to a functional threonine protease. Protein Sci. *15*, 1229–1238.

Bell, J.K., Goetz, D.H., Mahrus, S., Harris, J.L., Fletterick, R.J., and Craik, C.S. (2003). The oligomeric structure of human granzyme A is a determinant of its extended substrate specificity. Nat. Struct. Biol. *10*, 527–534.

Benkovic, S.J., and Hammes-Schiffer, S. (2003). A perspective on enzyme catalysis. Science *301*, 1196–1202.

Bittar, E.R., Caldeira, F.R., Santos, A.M., Günther, A.R., Rogana, E., and Santoro, M.M. (2003). Characterization of -trypsin at acid pH by differential scanning calorimetry. Braz. J. Med. Biol. Res. *36*, 1621–1627.

Bodi, A., Kaslik, G., Venekei, I., and Graf, L. (2001). Structural determinants of the half-life and cleavage site preference in the autolytic inactivation of chymotrypsin. Eur. J. Biochem. *268*, 6238–6246.

Bouchaud, J.-P., and Potters, M. (2004). Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management, Second Edition (Cambridge: Cambridge University Press).

Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., and Sauer, R.T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science *247*, 1306–1310.

Brumano, M.H., Rogana, E., and Swaisgood, H.E. (2000). Thermodynamics of unfolding of beta-trypsin at pH 2.8. Arch. Biochem. Biophys. *382*, 57–62.

Buck, M.J., and Atchley, W.R. (2005). Networks of coevolving sites in structural and functional domains of serpin proteins. Mol. Biol. Evol. *22*, 1627–1634.

Bush-Pelc, L.A., Marino, F., Chen, Z., Pineda, A.O., Mathews, F.S., and Di Cera, E. (2007). Important role of the cys-191 cys-220 disulfide bond in thrombin function and allostery. J. Biol. Chem. *282*, 27165–27170.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875–1882.

Chothia, C., and Lesk, A.M. (1982). Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. J. Mol. Biol. *160*, 309–323.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science *267*, 383–386.

Craik, C.S., Largman, C., Fletcher, T., Roczniak, S., Barr, P.J., Fletterick, R., and Rutter, W.J. (1985). Redesigning trypsin: alteration of substrate specificity. Science *228*, 291–297.

Darroch, J.N., and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics *43*, 1470–1480.

Datta, D., Scheer, J.M., Romanowski, M.J., and Wells, J.A. (2008). An allosteric circuit in caspase-1. J. Mol. Biol. *381*, 1157–1167.

Doolittle, R.F. (1996). Computer Methods for Macromolecular Seqeunce Analysis, Volume 266 (San Diego, CA: Academic Press).

Doyle, D.A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell *85*, 1067–1076.

Eisenmesser, E.Z., Bosco, D.A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. Science *295*, 1520–1523.

Ferguson, A.D., Amezcua, C.A., Halabi, N.M., Chelliah, Y., Rosen, M.K., Ranganathan, R., and Deisenhofer, J. (2007). Signal transduction pathway of TonB-dependent transporters. Proc. Natl. Acad. Sci. USA *104*, 513–518.

Filippakopoulos, P., Kofler, M., Hantschel, O., Gish, G.D., Grebien, F., Salah, E., Neudecker, P., Kay, L.E., Turk, B.E., Superti-Furga, G., et al. (2008). Structural coupling of SH2-kinase domains links Fes and Abl substrate recognition and kinase activation. Cell *134*, 793–803.

Fodor, A.A., and Aldrich, R.W. (2004). On evolutionary conservation of thermodynamic coupling in proteins. J. Biol. Chem. *279*, 19046–19050.

Fuentes, E.J., Der, C.J., and Lee, A.L. (2004). Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. J. Mol. Biol. *335*, 1105–1115.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Guinto, E.R., Caccia, S., Rose, T., Futterer, K., Waksman, G., and Di Cera, E. (1999). Unexpected crucial role of residue 225 in serine proteases. Proc. Natl. Acad. Sci. USA *96*, 1852–1857.

Halavaty, A.S., and Moffat, K. (2007). N- and C-terminal flanking regions modulate light-induced signal transduction in the LOV2 domain of the blue light sensor phototropin 1 from Avena sativa. Biochemistry *46*, 14001–14009.

Hamilton, N., Burrage, K., Ragan, M.A., and Huber, T. (2004). Protein contact prediction using patterns of correlation. Proteins *56*, 679–684.

Harper, S.M., Neil, L.C., and Gardner, K.H. (2003). Structural basis of a phototropin light switch. Science *301*, 1541–1544.

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

Hedstrom, L. (1996). Trypsin: a case study in the structural determinants of enzyme specificity. Biol. Chem. *377*, 465–470.

Hedstrom, L. (2002). Serine protease mechanism and specificity. Chem. Rev. *102*, 4501–4524.

Hedstrom, L., Perona, J.J., and Rutter, W.J. (1994). Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. Biochemistry *33*, 8757–8763.

Holm, L., and Sander, C. (1996). Mapping the protein universe. Science *273*, 595–603.

Huntington, J.A., and Esmon, C.T. (2003). The molecular basis of thrombin allostery revealed by a 1.8 A structure of the "slow" form. Structure *11*, 469–479.

John, D.M., and Weeks, K.M. (2000). van't Hoff enthalpies without baselines. Protein Sci. *9*, 1416–1419.

Kam, C.M., Hudig, D., and Powers, J.C. (2000). Granzymes (lymphocyte serine proteases): characterization with natural and synthetic substrates and inhibitors. Biochim. Biophys. Acta *1477*, 307–323.

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

Kullback, S. (1997). Information Theory and Statistics (New York: Dover Publications).

Kuriyan, J., and Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. Annu. Rev. Biophys. Biomol. Struct. *26*, 259–288.

Lapedes, A.S., Giraud, B.G., Liu, L., and Stormo, G.D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. In Statistics in Molecular Biology, F. Francoise Seillier-Moiseiwitsch, ed. (Providence, RI: American Mathematical Society), pp. 236–256.

Larson, S.M., Di Nardo, A.A., and Davidson, A.R. (2000). Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J. Mol. Biol. *303*, 433–446.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. Science *322*, 438–442.

Lee, S.Y., Banerjee, A., and MacKinnon, R. (2009). Two separate interfaces between the voltage sensor and pore are required for the function of voltage-dependent K(+) channels. PLoS Biol. *7*, e47.

Lee, W.S., Park, C.H., and Byun, S.M. (2004). Streptomyces griseus trypsin is stabilized against autolysis by the cooperation of a salt bridge and cation-pi interaction. J. Biochem. *135*, 93–99.

Lesk, A.M., and Chothia, C. (1982). Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. *160*, 325–342.

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

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

Martinez, J.C., and Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. Nat. Struct. Biol. *6*, 1010–1016.

McGrath, M.E., Vasquez, J.R., Craik, C.S., Yang, A.S., Honig, B., and Fletterick, R.J. (1992). Perturbing the polar environment of Asp102 in trypsin: consequences of replacing conserved Ser214. Biochemistry *31*, 3059–3064.

Mishra, P., Socolich, M., Wall, M.A., Graves, J., Wang, Z., and Ranganathan, R. (2007). Dynamic scaffolding in a G protein-coupled signaling system. Cell *131*, 80–92.

Naganathan, A.N., and Munoz, V. (2008). Determining denaturation midpoints in multiprobe equilibrium protein folding experiments. Biochemistry *47*, 6752–6761.

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

Ng, P.C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. *7*, 61–80.

Olmea, O., and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold. Des. *2*, S25–S32.

Olsson, A.Y., Lilja, H., and Lundwall, A. (2004). Taxon-specific evolution of glandular kallikrein genes and identification of a progenitor of prostate-specific antigen. Genomics *84*, 147–156.

Orengo, C.A., and Thornton, J.M. (2005). Protein families and their evolution-a structural perspective. Annu. Rev. Biochem. *74*, 867–900.

Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. Proteins (*Suppl 3*), 177–185.

Ota, N., and Agard, D.A. (2005). Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. J. Mol. Biol. *351*, 345–354.

Pasternak, A., Ringe, D., and Hedstrom, L. (1999). Comparison of anionic and cationic trypsinogens: the anionic activation domain is more flexible in solution and differs in its mode of BPTI binding in the crystal structure. Protein Sci. *8*, 253–258.

Perona, J.J., Craik, C.S., and Fletterick, R.J. (1993). Locating the catalytic water molecule in serine proteases. Science *261*, 620–622.

Perona, J.J., Hedstrom, L., Rutter, W.J., and Fletterick, R.J. (1995). Structural origins of substrate discrimination in trypsin and chymotrypsin. Biochemistry *34*, 1489–1499.

Peterson, F.C., Penkert, R.R., Volkman, B.F., and Prehoda, K.E. (2004). Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. Mol. Cell *13*, 665–676.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A., Guhr, T., and Stanley, H.E. (2002). Random matrix approach to cross correlations in financial data. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *65*, 066126.

Rawlings, N.D., and Barrett, A.J. (1994). Families of serine peptidases. Methods Enzymol. *244*, 19–61.

Rawlings, N.D., Morton, F.R., Kok, C.Y., Kong, J., and Barrett, A.J. (2008). MEROPS: the peptidase database. Nucleic Acids Res. *36*, D320–D325.

Riddle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I., and Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. Nat. Struct. Biol. *6*, 1016–1024.

Ruggles, S.W., Fletterick, R.J., and Craik, C.S. (2004). Characterization of structural determinants of granzyme B reveals potent mediators of extended substrate specificity. J. Biol. Chem. *279*, 30751–30759.

Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature *437*, 579–583.

Sadovsky, E., and Yifrach, O. (2007). Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated K+ channel. Proc. Natl. Acad. Sci. USA *104*, 19813–19818.

Shindyalov, I.N., Kolchanov, N.A., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. *7*, 349–358.

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

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043–1054.

Smock, R.G., and Gierasch, L.M. (2009). Sending signals dynamically. Science *324*, 198–203.

Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H., and Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature *437*, 512–518.

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

Thomas, D.J., Casari, G., and Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. Protein Eng. *9*, 941–948.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. *25*, 4876–4882.

Thornton, J.M., Orengo, C.A., Todd, A.E., and Pearl, F.M. (1999). Protein folds, functions and evolution. J. Mol. Biol. *293*, 333–342.

Wang, E.C., Hung, S.H., Cahoon, M., and Hedstrom, L. (1997). The role of the Cys191-Cys220 disulfide bond in trypsin: new targets for engineering substrate specificity. Protein Eng. *10*, 405–411.

Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A., and Bryant, S.H. (2000). Cn3D: sequence and structure views for Entrez. Trends Biochem. Sci. *25*, 300–302.

Wigner, E.P. (1967). Random matrices in physics. SIAM Rev. Soc. Ind. Appl. Math. *9*, 1–23.

Yu, H., Chen, J.K., Feng, S., Dalgarno, D.C., Brauer, A.W., and Schreiber, S.L. (1994). Structural basis for the binding of proline-rich peptides to SH3 domains. Cell *76*, 933–945.

Zarrinpar, A., Bhattacharyya, R.P., and Lim, W.A. (2003). The structure and function of proline recognition domains. Sci. STKE *2003*, RE8.

Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. *195*, 957–961.