Cell, Volume 138

Supplemental Data

Theory

Protein Sectors: Evolutionary Units of Three-Dimensional Structure

Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan

I. Supplemental Experimental Procedures

A. Multiple Sequence Alignments

For the S1A family, only sequences consisting of a single protease domain were included. Positions were truncated to the structure of rat trypsin (PDB 3TGI), yielding a final alignment of 1470 sequences and 223 positions that includes the digestive enzymes (e.g. trypsin, chymotrypsin, elastase), the immune cell proteases (tryptases, chymases, kallikreins, and granzymes), the coagulation enzymes (e.g. thrombin, factor X), the snake venom proteases, and the haptoglobins, non-catalytic members of the family that have evolved to bind free hemoglobin (Kurosky et al., 1980). Sequences were annotated with regard to catalytic specificity through examination of sequence file headers and literature survey, and for the purposes of this work, were binned into the following categories: "trypsin", "chymotrypsin", "kallikreins", "tryptases", "chymases", and "granzymes". The granzymes were further subdivided into the major types occurring in the alignment: "a", "k", "b", and "m". All sequences that did not unambiguously fall into these categories were declared "NA" for not annotated.

For the PDZ domain family, the alignment consisted of 240 sequences and 92 positions after truncation to the structure of the third PDZ domain of rat PSD95 (PDB 1BE9). For the PAS family, the alignment consisted of 1104 sequences and 123 positions after truncation to the structure of *A. sativa* LOV2 (PDB 2V0W). For SH2, 582 sequences and 79 positions after truncation to the Syp SH2 domain (PDB 1AYA). For SH3, 492 sequences and 52 positions after truncation to the Abl SH3 domain (PDB 2ABL).

B. Correlation Analysis

This section provides a short summary of the calculations used for the statistical coupling analysis. The supplementary discussion (section II) provides more detail, rationale, and further discussion regarding these methods.

1. Positional conservation: The conservation of an amino acid a at position i in a multiple sequence alignment is defined by $D_i^{(a)}$, the divergence (or relative entropy) of the observed frequency of a at i ($f_i^{(a)}$) from the background frequency of a in all proteins ($q^{(a)}$)(Cover and Thomas, 2006):

$$D_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{q^{(a)}} + \left(1 - f_i^{(a)}\right) \ln \frac{1 - f_i^{(a)}}{1 - q^{(a)}}.$$

$D_i^{(a)}$ is a non-linear function of $f_i^{(a)}$ that rises more and more steeply as $f_i^{(a)}$ approaches one. As a practical consequence, for all but the least conserved positions, the overall conservation of all amino acids at each position i is well approximated by $D_i^{(a_i)}$, the conservation of a_i, the most prevalent amino acid at that position (Fig. S1). We make use of this simplified "binary approximation" below.

2. SCA matrix: The basic principle of the SCA correlation matrix, $\tilde{C}_{ij}^{ab}$ is to weight the frequency-based correlations between positions i and j, $C_{ij}^{ab} = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}$, by a functional of their positional conservations $D_i^{(a)}$ and $D_j^{(b)}$:

$$\tilde{C}_{ij}^{ab} = \phi\left(D_i^{(a)}\right)\phi\left(D_j^{(b)}\right)C_{ij}^{ab}$$

Thus $\tilde{C}_{ij}^{ab}$ is a measure of the significance of observed correlations as judged by the conservation of the amino acids under consideration. Following earlier work, the weighting functions are chosen here to be gradients of positional conservation: $\phi = \partial D / \partial f$. The position-by-position SCA correlation matrix $\tilde{C}_{ij}$ (Fig. 1D) is constructed by invoking the binary approximation of the alignment: $\tilde{C}_{ij} = \phi\left(D_i^{(a_i)}\right)\phi\left(D_j^{(a_j)}\right)\left|C_{ij}^{a_i a_j}\right|$. The expressions for $D_i^{(a)}$, $\tilde{C}_{ij}^{ab}$, and $\tilde{C}_{ij}$ represent updated versions of measures of conservation and correlation reported previously for the SCA method (Lockless and Ranganathan, 1999; Suel et al., 2003).

3. Spectral cleaning: Due to statistical and historical noise, most correlations reported by $\widetilde{C}_{ij}$ are not functionally significant. A spectral decomposition of $\widetilde{C}_{ij}$ provides a way to partially sort out the different contributions to the correlations. The spectrum of $\widetilde{C}_{ij}$ is composed of 223 eigenvalues $\lambda_1 > \ldots > \lambda_{223}$, the lowest 218 of which can be attributed to statistical noise since randomized alignments retaining the same size and amino acid propensities at sites show eigenvalues of similar magnitude (Fig. S2A). Thus, only the top 5 modes of $\widetilde{C}_{ij}$ are interpreted. A similar approach motivated by random matrix theory (Wigner, 1967) is used in finance for defining significant correlations of stock performance based on limited time series of sampling returns (Bouchaud and Potters, 2004; Plerou et al., 2002).

Among the significant modes, the first mode has a distinctive property: it describes a "coherent" correlation of all positions. This is evident analytically since the first mode makes the dominant contribution to $\widetilde{C}_{ij}$ (Fig. S2A). As a first order approximation, $\widetilde{C}_{ij}$ can be written as $\tilde{C}_{ij}^{(1)} = S_i S_j \Big/ \sum_k S_k$ with $S_i = \sum_j \widetilde{C}_{ij}$; the matrix $\widetilde{C}_{ij}^{(1)}$ has only one non-zero mode $\lambda_1^{(1)} = \sum_i S_i^2 \Big/ \sum_i S_i$, with an associated eigenvector having for components $S_i \Big/ \left(\sum_k S_k^2\right)^{1/2}$. In other words, for SCA matrices with a dominant first mode, the first eigenvector should just report the net contribution of each position to the total correlation. Indeed, the first eigenvector of $\widetilde{C}_{ij}$ is well approximated by this equation (Fig. S2B), and $\lambda_1 = 28.6$, the first eigenvalue of $\widetilde{C}_{ij}$ is well approximated by $\lambda_1^{(1)} = 27.8$. Since each position contributes with the same sign to this first eigenvector, it corresponds to a coherent mode. Similar to how global fluctuations in the economy

coherently drive correlations between all stocks, historical noise is expected to produce coherent correlations between sequence positions (see example in Section II, the Supplemental Discussion).

4. Sector identification: The three sectors in the serine protease family are identified by examining the intermediate modes 2 to 5, and best visualized by the second and fourth eigenvectors ($|2\rangle$ and $|4\rangle$) of $\widetilde{C}_{ij}$ (Fig. S2-S3). The bra-ket notation is such that $|k\rangle$ denotes the k^{th} eigenvector and $\langle i|k\rangle$ the weight for position i along eigenvector k. The red sector is defined as the positions i for which $\langle i|2\rangle > \varepsilon$ and $\langle i|2\rangle > |\langle i|4\rangle|$, the blue sector as those for which $\langle i|2\rangle < -\varepsilon$ and $\langle i|2\rangle < -|\langle i|4\rangle|$, and the green sector as those for which $\langle i|4\rangle > \varepsilon$ and $\langle i|4\rangle > |\langle i|2\rangle|$ (Fig. S3F). The significance threshold $\varepsilon = 0.05$ is defined by examining eigenvector weights for 100 trials of randomizing the alignment (Fig. S2C-E). A few other positions show negative weights in $\langle i|4\rangle$ that are less well grouped. Unlike the other groups, these positions are further subdivided along the fifth eigenvector and are unlikely to represent a meaningful sector (Figs. S3-S4).

In general, a sector need not be simply associated with one eigenvector; instead, it could be defined by a linear combination of different eigenvectors. Statistical methods beyond spectral analysis may thus be more appropriate to define sectors. For example, techniques such as independent component analysis (ICA (Hyvarinen et al., 2001; Stone, 2004)) may be valuable. Preliminary analysis shows that ICA identifies the same three sectors in the S1A family as independent components (not shown). In addition, when strongly non-uniform distributions of sequences occur in an alignment, some of the top eigenvectors could represent "pseudo-sectors" (see Fig. S4 and Section II, the Supplemental Discussion); systematic elimination of these features of historical noise will require methods beyond those presented here.

5. The cleaned correlation matrix. The correlation matrix corresponding to the significant modes 2 to 4 of $\widetilde{C}_{ij}$ can formally be written $\tilde{C}' = \sum_{k=2}^{4} \lambda_k |k\rangle\langle k|$. Figure 1E shows the correlations in $\tilde{C}'$ for the 65 identified sector positions, with positions of the blue, green, and red sectors ordered by increasing values of $\langle i|2\rangle$, decreasing values of $\langle i|4\rangle$, and decreasing values of $\langle i|2\rangle$, respectively. As described in the Section II, the Supplemental Discussion, some negative correlations occur in $\tilde{C}'$ that are artifacts of the cleaning method and are ignored (Fig. S5).

C. Protein Expression and Purification

Proteases were expressed in a *Saccharomyces cerevisiae* system (strain DLM101α) where inactive enzymes are secreted into the culture medium. The culture medium is centrifuged at 7000g for 30 min to obtain cell-free supernatant. The supernatant is adjusted to pH 3.0 with 1 M HCL, gently stirred for 20 min at room temperature, and centrifuged for 120 min at 7000 g to pellet insoluble precipitates. Two to four ml of Toyopearl SP-650M cation-exchange resin is equilibrated in Buffer A (100 mM glacial acetic acid, 2 mM sodium acetate) and added to the supernatant and nutated for a minimum of one hour. The resin is allowed to settle and most of the supernatant is decanted. The remaining resin + solution is loaded onto a Biorad polyprep chromatography column, washed with at least 100 bed volumes of Buffer A, and bound protein are eluted with 5 ml steps of Buffer A adjusted to pH 5.0, 6.0, 7.0 and 8.0 with 200 mM Tris pH 8.0. Eluted proteins are dialyzed for >8hrs with enterokinase buffer (50 mM Tris pH 6.5, 10 mM $CaCl_2$). For mutants that do not self-activate, enterokinase light chain (NEB) is added until 50% or more of the protein is activated. Activation can be followed by SDS-PAGE. After activation, 1 to 3 mL of soybean trypsin inhibitor-agarose (Sigma-Alrdich) equilibrated in enterokinase buffer is added for at least one hour with nutating. The activated enzymes bind specifically to this resin. After binding, the resin is loaded into a

BioRad polyprep chromatography column and washed with 20-40 mL 50 mM Tris, pH 6.5 and 20-40 mL 50 mM Tris pH 6.5 + 0.5 M NaCl sequentially. The protein is eluted with 2-4 ml of 0.1 M formic acid (pH 2.2). Proteins are stored in this buffer at 4^{o} C.

D. Kinetic assays

Kinetic parameters of V_{max} and K_m were measured assuming pseudo-first order kinetics as previously described(Hedstrom et al., 1994). The substrate used was Suc-Ala-Ala-Pro-Lys-PNA (Bachem) dissolved in dimethylformamide (DMF) to 50mM. Enzymes hydrolyze this substrate releasing p-nitroaniline, which is detected by monitoring absorption at 410 nm (extinction coefficient of 10204 M^{-1} cm^{-1}). The enzyme reactions were done at 23^{o} C in 50 mM Hepes, 10 mM $CaCl_2$ and 100 mM NaCl, at a pH 8.0 (protease assay buffer, PAB). The total volume of reaction was 1 mL and the volume of substrate did not exceed 5%. A maximum of 20 ul of enzyme (in 0.1 M formic acid) was added to the reaction. In most cases, plots of initial velocity vs. substrate concentrations were fit to a hyperbola using non-linear regression to obtain K_m and V_{max} (Fig. S8a, S8b). R-square for all regressions was at least 0.9. To obtain k_{cat} (as V_{max}/active site concentration), active site concentration was measured by reacting the enzymes with 4-methylumbelliferyl p-guanidobenzoate (MUGB, Sigma-Aldrich), an enzyme inhibitor which releases a fluorescent compound, 4-methyl umbelliferone (4-MU) upon reacting with the enzyme. A standard curve of 4-MU was constructed to relate fluorescence counts to fluorophore concentration. Some enzymes did not react with MUGB and active site concentration was estimated by calculating the absorbance at 280 nm using the extinction coefficient of 33720 M^{-1} cm^{-1}. In some cases, the enzyme was not saturated with feasible concentrations of substrate so the approximation that K_m>>substrate concentration was used to calculate k_{cat}/K_m as the slope of the line of initial rate vs. substrate concentration. 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).

II. Supplemental Discussion on SCA/MDI Calculations

A. Measures of conservation

A multiple sequence alignment of M sequences of length L is represented by a binary array $x_{i,s}^{(a)}$, where $x_{i,s}^{(a)} = 1$ if sequence s has amino acid a at position i, and 0 otherwise ($s = 1,\ldots, M$ is for sequences, $i = 1,\ldots, L$ is for positions and $a = 1,\ldots, 20$ is for amino acids). A "binary approximation" is to consider only the most frequent amino acid a_i at position i ; the alignment is then represented by a binary array $x_{i,s}$ where $x_{i,s} = 1$ if sequence s displays the most frequent amino acid in the alignment at position i, and 0 otherwise (i.e., $x_{i,s} = x_{i,s}^{(a_i)}$).

As indicated in Section I, the Supplemental Experimental Methods, the measure of positional conservation $D_i^{(a)}$ is based on $f_i^{(a)}$, the observed frequency of amino acid a at position i, and $q^{(a)}$, the background probability of amino acid a. This background probability is an estimation of the mean frequency of a in all proteins and we take q = (0.073, 0.025, 0.050, 0.061, 0.042, 0.072, 0.023, 0.053, 0.064, 0.089, 0.023, 0.043, 0.052, 0.040, 0.052, 0.073, 0.056, 0.063, 0.013, 0.033), where amino acids are ordered according to the alphabetic order of their standard one-letter abbreviation. $f_i^{(a)}$ is computed as the number of sequences in the alignment having amino acid a at position i, divided by the total number of sequences, including those with a gap at i; it can also be written

$$f_i^{(a)} = \left\langle x_{i,s}^{(a)} \right\rangle_s ,$$

where $x_{i,s}^{(a)}$ is averaged over all M sequences s. Similarly, in the binary approximation, we consider $f_i^{(a_i)} = \left\langle x_{i,s} \right\rangle_s$. The position-specific conservation $D_i^{(a)}$, which measures the degree of deviation of $f_i^{(a)}$ from $q^{(a)}$, is derived from the probability $P_M[f_i^{(a)}]$ of observing $f_i^{(a)}$ in an alignment of M sequences, under the assumption that a has independent probability $q^{(a)}$ to be present in each sequence:

$$P_M[f_i^{(a)}] = \frac{M!}{\left(Mf_i^{(a)}\right)!\left(M(1-f_i^{(a)})\right)!}\left(q^{(a)}\right)^{Mf_i^{(a)}}\left(1-q^{(a)}\right)^{M(1-f_i^{(a)})}.$$

When M is large, the Stirling formula leads to the approximation

$$P_M[f_i^{(a)}] \approx e^{-MD_i^{(a)}}, \text{ with } D_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{q^{(a)}} + (1-f_i^{(a)}) \ln \frac{1-f_i^{(a)}}{1-q^{(a)}} .$$

The so-called relative entropy (Cover and Thomas, 2006) $D_i^{(a)}$ defines the positional conservation.

An overall conservation D_i taking into account the frequencies of all 20 amino acids can similarly be defined, but requires introducing a background probability for gaps. If γ represents the fraction of gaps in the alignment, a background probability distribution can be taken as $\overline{q}^{(0)} = \gamma$ for gaps, and $\overline{q}^{(a)} = (1-\gamma)q^{(a)}$ for the 20 amino acids. Denoting $f_i^{(0)} = 1 - \sum_{a=1}^{20} f_i^{(a)}$ the fraction of gaps at position i , we can then write the probability of observing jointly at position i the frequencies ($f_i^{(1)},\ldots,f_i^{(20)}$) of each of the 20 possible amino acids as

$$P_M\left[f_i^{(1)},\ldots,f_i^{(20)}\right] = \frac{M!}{\left(Mf_i^{(0)}\right)!\ldots\left(Mf_i^{(20)}\right)!}\left(\overline{q}^{(0)}\right)^{Mf_i^{(0)}}\ldots\left(\overline{q}_i^{(20)}\right)^{Mf_i^{(20)}} \approx e^{-MD_i}$$

where $D_i = \sum_{a=0}^{20} f_i^{(a)} \ln \frac{f_i^{(a)}}{\overline{q}^{(a)}}$ defines the overall conservation at position *i*.

The overall conservation D_i can be compared to $\overline{D}_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{\overline{q}^{(a)}} + (1 - f_i^{(a)}) \ln \frac{1 - f_i^{(a)}}{1 - \overline{q}^{(a)}}$, the equivalent of $D_i^{(a)}$, the positional conservation for amino acid a, when using the background probability distribution including gaps. As a general rule, we have $\overline{D}_i^{(a)} \leq D_i$ and in practice, for multiple sequence alignments, $\overline{D}_i^{(a)}$ is maximal for $a = a_i$, the most frequent amino acid at position *i*. Note that $\overline{D}_i^{(a_i)}$ and D_i are non-linear functions of $f_i^{(a)}$ that rise more and more steeply as $f_i^{(a)}$ approaches one. A consequence is that for all but the least conserved sites, the overall conservation D_i is well approximated by $\overline{D}_i^{(a_i)}$ (Fig. S1), which justifies the use of the binary approximation. In this approximation, we need not introduce a background probability for gaps and, therefore, we make use of $D_i^{(a_i)}$ rather than $\overline{D}_i^{(a_i)}$ in this work.

B. Measure of sequence similarity

Given a set *S* of positions, we define a similarity matrix between pairs of sequence s,t as

$$\Gamma_{st}^{(S)} = \left\langle x_{i,s}^{(a)} x_{i,t}^{(a)} \right\rangle_{a,i \in S} - \left\langle x_{i,s}^{(a)} \right\rangle_{a,i \in S} \left\langle x_{i,t}^{(a)} \right\rangle_{a,i \in S}$$

where averages are here made over all amino acids *a* and positions *i* in the set *S* under consideration. In Figure 6, *S* is taken to be either a sector or all the positions, and the sequences are represented in the one-dimensional space spanned by the first eigenvector of the similarity matrix $\Gamma_{st}^{(S)}$.

C. SCA calculations

In general, a covariance matrix reporting pairwise correlations between positions can be defined as

$$C_{ij}^{(ab)} = \left\langle x_{i,s}^{(a)} x_{j,s}^{(b)} \right\rangle_s - \left\langle x_{i,s}^{(a)} \right\rangle_s \left\langle x_{j,s}^{(b)} \right\rangle_s = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}$$

where $f_{ij}^{(ab)} = \left\langle x_{i,s}^{(a)} x_{j,s}^{(b)} \right\rangle_s$ represents the joint frequency of having *a* at position *i* and *b* at position *j*. The corresponding expression in the binary approximation is

$$C_{ij} = \left\langle x_{i,s} x_{j,s} \right\rangle_s - \left\langle x_{i,s} \right\rangle_s \left\langle x_{j,s} \right\rangle_s = f_{ij}^{(a_i a_j)} - f_i^{(a_i)} f_j^{(a_j)}.$$

SCA matrices can be obtained by weighting these covariance matrices by a function ϕ of the positional conservations $D_i^{(a)}$,

$$\widetilde{C}_{ij}^{(ab)} = \phi(D_i^{(a)}) \phi(D_j^{(b)}) C_{ij}^{(ab)},$$

or, in the binary approximation,

$$\widetilde{C}_{ij} = \phi(D_i^{(a_i)}) \phi(D_j^{(a_j)}) \left| C_{ij} \right|.$$

Although not essential, the absolute value taken in the last formula eliminates negative correlations that originate from alternative choices of amino acids at a position. Since our goal is to characterize positional correlations, the sign of amino acid-specific correlations is not considered in this work. The weights used here are

$$\phi(D_i^{(a)}) = \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} = \ln\left[\frac{f_i^{(a)}(1-q^{(a)})}{(1-f_i^{(a)})q^{(a)}}\right].$$

This choice of weights reflects the original principle of SCA to define correlations between positional conservations through a perturbation analysis on the sequence alignment (Lockless and Ranganathan, 1999). More precisely, if we introduce $D_{i,s}^{(a)}$, the positional conservation of amino acid *a* at position *i* for the alignment obtained by leaving out sequence *s*, the covariance matrix associated with this bootstrap procedure is

$$\hat{C}_{ij}^{(ab)} = \left\langle D_{i,s}^{(a)} D_{j,s}^{(b)} \right\rangle_s - \left\langle D_{i,s}^{(a)} \right\rangle_s \left\langle D_{j,s}^{(b)} \right\rangle_s .$$

In the limit of a large number *M* of sequences, expanding to first order in *1/M* this expression leads to $\hat{C}_{ij}^{(ab)} \approx \frac{1}{M^2}\tilde{C}_{ij}^{(ab)}$ (Efron and Tibshirani, 1994).

We also note that in previous implementations of the SCA method, a reduced matrix $\hat{C}_{ij}$ was defined from $\hat{C}_{ij}^{(ab)}$ by

$$\hat{C}_{ij} = \left(\sum_{a,b}\left(\hat{C}_{ij}^{(ab)}\right)^2\right)^{1/2} \approx \frac{1}{M^2}\tilde{C}_{ij}.$$

Within the range of validity of the binary approximation, this matrix corresponds to $\tilde{C}_{ij}$ and therefore yields equivalent results. A more detailed description of the SCA approach to measuring positional correlations will be reported elsewhere (O. Rivoire, S. Leibler, and R. Ranganathan). A MATLAB toolbox implementing the methods described here is available by request from the authors. The calculation of the SCA correlation matrix in this work is also described in the MATLAB script in section III of the supplementary information.

D. Spectral cleaning

Due to finite-size and phylogenetic effects (statistical and historical noises), most correlations reported by $\widetilde{C}_{ij}$ are not functionally significant. A spectral decomposition of $\widetilde{C}_{ij}$ offers a simple way to partially sort out the different contributions to the correlations. The spectrum of $\widetilde{C}_{ij}$, composed of 223 eigenvalues $\lambda_1 > \ldots > \lambda_{223}$, is shown in Fig. S2A. upper panel. The bulk of this spectrum, made of the 218 smallest eigenvalues, can be attributed to finite-size effects. Indeed, the same analysis performed on randomized alignments leads to the spectrum shown in the lower panel of Fig. S2A. The comparison of the two spectra indicates that only the top 5 modes of $\widetilde{C}_{ij}$ are informative. This "noise undressing" procedure applies more generally to any correlation matrix subject to statistical noise due to limited sampling; it has for instance been previously applied to the analysis of correlation of financial stocks, using time series of limited length. The approach is motivated by random matrix theory, which establishes, for several classes of noise, the existence of universal spectral distributions with a sharp threshold bounding the maximal eigenvalue. Note that the criterion invoked here to select a few significant modes fundamentally differs from the criterion usually invoked in principal component analysis where modes (principal components) are selected to explain a major fraction of the correlations, irrespectively of their origin. Here, the top 5 significant modes represent only about 20% of the total variance, consistent with the finding much of the total variance is dominated by statistical noise.

Among the significant modes, the first mode has a distinctive property: it describes a "coherent" correlation of all positions. This can be seen analytically by taking advantage of the fact that the first mode makes the dominant contribution to $\widetilde{C}_{ij}$. Indeed, as a first order approximation, $\widetilde{C}_{ij}$ can be approximated by

$$\widetilde{C}_{ij}^{(1)} = \frac{S_i S_j}{\sum_k S_k} \text{ with } S_i = \sum_j \widetilde{C}_{ij} .$$

The matrix $\widetilde{C}_{ij}^{(1)}$ has for only non-zero mode $\lambda_1^{(1)} = \sum_i S_i^2 \Big/ \sum_i S_i$, with an associated eigenvector having for components $S_i \Big/ \left(\sum_k S_k^2 \right)^{1/2}$. These expressions are good approximations of the first eigenvalue and eigenvector of the matrix $\widetilde{C}_{ij}$ (Fig. S2B). The first eigenvector thus reports the contribution of each position to the total correlation. Since each position contributes with the same sign to this first eigenvector (mathematically, a consequence of the Perron-Frobenius theorem), it corresponds to a global, coherent mode.

The origin of these coherent correlations may be purely historical. As a simple illustrative example, consider a situation where all sequences in an alignment derive independently from a common ancestral sequence, in absence of any selective pressure. The divergence of each sequence from the ancestral sequence is assumed to be described by a number ρ, with $0 < \rho < 1$, where ρ gives the probability for a position of the sequence to have conserved the original amino acid. ρ is itself assumed to be distributed over the sequences with a probability distribution $P(\rho)$ such that $\overline{\rho} = \int \rho P(\rho) d\rho > 1/2$, so that the dominant amino acid at each position in the alignment corresponds to the original amino acid in the ancestral sequence. Ignoring finite-size effects, the correlation matrix C_{ij} reads in the binary approximation $C_{ij} = \overline{\rho^2} - \overline{\rho}^2$, where the bar refers to an average with the distribution $P(\rho)$. In this simplistic model, since all positions are equivalent, they are associated with a common weight ϕ, and the SCA matrix is therefore a matrix made of identical elements, $\tilde{C}_{ij} = \phi^2 \left(\overline{\rho^2} - \overline{\rho}^2 \right)$. Such a constant matrix has a single non-zero mode (of order L, the number of positions), whose origin is purely historical.

In practice, we expect both historical and functional constraints to contribute to the first mode of the SCA matrix. Regardless, this coherent mode is not useful in defining the sectors, which correspond to non-coherent correlations of different groups of positions. Based on a similar logic, the first mode of the correlation matrix is disregarded when analyzing statistical interactions between stocks in finance; in this context, it is interpreted as the coherent response of all stocks to external economic factors. In addition, we note that historical constraints could also contribute partially to modes other than the first one (especially when the sampling of sequences is strongly non-uniform and segregated into subfamilies, see section F below). Previous studies have advanced various approaches for reducing the effect of non-uniform sampling in protein alignments; future work should evaluate these approaches for improving the signal to noise in the SCA matrix.

E. Sector identification

The identification of the sectors is based on the intermediate modes 2 to 5. Specifically, we identify here three sectors in the serine protease family based on the second and fourth eigenvectors, $|2\rangle$ and $|4\rangle$, of $\widetilde{C}_{ij}$ (we use here the bracket notation for representing eigenvectors: $|2\rangle$ thus denotes the second eigenvector, with

component along position *i* given by $\langle i|2\rangle$). The justification for using the modes 2 and 4 is provided in Figure S3. More precisely, the red sector is defined as the positions i for which $\langle i|2\rangle > \varepsilon$ and $\langle i|2\rangle > |\langle i|4\rangle|$, the blue sector as those for which $\langle i|2\rangle < -\varepsilon$ and $\langle i|2\rangle < -|\langle i|4\rangle|$, and the green sector as those for which $\langle i|4\rangle > \varepsilon$ and $\langle i|4\rangle > |\langle i|2\rangle|$. The threshold $\varepsilon = 0.05$ is chosen to separate significant weights along an eigenvector from statistical noise, and is obtained by comparison with an analysis of randomized alignments (Fig. S2C-E). Eigenvectors are in general defined only up to a multiplicative factor and are here normalized so that $\sum_i \langle i|k\rangle^2 = 1$, but the sign of the weights to which a sector is associated is arbitrary.

The particular eigenvector with which a sector is associated, while not arbitrary, has no fundamental significance. Thus, the green sector is found here along $|4\rangle$, but, in other implementations of the SCA, it may be found along $|3\rangle$ or $|5\rangle$. The group of positions with largest weights along an eigenvector can indeed comprise different sub-groups of positions not significantly correlated. Figure S3B ($|3\rangle$ vs $|2\rangle$) thus shows that the group of positions with largest positive weights along $|3\rangle$ is subdivided by $|2\rangle$ into two uncorrelated sub-groups, corresponding to the red and blue sectors. Figure S3C ($|4\rangle$ vs $|3\rangle$) shows that the group of positions with largest negative weights along $|3\rangle$ is similarly subdivided by $|4\rangle$ into two uncorrelated sub-groups, one of which corresponding to the green sector. A difference between the other sub-group, with $\langle i|4\rangle < -\varepsilon$, and the green sector, with $\langle i|4\rangle > \varepsilon$, is apparent in Figure S3E ($|5\rangle$ vs $|4\rangle$) which shows that $|5\rangle$ subdivides the sub-group with $\langle i|4\rangle < -\varepsilon$ but not the green sector, which consists of more correlated positions. Below, this sub-group is shown to be a "pseudo-sector", which unlike the blue, red, and green sectors, arises from a specific phylogenetic bias in the S1A alignment.

The same principles, when applied to stock correlations in the financial context, leads to the identification of economic sectors. Generally, in proteins as in finance, a sector needs not be simply associated with a eigenvector (it may for instance correspond to a linear combination of different eigenvectors). Statistical methods beyond spectral analysis may thus be more appropriate to define sectors. For example, upon the assumption that statistically independent sectors do exist, techniques such as independent component analysis (ICA, (Hyvarinen et al., 2001; Stone, 2004)) may be valuable. Preliminary analysis of the S1A family using ICA identifies the three sectors as independent components and is consistent with results reported here.

F. "Pseudo sectors"

For the statistical method followed here to lead to protein sectors, one requirement is that the sequences must be distributed sufficiently uniformly. While presenting a precise criterion of uniformity and providing solutions for cases where this condition is not met is beyond the scope of the present paper, the origin of the problem can be clearly illustrated on the example of the serine protease alignment.

A simple way to visualize how sequences are distributed is by projecting them along the top eigenvectors of the global similarity matrix, i.e., the matrix $\Gamma_{st}^{(S)}$ introduced in B, where *S* consists of all the positions. For our alignment of serine proteases, this representation, shown in Fig. S4A, reveals a small but distinct subfamily of sequences (in yellow), which comprises the snake venom proteases. The deviation from uniform sampling caused by the presence of this subfamily has a direct impact on the spectral property of the correlation matrix $\widetilde{C}_{ij}$: it is responsible for the "pseudo sector" corresponding to the subgroup shown as non-sector positions with

large negative weight along the third eigenvector of $\widetilde{C}_{ij}$, and large negative weight along the fourth eigenvector (Fig. S3).

To demonstrate this, we define the pseudo-sector as comprising positions i satisfying $\langle i|3\rangle > \varepsilon$, but not associated with either the red, blue or green sector; these positions are colored in magenta in Fig. S4B. Fig. S4C shows how the blue, red, green, and magenta sectors classify the sequences in relationship to the global phylogenetic bias that separates the snake proteases from the remainder of sequences. This analysis shows that the magenta sector uniquely separates the snake proteases from the rest of the sequences, a finding that implies that the magenta sector (but not the blue, red, and green sectors) is a "pseudo-sector" arising from non-uniform sampling. To further test the claim that the magenta sector specifically arises from the presence of the snake protease subfamily, we repeated the SCA and spectral analysis after removing this subfamily from the alignment. The result shows that while the blue, red, and green sectors remain intact, the magenta sector is now no longer identifiable in the significant eigenvectors of $\widetilde{C}_{ij}$ (Fig. S4D).

This example serves as an illustration of one of the limitations of the approach taken here to identify sectors from the eigenvectors of the correlation matrix $\widetilde{C}_{ij}$: when the sequences in the alignment are non uniformly distributed, and, more particularly, when distinct subfamilies of sequences are present, this approach can result in the identification of pseudo-sectors. However, the analysis of the S1A alignment also shows that such pseudo-sectors have distinct statistical properties, indicating that further methods may be designed to correct this problem. The development of such methods is a matter for future research, but two aspects of the solution seem clear: (1) simply eliminating a subfamily from the alignment is unlikely to be the most appropriate solution, as sequences forming a subfamily may contribute to identifying actual sectors, and (2) simply disregarding the positions attributed to a pseudo sector is also unlikely to be an appropriate solution, as pseudo sectors may correspond to artifactual subdivisions of actual sectors.

G. Representation of significant correlations (Fig. 1E)

To summarize the result of the above procedures, Figure 1E represents the correlations between the 65 identified sector positions captured by the significant modes 2 to 4 of $\widetilde{C}_{ij}$. The corresponding matrix can formally be written $\tilde{C}' = \sum_{k=2}^{4} \lambda_k |k\rangle\langle k|$. In Figure 1E, the sector positions are ordered with, first, the positions i of the red sector, ordered by decreasing values of $\langle i|2\rangle$, second, the positions i of the blue sector, ordered by increasing values of $\langle i|2\rangle$, and, third, the positions i of the green sector, ordered by decreasing values of $\langle i|4\rangle$. Finally, only the positive elements of $\tilde{C}'$ are represented.

$\tilde{C}'$ also contains some negative inter-sector correlations, mainly between the blue and red sectors. They reflect an artifact of a simplistic representation relying exclusively on the modes 2 to 4. Indeed, while the first mode is not required for identifying the sectors, it can account for parts of the functional correlations, which are not included in $\tilde{C}'$. Consistently with this argument, Figure S5 shows that the negative elements of $\tilde{C}'$ correspond to small elements of the original correlation matrix $\tilde{C}$ while, in contrast, the large positive elements of $\tilde{C}'$ correspond to large elements in $\tilde{C}$. Finding a more appropriate representation of the functional correlations is an issue beyond the scope of the present paper, which concentrates on the identification and description of sectors.

H. MDI calculations

Principles: The minimum discriminatory information (MDI,(Kullback, 1997)) method (see Methods) aims at generalizing the definition of positional conservation based on relative entropies to include correlations between positions. Its principles are completely distinct from the SCA method. It can be defined for an arbitrary number of amino acids and correlations of arbitrary order, although computational complexity limits the scope of the calculations (see below). The calculations presented here are made for pair-wise correlations between positions from the sectors, as identified by the spectral analysis of the $\widetilde{C}_{ij}$ matrix. For simplicity, we also restrict the presentation of the method to the binary approximation; a sequence is thus represented by a binary vector x with components x_i, where $x_i = 1$ if the sequence has the amino acid a_i at position *i*, and 0 otherwise.

Given the background distribution $Q(x) = \prod_i \left(q^{(a)}\right)^{x_i} \left(1 - q^{(a)}\right)^{1-x_i}$ and a set of frequencies *F* (whose elements have the form $f_i^{(a_i)}$ or $f_{ij}^{(a_i a_j)}$), the MDI method yields a probability distribution $P_F^*(x)$ whose marginals reproduce the frequencies in *F* and which is "minimally biased" with respect to *Q(x)*. The reproduction of the marginals precisely means that

$$\sum_{\{x_k\}_{k\neq i}} P_F^*(x_1,...,x_{i-1},1,x_{i+1},...,x_L) = f_i^{(a_i)}$$

when $f_i^{(a_i)}$ is in *F* and

$$\sum_{\{x_k\}_{k\neq i,j}} P_F^*(x_1,...,x_{i-1},1,x_{i+1},...,,x_{j-1},1,x_{j+1},...,x_L) = f_{ij}^{(a_i a_j)}.$$

when $f_{ij}^{(a_i a_j)}$ is in *F*.

The MDI approach selects for $P_F^*(x)$ the probability distribution *P(x)* which minimizes the relative entropy defined by

$$D(P\|Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}.$$

When *F* comprises only individual frequencies $f_i^{(a_i)}$, it can be shown that $P_F^*(x) = \prod_i \left(f_i^{(a_i)}\right)^{x_i} \left(1 - f_i^{(a_i)}\right)^{1-x_i}$ and that $D(P_F^* \,||\, Q) = \sum_i D_i^{(a_i)}$, where $D_i^{(a_i)}$ is the position-specific conservation at position i defined above. The entropic approach used for defining positional conservation can thus be seen as a particular case of the MDI approach.

When *F* consists of the pair-wise frequencies $f_{ij}^{(a_i a_j)}$, the optimal probability distribution $P_F^*(x)$ can be characterized as the only distribution of the form

$$P_F^*(x) = \frac{1}{Z} \exp\left(\sum_{i<j} J_{ij} x_i x_j + \sum_i h_i x_i \right)$$

where the parameters h_i, J_{ij} and *Z* are fixed by the constraints on the marginals of $P_F^*(x)$ associated with the $f_i^{(a_i)}$'s and $f_{ij}^{(a_i a_j)}$'s, and by the constraint that $\sum_x P_F^*(x) = 1$.

Computing the entropy $D(P_F^* \,||\, Q)$ for F including the $f_{ij}^{(a_i a_j)}$ for all pairs of positions of the serine protease family is a computationally intractable problem (the difficulty stems from the fact that the sequence space cannot be sampled exhaustively because of its enormous dimension, which, even in the binary approximation, is 2^{223}). We are, therefore, led to restrict the calculations to subsets of positions S (for which the associated set F consists of all the frequencies between pairs of positions in S). We use here a simple and exact method known as generalized iterative scaling (Darroch and Ratcliff, 2007), which can treat up to about 20 positions. More sophisticated or approximate algorithms could allow us to reach larger sizes, but including all positions is in any case out of reach and studying groups of small size is sufficient to exhibit the salient statistical features of the serine protease alignment.

Algorithm: The generalized iterative scaling algorithm (Darroch and Ratcliff, 2007) starts by setting all the parameters h_i and J_{ij} to $h_i^{(0)} = 0$ and $J_{ij}^{(0)} = 0$, and iteratively updates them according to the rules

$$h_i^{(t+1)} = h_i^{(t)} + \frac{2}{L(L+1)} \ln \frac{f_i^{(a_i)}}{\langle x_i \rangle_t}, \text{ and } J_{ij}^{(t+1)} = J_{ij}^{(t)} + \frac{2}{L(L+1)} \ln \frac{f_{ij}^{(a_i a_j)}}{\langle x_i x_j \rangle_t},$$

where $\langle x_i \rangle_t$ and $\langle x_i x_j \rangle_t$ are averages taken with the probability distribution $P_S^{(t)}(x) = \frac{1}{Z^{(t)}} \exp\left(\sum_{i<j} J_{ij}^{(t)} x_i x_j + \sum_i h_i^{(t)} x_i \right)$ and L is the number of positions under consideration ($Z^{(t)}$ is a normalization). These iterations are guaranteed to converge towards the exact value of the parameters. Given P_S^*, the probability distribution to which the iterations converge, the entropy for the group of positions S, D_S, is given by $D\left(P_S^* \| Q\right)$.

Entropies as measure of conservation: In the present context, entropies provide an estimate of the degree of conservation of a group of positions. Figure S6A thus depicts the entropies of different groups of positions, each consisting of the 5 positions with largest weight along the principal component of the SCA matrix defining the three sectors (for comparison, an additional group is represented in black, that includes the 5 positions with largest negative weight along $|4\rangle$). The green sector appears to be clearly more conserved than the red or blue sector, consistently with the fact that it is associated with the catalytic mechanism, which is the most conserved feature of the family.

Entropies as measure of statistical independence: Entropies can also be used to estimate the degree of statistical dependence between members of a group of positions S. This is done by comparing the entropy D_S of the group with the sum of positional conservations $\sum_{i \in S} D_i^{(a_i)}$. The difference $I_S = D_S - \sum_{i \in S} D_i^{(a_i)}$, which we call the correlation entropy, is necessarily non-negative, and quantifies the statistical interactions between positions. This part of the total entropy D_S is represented with a lighter color in Figure S6A. It appears to be larger for the red or blue sectors than for the group of positions with negative weights along $|4\rangle$, represented in black, although this group has a total entropy D_{black} of same order of magnitude than D_{red} and D_{blue}. This observation indicates that positions belonging to this group are statistically more independent than those forming the red and blue sector, which supports the view that they should not define a sector.

Entropies can finally be used to estimate the degree of statistical independence between different groups of positions. In Figure S6B, we computed the entropy $D_{red+blue+green}$ for the group of 15 positions comprising the same 5 positions from each sector than in Figure S6A. The difference $\Delta = D_{red+blue+green} - D_{red} - D_{blue} - D_{green}$, represented in white in Figure S6B, quantifies the statistical interdependence of the sectors. It is to be compared

with the correlation entropy for these 15 positions, $I_{red+blue+green} = D_{red+blue+green} - \sum_i D_i^{(a_i)}$. The latter quantity is represented in Figure 2D where it is decomposed into the 3 parts measuring the respective intra-sector statistical dependences of each sector, I_{red}, I_{blue} and I_{green} (respectively in red, blue and green), and a part measuring the inter-sector statistical dependence $\boldsymbol{\Delta}$ (in white), corresponding to $I_{red+blue+green} = I_{red} + I_{blue} + I_{green} + \Delta$. It is clear that the inter-sector contribution represents only about 20% of the total; as a comparison, when forming other groups with the same positions, the inter-group contribution is found to account for about 60% of the total (these groups were randomly formed with the only constraint that no more than 3 of the 5 positions of a same sector are assigned to the same group).

III. Supplemental MATLAB Script for SCA Calculations

The following script details how the calculations and figures related to the identification of the sectors by SCA can be reproduced using standard MATLAB functions. The only requirements are 'msa_serprot.mat', a mat-file containing the multiple sequence alignment of the S1A serine protease family, and the MATLAB Image Processing toolbox, for representing the SCA matrix. The script follows the sections A-E of the preceding supplementary notes. MATLAB codes reproducing the calculations presented in sections F-G are available upon request. The S1A alignment is available for download from the Ranganathan lab website, and a MATLAB toolbox for general usage of SCA methods is available by request from the authors (rama.ranganathan@utsouthwestern.edu).

```
%% A. Measures of conservation

load msa_serprot
% loads 'msa', an alignment of the S1A serine protease family having the
% form of a 1470*223 array, where each line corresponds to a different
% sequence. Amino acids are represented by their standard one-letter
% abbreviation and gaps by '-':
Code_aa='ACDEFGHIKLMNPQRSTVWY-';

[N_seq,N_pos]=size(msa);
% N_seq gives the number of sequences, N_pos the number of positions.

% Background probabilities:
freq_bg=[.073 .025 .050 .061 .042 .072 .023 .053 .064 .089...
          .023 .043 .052 .040 .052 .073 .056 .063 .013 .033];

% Frequencies of amino acids at given positions:
freq=zeros(21,N_pos);
for a=1:21, freq(a,:)=sum(msa==Code_aa(a))./N_seq; end
% freq(a,i) gives the frequency of amino acid a at position i.
% (gaps are treated as a 21st amino acid for latter convenience)

% Prevalent amino acid at each position:
[freq_bin,prev_aa]=max(freq(1:20,:));
% Code_aa(prev_aa(i)) gives the prevalent amino acid at position i, and
% freq_bin(i) its frequency.

% Simplified alignment in the binary approximation:
msa_bin=1.*(msa==repmat(Code_aa(prev_aa),N_seq,1));
% for each position (column of the array msa_bin), the prevalent amino acid
% is represented by '1', and all other amino acids, including gaps, by '0'.

% Background probabilities for the prevalent amino acids:
freq_bg_bin=freq_bg(prev_aa);

% Relative entropies in the binary approximation:
D_bin=freq_bin.*log(freq_bin./freq_bg_bin)...
      +(1-freq_bin).*log((1-freq_bin)./(1-freq_bg_bin));

% Fig. 1A: relative entropies.
figure(1); bar(1:N_pos,D_bin);axis([0 N_pos+1 0 4]);
xlabel('positions');ylabel('D_i^{(a_i)}');

% A histogram of relative entropies.
figure(2); hist(D_bin,N_pos/5);
xlabel('D_i^{(a_i)}');ylabel('number');

% Fraction of gaps:
frac_gaps=sum(freq(21,:))/N_pos;

% Background probabilities accounting for gaps:
freq_bg_wg=[(1-frac_gaps)*freq_bg frac_gaps];
freq_bg_bin_wg=freq_bg_wg(prev_aa);

% Relative entropy in the binary approx with above background probability:
D_bin_wg=freq_bin.*log(freq_bin./freq_bg_bin_wg)...
         +(1-freq_bin).*log((1-freq_bin)./(1-freq_bg_bin_wg));

% Overall conservation
```

```
D_glo=zeros(1,N_pos);
for i=1:N_pos,
    for a=1:21
        if(freq(a,i)>0)
            D_glo(i)=D_glo(i)+freq(a,i)*log(freq(a,i)/freq_bg_wg(a));
        end
    end
end
% (when freq(a,i)=0, freq(a,i)*log(freq(a,i)) is to be considered =0,
%  since x*log(x)->0 for x->0)

% Fig. S1: validity of the binary approximation
figure(3); plot(D_glo,D_bin_wg,'o');
xlabel('Overall conservation');
ylabel('Conservation in the binary approximation');


%% B. SCA calculations

% Correlation matrix in the binary approximation:
freq_pairs_bin=msa_bin'*msa_bin/N_seq;
C_bin=freq_pairs_bin-freq_bin'*freq_bin;

% Weights (defined as gradients of relative entropy):
W=log(freq_bin.*(1-freq_bg_bin)./(freq_bg_bin.*(1-freq_bin)));

% SCA matrix:
C_sca=(W'*W).*abs(C_bin);

% Fig. 1D: representation of the SCA matrix
figure(4); imshow(C_sca,[0 .5]);colormap(jet);


%% C. Spectral cleaning

% Spectrum of the SCA matrix
[eigvect_unsorted,lambda_unsorted]=eig(C_sca);
[lambda,lambda_order]=sort(diag(lambda_unsorted),'descend');
eigvect=eigvect_unsorted(:,lambda_order);
% (the eigenvector are ordered for future convenience)

% A randomization is performed at the level of the alignment:
% the amino acids are permuted between the sequences independently
% for each position.
% The positional conservations are therefore preserved.
N_samples=100; N_ev=5;
lambda_rnd=zeros(N_samples,N_pos);
eigvect_rnd=zeros(N_samples,N_pos,N_ev);
for s=1:N_samples
    msa_bin_rnd=zeros(N_seq,N_pos);
    for pos=1:N_pos
        perm_seq=randperm(N_seq);
        msa_bin_rnd(:,pos)=msa_bin(perm_seq(:),pos);
    end
    freq_pairs_bin_rnd=msa_bin_rnd'*msa_bin_rnd/N_seq;
    C_bin_rnd=freq_pairs_bin_rnd-freq_bin'*freq_bin;
    C_sca_rnd=(W'*W).*abs(C_bin_rnd);
    [eigvect_unsorted,lambda_unsorted]=eig(C_sca_rnd);
    [lambda_sorted,lambda_order]=sort(diag(lambda_unsorted),'descend');
    lambda_rnd(s,:)=lambda_sorted;
    eigvect_rnd(s,:,:)=eigvect_unsorted(:,lambda_order(1:N_ev));
end
% Note that 'freq_bin' and 'W' are not affected by the randomization.

% Fig. S2A: comparison of spectra
figure(5);
subplot(2,1,1); [yhist,xhist]=hist(lambda,N_pos); bar(xhist,yhist,'k');axis([0 30 0 35]);
xlabel('eigenvalues (actual alignment)');ylabel('number');
[n]=hist(lambda_rnd(:),xhist);
subplot(2,1,2); bar(xhist,n/N_samples,'k'); axis([0 30 0 35]);
xlabel('eigenvalues (randomized alignments)');ylabel('number');
% (the histogram from randomized alignments is normalized for comparison)


% Fig. S2B: Interpretation of the first mode
figure(6);
```

```
plot(eigvect(:,1),sum(C_sca)/(sum(sum(C_sca).^2))^(1/2),'o');
axis([0 .25 0 .25]);
xlabel('<i |1> (first eigenvector of C\_ sca)');
ylabel('\Sigma_j C\_ sca_{ij} (normalized)');

% Fig. S2C-E: Threshold for the weights on eigenvectors 2 to 4
figure(7);
for k=2:4
    subplot(2,3,k-1);
    hist(eigvect(:,k),N_pos); axis([-.25 .25 0 5]);
    xlabel(['<i |' num2str(k) '>']); ylabel('number');
    subplot(2,3,k+2);
    z=eigvect_rnd(:,:,k);[n,x]=hist(z(:),N_pos);
    plot(x,n/N_samples,'r'); axis([-.25 .25 0 5]);
    xlabel(['<i |' num2str(k) '> (random)']); ylabel('number');
end

% Definition of the noise threshold:
threshold=.05;

% Definition of sector positions:
sec_red=find(eigvect(:,2)>max(threshold,abs(eigvect(:,4))));
sec_blue=find(eigvect(:,2)<-max(threshold,abs(eigvect(:,4))));
sec_green=find(eigvect(:,4)>max(threshold,abs(eigvect(:,2))));

% Cleaned SCA matrix:
C_clean=zeros(N_pos,N_pos);
for k=2:4, C_clean=C_clean+lambda(k)*eigvect(:,k)*eigvect(:,k)'; end

% Ordering of sector positions:
[x,order]=sort(eigvect(sec_blue,2)); sec_blue_ord=sec_blue(order);
[x,order]=sort(-eigvect(sec_green,4)); sec_green_ord=sec_green(order);
[x,order]=sort(-eigvect(sec_red,2)); sec_red_ord=sec_red(order);

% Fig. 1E: cleaned SCA matrix
sec_all_ord=[sec_blue_ord; sec_green_ord; sec_red_ord];
figure(8); imshow(C_clean(sec_all_ord,sec_all_ord),[0 .25]);colormap(jet);

% Negative elements in the cleaned SCA matrix
figure(9); imshow(C_clean(sec_all_ord,sec_all_ord),[-.5 .5]);colormap(jet);

% Fig. S4: relation between cleaned and original correlations
figure(10);
plot(C_sca(:),C_clean(:),'o');
xlabel('elements of the original SCA matrix');
ylabel('elements of the cleaned SCA matrix (based on eigenvectors 2-4)');


%% D. Sector indentification

% Fig. S3: representation of the significant eigenvectors
modes=[2 3; 2 4; 3 4; 4 5];
figure(11);
for k=1:4
    subplot(2,2,k); x=modes(k,1); y=modes(k,2);
    plot(eigvect(:,x),eigvect(:,y),'ok',...
         eigvect(sec_blue,x),eigvect(sec_blue,y),'ob',...
         eigvect(sec_green,x),eigvect(sec_green,y),'og',...
         eigvect(sec_red,x),eigvect(sec_red,y),'or');
    xlabel(['<i |' num2str(x) '>']);ylabel(['<i |' num2str(y) '>']);
    axis([-.4 .4 -.4 .4]);
end
```

Supplemental References

Bouchaud, J.-P., and Potters, M. (2004). Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management, 2 edn (Cambridge University Press).

Cover, T.M., and Thomas, J.A. (2006). Elements of Information Theory, 2 edn (Wiley-Interscience).

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. (2007) Generalized Iterative Scaling For Log-Linear Models. The Annals of Mathematical Statistics *43*, 1470-1480.

Efron, B., and Tibshirani, R.J. (1994). An Introduction to the Bootstrap, 1 edn (Chapman and Hall/CRC).

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

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.

Hyvarinen, A., Karhunen, J., and Oja, E. (2001). Independent Component Analysis (New york, John Wiley and Sons, Inc.).

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

Kurosky, A., Barnett, D.R., Lee, T.H., Touchstone, B., Hay, R.E., Arnott, M.S., Bowman, B.H., and Fitch, W.M. (1980). Covalent structure of human haptoglobin: a serine protease homolog. Proc Natl Acad Sci U S A *77*, 3388-3392.

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

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.

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.

Stone, J.V. (2004). Independent Component Analysis: A Tutorial Introduction (The MIT Press).

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.

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.

Wigner, E.P. (1967). Random Matrices in Physics. Siam Review *9*, 1-23.

Red Sector	Blue Sector	Green Sector
17	21	19
161	26	33
172	46	42
176	52	43
177	68	55
180	69	56
183	71	57
187	77	58
188	80	102
189	81	141
191	104	142
192	105	184
213	108	194
215	118	195
216	123	196
220	124	197
221	136	198
226	153	199
227	157	213
228	201	214
	210	216
	229	225
	237	
	242	
	245	

Table S1: Sector composition; by convention, residue numbering is per bovine chymotrypsin. See Supplemental Experimental Procedures for specific criteria for determining sectors.

		Thermal stability	Catalytic parameters		
		Tm (Kelvin)	Km (M^{-1})	kcat (s^{-1})	kcat/Km (M^{-1} s^{-1})
WT	Rat trypsin	325.58 (+/- 0.71)	6.6E-05 (+/- 4.3E-06)	27 (+/- 0.53)	4.1E+05 (+/- 2.8E+04)
Blue Sector	M104A	310.65 (+/- 0.92)	6.7E-05 (+/- 8.2E-06)	14 (+/- 0.76)	2.1E+05 (+/- 2.9E+04)
	L105A	315.10 (+/- 0.70)	1.4E-04 (+/- 1.9E-05)	38 (+/- 2.37)	2.7E+05 (+/- 4.1E+04)
	Q210A	319.85 (+/- 0.26)	1.2E-04 (+/- 1.9E-05)	43 (+/- 3.02)	3.4E+05 (+/- 5.8E+04)
	T229A	317.55 (+/- 1.74)	1.1E-04 (+/- 1.5E-05)	45 (+/- 2.93)	4.1E+05 (+/- 6.2E+04)
	P124A	318.58 (+/- 0.12)	1.0E-04 (+/- 1.0E-05)	40 (+/- 1.20)	4.0E+05 (+/- 4.2E+04)
	C157A	310.02 (+/- 1.19)	8.7E-05 (+/- 9.0E-06)	60 (+/- 2.09)	6.9E+05 (+/- 7.6E+04)
	C136A	284.0 (+/- 0)	7.7E-05 (+/- 1.2E-05)	19 (+/- 1.20)	2.5E+05 (+/- 4.0E+04)
	M104A,L105A	310.75 (+/- 0.69)	2.5E-05 (+/- 6.4E-06)	18 (+/- 1.19)	7.4E+05 (+/- 2.0E+05)
	M104A,Q210A	306.60 (+/- 0.76)	6.2E-05 (+/- 9.0E-06)	35 (+/- 1.78)	5.6E+05 (+/- 8.6E+04)
	M104A, T229A	306.92 (+/- 0.25)	8.2E-05 (+/- 1.9E-05)	50 (+/- 4.81)	6.1E+05 (+/- 1.5E+05)
	L105A, Q210A	310.23 (+/- 0.57)	1.4E-04 (+/- 2.4E-05)	24 (+/- 1.91)	1.7E+05 (+/- 3.1E+04)
	L105A, T229A	310.12 (+/- 0.12)	1.3E-04 (+/- 1.0E-05)	37 (+/- 1.31)	2.7E+05 (+/- 2.3E+04)
	Q210A, T229A	311.32 (+/- 0.32)	1.7E-04 (+/- 2.3E-05)	65 (+/- 4.52)	3.8E+05 (+/- 5.7E+04)
Red Sector	G216A	324.15 (+/- 1.15)	7.9E-03 (+/- 1.7E-03)	72 (+/- 11.14)	9.1E+03 (+/- 2.4E+03)
	G226A	326.52 (+/- 1.24)	8.1E-03 (+/- 1.1E-03)	4 (+/- 0.04)	4.8E+02 (+/- 6.4E+01)
	C191A	322.05 (+/- 0.20)	5.0E-03 (+/- 1.8E-04)	1 (+/- 0.02)	1.6E+02 (+/- 7.5E+00)
	D189A	324.12 (+/- 1.96)	N/A	N/A	1.1E+01 (+/- 5.7E-01)
	V183A	324.85 (+/- 0.89)	8.1E-05 (+/- 9.2E-06)	23 (+/- 0.89)	2.8E+05 (+/- 3.4E+04)
	Hswap	327.65 (+/- 0.20)	N/A	N/A	1.1E+01 (+/- 6.5E-01)
Blue-Red Sector	G216A, Q210A	319.32 (+/- 1.19)	8.3E-03 (+/- 2.5E-03)	51 (+/- 13.67)	6.2E+03 (+/- 2.5E+03)
	G216A, C157A	309.02 (+/- 0.60)	4.6E-03 (+/- 1.9E-03)	16 (+/- 7.47)	3.5E+03 (+/- 2.2E+03)
Non Sector	Y29A	321.22 (+/- 0.72)	9.1E-05 (+/- 7.1E-06)	47 (+/- 1.07)	5.2E+05 (+/- 4.2E+04)
	Q30A	325.38 (+/- 0.61)	1.1E-04 (+/- 2.3E-05)	39 (+/- 3.02)	3.7E+05 (+/- 8.2E+04)
	K230A	321.75 (+/- 0.71)	1.8E-04 (+/- 3.2E-05)	45 (+/- 4.14)	2.6E+05 (+/- 5.2E+04)

Table S2: The kinetic parameters and thermal stability for enzymes shown in Figure 5 with standard errors in parenthesis. Each measurement represents between three and five independent experiments.

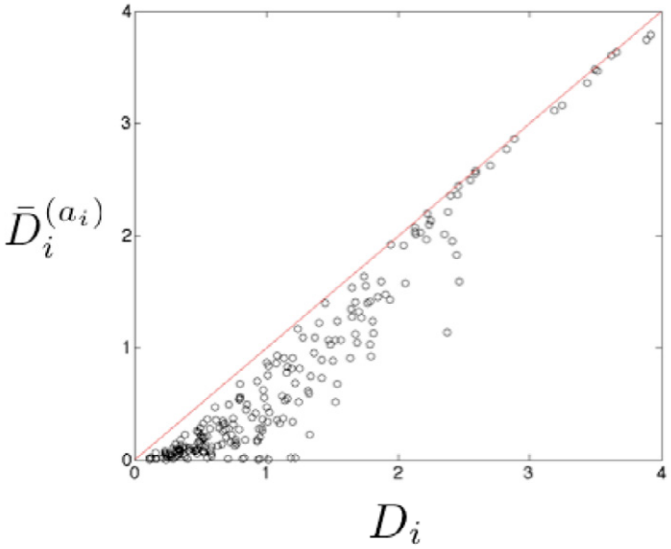

Figure S1: Close agreement between the overall relative entropy of positions (D_i) compared with $\overline{D}_i^{(a_i)}$, the value for just the most prevalent amino acid in the multiple sequence alignment of 1470 members of the trypsin family of serine proteases. These data provide the basis for the simplified binary approximation used in this work.

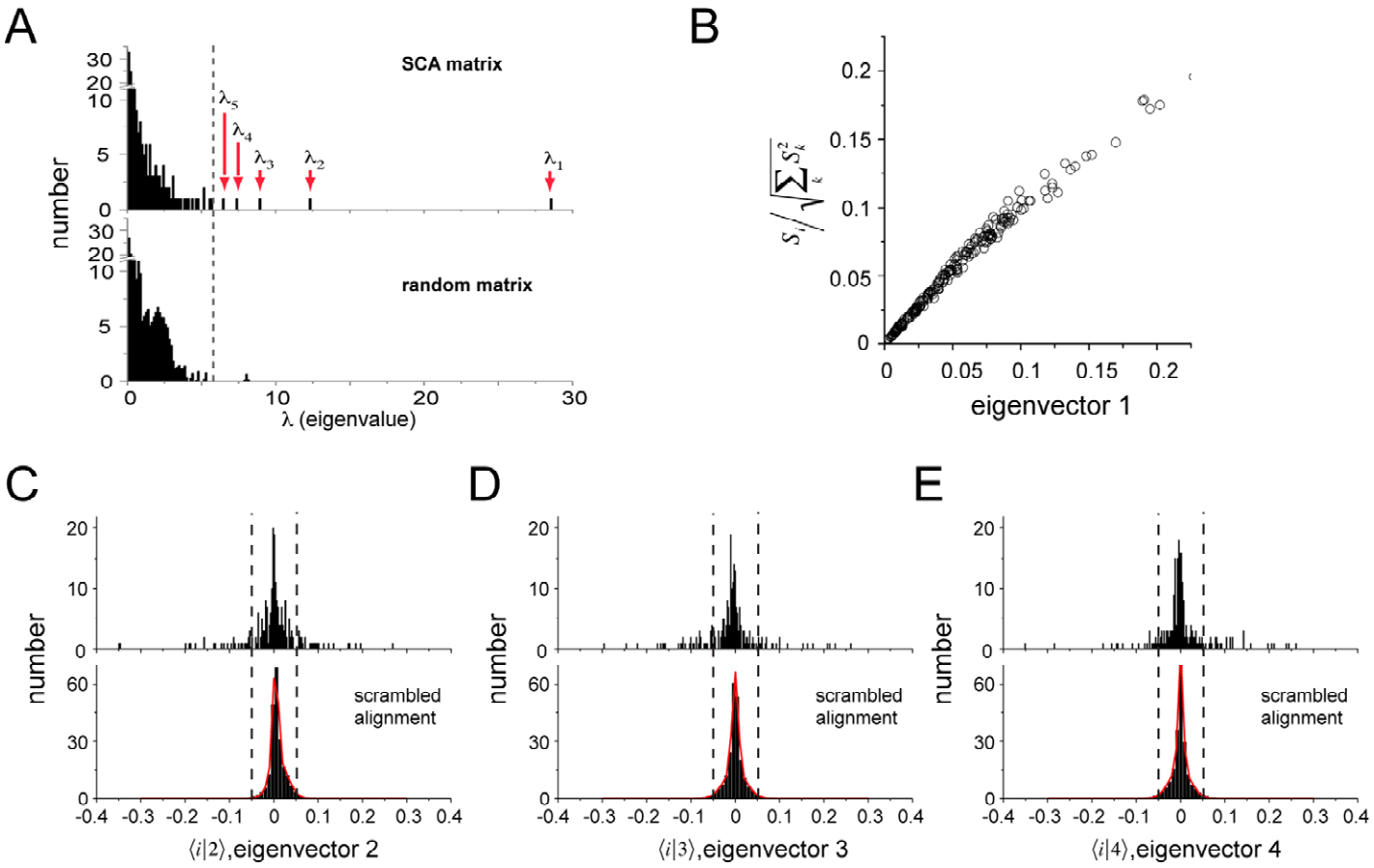

Figure S2: A, Eigenvalue spectra for the $\tilde{C}_{ij}$ matrix corresponding to the S1A serine protease family (top panel) and for a hundred trials for randomizing the S1A sequence alignment (bottom panel). The randomization process scrambles the order of amino acids in each alignment column independently; thus amino acid frequencies at positions are never changed. This analysis shows that the bulk of the spectrum (comprising the lowest 218 out of 223 total eigenvalues) can be attributed to limited sampling of sequences. **B,** A scatter plot of the first mode of the $\tilde{C}_{ij}$ matrix against the net contribution of each position to the total correlation. As described in the Supplemental Experimental Procedures and Supplemental Discussion, this relationship is expected for SCA matrices with a dominant first mode. **C-E**, The distribution of positional weights for eigenvectors 2-4 of the $\tilde{C}_{ij}$ matrix (upper panels). The bottom panels show the average distribution of position weights for 100 trials of randomization of the S1A multiple sequence alignment. The randomization procedure shuffles the order of amino acids at each position in the alignment, a process that eliminates all correlations between positions while preserving the amino acid frequencies at each position. The randomized eigenvector distributions are minimally well-fit by a two Gaussian model (r^2>0.95 in each case, in red). The dashed lines at +/- 0.05 for each eigenvector represent roughly two standard deviation limits for the broader Gaussian and are used as significance thresholds for determining sector compositions (see Fig. S3).

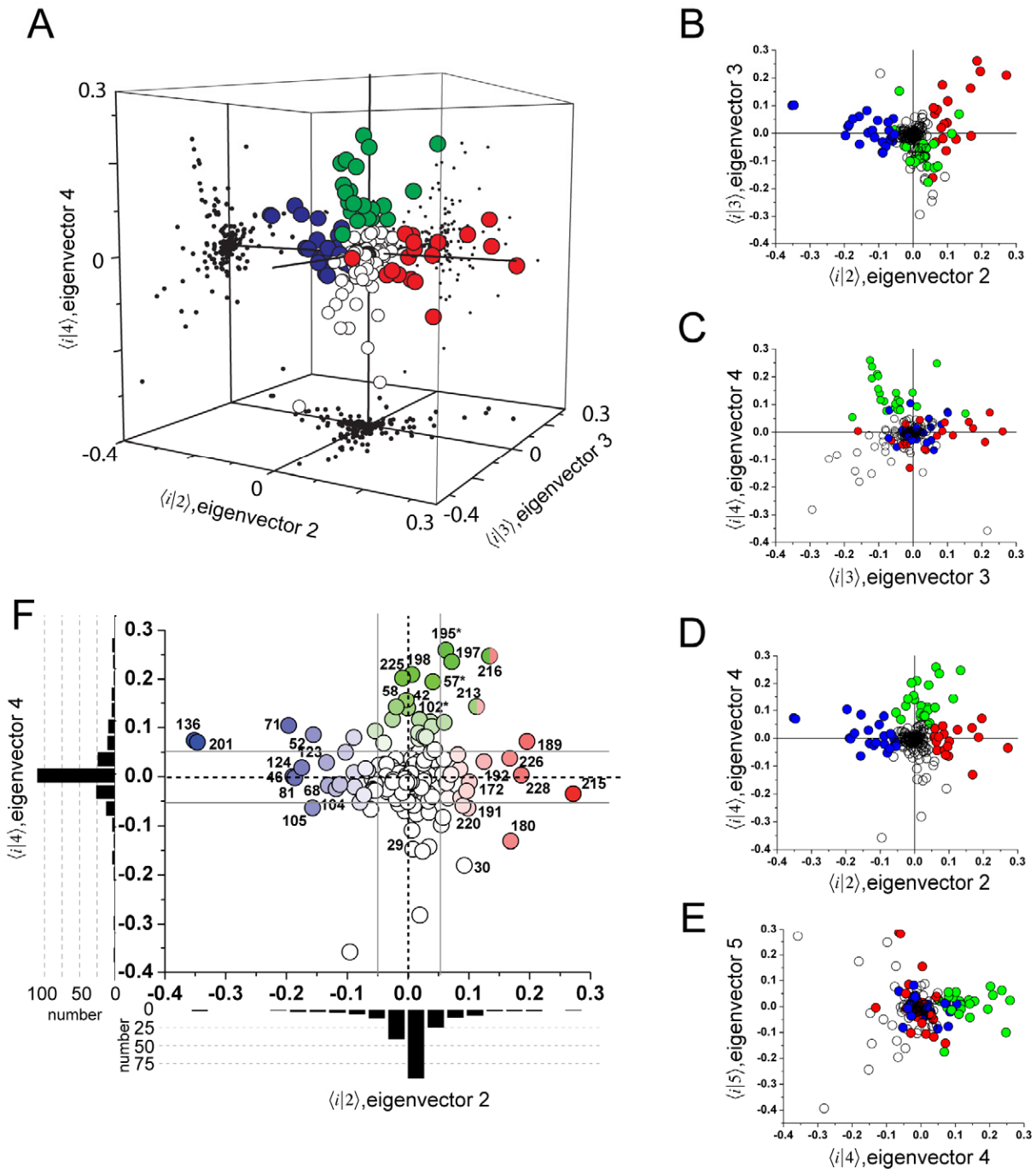

Figure S3: Residue weights along eigenvectors 2-4 of the SCA correlation matrix. **A**, a three dimensional scatter plot of these eigenvectors shows a pattern of residue weights in which most of the 223 positions in the alignment contribute little and cluster near the origin, and the three sectors (colored red, blue and green), each comprising a small fraction of total residues, emerge as distinct groups along characteristic directions. Together with the two dimensional projections of all pairs of these eigenvectors (**B-D**, and black dots, panel **A**), these data show (**B**) that the red and blue sectors separate along the second principal component, (**C**) that the green sector begins to emerge along eigenvector 3 and is further separated from other positions with significant eigenvector 3 weights (in white) along eigenvector 4, and (**D**) that the plot of eigenvectors 2 and

4 provides the clearest basis for sector identification. **E**, a plot of eigenvectors 4 and 5 shows that while the green sector remains intact, the group of white residues also separating along eigenvectors 3 and 4 are yet further subdivided along eigenvector 5, arguing that these are unlikely to represent a unique sector. The finer subdivisions of these positions along the lower eigenvectors may have functional meaning, but given the first order approach in this work of just studying the statistically obvious eigenvectors, is not considered here. **F**, Detailed plot of residue weights for eigenvectors 2 and 4, with labels for positions making large contributions to each sector. The color gradients reflect the weight along the respective eigenvectors for comparison with Fig. 3A-C, and the starred residues represent the catalytic triad. The gray lines reflect the average range of weights for 100 trials of alignment randomization (see Supplemental Discussion and Fig. S2 C-E).

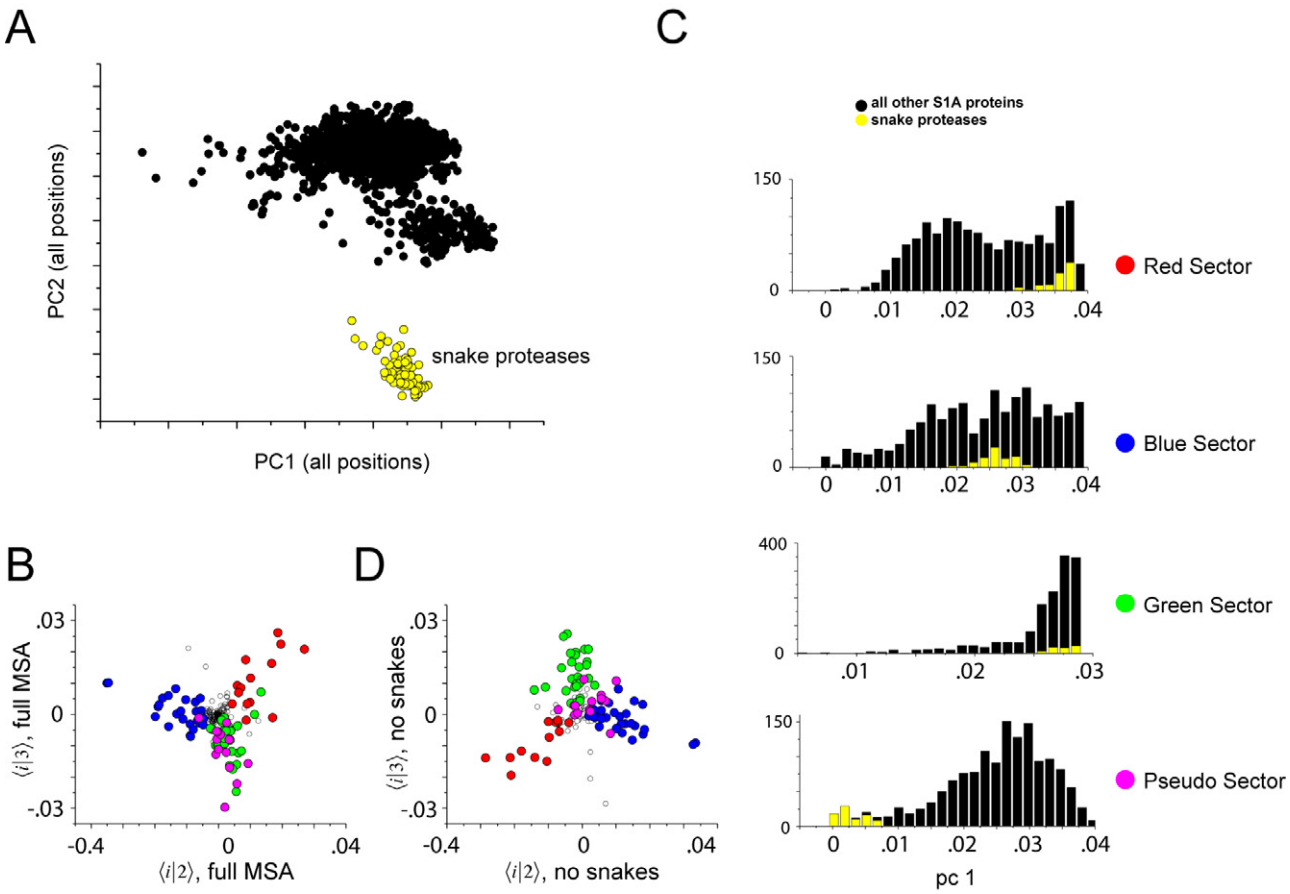

Figure S4: A pseudo-sector in the S1A family. **A**, S1A sequences projected on the two top eigenvectors of the global similarity matrix $\Gamma_{st}^{(S)}$ (section B), where *S* consists of all the positions. The first two eigenvectors capture much of the variance in the similarity matrix, and so this analysis provides a reasonable mapping of the sequence relationships between S1A proteins. This analysis shows that while most S1A proteins are roughly uniformly distributed, a small, distinct subfamily of sequences exists (in yellow), which comprises the snake venom proteases. **B**, sequence position weights for eigenvectors 2 and 3 of the $\tilde{C}_{ij}$ matrix (as in Fig. 3B) but with the pseudo-sector positions colored in magenta. **C**, histograms of all the

1470 sequences in the S1A alignment projected along the first principal component of the similarity matrix calculated based on the blue, red, green, and magenta sectors, as labeled. The snake proteases are shown in yellow; this analysis shows that while the blue, red, and green sectors do not effectively classify sequences by this globally distinct subfamily, the magenta sector does. This identifies the magenta sector as a pseudo-sector – a group of positions that likely emerges as correlated due to historical noise. **D**, The positional weights along eigenvectors 2 and 3 of the $\tilde{C}_{ij}$ matrix calculated after removal of the snake proteases. This shows that while the blue, red, and green sectors are intact, the magenta sector is no longer evident, a finding that reinforces the notion that this pseudo-sector emerges solely due to the presence of the distinct clade of snake proteases.

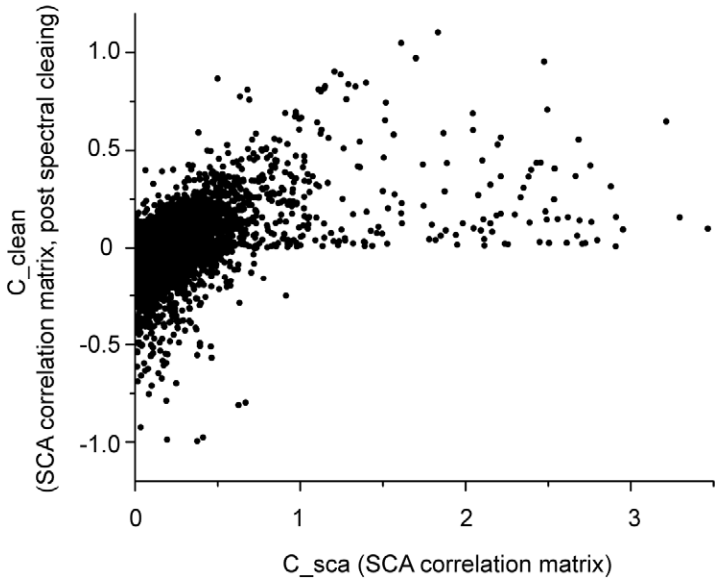

Figure S5: A scatterplot of correlations in the initial SCA correlation matrix ($\tilde{C}_{ij}$, Fig. 1D, called C_sca in the MATLAB script) against those in the correlation matrix computed after spectral "cleaning" to remove eigenvalues 1 and 5-223 ($\tilde{C}'_{ij} = \sum_{k=2}^{4} \lambda_k |k\rangle\langle k|$, called C_clean in the MATLAB script). Post cleaning, the correlations matrix shows negative correlations that arise exclusively from weakly correlated position pairs in the initial correlation matrix. These negative correlations are artifactual in nature due to complete removal of the first eigenvalue and are not shown in representation of the cleaned SCA correlation matrix (Fig. 1E).

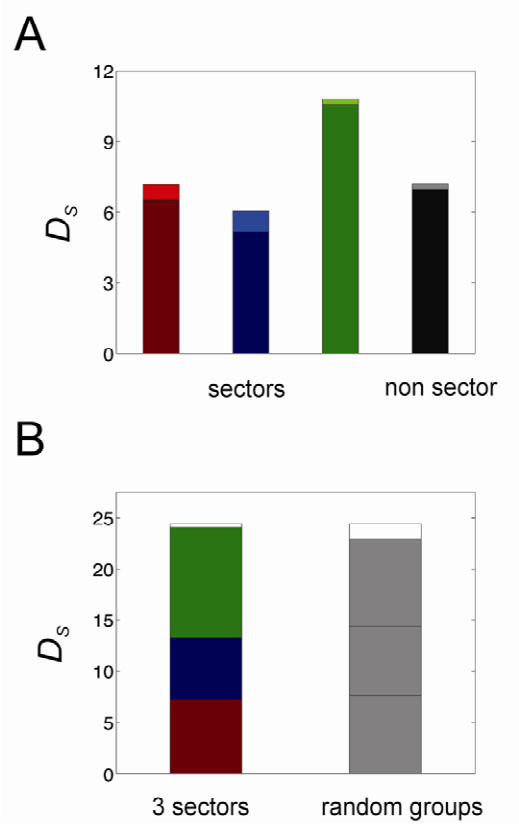

Figure S6: A, Entropies D_S for S composed of the 5 positions with largest positive weights along eigenvector 2 (red sector, in red), negative weights along eigenvector 2 (blue sector, in blue), positive weights along eigenvector 4 (green sector, in green) or negative weights along eigenvector 4 (non-sector, in black). D_S is the sum of position-specific entropies, $\sum_{i \in S} D_i^{(a_i)}$ (darker color), and of a correlation entropy I_S (lighter color). **B,** Entropy D_S for S composed of the 15 positions consisting of the 5 positions with largest positive weights along eigenvector 2 (red sector), negative weights along eigenvector 2 (blue sector) and positive weights along eigenvector 4 (green sector). D_S is the sum of the sector-specific entropies shown in A, $D_{red} + D_{blue} + D_{green}$ (in red, blue and green), and of an inter-sector correlation entropy Δ (in white). For comparison, the results of the decomposition of D_S in 3 groups randomly formed with the same 15 positions, is also shown.

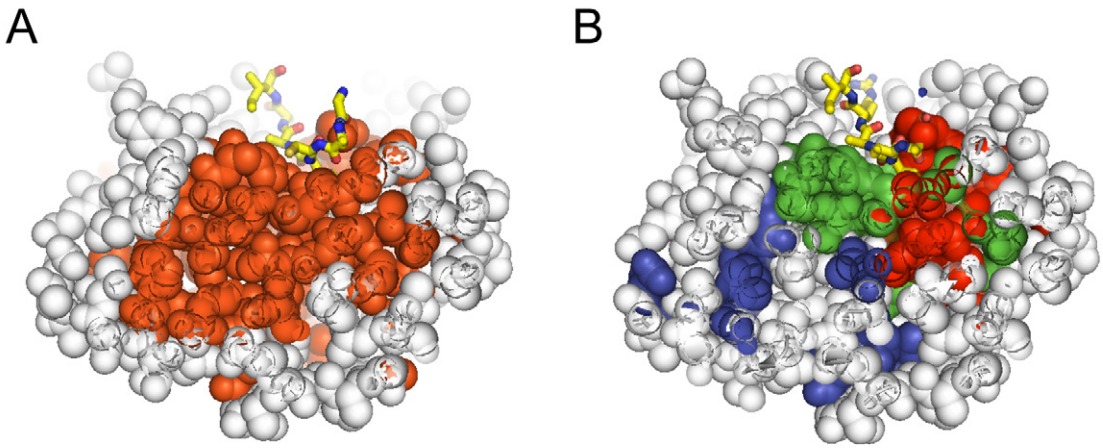

Figure S7: A slice through the core of rat trypsin, with residues colored by positional conservation (**A,** reproduction of Fig. 1C), or sector composition (**B,** defined in Fig. 3, and Figs. S2-S3). The data show that while positional conservation follows the simple rule that conserved residues tend to be located within the protein core and at functional surfaces, the analysis of the pattern of correlated conservation of sequence positions uniquely reveals the decomposition of conservation into sectors. Note that while most sector residues are also at least moderately conserved, not all conserved residues contribute to sectors. Taken together, this suggests a heterogeneous pattern of conservation within the protein core in which some conserved residues act more independently and idiosyncratically within members of a protein family while others act cooperatively and more systematically.

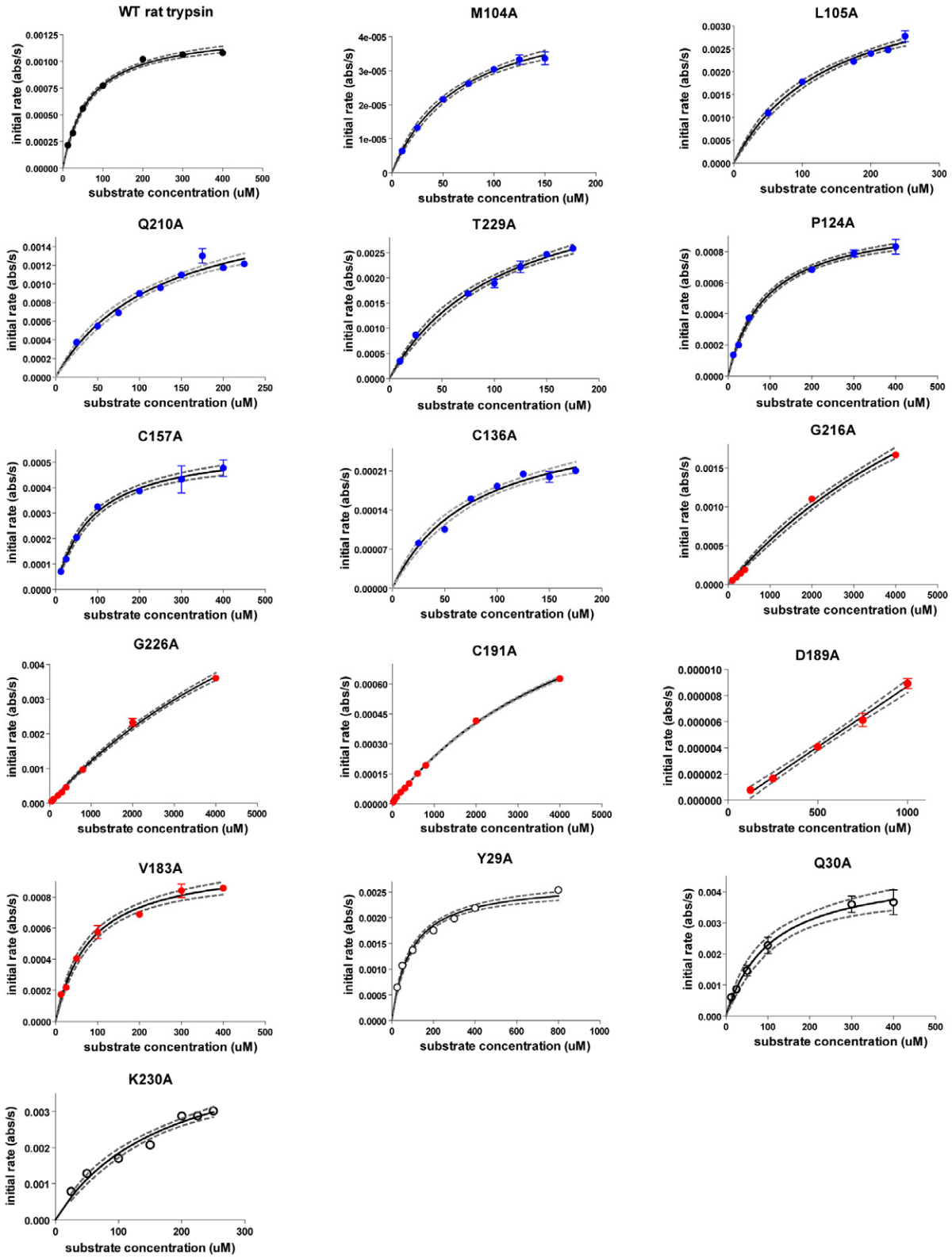

Figure S8a: Kinetic analysis of single alanine mutants in rat trypsin. Shown are plots of initial velocities as a function of substrate concentration; experiment conditions and details of the assay are described in the methods section. Symbols are color coded according to the following scheme: black (wild-type), red (red sector mutants), blue (blue sector mutants), white (non-sector mutants).

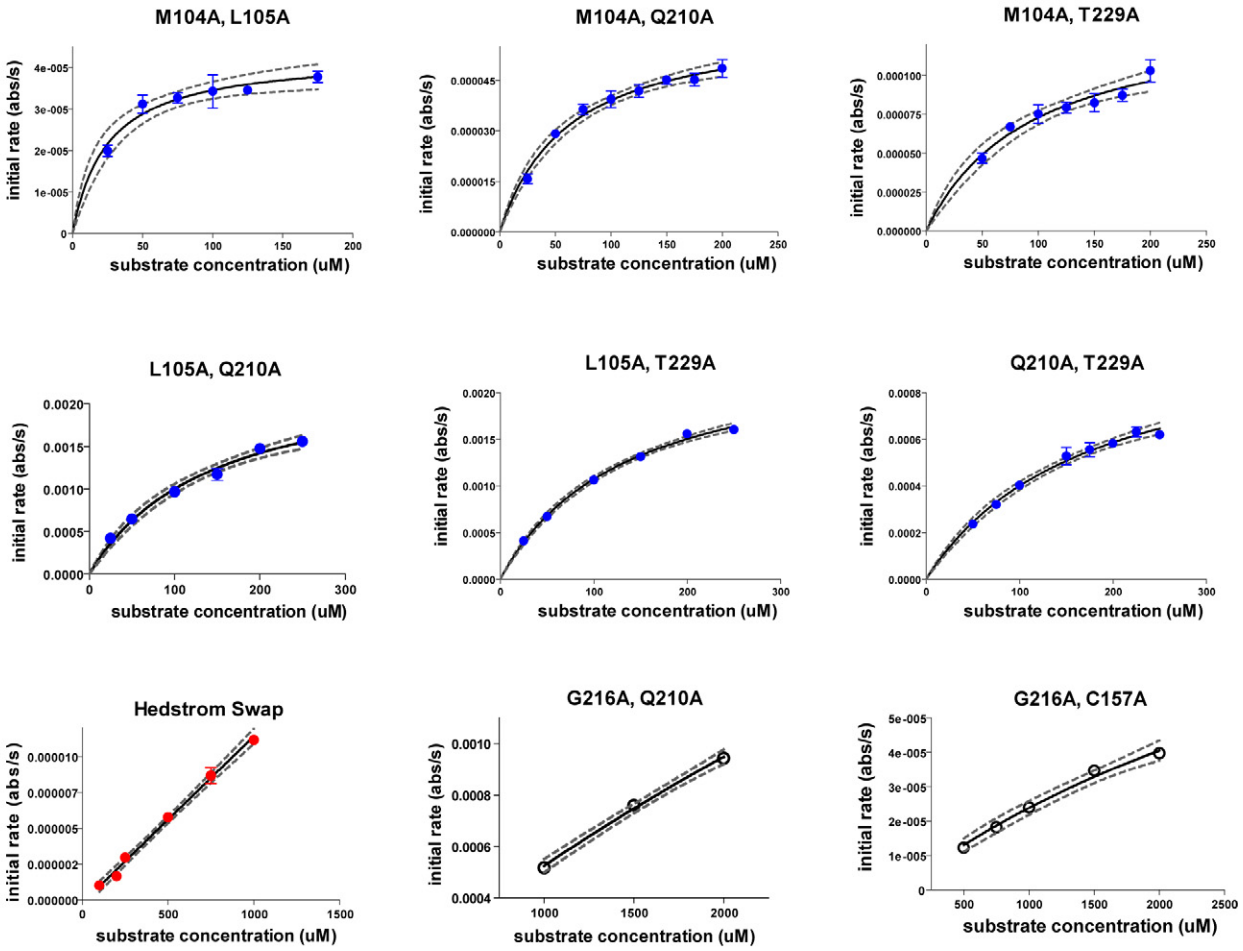

Figure S8b: Kinetic analysis of double alanine mutants in rat trypsin. Plots and color coding are as described in the Fig. S8a legend, except that white symbols in panels B-C represent double mutants as indicated in Fig. 5C

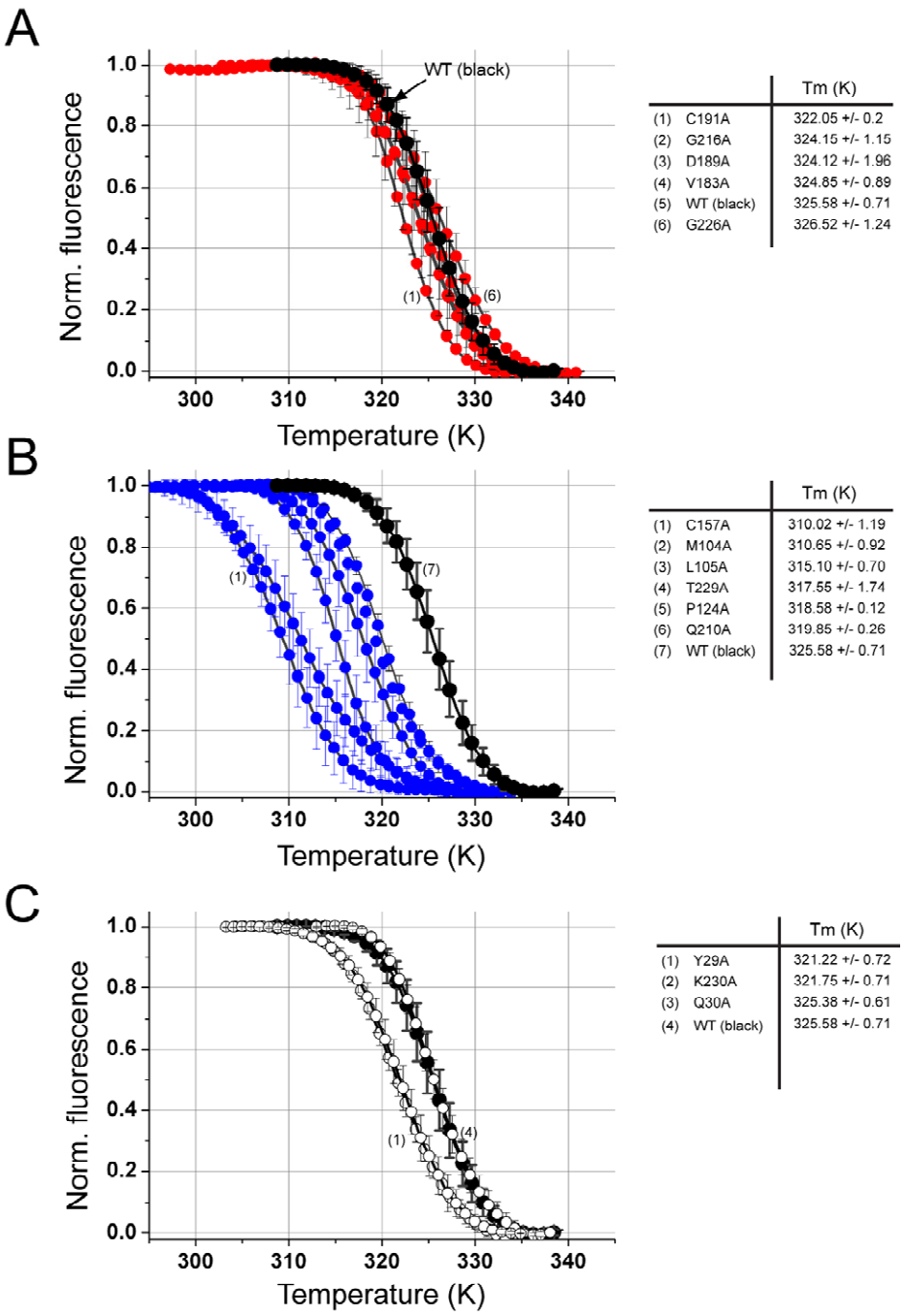

Figure S9a: Thermal denaturation data for single alanine mutants in rat trypsin. **A**, Mutants in the red sector (red curves) show denaturation profiles and associated T_ms (table to right, and in order of appearance in the graph left to right) that are similar to that of wild-type (in black). **B**, In contrast, mutations in the blue sector (blue curves) show T_ms (table to right, and in order left to right) that are significantly lower than wild type (in black). **C**, Mutants in non sector positions (in white) show T_ms that are similar to that of wild type (in black). For clarity, symbols are plotted every 20 data points.

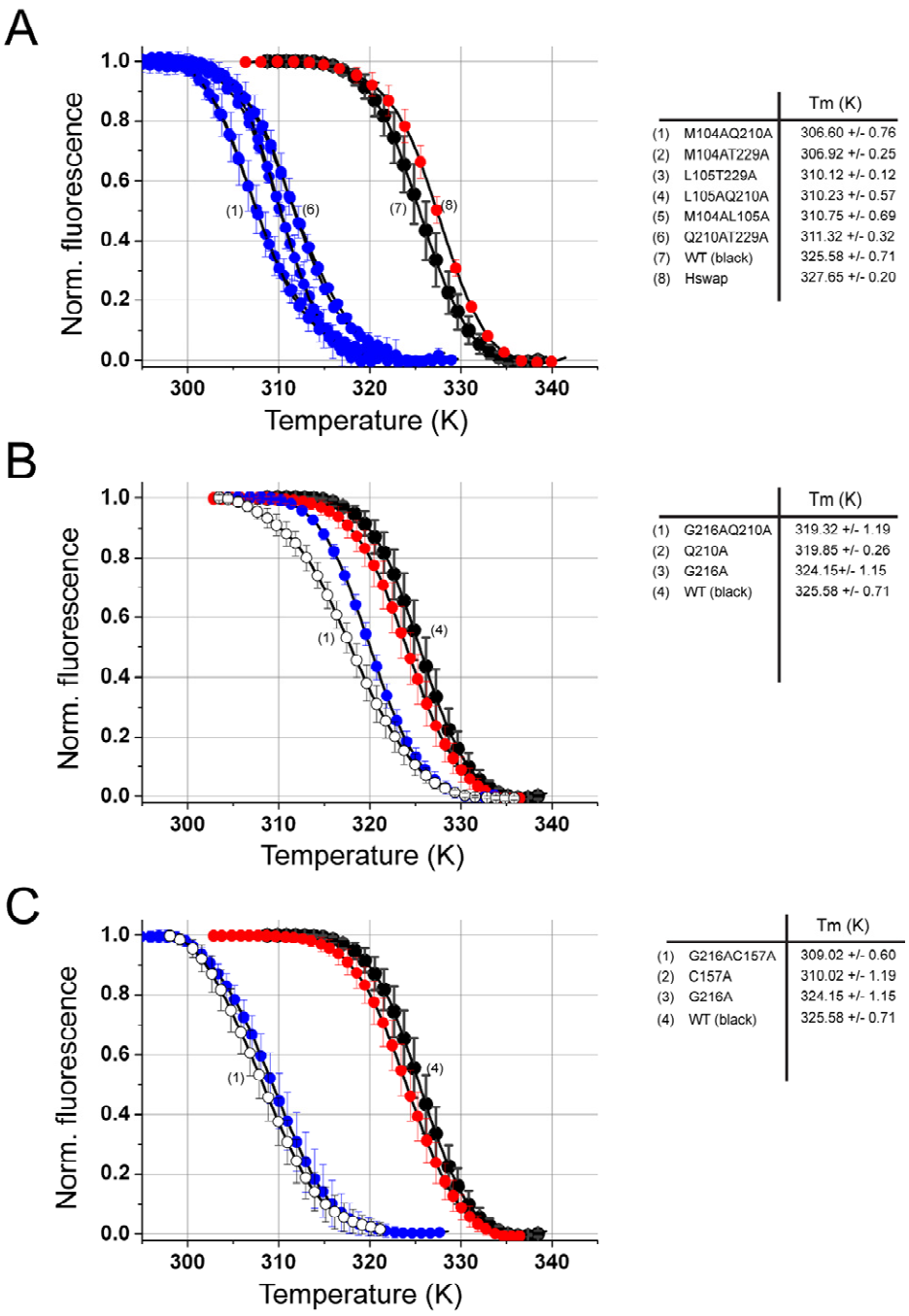

Figure S9b: Thermal denaturation data for double alanine mutants in rat trypsin. **A**, Double mutants within the blue sector (blue curves) show T_ms (table to right, in numbered order of appearance in graph left to right) that differ significantly from that of wild type (in black) and that are generally non additive with regard to the degree of destabilization for the single mutants (compare with Fig. S9a, panel A). In contrast, a multiple mutant in red sector residues (Hswap, red curve) shows a T_m close to that of wild type trypsin. **B-C**, Double mutants between red and blue sectors (in white) show destabilization relative to wild type (in black) that is nearly additive with regard to their associated single mutants (in red and blue). For clarity, symbols are plotted every 20 data points.

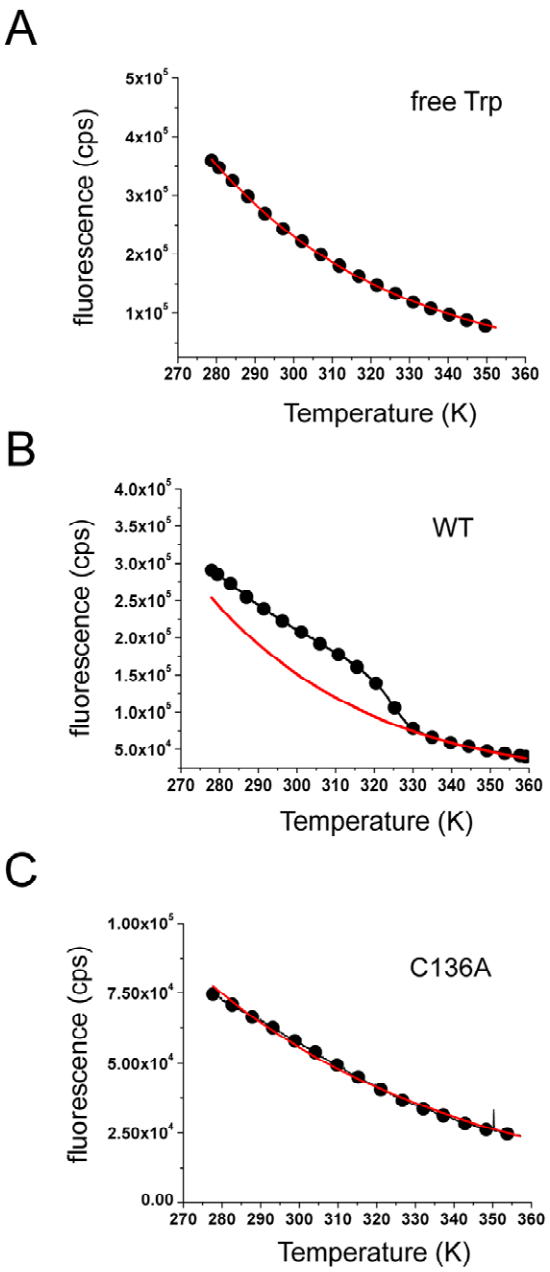

Figure S10: Thermal denaturation in C136A shows no unfolding transition under the experimental conditions used in this work. **A**, Intrinsic temperature dependence of tryptophan fluorescence is well fit by a single exponential function. **B**, Raw data for thermal denaturation of wild-type rat trypsin shows a clear sigmoidal transition corresponding to protein unfolding followed by a post-transition region that is well fit to a single exponential. **C**, Thermal denaturation in C136A shows no discernable transition, and is approximated throughout by a single exponential function.

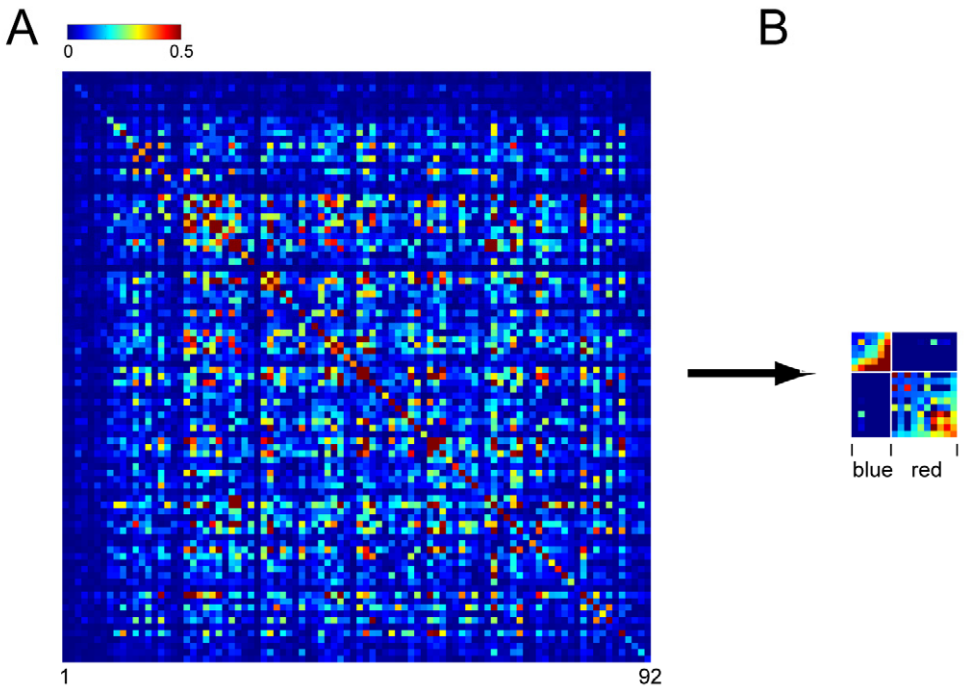

Figure S11: Sectors in the PDZ (PSD95/Dlg1/ZO-1) family of protein interaction modules. **A**, the SCA correlation matrix ($\tilde{C}_{ij}$) for an alignment of 240 PDZ domains (92 x 92 sequence positions), and **B**, the SCA matrix after removal of statistical noise and of global, coherent correlations ($\tilde{C}'_{ij} = \sum_{k=2}^{3} \lambda_k |k\rangle\langle k|$), and trimming to the 16 positions that show significant weights in the remaining eigenvectors 2 and 3. The 16 positions form two sectors, labeled red and blue.

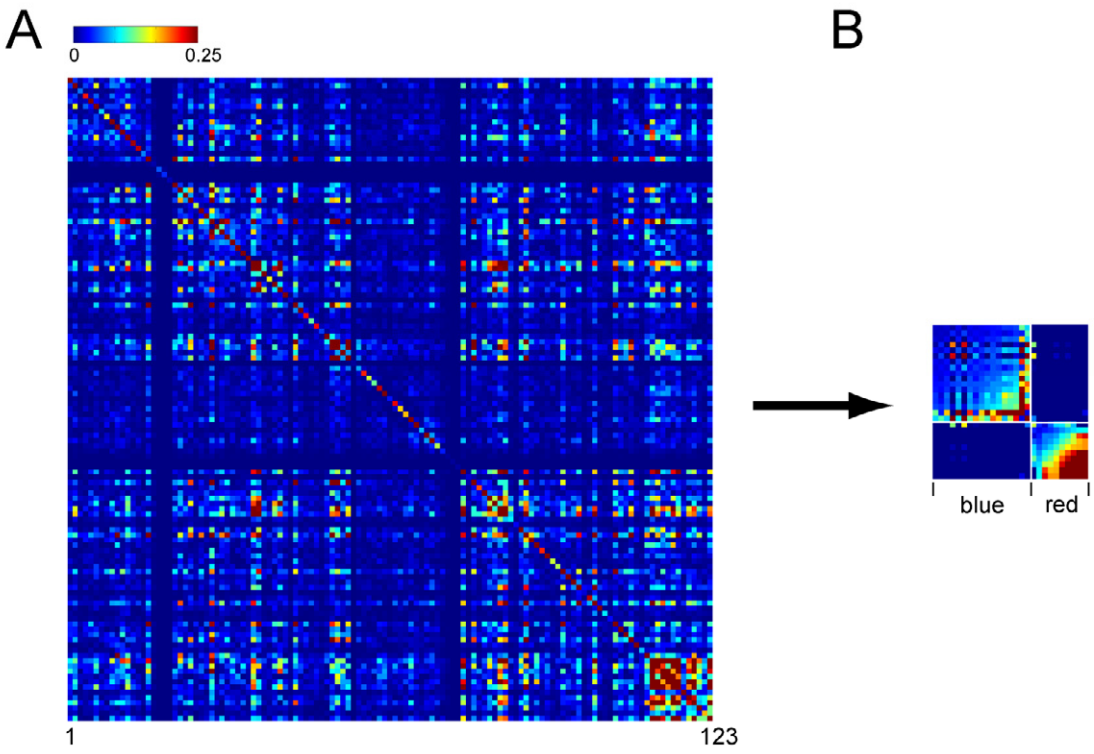

Figure S12: Sectors in the PAS (Per/Arnt/Sim) family of allosteric signaling modules. **A**, the SCA correlation matrix ($\tilde{C}_{ij}$) for an alignment of 1104 PAS domains (123 x 123 sequence positions), and **B**, the SCA matrix after removal of statistical noise and of global, coherent correlations ($\tilde{C}'_{ij} = \sum_{k=2}^{3} \lambda_k |k\rangle\langle k|$), and trimming to the 27 positions that show significant weights in the remaining eigenvectors 2 and 3. The 27 positions form two sectors, labeled red and blue.

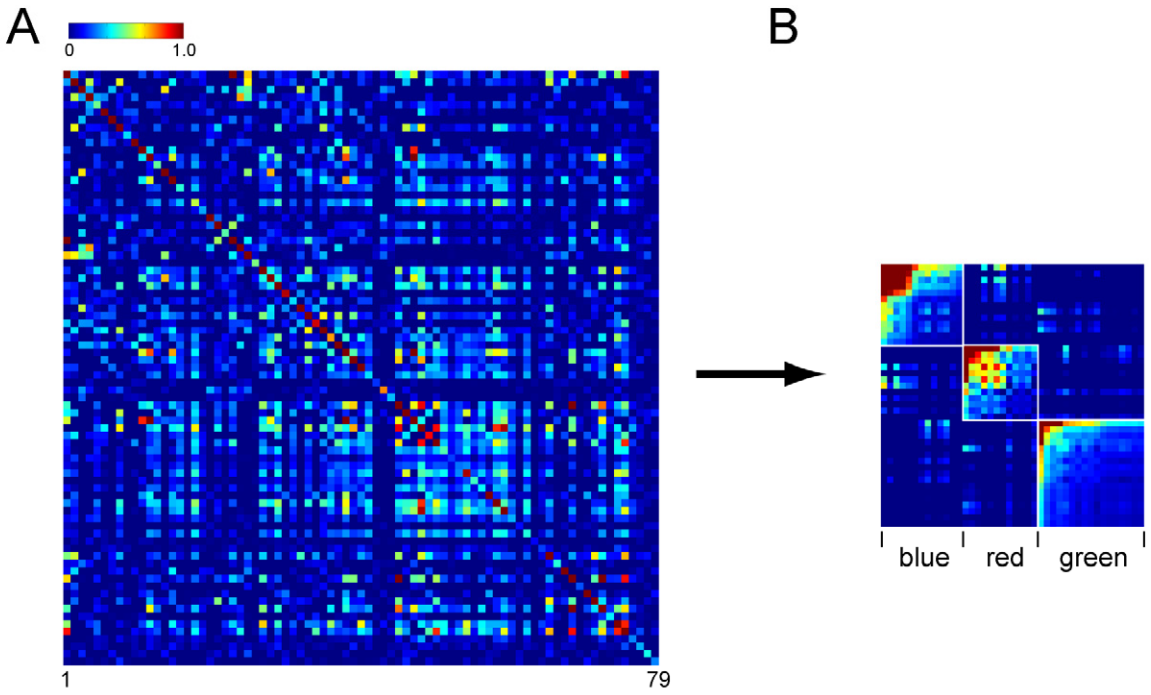

Figure S13: Sectors in the SH2 family of protein interaction modules. **A**, the SCA correlation matrix ($\tilde{C}_{ij}$) for an alignment of 582 SH2 domains (79 x 79 sequence positions), and **B**, the SCA matrix after removal of statistical noise and global, coherent correlations ($\tilde{C}'_{ij} = \sum_{k=2}^{3} \lambda_k |k\rangle\langle k|$), and trimming to the 42 positions that show significant weights in eigenvectors 2 and 3. The 42 positions comprise three sectors, labeled blue, red, and green.

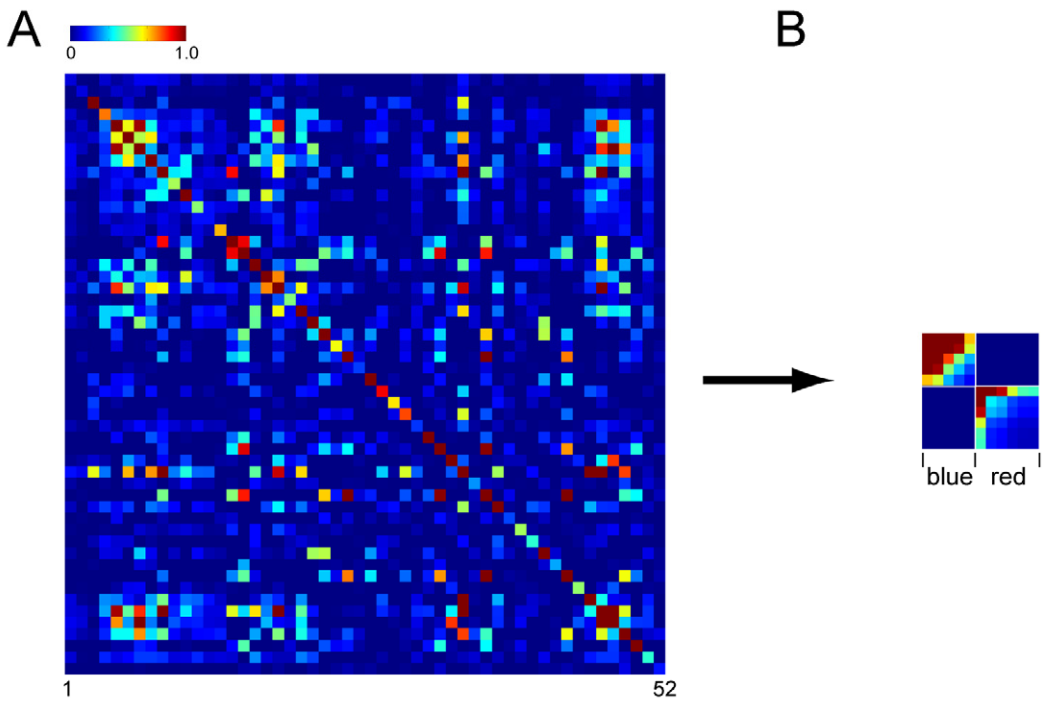

Figure S14: Sectors in the SH3 family of protein interaction modules. **A**, the SCA correlation matrix ($\tilde{C}_{ij}$) for an alignment of 492 SH3 domains (52 x 52 sequence positions), and **B**, the SCA matrix after removal of statistical noise and global, coherent correlations ($\tilde{C}'_{ij} = \lambda_2 |2\rangle\langle 2|$), and trimming to the 11 positions that show significant weights in eigenvector 2. The 11 positions comprise two sectors, labeled blue and red.