Supplementary Material to
Identifying relevant positions in proteins by Critical Variable Selection

Silvia Grigolon, Silvio Franz, Matteo Marsili

I. STATISTICAL COUPLING ANALYSIS IN A NUTSHELL

We here give an overview of Statistical Coupling Analysis as applied to our datasets.

Let us consider a MSA as an ensemble of N sequences $\bar{s}^{\alpha} = \{s_1^{\alpha}, ..., s_L^{\alpha}\}$ of length L where each s_i^{α} ($\alpha = 1, ..., N$ and $i = 1, ..., L$) represents either an amino acid or a gap or an uncertain letter as well and can then take $q = 21$ values. A first measure of conservation throughout the dataset is given by the frequency of the amino acid a at position i, i.e.:

$$f_i^a \equiv \frac{1}{N} \sum_{\alpha=1}^{N} \delta_{a,s_i^{\alpha}}.$$

Pair - frequencies can be also defined in a straightforward manner, as:

$$f_{ij}^{ab} \equiv \frac{1}{N} \sum_{\alpha=1}^{N} \delta_{a,s_i^{\alpha}} \delta_{b,s_j^{\alpha}},$$

that gives a measure of the simultaneous appearance of the amino acids a and b respectively at positions i and j. The *correlation matrix* C_{ij}^{ab} in such defined model will be then:

$$C_{ij}^{ab} \equiv f_{ij}^{ab} - f_i^a f_j^b, \tag{1}$$

being a $qL \times qL$ matrix. In [1], a further quantity, ϕ_i^a, called *positional information* was introduced. This is aimed at highlighting highly conserved positions with respect to the background amino acids frequencies within the correlation matrix. Let us define the background frequency of the $a-$th amino acid as $\nu^a = \frac{1}{L} \sum_{i=1}^{L} f_i^a$. The bias of a site i towards one particular amino acid with respect to the background can be quantified by the Kullback - Leibler divergence, $D_{KL}(f_i||\nu) = \sum_{a=1}^{q} f_i^a \quad \log(\frac{f_i^a}{\nu^a})$. The *positional information* is defined as $\phi_i^a \equiv \frac{\partial D_{KL}(f_i||\nu)}{\partial f_i^a}$.

In [1] it was suggested to rescale the correlation matrix C_{ij}^{ab} taking into account the positional information as follows:

$$\tilde{C}_{ij}^{ab} = (f_{ij}^{ab} - f_i^a f_j^b)\phi_i^a \phi_j^b. \tag{2}$$

In order to avoid singularities due to the presence of the logarithm in the Kullback-Leibler divergence, we used pseudo counts to regularise frequencies [2, 3], i.e., adding at each position a fictive count.

In addition, due to sampling biases [4–6], the dataset is not spatiotemporally homogenous and with an overabundance of some specific very similar sequences. To limit this bias, we have reweighed sequences by collapsing those overlapping at least of the 90%, which we will refer to as *similarity threshold* σ. We verified that values $0.9 \leq \sigma \leq 1$ does not change sensibly the results for the families we analysed. In the following we shall call the number of effective sequences M_{eff}.

Regularisation and reweighing can be expressed in a compact manner for single-site and pair frequencies as follows:

$$f_i^a = \frac{1}{M_{eff} + \lambda M_{eff}} \Big(\frac{\lambda}{q} + \sum_{\alpha=1}^{M_{eff}} \delta_{a,s_i^{\alpha}} \Big) \tag{3}$$

and

$$f_{ij}^{ab} = \frac{1}{M_{eff} + \lambda M_{eff}} \Big[\frac{\lambda}{q} \delta_{ij} \delta_{ab} + \frac{\lambda}{q^2} (1 - \delta_{ij}) + \sum_{\alpha=1}^{N} \delta_{a,s_i^{\alpha}} \delta_{b,s_j^{\alpha}} \Big], \tag{4}$$

where $\lambda = 1$ is the pseudo count.

The regularised and reweighed $\tilde{C}_{ij}^{ab}$ is still a $qL \times qL$ matrix: to reduce it to a $L \times L$ matrix, we used the so-called *Frobenius norm*, i.e.:

$$\bar{C}_{ij} = \sqrt{\sum_{a,b=1}^{q} \tilde{C}_{ij}^{ab\ 2}}, \tag{5}$$

$\bar{C}_{ij}$ is now a $L \times L$ symmetric matrix. Such a matrix does not show the usual properties of a typical correlation matrix: its diagonal elements are indeed not unitary and this is due to the rescaling procedure aimed at highlighting the conservation at each position. Note that this reflects the main aim of the original SCA, i.e., to take into account at the same time both pairwise correlations and positional conservation. To perform SCA, as we previously discussed, one must compare the spectral properties (i.e., eigenvalues and eigenvectors' components) of the $\bar{C}_{ij}$ with those of the correlation matrix got from the reshuffled dataset. Data reshuffling is performed constraining on the single amino acid frequency at each site, i.e., randomly exchanging two different amino acids at the same position i. The procedure to compute the correlation matrix is exactly the same as before and we call the random matrix $\mathcal{M}_{ij}$.

As introduced in the Main Text, to figure out whether some relevant information is enclosed in the dataset, one has firstly to compare eigenvalues' distributions relatively to the $\bar{C}_{ij}$ with those of $\mathcal{M}_{ij}$. We expect $\mathcal{M}_{ij}$'s eigenvalues distribution to be Marchenko-Pastur like [7], i.e., a bulk of very small eigenvalues and short tails. Fig. S1 shows at least four eigenvalues (black blocks) of the $\bar{C}_{ij}$ computed for the PF00072 rising out of the random bulk (orange blocks). The first highest eigenvalue has not been shown since it is a consequence of the phylogenetic history characterising the dataset [1] and of the use of pseudo counts and it will not be taken into account for sectors selection. In order to identify *sectors*, we sticked with the second and third highest eigenvalues, λ_2 and λ_3 (Fig. S1) and their associated eigenvectors, $|2\rangle$ and $|3\rangle$ (Fig.S2). The aim is to display those sites giving *signals* along these directions, i.e., having a projection along the two eigenvectors significantly higher than the random one. Commonly, one defines a discrimination threshold ϵ to distinguish the randomness in the eigenvectors' components from actual biologically relevant signals [1]. As our aim consists of mainly comparing the relevant sites identified by CVS with those identified by SCA, we first introduced a measure of relevance in SCA as well. Let us consider the eigenvectors associated with the second and third highest eigenvalues of the correlation matrix $\bar{C}_{ij}$ and the projection of the correlation matrix along these directions (Fig.S3). As shown in Fig.S2, most of the randomness will be localised around small values of the eigenvectors' components. In turn, actual relevant signals can be detected far from this random bulk. We thus defined the relevance of the $i-th$ position as the distance of the $i-th$ point in the plane spanned by the eigenvectors associated to the previously mentioned highest eigenvalues. The most relevant sites will be then the most far from the origin of this plane.

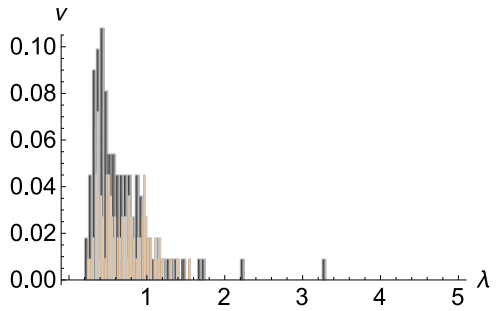

FIG. S1: Eigenvalues distribution for the actual correlation matrix (orange blocks) and the random matrix obtained from the reshuffled dataset (black blocks).

To compare CVS results to those obtained by SCA, we then sorted both lists of sites according to the relevance definition in each of these methods and we computed the overlap between the two lists by considering the top n positions. As discussed in the Main Text, for small values of n the overlap is very small, thus the two methods actually give very different results. Increasing n, the overlap increases and becomes quite different from the random one, till reaching a maximum for values of $n \simeq L/2$ (Fig.4a). We thus chose this value of n for our comparisons shown in the Main Text.

For completeness, we also defined sectors for the SCA results, following the same clustering procedure as in [1]. Note however that here no discrimination threshold is imposed but sites are sorted according to their distance in the plane spanned by the principal components. For PF00072 , four sectors have been identified by grouping the positions in the plane spanned by $|3\rangle$ and $|2\rangle$ in the following way:

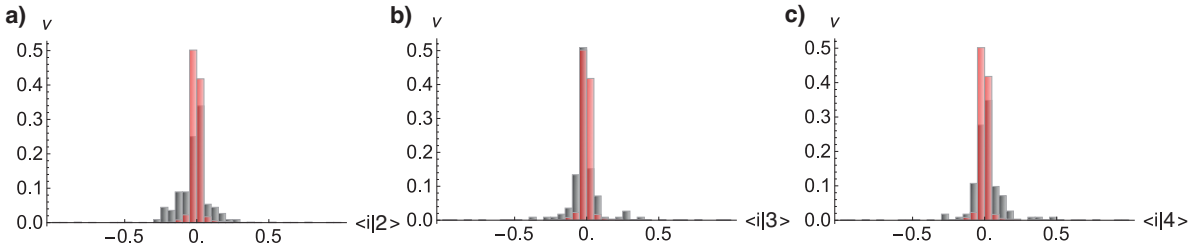

FIG. S2: Histograms of eigenvectors components frequencies (second eigenvectors in a), third in b) and fourth in c)), ν, relatively to the correlation matrix $\bar{C}_{ij}$ (black blocks) and to the random one (red blocks).

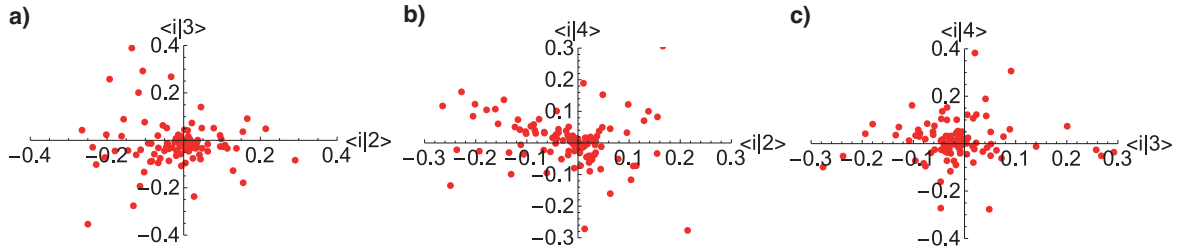

FIG. S3: Eigenvectors components for the matrix $\bar{C}_{ij}$. While in a) one can still see a clusters structure, in the others the eigenvectors' components are much noisier.

- the first sector is identified by all those positions $\langle i|2\rangle > 0$ and $|\langle i|2\rangle| > |\langle i|3\rangle|$;
- the second sector is identified by all those positions $\langle i|2\rangle < 0$ and $|\langle i|2\rangle| > |\langle i|3\rangle|$;
- the third sector is identified by all those positions $\langle i|3\rangle > 0$ and $|\langle i|2\rangle| < |\langle i|3\rangle|$;
- the fourth sector is identified by all those positions $\langle i|3\rangle < 0$ and $|\langle i|2\rangle| < |\langle i|3\rangle|$.

The obtained sectors are shown in Fig. S5 and their meaning is discussed in the Main Text.

Fig. S5 plots the first 50 points in this list for PF00072 in the space spanned by the 2nd and 3rd principal components (as discussed in [1], the largest eigenvalue should not be considered since it is a signature of the phylogenetic history of the dataset). Performing a clustering procedure onto this set, one finally gets groups of mostly correlated sites, usually called *sectors* in the literature [1, 8]. For PF00072 , four sectors can be identified, corresponding to functional domains on the tertiary structure. Within CVS instead no notion of sectors has been defined yet: we thus stick with the only notion of relevance given by the counts discussed in Main Text.

In our analysis we just sticked with the two highest eigenvalues of the correlation matrix. Yet, as pointed out in [8], smaller eigenvalues can still give signals about some positions. However, we found that CVS recovers most of these positions found to be relevant along eigenvectors associated to smaller eigenvalues.

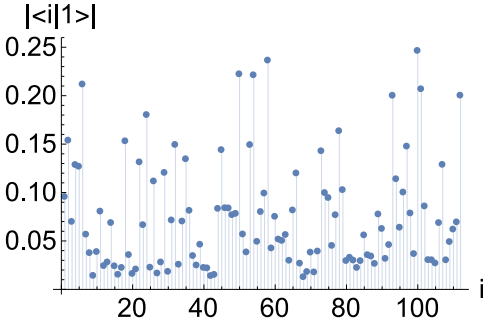

FIG. S4: Absolute value of the first eigenvector components, $|\langle i|1\rangle|$, plotted along the sequence. Notice that most of them are at least higher than 0.025.

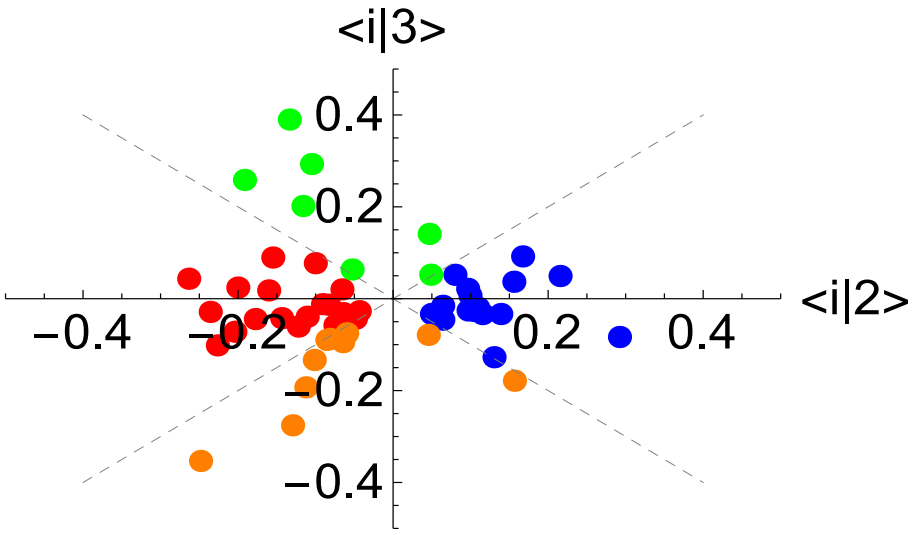

FIG. S5: Sectors identified by clustering the positions along the protein in the plane spanned by the principal components. The clustering procedure has been performed as explained in the Text.

II. DIRECT COUPLING ANALYSIS IN A NUTSHELL

Direct contact prediction from MSAs has recently been subject of intense research. Here we focus on the well-known method of Direct Coupling Analysis. As discussed in the Introduction of Main Text, many approaches have been proposed so far some aimed at minimising the detection of false positives while some others at weighing the gaps introduced by the alignments. Hereafter, we will refer to [2] where they introduced and tested the so-called naive Mean Field approach (nMFDCA) to infer the interactions between the amino acids.

The ansatz of DCA methods is that each sequence is the outcome of a Boltzmann - distribution, $P(\vec{s})$, obtained from a maximum entropy principle under the constraints that marginal distributions must match the experimental ones, i.e.:

$$P(s_i = a) \equiv f_i^a$$
and
$$P(s_i = a, s_j = b) \equiv f_{ij}^{ab}.$$

This allows the introduction of a q-state Potts' hamiltonian, $\mathcal{H}$, given by:

$$\mathcal{H}[\vec{s}] = -\sum_{i<j} J_{ij}(s_i, s_j) - \sum_i h_i(s_i). \tag{6}$$

The model we are aimed at fitting data with is then a 21-state Potts' model.

Since for each site i frequencies sum up to 1, being there L constraints for each site, the model has $(q-1)L$ free parameters which can be inferred exploiting the Plefka's expansion of the Gibbs' free energy generalized to the q-state Potts' model: it relates couplings $J_{ij}(s_i, s_j)$ to the $(q-1)L \times (q-1)L$ correlation matrix $C_{ij}(s_i, s_j)$ [2, 9]. The correlation matrix $C_{ij}(s_i, s_j)$ is defined as in SCA other than the rescaling with positional conservation [10]. One can check that from the Plefka's expansion it follows that:

$$J_{ij}(s_i, s_j) = -(C_{ij}(s_i, s_j))^{-1}. \tag{7}$$

However, to ensure matrix inversion, one must neglect the $q-$th degree-of-freedom because of the L frequency constraints we discussed before. Here, the use of pseudo counts is fundamental in order to avoid singularities due to positional under sampling.

This allows to obtain a regular $J_{ij}(s_i, s_j)$ matrix whose dimensions are $(q-1)L \times (q-1)L$: to perform a dimensional reduction on the couplings and turning again to a $L \times L$ matrix, one can introduce again the $q-th$ degree-of-freedom (as a null column/row) and then to standardize the couplings in the following way:

$$\tilde{J}_{ij}(s_i, s_j) = J_{ij}(s_i, s_j) - \mu_{ij}(s_i) - \mu_{ij}(s_j) + \mu_{ij}, \tag{8}$$

where $\mu_{ij}(s_i) = \frac{1}{L}\sum_{s_j=1}^{q} J(s_i,s_j)_{ij}$ (analogously for $\mu_{ij}(s_j)$) and $\mu_{ij} = \frac{1}{L^2}\sum_{s_i,s_j=1}^{q} J_{ij}(s_i,s_j)$ and then take the Frobenius norm as previously defined. The Frobenius norm computed on this new couplings matrix is called the *F–score*, defined as:

$$F_{ij} \equiv ||\tilde{J}_{ij}(s_i,s_j)||_{s_i,s_j} = \sqrt{\sum_{s_i,s_j=1}^{q} J_{ij}(s_i,s_j)^2}. \tag{9}$$

The F-score turns out to have zero elements on the diagonal, i.e., zero self-couplings, and a better highlight of structures within the coupling matrix [8].

[1] N. Halabi, O. Rivoire, S. Leibler and R. Ranganathan, *Cell*, 2009, **138**, 774–86.
[2] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt, *PNAS*, 2011, **108**, E1293–1301.
[3] B. Lunt, H. Szurmant, A. Procaccini, J. A. Hoch, T. Hwa and M. Weigt, *Methods Enzymol*, 2010, **471**, 17–41.
[4] F. Ozsolak and P. M. Milos, *Nat Rev Genet*, 2011, **12**, 87–98.
[5] E. R. Mardis, *Nature*, 2011, **470**, 198–203.
[6] I. Pagani, K. Liolios, J. Jansson, I. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz and N. C. Krypides, *Nucleic Acids Res*, 2012, **40**, D571–79.
[7] V. A. Marchenko and L. A. Pastur, *Mat. USSR Sb.*, 1967, **1**, 457–83.
[8] S. Cocco, R. Monasson and M. Weigt, *PLoS Comput Biol*, 2013, **9**, e1003176.
[9] T. Plefka, *J. Phys. A.: Math. Gen.*, 1982, **15**, 1971.
[10] O. Rivoire, *Phys. Rev. Lett.*, 2013, **110**, 178102.