Supporting Information

Rausell et al. 10.1073/pnas.0908044107

SI Text

SI Materials and Methods

A. Details of the filtering of the Pfam-A alignments are as follows:

1. Alignments contained only eukaryotic sequences with evidence at the "protein" or "transcript" level from UniProtKB/Swiss-Prot (1) (release 54.6).
2. To retrieve the most informative area of the alignment, sequences and positions were restricted to less than 30% gaps. This selection was done recursively as follows:
   for ($i = 70$; $i \geq 30$; $i = i - 10$) {keep sequences with (% of gaps $\leq i$) based on positions with (% of gaps $\leq i$)}
   For each "*i*," the original sequences were recovered.
   Note that sequences were not trimmed in any way so that their ungapped version at this point remained true to the original Pfam sequence.
3. Redundancy at 95% sequence identity was removed, calculated by means of Cd-hit (2). For each redundant cluster, a representative sequence was selected maximizing the following scoring:
   - Sequences with PDB: +4.
   - Sequences with EC number: +2.
   - Sequences from HUMAN: +1.

In case of a tie within a redundant cluster, the largest sequence was selected.

4. Outliers were removed and they were defined as sequences with less than 40% identity against any other within the alignment.
5. Gappy positions (>10% gaps) were removed.
6. Alignments with less than 12 sequences or less than 25 positions were ignored. No upper limits were imposed regarding the number of sequences or positions.

B. Details of the protocol followed to define both the subfamilies and the specificity determining positions within a multiple sequence alignment. In this section, we provide the full mathematical details of the protocol followed to define both protein subfamilies and specificity determining positions (SDPs). In order to make such a description more clear to a nonexpert reader, we start with an intuitive explanation referring to the corresponding technical subsections.

In this work, the analysis of the multiple sequence alignments (MSAs) was carried out under the framework provided by multiple correspondence analysis (MCA). This methodology is based on correspondence analysis (3), a multivariate descriptive statistical technique that can be viewed as equivalent to a principal component analysis (PCA) when dealing with qualitative/binary data, such that proper quantitative values can be assigned to them (3, 4). MCA provides a meaningful vectorial representation of both proteins and residues on a space defined by the independent sources of variation within the MSA. This approach allows uncovering groups of proteins linked to groups of residues that will be later considered as protein subfamilies and SDPs, respectively.

First, the MSA is coded as a binary matrix (*Binary coding of a multiple sequence alignment*), which denotes an initial vectorial representation of both sequences and residues. The MCA performs a transformation of the coordinate system where the initial vectors are represented, into the so-called principal axes (*Definition of the principal axes and their associated explained variance*). The main property of these new "principal axes" is that they are uncorrelated, and, as a consequence, they can be interpreted as independent sources of variation within the initial MSA. Next, sequences and residues coordinates in the principal axes are obtained (*Projection of the sequences within the MSA onto the principal axes* and *Projection of the residues within the MSA onto the principal axes*, respectively). In this new vectorial representation of sequences and residues, distances between any pair of sequences and between any pair of positions account for Chi-squared distances. Chi-squared distances are suitable for binary/qualitative data such as the previously described coding for both sequences and residues (*Binary coding of a multiple sequence alignment*). Importantly, in these new coordinate systems, sequences and positions can be compared through the so-called "MCA pseudovaricentric relationships." According to these relationships, the center of masses of any group of sequences points to those residues particularly associated to them (*Pseudovaricentric relationships between the projected sequences and the projected residues*).

Additionally, the MCA framework allows one to obtain a favorable balance between relevant and noisy information within the initial MSA. This balance is achieved by the analysis of the amount of information provided by each new principal axis. In order to evaluate which axes can be considered relevant, the incremental information they convey is statistically assessed (*Selection of the number of informative principal axes*).

The *k*-means algorithm is used to obtain an automatic clustering of sequences in the selected dimensions. Clustering solutions are gathered for a prespecified number of groups ranging from two to one-fourth of the number of proteins (with a maximum of 50). Then, the clustering solution maximizing the ratio of the intercluster distances over the intracluster distances is selected. This procedure automatically identifies the putative groups of proteins, which are regarded as different subfamilies within the MSA (*Clustering of the projected sequences in the selected axes to define protein subfamilies*).

Finally, the centers of masses of these groups of proteins are then used to detect those residues particularly associated to them (*Assignment of residues to protein subfamilies*). SDPs are identified as those that better fit such a grouping (*SDPs definition*).

B.1. Binary coding of a multiple sequence alignment.

Given a MSA of N sequences and L positions, a data matrix $W_{N\times Q}$ of dimensions $N \times Q$ (where $Q = 21L$) is built, representing each position "l" in the alignment as a complete disjunctive category with 21 different modalities (representing the 20 amino acid types plus the gap), encoding the presence of a modality with "1" and its absence with "0."

Columns in W without a "1" are removed for subsequent consistency without a loss of generality, resulting in a matrix $X_{N\times P}$ of dimensions $N \times P$, where $P < Q$.

B.2. Definition of the principal axes and their associated explained variance.

The data matrix X defined above, with the general term x_{np}, allowed us to define the following frequencies:

$$x_{nS} = \sum_{p=1}^{P} x_{np}; \qquad x_{Sp} = \sum_{n=1}^{N} x_{np}; \qquad x_{SS} = \sum_{n=1}^{N}\sum_{p=1}^{P} x_{np};$$

$$f_{nS} = x_{nS}/x_{SS}; \qquad f_{Sp} = x_{Sp}/x_{SS}; \qquad f_{np} = x_{np}/x_{SS}.$$

Let $Y_{N\times P}$ be the matrix with the general term $y_{np} = f_{np}/(f_{Sp}\sqrt{f_{nS}})$. Let $Z_{N\times P}$ be the matrix with the general term

$z_{np} = f_{np}/\sqrt{f_{Sp} f_{nS}}$ and $Z^T_{P\times N}$ its transpose. Let $V_{N\times P}$ represent the matrix displaying by columns the eigenvectors of ZZ^T.

The space generated by the eigenvectors of ZZ^T provides an orthogonal decomposition of the sequences versus the residue-positions association between its sources of variation. For the sake of consistency and the improvement of the measure of fit, adjustments of eigenvalues are made a posteriori (5): Let λ_i be the ith nonnull eigenvalue of matrix ZZ^T. The trivial solution $\lambda = 1$ and the eigenvalues fulfilling $\lambda_i < (1/L^2)$, as well as its associated eigenvectors, are disregarded in the analysis. Therefore, $V_{N\times K}$ becomes $V_{N\times J}$, where J is the number of remaining eigenvalues. The eigenvalues are adjusted as follows:

$$\lambda_j^{\text{adj}} = \left(\frac{L}{L-1}\right)^2 \left(\lambda_j - \frac{1}{L}\right)^2.$$

Then, the percentage of total variance explained by each associated eigenvector can be calculated as

$$\frac{\lambda_j^{\text{adj}}}{\sum_{j=1}^{J} \lambda_j^{\text{adj}}} \cdot 100.$$

B.3. Projection of the sequences within the MSA onto the principal axes.

Let $R_{N\times N}$ be the diagonal matrix of dimensions $N \times N$ with the general term $r_{nn} = \frac{1}{\sqrt{f_{nS}}}$. Let $T_{P\times P}$ be the diagonal matrix of dimensions $P \times P$ with the general term $t_{pp} = \frac{1}{\sqrt{f_{Sp}}}$. Let $D_{J\times J}$ be the diagonal matrix of dimensions $J \times J$ with the general term $d_{jj} = \sqrt{\lambda_j}$ and $D'_{J\times J}$ its inverse. Let $D^{\text{adj}}_{J\times J}$ be the diagonal matrix of dimensions $J \times J$ with the general term $d^{\text{adj}}_{jj} = \sqrt{\lambda_j^{\text{adj}}}$.

The projection of the N sequences onto the space generated by the $V_{N\times J}$ eigenvectors of ZZ^T can be calculated as

$$\Theta^{\text{ppal}}_{N\times J} = \sqrt{N}(V_{N\times J} D^{\text{adj}}_{J\times J}),$$

where the general term $\theta^{\text{ppal}}_{nj}$ of $\Theta^{\text{ppal}}_{N\times J}$ accounts for the "principal coordinates" of the sequence "n" in the principal axis "j." Their so-called "standard coordinates" $\theta^{\text{stdr}}_{nj}$ are calculated as

$$\Theta^{\text{stdr}}_{N\times J} = \sqrt{N}(V_{N\times J} D'_{J\times J} D^{\text{adj}}_{J\times J}).$$

B.4. Projection of the residues within the MSA onto the principal axes.

In an analogous way, the projection of the P residue positions onto the space generated by the $V_{N\times J}$ eigenvectors of ZZ^T can be calculated as

$$\Psi^{\text{ppal}}_{P\times J} = T_{P\times P}(Z'_{P\times N} V_{N\times J}) D'_{J\times J} D^{\text{adj}}_{J\times J},$$

where the general term ψ^{ppal}_{pj} of $\Psi^{\text{ppal}}_{P\times J}$ accounts for the principal coordinate of the residue position "p" in the principal axis j.

B.5. Pseudovaricentric relationships between the projected sequences and the projected residues.

Let C be a number of additional features for which the N sequences within the MSA can be evaluated in binary terms. That is, we can build a binary matrix $\Phi_{N\times C}$ with general term φ_{nc}, where $\varphi_{nc} = 1$, if sequence n has a feature c or otherwise $\varphi_{nc} = 0$. That is,

$$\varphi_{nc}\left\{\begin{matrix}\varphi_{nc} = 1 \Leftrightarrow n \in c; n\{1,\ldots,N\}, c\{1,\ldots,C\}; \forall\, c\, \exists n/\varphi_{nc} = 1\} \\ \varphi_{nc} = 0 \Leftrightarrow n \notin c\end{matrix}\right\}.$$

Let $H_{C\times C}$ be the diagonal matrix of dimensions $C \times C$ with the general term $h_{cc} = 1/\sum_{n=1}^{N} \varphi_{nc}$.

Because of the so-called "pseudovaricentric relationships" in the MCA, it can be demonstrated that projecting a feature c, coded as a column in $\Phi_{N\times C}$, as a supplementary point in principal coordinates onto the space generated by the J eigenvectors of ZZ^T displayed in $V_{N\times J}$, and together with the P residue positions, is equivalent to assessing the center of masses of the n sequences having the feature c from the standard coordinates $\theta^{\text{stdr}}_{nj}$ of such sequences. The principal coordinates $\omega^{\text{ppal}}_{cj}$ of the feature c onto the principal axes j are calculated as

$$\Omega_{C\times J} = ((R_{N\times N}\Phi_{N\times C} H_{C\times C})^T V_{N\times J}) D'_{J\times J} D^{\text{adj}}_{J\times J}.$$

We can summarize these concepts by the MCA providing the orthogonal decomposition of the sources of variation within the initial MSA. These sources are represented by each of the principal axes (eigenvectors) coded by $V_{N\times J}$, and they can be prioritized by ranking their associated eigenvalues in descending order. Sequences and residue positions are represented in this space through their principal coordinates $\Phi^{\text{ppal}}_{N\times J}$ and $\Psi^{\text{ppal}}_{P\times J}$, respectively. As a main property, the eigenvalue associated to each principal axes is equal to the variance of the principal coordinates in that axis, for both sequences and residue positions. In this space, the Euclidean distance between any pair of sequences, based on the principal coordinates, and between any pair of residue positions accounts for their Chi-squared distances. Chi-squared distances are suitable to compare individual elements described by means of qualitative variables. Furthermore, sequences versus positions can be compared in this space through the so-called pseudovaricentric relationships, whereby the center of masses of the standard coordinates of any set of sequences is equivalent to the principal coordinates of a residue position perfectly matching the pattern of those sequences' presence/absence. Hence, a set of sequences can be translated into the principal coordinates of the residue positions, which is directly comparable in terms of Euclidean distances (in turn accounting for the chi-squared distances).

B.6. Selection of the number of informative principal axes.

Once the total variability within the initial MSA has been decomposed into orthogonal sources of variation and ranked in decreasing order of importance, the next step is to evaluate how many of them can be considered relevant. This step is achieved on the basis of a nonparametric Wilcoxon rank sum test (6). The goal is to assess the statistical confidence of the increase of information when considering one more dimension. The first I axes to be considered are calculated as

$$I = \left\{\begin{matrix}\max i \in [2,J]/\text{Prob}_{\text{Wilcoxon}}\left(\left[\sum_{j=1}^{i-1}\left(\theta^{\text{ppal}}_{n\times j}\right)^2, \right.\right. \\ \left.\left. \forall\, n \le N\right] < \left[\sum_{j=1}^{i}\left(\theta^{\text{ppal}}_{n\times j}\right)^2, \forall\, n \le N\right]\right) < 0.01 \\ 1\end{matrix}\right\}.$$

For practical reasons, in this work we limited "I" to be less than or equal to "10". The rationale for comparing both previous distributions is based on the relationship between the variability explained in a space defined by I axes and the sum of its associated eigenvalues, where

$$\sum_{j=1}^{I} \lambda_j^{\text{adj}} = \sum_{n=1}^{N}\sum_{j=1}^{I} (\theta^{\text{ppal}}_{n\times j})^2.$$

This means of selecting the optimal number of dimensions can be viewed as a way to strike the most favorable balance between the relevant and the noisy information.

B.7. Clustering of the projected sequences in the selected axes to define protein subfamilies.

To obtain an automatic clustering of sequences in the selected dimensions, we used the k-means algorithm implemented in ref. 7. We performed the clustering of the coordinates coded in $\Theta^{\text{ppal}}_{N\times J}$, which was run for a prespecified number "g" of groups using the following parameters: dist $= e$; method $= a$; npass $=$ max $(N \times 10;\ 500)$. We considered a solution for a given number "g" of groups only if the clustering solution was found at least 5% of the times that the algorithm was run (ifound $\geq$ npass $* 0.05$). The optimal solution was that which maximized the CH_{index} (8). The CH_{index} measures the ratio of the average intercluster simple deviations over the average intracluster simple deviations. The optimal clustering and the corresponding number G of groups are selected as follows:

$$G = g \in \left[2, \min\left(\frac{N}{4}; 50\right)\right] / \max_{g} \text{CH}_{\text{index}} = \\ = \frac{\sum_{i=1}^{g} d(M_i, M_{\text{total}})/(g-1)}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} d(M_i, s_i^j)/(N-1)},$$

where g is the number of clusters, N is the total number of sequences, n_i is the number of sequences in the cluster i, M_i is the center of masses of the cluster i, M_{total} is the center of masses of all the sequences (which upon MCA will be the origin of the coordinates "0"), s_i^j is the sequence j of the cluster i, and $d(\,,\,)$ accounts for the Euclidean distance of the coordinates coded in $\Theta^{\text{ppal}}_{N\times J}$. When $N/4$ is not an exact division, it is rounded to the higher integer. The optimal G clusters of sequences were taken as the subfamily composition within the MSA.

B.8. Assignment of residues to protein subfamilies.

Each subfamily within the MSA can be regarded as a feature for which the N sequences within the MSA are evaluated in binary terms, that is, "1" if sequence n belongs to that subfamily or "0" otherwise. This coding leads to the general binary matrix $\Phi_{N\times C}$ previously defined, where the C additional features are built from each of the ways that a number of γ groups can be taken together from among the G number of optimal clusters, regardless of order, that is, $\binom{G}{\gamma}$ from $\gamma = 1$ to $\gamma = G$. Therefore, the total number of additional features C will be

$$C = \sum_{\gamma=1}^{\gamma=G} \binom{G}{\gamma} = \sum_{\gamma=1}^{\gamma=G} \frac{G!}{(G-\gamma)!\gamma!}.$$

As explained before, once $\Phi_{N\times C}$ has been defined, the principal coordinates of each feature c on the principal axis j are given by $\Omega_{C\times J}$, and these coordinates can be directly compared through Euclidean distances against the principal coordinates $\Psi^{\text{ppal}}_{P\times J}$ of the residue positions in the same axis. To partition the P residue positions according to the C groupings of sequences coded in $\Phi_{N\times C}$, we assigned each residue position p to the nearest grouping c, that is,

$$p \in c \Leftrightarrow \min_{c} d(\overline{\omega_p}, \overline{\psi_c^{\text{ppal}}}).$$

B.9. SDPs definition.

The set of residue positions p associated to c were then ranked in increasing order of distance, which provides a ranking of the fitness of the positions to the assigned feature. Residue positions p accounting for gaps at this point were disregarded from the analysis, and they were not considered in the rankings. For each grouping c, we considered a subset c' of the best ranked positions p' defined by a threshold $\beta \in [0, 1]$ as a fraction of the total set, as follows:

$$c' \subset c / \forall\, c'' \subset c' / |c''| = \beta \cdot |c| :\ \max \text{rank}(\forall\, p'' \in c'') = \\ = \max \text{rank}(\forall\, p' \in c') < \min \text{rank}(\forall\, p \in \overline{c'}).$$

When $\beta \cdot |c|$ was not an exact division, it was rounded to the higher integer.

Let $\Delta^{\text{ppal}}_{L\times C}$ be a binary matrix with general term δ_{lc} as follows:

$$\delta_{lc} \left\{ \begin{array}{l} \delta_{lc} = \text{rank}(p) \ \Leftrightarrow\ p \in c' \subset c \wedge p \in l \\ \delta_{jc} = 0 \end{array} \right\}.$$

We defined SDPs as those "l" whose residues have an average rank equal to or less than α upon a complete disjunctive partition of the N sequences within the MSA, that is,

$$l \in \text{SDPs} \Leftrightarrow \left\{ \begin{array}{l} \forall\, k \in C / \delta_{lk} > 0 :\ k \in K \wedge \sum_{k} \overline{\varphi_k} = \vec{1} \wedge \frac{\sum_{k}^{\delta_{lk}}}{|K|} \leq \alpha \\ \forall\, i,j \in C / \delta_{li} > 0 \wedge \delta_{lj} > 0 \wedge i \neq j :\ \overline{\varphi_i} \cdot \overline{\varphi_j} = 0 \end{array} \right\}.$$

In this work, we established $\alpha = 10$ and disregarded any $c \in C / |\overline{\varphi_c}| \leq 3$ with reference to both its associated residue positions c' and the definition of a complete disjunctive partition.

C. Additional details concerning the analysis of the functional organization of protein subfamilies.

C.1. EC set.

We defined classes of proteins within a MSA as those sharing the first two digits of the EC code (e.g., EC 2.11./EC 2.10./EC 1.4. are all different classes), while "groups" were defined as the sets within classes sharing the four EC digits (EC 2.11.7.5/EC 2.11.7.8/EC 2.11.3.1 are all different groups within the same class). Pfam families with partially overlapping EC classes or partially overlapping EC groups within a class were disregarded. Only one reference EC class was chosen for a Pfam case: the one with the largest number of sequences and then that with the most groups. For the analysis we considered those families fulfilling the following requirements: (*i*) more than one "group with at least three sequences"; (*ii*) at least 25% of the sequences should be included in an EC group; and (*iii*) no group contains more than 80% of the sequences with a EC group assigned.

C.2. Interaction set.

Pairs of interacting yeast and human proteins were taken from the Database of Interacting Proteins core datasets of "small scale" experiments (9). We extracted "manually drawn" pathways for yeast and humans from the Kyoto Encyclopedia of Genes and Genomes (10) pathway database. BioGRID (release 2.0.49) (11) was used as a general repository for interaction datasets in yeast and humans that considers high-throughput experiments. Experimentally determined subcellular localization was extracted from the MIPS database (12) (on 14-11-2005) for yeast, and from the eukaryotic subcellular localization database (eSLDB) for yeast and humans (13) (on 21-1-2009). The subcellular locations considered and their mapping between fields for both databases, were as shown in Table S1.

D. Structural definition of ligand binding sites and protein-protein binding sites. All the sequences within a MSA for which structural information was available were aligned against their PDB crystal structures [as detected by the Macromolecular Structure

Database (14)]. We only considered PDBs stemming from x-ray diffraction experiments with a resolution below 3 Å and fewer than 10 missing atoms. Blast2Seq (15) was used to obtain the corresponding pairwise alignments.

From these structures, those with a structural domain assigned by the Structural Classification of Proteins (SCOP) (16), with at least an 80% alignment overlap with the corresponding defined Pfam domain, were selected.

Functional residues from these structures were gathered as follows:

- Ligand binding residues were retrieved from FireDB (17).
- Protein-protein binding sites were defined following the approach of Valdar and Thornton (18), implemented in http://www.biochem.ucl.ac.uk/bsm/valdarprograms/pdb_defineface.html. To this end, we used the biological units provided by the Resource for Studying Biological Macromolecules PDB (19) (on 03-12-2008; ftp://ftp.rcsb.org/pub/pdb/data/biounit/coordinates) to eliminate the crystallographic interfaces. BioUnit PDB files were filtered through the program implemented at http://www.biochem.ucl.ac.uk/bsm/valdarprograms/pdb_fixcoord.html, before using pdb_defineface under the following parameters: *–rm_noalpha –het_nonaminos –unhet_aminos*. Homomeric and heteromeric interaction interfaces were assessed in a pairwise manner; for instance: Given a BioUnit PDB file containing three chains A, B, and C, the interaction interface of A was assessed as the sum of the interaction interfaces of A with B and A with C. Only Pfam residues within SCOP limits were considered to define interaction interfaces, unless the length of the PDB chain outside of those limits was shorter than 15 residues.

The functional residues gathered (binding site and interaction interface) were then mapped onto one of these structures, which was selected as the optimal representative of the whole family. This selection was performed through the infrastructure implemented in FireDB. In FireDB, all PDB chains are clustered at 97% of sequence identity, each group being represented by a so-called "master sequence." A master sequence was first selected as the family representative because it maximizes the following parameters:

1. The total sum of "% SDPs" plus "% of binding site residues along the PDBs and within the alignment" it covers;
2. alignment coverage (disregarding gappy columns at 10% level);
3. % sequence identity with the corresponding Pfam sequence.

Then, a PDB chain from the "representative master sequence" was selected as the "representative PDB chain" for the family, which maximized the following parameters:

1. The total sum of "% SDPs" plus "% of binding site residues along the PDBs and within the alignment" it covers;
2. x-ray diffraction resolution.

E. Analysis of protein subfamilies and SDPs identified by several methods. In this section, we aim to further support the generality of the conclusions drawn in this work. To do so, we reproduce the key analyses of our study using four state-of-the-art methods with an accessible stand-alone software: evolutionary trace (ET) (20), combinatorial entropy optimization (CEO) (21), mutational behavior (MB) (22), and PCA (23). In PCA, the MSA decomposition reported by Casari et al. is followed by the same protocol described above (sections B.6–B.9) in order to automatize the detection of protein subfamilies and SDPs.

The previous methods are all sequence-based (they use a MSA as an input and do not require structural information) and unsupervised (they do not require an external functional classification). Table S4 summarizes their main characteristics. Such methods are representative of the major approaches in the field. We applied all these methods to the same collection of Pfam alignments described in *Methods* (*Dataset of protein families*).

Fig. S2 shows the correspondence between subfamilies, EC groups, and specific interactors. Results for PCA and CEO methods are represented in the receiver operating characteristic (ROC) space (see *Analysis of the functional organization of protein subfamilies* in *Methods*). The comparison of these plots to those corresponding to MCA (Fig. 1) supports the global tendency of protein subfamilies to agree with both EC and interaction-based classifications. Notwithstanding, minor differences among the performances of the methods can be observed: CEO tends to be more specific than MCA and PCA at the expense of being less sensitive. Interestingly, considered as a whole, MCA performs qualitatively better according to both functional classifications.

Additionally, a functional enrichment analysis for the SDPs produced by the different methods was performed, in the same way as described in *Methods* (*Enrichment tests*). The stand-alone programs of ET and CEO did not provide a threshold to define a set of SDPs but a ranking of all the positions within the MSA. In order to make them comparable, we selected their first n ranked positions for each family, where n is the respective average number of SDPs predicted by the MCA, PCA, and MB methods. This criterion is intended to retrieve a comparable number of predictions for the methods, although it might not be the optimal selection for any of the individual methods.

Table S3 shows the functional enrichment p values obtained for the positions produced by the different methods. The general association between SDPs and both ligand-binding sites and interfaces is reproduced by all the methods. The procedure used to retrieve a comparable number of SDPs (see above) might be causing the poor performance of CEO in this analysis. Further study of CEO performance is, however, beyond the scope of this work.

ET results are highly enriched in all types of functional residues, to a level similar to that of conserved positions. These results are coherent with the more comprehensive definition of "evolutionary relevant positions" considered by ET. Indeed, 76% of the positions within the ET set were conserved positions, in contrast to the less than 1% for the other methods. This overlap led us to generate a new set of ET predictions (referred as ET-ConsFree in Table S3) by filtering out conserved positions from its rankings. In this case, results become gradually more similar to those of the other methods.

Taken together, the figures shown by PCA, MB, and ET-ConsFree support the general association between SDPs and both ligand-binding sites and interfaces. Furthermore, the observed general trends point to a consistent association of SDPs to heteromeric interfaces, whereas for the homomeric and intra-protein interfaces this association remains inconclusive.

Importantly for this study, the results obtained with our approach are qualitatively similar to those produced by the other methods. Moreover, the capacity of MCA to produce a simultaneous classification of subfamilies and residues is particularly adequate for the proposed analysis.

1. Uniprot Consortium (2009) The Universal Protein Resource (UniProt) 2009 *Nucleic Acids Res* 37:D169–174.
2. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
3. Greenacre MJ (1984) *Theory and Application of Correspondence Analysis* (Academic, London).
4. Lebart L, Morineau A, Warwick KM (1984) *Multivariate Descriptive Statistical Analysis* (Wiley, New York).
5. Greenacre M & Blasius J (2006) *Multiple Correspondence Analysis and Related Methods* (Springer, Berlin).
6. Miller I, Miller M (1998) *Mathematical Statistics* (Prentice Hall International, London).

7. de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20:1453–1454.
8. Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat* 3:1–27.
9. Xenarios I, et al. (2002) DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 30:303–305.
10. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32: D277–280.
11. Stark C, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34:D535–539.
12. Mewes H, Albermann K, Heumann K, Liebl S, Pfeiffer F (1997) MIPS: A database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* 25:28–30.
13. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2007) eSLDB: Eukaryotic subcellular localization database. *Nucleic Acids Res* 35: D208–212.
14. Velankar S, et al. (2005) E-MSD: An integrated data resource for bioinformatics. *Nucleic Acids Res* 33:D262–265.
15. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
16. Andreeva A, et al. (2004) SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res* 32:D226-229.
17. Lopez G, Valencia A, Tress M (2007) FireDB—A database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35:D219–223.
18. Valdar WS, Thornton JM (2001) Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins* 42:108–124.
19. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
20. Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 336:1265–1282.
21. Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8:R232.
22. del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues *J Mol Biol* 326:1289–1302.
23. Casari G, Sander C, Valencia A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol* 2:171–178.
24. Dessimoz C, Boeckmann B, Roth AC, Gonnet GH (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res* 34(11):3309–3316.

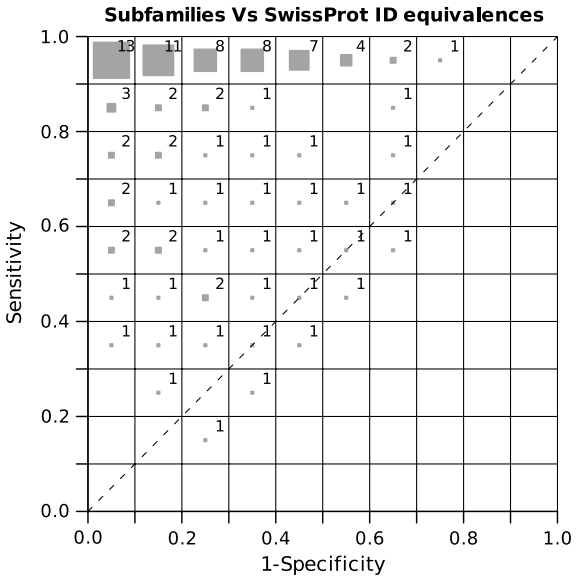


Fig. S1. Correspondence between the subfamilies and the SwissProt ID equivalences represented in the ROC space, where the distribution of the families is shown as a bidimensional histogram. The size of the colored boxes in each bin of the ROC space represents the percentage of protein families they contain, whereas the number shows the actual percentage. For the sake of simplicity, percentage values are rounded to the nearest integer (so that they may not add up to 100). For the analysis we considered proteins of the same ID group as those for which the first part of the ID coincides (i.e., OAT_HUMAN, OAT_MOUSE, OAT_YEAST, etc.). These labels correspond to abbreviations of the protein/gene names, and they are the same for the orthologous sequences according to SwissProt. The set analyzed contains 799 Pfam families fulfilling the following requirements: (*i*) more than one ID group with at least three sequences; (*ii*) at least 25% of the sequences are included in one group; and (*iii*) no group contains more than 80% of the sequences with an ID group assigned. This analysis aims to complement those performed through the EC and Interaction labels. The idea is to exploit the functional information implicit in the SwissProt ID names, as previously performed in other studies (e.g., ref. 24). The plot shows the consistency of the subfamily classification for a large set of families in the context of functionally coherent sets of orthologues. This consistency can be observed in the tendency of subfamilies to gather a variable number of ID groups without splitting them (*Upper*).

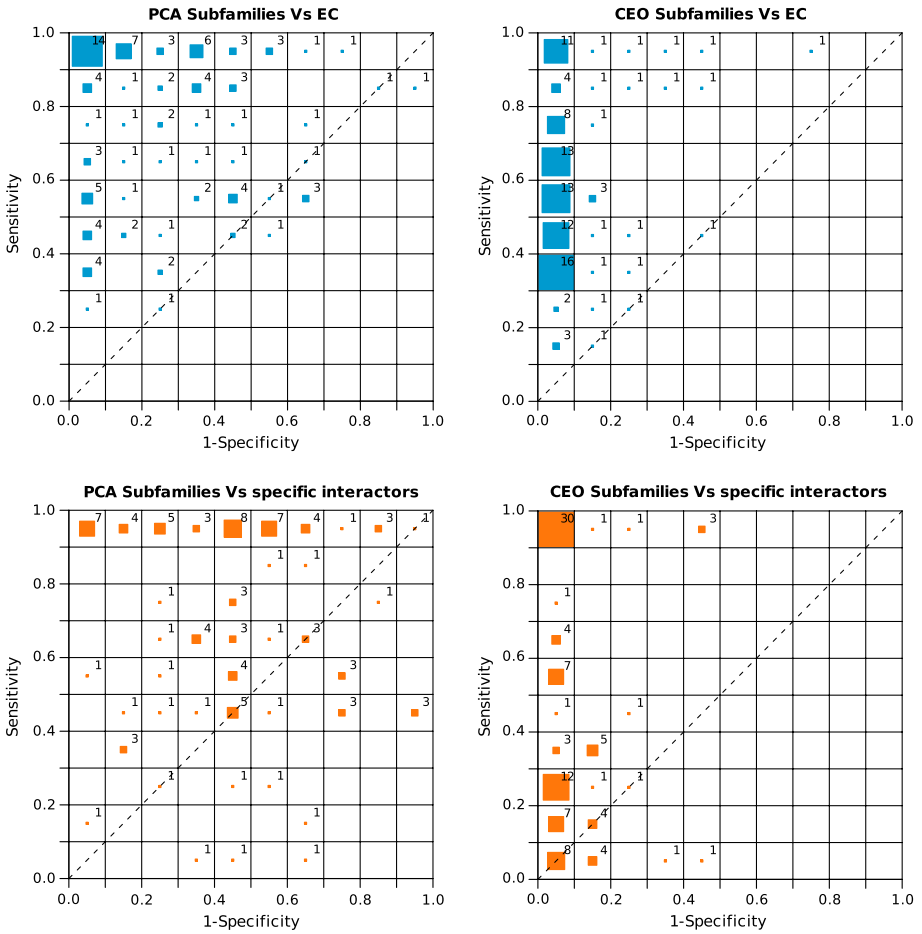


Fig. S2. Correspondence between the different subfamilies obtained by PCA (*Left*) and CEO (*Right*), and the EC groups (*Top*) and specific interactors (*Bottom*) for each protein family represented in the ROC space, where the distribution of the families is shown as a bidimensional histogram. The size of the colored boxes in each bin of the ROC space represents the percentage of protein families they contain (up to a maximum size representing 15%), whereas the number shows the actual percentage. For the sake of simplicity, percentage values are rounded to the nearest integer (so that they may not add up to 100).

	Q14186	Q14188	Q01094	Q14209	O00716	Q16254	Q15329	Q13547	P35232	P38398	P19838	P20226	P28749	P06400
Q14186			+	+		+	+		-	-	-	-	+	+
Q14188			+	+	+	+			-				+	+
Q01094	+	+				-	-	+	+	+	+	+		+
Q14209	+	+				-			-	-		-		+
O00716	+					-	-		-	-		-		+
Q16254	+	+	-	-	-		-	+			-	-	+	+

Swiss AC	Swiss ID	Name
Q14186	TFDP1_HUMAN	Transcription factor Dp-1
Q14188	TFDP2_HUMAN	Transcription factor Dp-2
Q01094	E2F1_HUMAN	Transcription factor E2F1
Q14209	E2F2_HUMAN	Transcription factor E2F2
O00716	E2F3_HUMAN	Transcription factor E2F3
Q16254	E2F4_HUMAN	Transcription factor E2F4
Q15329	E2F5_HUMAN	Transcription factor E2F5
Q13547	HDAC1_HUMAN	Histone deacetylase 1
P35232	PHB_HUMAN	Prohibitin
P38398	BRCA1_HUMAN	Breast cancer type 1 susceptibility protein
P19838	NFKB1_HUMAN	Nuclear factor NF-kappa-B p105 subunit
P20226	TBP_HUMAN	TATA-box-binding protein
P28749	RBL1_HUMAN	Retinoblastoma-like protein 1
P06400	RB_HUMAN	Retinoblastoma-associated protein

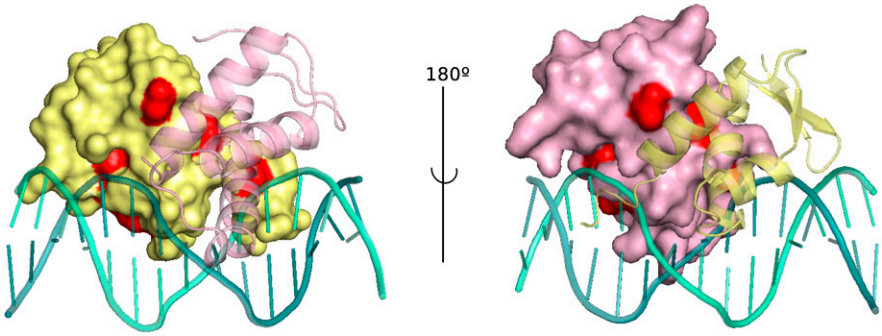


Fig. S3. Results for the E2F/TDP family of transcription factors (PF02319). Top table: Full list of the positive and negative interactions extracted for this family (see *Methods*). Middle table: Swissprot IDs and names of the interacting proteins. Bottom panel: SDPs (*Red Surface*) mapped to the heterodimeric structure of human E2F4 (*Yellow*) - DP2 (*Pink*) bound to DNA.

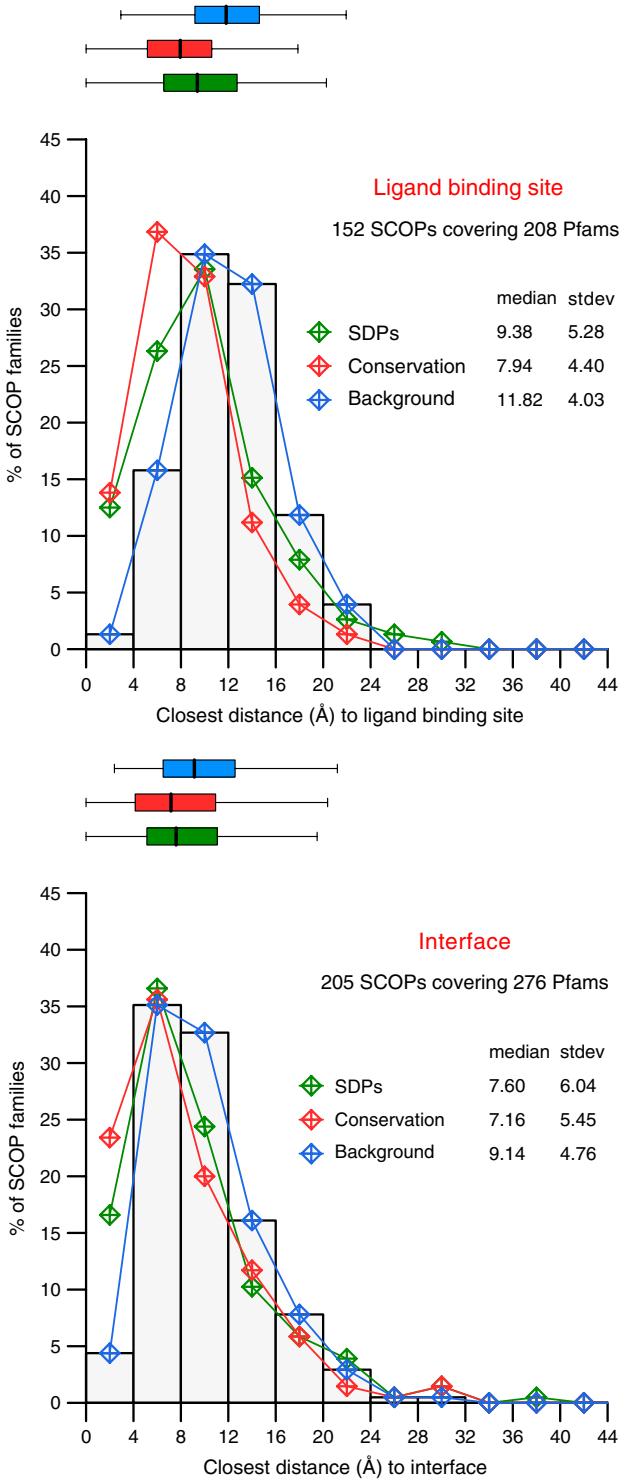


Fig. S4. Distribution of the C_β-C_β atom distances from SDPs (*Green*), from conserved positions (*Red*), and in the background (*Blue*) to the ligand binding sites (*Upper*) and interaction surfaces (*Lower*), as averaged per family and per structurally redundant group (see *Methods*).

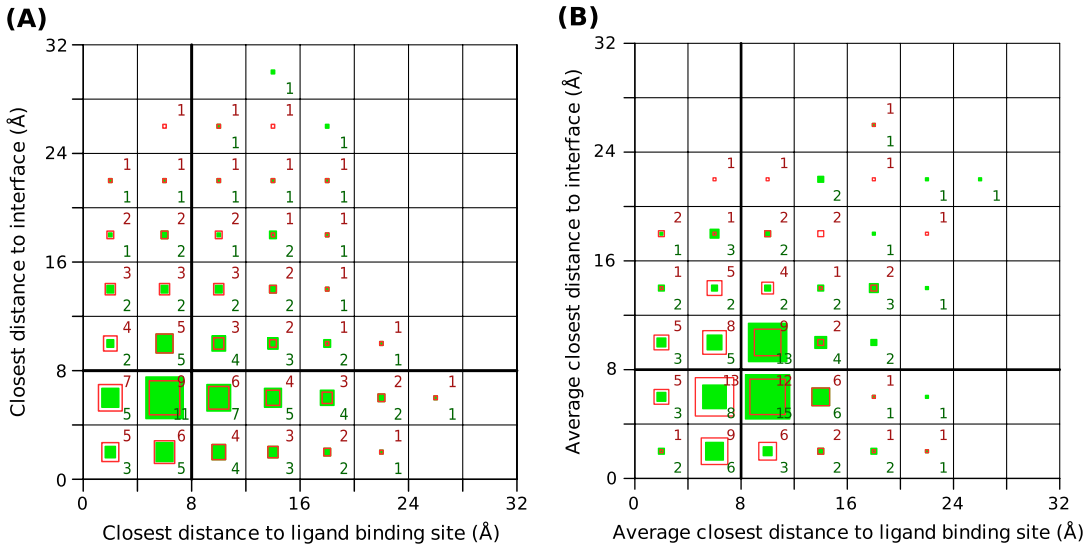
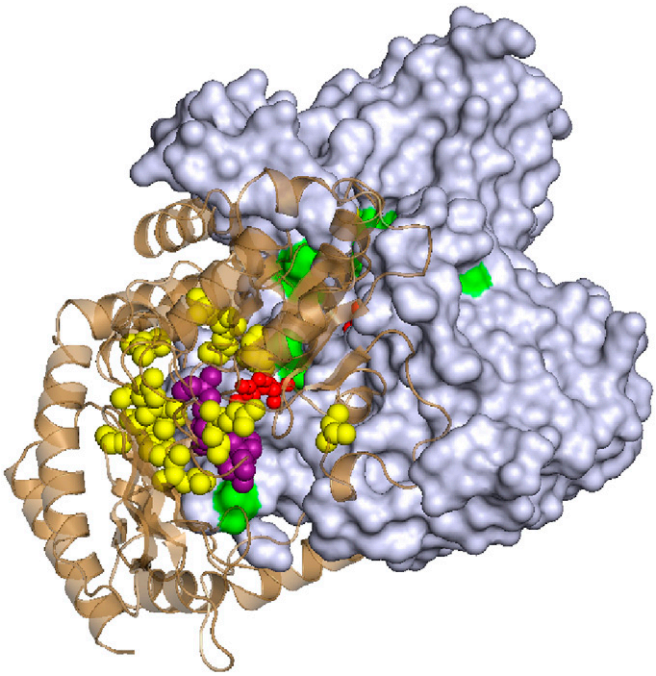


Fig. S5. Bidimensional histograms representing the joint distribution of the closest distances from the SDPs (*Green*) and the conserved positions (*Red*) to ligand binding sites (*x* axis) and the interfaces (*y* axis). (*A*) Joint distribution of the distances in the 168 Pfam families that contain both types of functional regions. The size of the colored boxes represents the percentage of SDPs and conserved positions of each bin in the histogram, whereas the number shows the actual percentage. (*B*) Joint distribution where distances in (*A*) have been averaged per family and per structurally redundant group. The size of the colored boxes represents the percentage of structurally nonredundant groups of each bin in the histogram, whereas the number shows the actual percentage. For the sake of simplicity, percentage values are rounded to the nearest integer in both histograms (so that they may not add up to 100).

Only SITE	OVERLAP	Only INTE			
+	+	+	9.1%	19.4%	41.2%
+	+	-	1.8%		
-	+	+	4.8%		
-	+	-	3.6%		
+	-	+	21.8%		
+	-	-	15.8%		
-	-	+	23.6%		
-	-	-	19.4%		
48.5%	19.4%	59.4%			
57.0%					
	64.8%				
80.6%					

Fig. S6. Conditional cumulative percentages of families in which at least one SDP was part of (*i*) the ligand binding site, (*ii*) the protein interaction region, or (*iii*) both (functional overlap). Percentages are assessed for the 168 Pfam families containing both types of regions.



Movie S1. The homodimeric structure of the human ornithine aminotransferase (PDB 1oat), a class III aminotransferase, bound to Pyridoxal-5′-phosphate (*Red Spheres*). The two subunits of the complex are shown as brown cartoons and gray surface. SDPs are highlighted in a yellow/violet space fill and with a green surface, respectively.

Movie S1 (AVI)

Table S1. Equivalences between MIPS and eSLDB subcellular localization categories

MIPS field	eSLDB field
701 extracellular	Extracellular
705 bud	(no equivalent category)
710 cell wall	Cell wall
715 cell periphery	(no equivalent category)
720 plasma membrane	(no equivalent category)
725 cytoplasm	Cytoplasm
730 cytoskeleton	Cytoskeleton
735 ER	Endoplasmic reticulum
740 golgi	Golgi
745 transport vesicles	Vesicles
750 nucleus	Nucleus
755 mitochondria	Mitochondrion
760 peroxisome	Peroxisome
765 endosome	Endosome
770 vacuole	Vacuole
775 microsomes	Lysosome

Table S2. Datasets of protein families analyzed according to the functional and structural information gathered

Total number of Pfam families			1262
Within EC set		149	
Within Interaction set		72	
Within SCOP set AND (ligand binding site OR interface)		316 (230)	
With ligand binding site AND interface		168 (127)	
With ligand binding site		208 (152)	
With interface		276 (205)	
With Intra interface	170 (130)		
With Homo interface	171 (137)		
With Hetero interface	87 (73)		

Numbers represent the number of Pfam families in each subset with the corresponding number of nonredundant SCOP families in parentheses.

Table S3. Results of the Wilcoxon rank sum tests evaluating the "enrichment" at the annotated sites of the SDPs retrieved by different methodologies

	Site & total interface	Site	Total interface	Hetero	Homo	Intra
p value						
Cons	1.92E-27	4.73E-20	1.92E-09	5.32E-02	9.09E-03	4.29E-07
ET	5.55E-26	2.85E-22	4.87E-11	2.86E-04	1.79E-02	3.59E-07
ET-Consfree	7.17E-15	7.04E-13	4.85E-06	8.24E-04	7.09E-02	3.23E-03
MCA	1.67E-05	4.25E-04	1.89E-02	1.75E-02	5.13E-01	4.99E-01
PCA	3.91E-05	3.82E-03	2.59E-02	7.50E-02	5.66E-01	4.51E-01
MB	5.59E-06	1.63E-02	3.03E-04	1.79E-01	1.40E-01	4.32E-02
CEO	8.27E-02	7.64E-01	1.64E-01	3.05E-01	5.02E-01	4.32E-01
Median difference						
Cons	15.25%	14.60%	6.34%	1.84%	1.99%	6.76%
ET	13.49%	13.72%	6.12%	4.82%	1.69%	5.56%
ET-Consfree	8.54%	7.94%	3.74%	4.08%	1.21%	2.30%
MCA	5.32%	3.97%	2.00%	4.77%	-0.01%	0.00%
PCA	4.48%	2.80%	1.65%	1.94%	-0.09%	0.17%
MB	3.62%	1.73%	2.54%	1.33%	0.81%	1.37%
CEO	1.29%	-0.56%	0.85%	0.66%	0.00%	-0.71%

Table S4. Main features of the methods used to retrieve protein subfamilies and SDPs.

Method	Major methodological category	Subfamily detection	Preestablished threshold to define SDPs	Reference
MCA	Multivariate analysis	Yes	Yes	This work
PCA	Multivariate analysis	Yes	Yes	(23); this work
MB	Positional correlation with the global family variability	No	Yes	(22)
ET	Tree-guided & entropy-based	No	No	(20)
CEO	Entropy-based	Yes	No	(21)