Supplementary information S1 (boxes): Technical information on co-evolution methods

Box 1 –Methods to predict interactions between residues based on the similarity of their patterns of variation.

a.1 Correlated mutations

Correlated mutations are usually computed as the Pearson's correlation coefficient between the patterns of amino acid similarity of two positions in a MSA (often based on Maxhom-McLachlan[1]).

a.2 CAPS

The probability of including more drastic amino acid changes depends on the time of divergence. CAPS[2] corrects the values of the amino acid substitution matrix (BLOSUM) considering the corresponding time of sequence divergence estimated as the ratio of synonymous nucleotide substitutions.

Correlation of the within position variability of these values between positions is used to measure co-variation and the corresponding Z-score is calculated. This analysis is repeated after the removal of every well-defined clade in the protein tree. Only positions that co-vary consistently in the different sub-alignments are considered not to be caused by the phylogenetic background.

a.3 Mutual information

In the simplest case[3], coupling between pairs of positions in a MSA is estimated according to their mutual information, calculated as the Shannon's entropy of a position given the other position in the alignment. In this case, the coupling is based on the similarity in the presence/absence of the distinct residues in the corresponding positions of each sequence.

a.4 Mutual Information removing the inter-sequence divergence background

Mutual information tends to assign high values for pairs of amino acids that vary according to the phylogenetic background, providing an uncontrolled mixture of inter-residue couplings, residues specific to subfamilies (SDPs) and spurious results.

MIp[4] is a good example of the general idea of correcting for background divergence, although it has been implemented in other forms. In this case the background of inter-sequence divergence is calculated as the average product correction (APC):

$$APC(a,b) = \frac{MI(a,\bar{x})\,MI(b,\bar{x})}{\overline{MI}},$$

where $MI(a,\bar{x})$ (or $MI(b,\bar{x})$) is the average mutual information between position a (or b) with respect to the rest of the positions in the alignment, and $\overline{MI}$ is the average mutual information between all the positions.

The corrected mutual information value (*MIp*) is calculated as:

$$MIp = MI(a,b) - APC(a,b)$$

Interestingly, this correction also corrects the natural tendency of mutual information to penalize co-variation of low entropic (*i.e.* conserved) positions[4].

a.5 Direct Coupling Analysis (DCA)

Previous methods consider that co-variation of a pair of residues is independent of the other positions in the alignment. However, inter-position co-variation is at least partially determined by the co-variation with other residues. DCA[5] deals with this situation by estimating a global statistical model of the alignment accounting for inter-position coupling and position-based amino acid biases.

$$P(A_1,...,A_N) = \frac{1}{Z}\exp\left\{-\sum_{i<j} e_{ij}(A_i,A_j) + \sum_i h_i(A_i)\right\}$$

In this case, $e_{ij}(A_i,A_j)$ is the inter-residue direct coupling between amino acid A_i at position i and amino acid A_j at position j, and where $h_i(A_i)$ is the local bias of the position i for the amino acid A_i.

This optimization problem was originally solved using a message-passing approach[5] and more recently, using a much more efficient analytical heuristic approach[6]. The resulting model contains the set of direct statistical couplings used to calculate the Direct Information (DI) measure, which is the mutual information from the two positions induced by the estimated direct coupling.

Box 2 – Methods to detect sites of binding specificity based on the simultaneous variation of groups of positions in protein families.

b.1 EvolutionaryTrace (ET)

In its first implementation[7], ET generated clusters of sequences with a minimum identity between them, and then SDPs were defined as positions containing different invariant amino acids for the different clusters. These SDPs are ranked according to the number of clusters needed to define them, and in this way, the highest ranked positions correspond to the fully conserved ones that cannot be considered as true SDPs.

In the current implementation[8], rvET (real value ET), the splits are based on the nodes of a pre-calculated tree, and the condition of complete invariability has been soften by including the value of the weighted added entropy to all the possible groups defined according to the tree. Thus, rvET assigns empirical (or manual) weights according to the node and the groups in order to provide additional flexibility to the analysis, and typically to focus on those conservations implying the master protein.

b.2 Mutual Information

Mutual information (see Box1) has been also used for detecting SDPs[9,10]. In these methodologies, the distributions of the different amino acids are compared to the distribution of the respective sequences within a pre-defined subfamily classification. These classifications can be based on groups of orthologues in a protein family[9] or, in a similar way to ET[7] (see above), obtained from tree-based partitions[10]. In both cases, the significance of a given SDP can be defined from the Z-score obtained from the distribution of mutual information values.

b.3 Mutational Behaviour (MB)

Mutational Behaviour[11] calculates the Pearson's correlation coefficient between the patterns of amino acid similarities of every position and the sequence distance matrix

of the MSA. In this case, sequence distances are calculated as the average value of the amino acid similarities.

b.4 SequenceSpace

SequenceSpace[12] performs a PCA of the MSA. In this case, the MSA is coded as a binary matrix where each column represents the presence (1) or absence (0) of each of the amino acids (and the gap) at every position.

PCA results in an orthogonal decomposition of the sequences against the residue-position association (e.g. Alanine in position 20). The obtained principal components are ordered according to the total variance explained, with the first component being the most informative. The results of the PCA can be used to represent sequences and residue positions on equivalent spaces of a small number of informative dimensions. In this way, the directions of the observed clusters of sequences (subfamilies) allow the representative residue-positions populating equivalent regions of their vectorial space to be identified. In its original implementation, SequenceSpace requires human manipulation and visual inspection to select the axes, subfamilies and representative SDPs.

b.5 S3det

S3det[13] follows a similar philosophy to SequenceSpace. It performs a Multiple Correspondence Analysis (MCA) of the MSA to obtain the MCA-derived "principal axes" that are more appropriate to deal with categorical data, such as patterns of amino acids. The distances in spaces defined by these "principal axes" account for the Chi-squared distances and they are suitable for the binary encoding of the MSA. Furthermore, S3det can represent sequences and residues-positions simultaneously in the equivalent spaces defined by the *n* first "principal axes", without the need for additional manipulations.

S3det performs successive Wilcoxon rank sum tests to assess the inclusion of additional dimensions, and a blind clustering protocol to determine the optimal subfamily definition as the most robust solutions from a large number of *k-means* runs

to cluster the space of the sequences. Then SDPs are automatically detected as those whose residue-position vectors are the closest to a subset vectors defining a completely disjunctive partition of the alignment. These SDPs are conserved within all subfamilies and they are different between at least two groups of them.

Box 3 – Methods to detect networks of interacting residues based on concomitant evolutionary constraints.

c.1 SCAold

The original implementation of Statistical Coupling Analysis[14] (SCAold) was proposed to analyse statistical evolutionary interactions in a thermodynamic framework. In this context, the conservation of a site was defined as

$$\Delta G_i^{stat} = kT * \sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$$

where kT* is an arbitrary constant, P_i^x is the probability (frequency) of an amino acid x at position i, and P_{MSA}^x in the whole MSA. In this framework, the statistical dependence of the amino acid distribution at position i of a given perturbation dj at the position j can be calculated as

$$\Delta\Delta G_{i,j}^{stat} = kT * \sqrt{\sum_x \left(\ln \frac{P_{i|\delta j}^x}{P_{MSA|\delta j}^x} - \ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$$

where $P_{i|\delta j}^x$ and $P_{MSA|\delta j}^x$ correspond to the probability of finding the residue x at position i, conditioned by the perturbation dj at position j. In practical terms a perturbation at j is defined as the presence of a given amino acid in this position.

This perturbation-based framework allows the co-evolution between two positions to be decomposed into the parallel patterns of conservation associated to each amino acid. In this sense, integration of all the perturbations would lead to a correlated mutations-like scenario. However, this method originally focused on conserved amino acids at key positions known to be responsible for the functional specificity of the families studied[14,15]. This scenario is related to the detection of SDPs, where the subfamily definition is used to explore functional specificity-associated conservation.

c.2 SCAnew

SCAnew[16] calculates a conservation-weighted correlation matrix that corresponds to the co-evolution between every pair of positions. This weighted correlation highlights the correlations between conserved residues, part of the rationale behind SCAold. This matrix is calculated as

$$\tilde{C}_{ij}^{(ab)} = \phi_i \phi_j C_{ij}^{(ab)}$$

where ϕ_i and ϕ_j correspond to the gradient of relative entropy of the most prevalent amino acid at positions i and j with respect to a background distribution obtained from a non-redundant sequence database, and $C_{ij}^{(ab)}$ is the co-variance matrix.

Spectral decomposition of the weighted-correlation matrix is then carried out and this can be considered as a conservation-biased PCA that is similarly used to define a space of "principal axes". SCAnew randomizes the MSA in order to determine which dimensions are significantly informative. In addition, it removes the first dimension as the one likely to most clearly reflect the phylogenetic signal. As a result, groups of co-evolving residues or "protein sectors" are detected as those with a long projection on one of these informative "principal axes".

SCAnew only uses the *space of residues*, although this space is actually connected with an equivalent *space of sequences* comparable to those used in SequenceSpace and S3det. As these spaces tend to separate different subfamilies, longer projections on the axes tend to reflect SDPs for a given subfamily (or combinations of them). Thus SCAnew, at least conceptually, can be considered to detect groups of residues specifically conserved in the same subfamily (or combinations of them), a definition that clearly overlaps that of SDPs. However, as no direct comparison of these methods is available, it is hard to define the actual level of overlap between them in practical terms.

Box 4 –Methods to identify protein interactions based on the parallel evolution of their sequences.

d.1 Mirrortree.

The original *mirrortree* approach[17] quantifies the similarity between two phylogenetic trees as the Pearson's correlation coefficient between the corresponding inter-leaf distances. These distances can be obtained directly from the pairwise comparison of the sequences of orthologues aligned in the MSA or extracted from the previously constructed corresponding phylogenetic trees.

d.2 Tol-mirrortree

Phylogenetic background sequence divergence due to species evolution can detect spurious tree similarities. Tol-mirrortree[18] addresses this issue by explicitly subtracting the expected contribution of the sequence drift (estimated from a 16S rRNA tree) from the inter-leaf distances. Then, Tol-mirrortree quantifies the similarity between the two corrected matrixes as the Pearson's correlation coefficient between them.

d.3 ContextMirror

ContextMirror[19] uses mirrortree's output to define the co-evolutionary profile of each protein as a vector containing the correlation values of its tree with every other tree of the proteins in a reference species. The similarities between every pair of co-evolutionary profiles are quantified as its Pearson's correlation coefficient. Partial correlation coefficients of each protein pair given every third protein were calculated. Then for every protein pair the rest of proteins are sorted according to their contribution to the correlation of the pair. In such a way the first element of a list corresponds to the maximum pair correlation unexplained by any third protein. Different levels of pair co-evolutionary specificity are defined by descending in these lists.

d.4 MatrixMatchMaker (MMM).

The starting point for MMM[20] is also a pair of distance matrices for the two families of interest. In this case, the matrices can have different dimensions (a distinct number paralogues). The method looks for the largest sub-matrixes of these two initial matrices representing the two sub-trees of highest similarity. For this, each possible triplet of inter-leaf distances in the first matrix is compared to the second matrix in a search for a triplet with similar distance ratios. Two distance ratios, such as AB/WX and AC/WY, are considered similar if the parameter |(AB x WY) - (WX x AC)| / [(AB x WY) + (WX x AC)] is lower than a pre-defined threshold. Once a pair of compatible triplets is found with this procedure, new leaves are recursively added to them as long as the triplets formed between these leaves and all possible pairs in their corresponding growing sets fulfil the aforementioned criteria. Finally, a global score is calculated that considers the number of coupled sequences and the RMSD of their distances.

d.5 Similarity of phylogenetic profiles.

In the first versions of the "phylogenetic profiling" method[21], the species distribution of a family of orthologues was represented as a binary vector coding the presence (1) or absence (0) of each orthologue of the protein in the corresponding species. Defined in this way, different measures of similarity between binary vectors can be used, such as mutual information or the Hamming distance.

More recent definitions of "phylogenetic profile" encode the sequence similarity between the orthologue in the corresponding species and that of an organism of reference. For example, Date and Marcotte[22] use a sequence similarity measure based on the BLAST E-value. Values of this measure are then segregated into discrete bins, and the similarity between the phylogenetic profiles is quantified as their mutual information.

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

2. Fares, M. A. & Travers, S. A. A. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* **173**, 9–23 (2006).
3. Korber, B. T., Farber, R. M., Wolpert, D. H. & Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7176–7180 (1993).
4. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).
5. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
6. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–301 (2011).
7. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
8. Mihalek, I., Res, I. & Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **336**, 1265–1282 (2004).
9. Mirny, L. A. & Gelfand, M. S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **321**, 7–20 (2002).
10. Kalinina, O. V., Gelfand, M. S. & Russell, R. B. Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics* **10**, 174 (2009).
11. del Sol Mesa, A., Pazos, F. & Valencia, A. Automatic Methods for Predicting Functionally Important Residues. *J. Mol. Biol.* **326**, 1289–1302 (2003).
12. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178 (1995).
13. Rausell, A., Juan, D., Pazos, F. & Valencia, A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1995–2000 (2010).
14. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
15. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59–69 (2003).
16. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).
17. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614 (2001).
18. Pazos, F., Ranea, J. A. G., Juan, D. & Sternberg, M. J. E. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–1015 (2005).
19. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl.*

Acad. Sci. U.S.A. **105**, 934–939 (2008).
20. Tillier, E. R. M. & Charlebois, R. L. The human protein coevolution network. *Genome Res.* **19**, 1861–1871 (2009).
21. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285–4288 (1999).
22. Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062 (2003).