Note 109: A summary of SCA calculations

Rama Ranganathan[1] and Olivier Rivoire[2]
[1] *The Green Center for Systems Biology, and Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas, USA.*
[2] *Laboratoire Interdisciplinaire de Physique, CNRS & Université Joseph Fourier, Grenoble, France.*

(Dated: September 26, 2011)

This document provides a summary of the current implementation of the statistical coupling analysis (SCA) method. The conceptual goal of SCA is to quantitatively parameterize the statistical patterns encoded in ensembles of proteins that share a common evolutionary origin. The idea is that this statistical analysis will provide a necessary foundation for understanding the physical mechanism and evolutionary origins of natural proteins.

I. PRELIMINARIES - MULTIPLE SEQUENCE ALIGNMENT AND FREQUENCIES

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 ($s = 1, \dots, M$ is for sequences, $i = 1, \dots, L$ is for positions and $a = 1, \dots, 20$ is for amino acids). The frequency $f_i^{(a)}$ of an amino acid a at position i is the number of sequences in the alignment having amino acid a at position i ($M_i^{(a)}$), divided by the total number of sequences, or equivalently, the average value of $x_{i,s}^{(a)}$ over all sequences s:

$$f_i^{(a)} = \frac{M_i^{(a)}}{M} = \langle x_{i,s}^{(a)} \rangle_s. \tag{1}$$

As described below in section II, the conservation of each position in the multiple sequence alignment is measured by the divergence of the observed frequency $f_i^{(a)}$ of amino acid a at position i from the background probability $q^{(a)}$ of amino acid a. This background probability is computed from the mean frequency of amino acid a in all proteins in the non-redundant database. Specifically,

$$\begin{aligned} q = \; & (0.073, 0.025, 0.050, 0.061, 0.042, 0.072, 0.023, 0.053, 0.064, 0.089, \\ & 0.023, 0.043, 0.052, 0.040, 0.052, 0.073, 0.056, 0.063, 0.013, 0.033), \end{aligned}$$

where amino acids are ordered according to the alphabetic order of their standard one-letter abbreviation.

Some calculations also require introducing a background probability for gaps. If γ represents the fraction of gaps in the alignment, a background probability distribution can be taken as $\bar{q}^{(0)} = \gamma$ for gaps, and $\bar{q}^{(a)} = (1-\gamma)q^{(a)}$ for the 20 amino acids. A practical strategy is to truncate alignments to sequence positions with a frequency of gaps $f_i^{(0)}$ no greater than 0.2; this prevents trivial over-representation of gaps in a sequence alignment and ensures calculations are made only at largely non-gapped sequence positions. Alternatively, one can truncate the alignment to the positions present in an atomic structure of a representative member of the protein family.

II. POSITION-SPECIFIC CONSERVATION - FIRST ORDER STATISTICS

The conservation of amino acid a at position i, considered independently of other positions, is measured by the statistical quantity $D_i^{(a)}$, the so-called Kullback-Leibler relative entropy (1) of $f_i^{(a)}$ given $q^{(a)}$. This definition is derived from the probability $P_M[f_i^{(a)}]$ of observing $f_i^{(a)}$ in an alignment of M sequences given a background probability $q^{(a)}$:

$$P_M[f_i^{(a)}] = \frac{M!}{(Mf_i^{(a)})!(M(1-f_i^{(a)})!} (q^{(a)})^{Mf_i^{(a)}} (1-q^{(a)})^{M(1-f_i^{(a)})}. \tag{2}$$

When M is large (the relevant limit for SCA), the Stirling formula leads to the approximation

$$P_M[f_i^{(a)}] \simeq e^{-MD_i^{(a)}}, \quad \text{with} \tag{3}$$

$$D_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{q^{(a)}} + (1-f_i^{(a)}) \ln \frac{1-f_i^{(a)}}{1-q^{(a)}}. \tag{4}$$

The value of $D_i^{(a)}$ indicates how unlikely the observed frequency of amino acid a at position i would be if a occurred randomly with probability $q^{(a)}$ - a definition of position-specific conservation. Note that $D_i^{(a)} = 0$ only when $f_i^{(a)} = q^{(a)}$ and increases more and more steeply as $f_i^{(a)}$ deviates from $q^{(a)}$, a result consistent with intuition that a measure of conservation should non-linearly represent the divergence of the observed distribution of amino acids from their randomly expected values.

What is an appropriate number of sequences to carry out SCA? A more precise relation between the probability $P_M[f_i^{(a)}]$ and the relative entropy $D_i^{(a)}$ is

$$-\frac{1}{M}\ln P_M[f_i^{a)}] = D_i^{(a)} + \frac{\ln M}{2M} + O\left(\frac{1}{M}\right). \tag{5}$$

The values of $D_i^{(a)}$ are typically of order 1-3 (the scale is given by $\ln 20 \approx 3$), so the corrective term $\ln M/(2M)$ can be neglected when M is of order of 100 sequences or greater ($M = 100$ corresponds to $\ln M/(2M) \approx 0.02$). This gives a lower bound on the size of alignments appropriate for SCA studies.

Equation 4 gives the conservation of each amino acid a at each position i. An overall positional conservation D_i taking into account the frequencies of all 20 amino acids can also be defined, but requires introducing a background probability for gaps (see Sec. I). Denoting $f_i^{(0)} = 1 - \sum_{a=1}^{20} f_i^{(a)}$ the fraction of gaps at position i, we can write the probability of jointly observing the frequencies $(f_i^{(1)}, \dots, f_i^{(20)})$ of each of the 20 possible amino acids at position i as

$$P_M[f_i^{(1)}, \cdots, f_i^{(20)}] = \frac{M!}{(Mf_i^{(0)})! \cdots (Mf_i^{(20)})!}(\bar{q}^{(0)})^{Mf_i^{(0)}} \cdots (\bar{q}^{(20)})^{Mf_i^{(20)}} \simeq e^{-MD_i} \tag{6}$$

where $D_i = \sum_{a=0}^{20} f_i^{(a)} \ln\left(f_i^{(a)}/\bar{q}^{(a)}\right)$ defines the overall conservation at position i.

III. CORRELATED CONSERVATION - SECOND ORDER STATISTICS

A. General Principles

Given an alignment $x_{i,s}^{(a)}$, a covariance matrix reporting pair-wise correlations between amino acids at positions is defined as

$$C_{ij}^{(ab)} = \langle x_{i,s}^{(a)} x_{j,s}^{(b)} \rangle_s - \langle x_{i,s}^{(a)} \rangle_s \langle x_{j,s}^{(b)} \rangle_s = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}, \tag{7}$$

where $f_{ij}^{(ab)} = \langle x_{i,s}^{(a)} x_{j,s}^{(b)} \rangle_s$ represents the joint frequency of having a at position i and b at position j. However, a fundamental principle of SCA is to compute correlations not for the raw alignment $x_{i,s}^{(a)}$, but for a weighted alignment

$$\tilde{x}_{i,s}^{(a)} = \phi_i^{(a)} x_{i,s}^{(a)}, \tag{8}$$

where $\phi_i^{(a)}$ is a functional of the positional conservation $D_i^{(a)}$. That is, the alignment $\tilde{x}_{i,s}^{(a)}$ now contains entries that are not merely binary (that is, 0 or 1), but that quantitatively represent the significance of observing each amino acid in each sequence in the alignment. Replacing $\tilde{x}_{i,s}^{(a)}$ for $x_{i,s}^{(a)}$ in Eq. (7), it is straightforward to see that the result is a weighted correlation matrix which now reports the significance of correlations as judged by the degree of conservation of the underlying amino acids:

$$\tilde{C}_{ij}^{(ab)} = \phi_i^{(a)} \phi_j^{(b)} C_{ij}^{(ab)}. \tag{9}$$

B. Choice of weights

Equation (9) gives the general definition of the SCA correlation tensor, but what specific form should the weighting function ϕ take? The approach taken in SCA (versions 3.0 and greater) is to consider the effect on the conservation of each position i upon removing each sequence s. The idea is that this "perturbation" will provide an estimate of the significance of each amino acid at each position in the alignment by its impact on the measure of conservation used (here, the relative entropy D). To develop this formally, let $M_i^{(a)}$ be the number of sequences with amino acid

a at position i, and M be the total number of sequences. When sequence s is left out, the frequency $f_i^{(a)} = M_i^{(a)}/M$ becomes

$$f_{i,s}^{(a)} = \frac{M_i^{(a)} - x_{is,}^{(a)}}{M-1} = \left(1 + \frac{1}{M}\right) f_i^{(a)} - \frac{x_{i,s}^{(a)}}{M} + O\left(\frac{1}{M^2}\right), \tag{10}$$

where we remind that $x_{i,s}^{(a)} = 1$ if sequence s has amino acid a at position i, and 0 otherwise. In the limit of large number of sequences M, expanding $D_i^{(a)}$ viewed as a function of $f_{i,s}^{(a)}$, to first order in $1/M$ leads to

$$D_{i,s}^{(a)} \approx \hat{D}_i^{(a)} - \frac{x_{i,s}^{(a)}}{M} \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}}, \tag{11}$$

where $\hat{D}_i^{(a)}$ is the relative entropy $D_i^{(a)}$ with $f_i^{(a)}$ replaced by $(1 + 1/M)\, f_i^{(a)}$. Ignoring the scaling factor of $1/M$ (or, equivalently, rescaling the perturbation in conservation $D_{i,s}^{(a)} - \hat{D}_i^{(a)}$ by M to be independent of alignment size), we find that this perturbation approach indicates a weighting function ϕ for the alignment that is the gradient of relative entropy:

$$\phi_i^{(a)} = \left|\frac{\partial D_i^{(a)}}{\partial f_i^{(a)}}\right| = \left|\ln\left[\frac{f_i^{(a)}(1 - q^{(a)})}{(1 - f_i^{(a)})q^{(a)}}\right]\right|. \tag{12}$$

The absolute value of the gradient is taken to ensure positive weights[1]. Applying these weights to the general definition given in equation 9), we have the specific form used for the SCA correlation tensor in versions 3.0 and greater of the SCA toolbox:

$$\tilde{C}_{ij}^{(ab)} = \left|\frac{\partial D_i^{(a)}}{\partial f_i^{(a)}}\right| \left|\frac{\partial D_j^{(b)}}{\partial f_j^{(b)}}\right| C_{ij}^{(ab)}. \tag{13}$$

It is important to understand the nature of the function ϕ in controlling the patterns of correlations emerging from the weighted alignment $\tilde{x}_{i,s}^{(a)}$. The weights chosen by the sequence perturbation approach described above (a version of so-called "jackknife resampling") have the property of rising steeply as the frequencies of amino acids $f_i^{(a)}$ approach one. As a consequence, these weights will damp correlations in C_{ij}^{ab} arising from weakly conserved amino acids (the gradient of $D_i^{(a)}$ approaches zero as $f_i^{(a)} \to q^{(a)}$), and will emphasize conserved correlations. In essence, this function imposes a particular mathematical form on the basic underlying principle of SCA that the functional relevance of correlations should scale with their conservation. In this regard, different weighting functions are possible (1) if mathematical formalisms other than the Kullback-Leibler entropy are used for defining positional conservation, or (2) if other approaches than the particular sort of sequence perturbation analysis described here are used for determining weights. For example, the SCAv.5.0 MATLAB toolbox includes the possibility of using Rényi relative entropies (a generalization of the Kullback-Leibler entropy) or indeed even arbitrary user-defined functions of $f_i^{(a)}$ and $q^{(a)}$, as a measure of conservation. In addition, as shown in section IV.B, the original implementation of SCA (versions 1-2) involved a different perturbation technique that leads to a somewhat different weighting function. Regardless of these differences, the salient point is that all of these approaches are variations on the general principle in SCA of a weighted correlation (Eq. (9)).

In general, the problem of applying more complex conservation functions and associated weights is deeply connected with fundamentally understanding the nature of the evolutionary process that generates the observed amino acid distributions and the nature of our sampling of these distributions in our databases of available sequences. These are important future research goals, but in the absence of such deeper understanding the current approach in SCA of using the Kullback-Leibler entropy as a measure of conservation and gradients of this entropy function as weights represents a simple and analytically well-defined implementation of the general principles of this method.

[1] The absolute value is to ensure positive weights, but since $\partial D_i^{(a)}/\partial f_i^{(a)}$ is negative only for amino acids that are irrelevant at position i, i.e., for which $f_i^{(a)} < q^{(a)}$, it is not essential.

Distinct from the principle that the relevance of correlations should scale with conservation, the weighting function also contributes to separating signal (the true evolutionary constraints) from correlation noise due to finite and biased sampling in practical sequence alignments. For example, one way that weakly conserved amino acids are expected to show strong correlations in C_{ij}^{ab} is due to the existence of small clades of historically related sequences (bias in sampling) in which correlations are more due to the lack of sufficient time to diverge rather than due to functional constraints. In addition, when the number of sequences M is on the same order as the number of positions L, we expect significant correlation noise due to limitations in sampling. The non-linear weighting function used in SCA has the effect of minimizing the effect of these noise sources in interpretation of correlations.

C. Reduction to positional correlations

The first step in analysis of the SCA correlation tensor is to reduce the four-dimensional array of L positions × L positions × 20 amino acids × 20 amino acids to a $L \times L$ matrix of positional correlations. Indeed, the goal in SCA is to identify collectively evolving groups of positions (”protein sectors”) whose correlations are properties of the whole family (or of functional subfamilies, see below) regardless of the amino acids by which their correlation is identified.

How should we carry out the dimension reduction of the SCA correlation tensor to a matrix of positional correlations? One empirical property of the tensor provides a straightforward approach: analysis of the 20×20 amino acid correlation matrices for fixed pairs of positions (i, j) shows that these matrices have approximately rank one (**?**). That is, the information content in these matrices can be captured in a single scalar value. To explain this, we carry out the so-called singular value decomposition of $\tilde{C}_{ij}^{(ab)}$ for each (i, j):

$$\tilde{C}_{ij}^{ab} = \sum_{c=1}^{20} P_{ij}^{ac} \lambda_{ij}^{c} Q_{ij}^{cb}. \tag{14}$$

In this decomposition, each 20×20 amino acid correlation matrix for each (i, j) is written as a sum of products of three 20×20 matrices: λ, a diagonal matrix of singular values that indicate the quantity of variance in $\tilde{C}_{ij}^{(ab)}$ captured, and P and Q, matrices of singular vectors whose columns give weights for the combination of amino acids at positions i and j that are associated with each singular value. Interestingly, the singular value decomposition of $\tilde{C}_{ij}^{(ab)}$ for each (i, j) has the property that $\lambda_{ij}^{1} \gg \lambda_{ij}^{c}$ for $c \neq 1$. That is, the information in the amino acid correlation matrix for each pair of positions can be represented by one number, the top singular value (also known as the ”spectral norm”):

$$\tilde{C}_{ij}^{(ab)} \simeq P_{ij}^{a1} \lambda_{ij}^{1} Q_{ij}^{1b}. \tag{15}$$

Thus, a matrix of positional correlations $\tilde{C}_{ij}$ can be defined simply by taking the spectral norm of $\tilde{C}_{ij}^{(ab)}$ for each pair (i, j) of positions:

$$\tilde{C}_{ij} = \lambda_{ij}^{1}. \tag{16}$$

This is one definition of the SCA positional correlation matrix, and is essentially identical to that computed in versions of the SCA Toolbox from 2.0 to 4.5 (for slight technical differences from these earlier versions see section IV.C). It is also returned as an option by the SCA v5.0 MATLAB program sca5.m. Below, we will show another empirical property of the SCA correlation tensor that will permit definition of a $\tilde{C}_{ij}$ matrix using a different mathematical approach that is used by default in SCAv5.0.

Two further notes with regard to this dimension reduction step: (1) Since $\tilde{C}_{ij}^{ab} = \tilde{C}_{ji}^{ba}$, $Q_{ij}^{cb} = P_{ji}^{bc}$, we can simplify Eq. (15) to $\tilde{C}_{ij}^{(ab)} \simeq P_{ij}^{a1} \lambda_{ij}^{1} P_{ji}^{b1}$. (2) Also, since $(P_{ij}^{a1})^{-1} = P_{ij}^{1a}$, we note that this reduction corresponds to $\tilde{C}_{ij} \simeq \sum_{a,b} P_{ij}^{1a} \tilde{C}_{ij}^{ab} P_{ji}^{1b}$.

D. The alignment projection approach to SCA

In section III.C above, we indicated how for each pair of positions (i, j), we can reduce the 20×20 matrix of amino acid correlations by singular value decomposition to just the top singular value. The singular vectors corresponding to the top singular value (P_{ij}^{a1} and P_{ji}^{1b} for positions i and j, respectively) contain the weights for the amino acids at these positions that contribute to the top singular value for each (i, j). Interestingly, study of these top singular vectors shows another empirical finding about the SCA correlation tensor. For a given position i, we find that P_{ij}^{1a} is approximately independent of j. That is, the amino acids by which a position i makes correlations with other positions j is nearly the same, and therefore is essentially a property of just position i taken independently. This is a non-trivial finding and suggests a simple but powerful approach for SCA in which we reduce the alignment itself directly from the $M \times L \times 20$ three-dimensional weighted tensor ($\tilde{x}_{i,s}^{a}$) introduced in equation 8 to a $M \times L$ two-dimensional weighted matrix ($\tilde{x}_{i,s}$). That is, we can use the collection of top singular vectors for each position i averaged over all j ($\bar{P}_i^a$) as a "projection matrix" to reduce the amino acid dimension of the alignment (see appendix, section IV.F):

$$\tilde{x}_{i,s} = \sum_a \bar{P}_i^a \tilde{x}_{si}^a \tag{17}$$

In practice, the projection matrix can be obtained directly from the weighted frequencies of amino acids at positions in the alignment. This makes sense; the finding that the top singular vectors of amino acid correlations for position i are independent of j implies that the average singular vector should be just a property of the amino acid distribution at site i. The projection matrix can be written as:

$$\bar{P}_i^a = \frac{\langle \tilde{x}_{si}^a \rangle_s}{\left(\sum_b \langle \tilde{x}_{si}^b \rangle_s^2\right)^{1/2}} = \frac{\phi_i^a f_i^a}{\left(\sum_b (\phi_i^b f_i^b)^2\right)^{1/2}}. \tag{18}$$

Three quantities can be computed directly from the now projected alignment matrix $\tilde{x}_{i,s}$:

(1) a SCA positional correlation matrix (written as $\tilde{C}_{ij}^P$, to distinguish it formally from $\tilde{C}_{ij}$ defined above):

$$\tilde{C}_{ij}^P = \left| \frac{1}{M} \sum_s \tilde{x}_{si} \tilde{x}_{sj} - \frac{1}{M^2} \left(\sum_s \tilde{x}_{si} \right) \left(\sum_s \tilde{x}_{sj} \right) \right|, \tag{19}$$

(2) a SCA sequence correlation matrix ($\tilde{C}_{st}^S$) that represents the statistical groupings of sequences into subfamilies (if any):

$$\tilde{C}_{st}^S = \left| \frac{1}{L} \sum_i \tilde{x}_{si} \tilde{x}_{ti} - \frac{1}{L^2} \left(\sum_i \tilde{x}_{si} \right) \left(\sum_i \tilde{x}_{ti} \right) \right|. \tag{20}$$

It is important to note that this sequence correlation matrix represents relationships between sequences where positions are weighted by the conservation-based weighting function ϕ; thus this mapping of sequence space is more likely to reveal functional distinctions between sequence subfamilies rather than just historical relationships.

(3) a matrix (Π) that provides a mapping between these two spaces. To explain the Π matrix, we note that the projected alignment $\tilde{x}_{i,s}$ can be written by singular value decomposition as $\tilde{x} = \sum_{n=1}^{\min(L,M)} |U_n\rangle \lambda_n \langle V_n|$. This provides a mapping from the space of positional correlations to the space of sequence correlations by:

$$\Pi = \sum_n |U_n\rangle \langle V_n| \tag{21}$$

That is, if we detect collectively evolving groups of amino acid positions (sectors) by analysis of the $\tilde{C}_{ij}^P$ matrix, then Π provides a mapping that can test whether these sectors are associated with the divergence of specific subfamilies in the alignment. In addition, the matrix $\Pi^\top$ provides the inverse mapping - from the space of sequence correlations to the space of positional correlations. With this inverse mapping, it is possible to use prior knowledge of functional divergence in the sequence alignment to target identification of a sector that is responsible for this divergence. We will show the technical details of these mappings below in section III.F.

E. Spectral decomposition/Independent component analysis

The process of identifying sectors from the $\tilde{C}_{ij}^P$ matrix begins with spectral (or eigenvalue) decomposition. The motivation is that the presence of significant correlations between positions in the $\tilde{C}_{ij}^P$ matrix indicates that treating the amino acid positions as the basic functional units of proteins is not the most informative representation. Instead, we should seek a reparameterization of the protein in which the units are collective groups of amino acids that coevolve per the positional correlation matrix (the "sectors"). Eigenvalue decomposition is the simplest first step in achieving this reparameterization. This decomposition is always available for any square positive semi-definite matrix (such as $\tilde{C}_{ij}^{P}$) and represents the matrix as a product of three matrices:

$$\tilde{C^P} = \tilde{V}\tilde{\Delta}\tilde{V}^{\top}, \tag{22}$$

where $\tilde{\Delta}$ is a diagonal matrix of so-called eigenvalues and $\tilde{V}$ is a matrix whose columns contain the associated eigenvectors. The eigenvectors contain the weights for linearly combining amino acid positions into new variables ("eigenmodes") that are now de-correlated, and the associated eigenvalues indicate the magnitude of the information in $\tilde{C}_{ij}^{P}$ that is captured.

The essence of sector identification is to study the pattern of positional contributions to the statistically significant top eigenmodes of the $\tilde{C}_{ij}^{P}$ matrix. To determine the number k of eigenmodes that are significant, we compare the spectrum of $\tilde{C}_{ij}^{P}$ with that derived from randomized alignments where the amino acids are scrambled independently at each position. In such randomized alignments, the eigenmodes reflect only the spurious correlations that are possible due to finite sampling in the alignment and provides a basis for a significance cutoff for eigenvalue magnitudes. As pointed out earlier, when the number of sequences M in the alignment is not large compared to the number of positions L, we expect the majority of weak correlations to be accountable by finite sampling considerations; indeed, we find that typically just the top few eigenmodes of the $\tilde{C}_{ij}^{P}$ matrix are significant. These top modes can be examined for distinct groups of correlated amino acid positions to define sectors.

However, different eigenvectors are not expected to directly represent statistically independent sectors. Instead, if independent sectors exist for a particular protein family, they will generally correspond to groups of positions emerging along combinations of eigenvectors. The reason is due to the fact that decorrelation of positions (by diagonalizing the SCA correlation matrix - the essence of eigenvalue decomposition) is a weaker criterion than achieving statistical independence (which requires absence of not only pairwise correlations, but lack of any higher order statistical couplings). In other words, if the non-independence of a set of variables is not completely captured in just their pairwise correlations, then just the linear combination of these variables indicated by eigenvectors of the correlation matrix cannot be assumed to produce statistically independent transformed variables (xx).

Independent component analysis (ICA) - an extension of spectral decomposition - is a heuristic method designed to transform the k statistically significant top eigenmodes of a correlation matrix into k maximally independent components through an iterative optimization process. In this process, the k top eigenvectors of $\tilde{C}^{P}$, written as $|\tilde{V}_1\rangle, \ldots, |\tilde{V}_k\rangle$, are rotated by a $k \times k$ so-called "unmixing" matrix W^P to to yield k maximally independent components $|\tilde{V}_1^p\rangle, \ldots, |\tilde{V}_k^p\rangle$:

$$\tilde{V}_{1\ldots k}^{p} = W^P \tilde{V}_{1\ldots k} \tag{23}$$

We call this process "posICA" to indicate that ICA is carried out on the eigenvectors of the SCA positional correlation matrix. In principle, this linear transformation of eigenvectors should help to better define independent sectors (if such exist in the protein family under study) as groups of positions now projecting along the transformed axes - the independent components (ICs) of position space ($\tilde{V}^p$).

It is also possible to apply ICA to the top eigenvectors of the SCA sequence correlation matrix $\tilde{C}^{S}$ (equation 20). Given a spectral decomposition $\tilde{C}^{S} = \tilde{U}\tilde{S}\tilde{U}^{\top}$, we can use ICA to derive a different unmixing matrix W^S that can be used to rotate the k top eigenvectors of $\tilde{C}^{S}$, written as $|\tilde{U}_1\rangle, \ldots, k\tilde{U}_k\rangle$, to yield k maximally independent components $|\tilde{U}_1^s\rangle, \ldots, |\tilde{U}_k^s\rangle$:

$$\tilde{U}_{1\ldots k}^{s} = W^S \tilde{U}_{1\ldots k} \tag{24}$$

We call this process "seqICA" to indicate that ICA is carried out on the eigenvectors of the SCA sequence correlation matrix. This transformation should provide a better description of sequence subfamilies as groups of sequences emerging largely orthogonally along the independent components of sequence space ($\tilde{U}^s$).

How do we technically obtain the W matrices? Various implementations of ICA can be used that apply different measures of independence and different algorithms for optimizing them. We use one of the simplest implementations of ICA, proposed in Ref. (2) with modifications introduced in Ref. (3) (the results should however be robustly recovered when using other algorithms for ICA). For posICA the input of the algorithm is the $k \times L$ matrix Z whose rows correspond to $|\tilde{V}_1\rangle, \ldots, |\tilde{V}_k\rangle$ while for seqICA it is the $k \times M$ matrix Z whose rows correspond to $|\tilde{U}_1\rangle, \ldots, |\tilde{U}_k\rangle$. The algorithm iteratively updates the unmixing matrix W, starting from the $k \times k$ identity matrix $W = I_k$, with increments ΔW given by

$$\Delta W = \epsilon \left(I_k + \left(1 - \frac{2}{1 + \exp(-WZ)} \right) (WZ)^{\top} \right) W. \tag{25}$$

The parameter ϵ is a learning rate that has to be sufficiently small for the iterations to converge. The iterations lead to W^s for seqICA and W^p for posICA. The vectors $|\tilde{V}_j^p\rangle$ are obtained by applying W^p to $|\tilde{V}_j\rangle$, and the vectors $|\tilde{U}_n^s\rangle$ by applying W^s to $|U_n\rangle$; these vectors are normalized to unit norm.

Upon ICA rotation, sector positions should correspond to positions i contributing significantly to one of the independent components $\langle i|\tilde{V}_j^p\rangle > \eta_i$, where η_i is a threshold of significance obtained empirically. Similarly, sequence subfamilies should correspond to sequences s contributing significantly to one of the independent components $\langle s|\tilde{U}_n^s\rangle > \eta_s$ where η_s represents an empirical cutoff.

F. Mapping between sequence and positional correlations

As described in Eq. (21), the Π matrix provides a mapping between the space of positional correlations in $\tilde{C}^P$ (which defines sectors) and the space of sequence correlations in $\tilde{C}^S$ (which defines functional subfamilies). We can apply Π following posICA to the positional independent components $|\tilde{V}_1^p\rangle, \ldots, |\tilde{V}_k^p\rangle$ to produce $|\tilde{U}_1^p\rangle, \ldots, |\tilde{U}_k^p\rangle$, the sequence space mapped onto the positional ICs:

$$\tilde{U}_{1\ldots k}^p = \Pi \tilde{V}_{1\ldots k}^p \tag{26}$$

If independent sectors are identified in $\tilde{V}^p$, then this mapping can indicate how sectors control relationships between sequences in $\tilde{U}^p$. For example, in the S1A serine proteases (xx), we find evidence for three quasi-independent sectors from study of $\tilde{V}^p$, and show that each of these sectors is associated with an independent functional classification of sequences.

We can also make the inverse mapping in which we use the Π matrix following seqICA to make a mapping from the sequence independent components $|\tilde{U}_1^s\rangle, \ldots, |\tilde{U}_k^s\rangle$ to produce $|\tilde{V}_1^s\rangle, \ldots, |\tilde{V}_k^s\rangle$, the positional correlation space mapped onto the sequence ICs:

$$\tilde{V}_{1\ldots k}^s = \Pi^\top \tilde{U}_{1\ldots k}^s. \tag{27}$$

If analysis of $\tilde{U}^s$ shows interesting functional separations of sequences comprising the protein family, then this mapping can indicate whether a distinct sector is responsible for this functional divergence. For example, in the Hsp70/110 family of molecular chaperones, study of $\tilde{U}^s$ indicates a clear separation of the allosteric Hsp70 proteins and the non-allosteric Hsp110 proteins alone one independent component. Interestingly, the corresponding component in $\tilde{V}^s$ reveals the sector that is associated with this functional divergence.

Note that we describe the usage of the Π matrix to map between the independent components of the positional ($\tilde{C}^P$) and sequence ($\tilde{C}^S$) correlation matrices, but the same mapping can also be made just to the eigenvectors of $\tilde{C}^P$ and $\tilde{C}^S$. Indeed, in cases where multiple independent sectors are not evident, sector identification and sequence space mappings are conducted directly from the eigenvalue decompositions without application of ICA. In this case, we define $\tilde{U}' = \Pi\tilde{V}$ as the sequence space mapping from the eigenvectors of the $\tilde{C}^P$ matrix (Eq. 22), and $\tilde{V}' = \Pi^\top\tilde{U}$ as the position space mapping from the eigenvectors of the $\tilde{C}^S$ matrix.

IV. APPENDIX

This section mostly provides information about relationships with previous descriptions of the SCA approach. The original implementation of SCA defined conserved correlations through a specific type of perturbation analysis on the sequence alignment (MATLAB SCA Toolbox 1.5, Sec. **??**), and this and more recent implementations described dimension reduction of the SCA correlation tensor through a formally different (though practically near-identical) type of matrix norm. Here, we explain these technical differences, and provide a more detailed explanation of the calculation of the projection matrix (equation 18).

A. Equivalence with previous definitions of conservation

$D_i^{(a)}$ is equivalent to measures of positional conservation introduced in previous reports of the SCA method. In essence, $D_i^{(a)}$ is the asymptotic limit for large M for $\Delta G_i^{\text{stat},a}$ (MATLAB SCA Toolbox v1.0, as reported in Refs. (4–7)), and $\Delta E_i^{\text{stat},a}$ (SCA Toolbox v1.5, as reported in Ref. (8)):

$$\Delta G_i^{\text{stat},a} = \Delta E_i^{\text{stat},a} = -\frac{1}{M} \ln P_M[f_i^{(a)}] \simeq D_i^{(a)}. \tag{28}$$

The pre-factor $-1/M$ scales the positional conservation parameter for alignments of different size, and represents the statistical unit of conservation symbolically indicated by kT^* or γ^* in previous works.

B. The original SCA method

The implementation of the SCA method introduced originally in Ref. (4) was based on a perturbation to the amino acid distribution at one test site i to measure the difference in position-specific conservation of each amino acid at a second site j. In general, the perturbation consisted of restricting the test site to the most prevalent amino acid a_i, a manipulation that extracts a sub-alignment with size equal to $f_i^{(a_i)}M$. For test sites in which sub-alignments retained sufficient size and diversity to be globally representative of the full alignment (i.e., $f_i^{(a_i)}M > 100$ sequences), a difference conservation value was calculated:

$$\Delta\Delta G_{j,i}^{\mathrm{stat},b,a_i} = \Delta\Delta E_{j,i}^{\mathrm{stat},b,a_i} = -\frac{1}{M}\left[\ln\left(P_M\left[f_j^{(b)}\right]\right) - \ln\left(P_M\left[f_{j|i}^{(b)|a_i}\right]\right)\right], \tag{29}$$

where $f_{j|i}^{(b)|a_i}$ is the frequency of amino acid b in the sub-alignment obtained by retaining only the sequences having a well represented amino acid a_i at position i. $\Delta\Delta G_{j,i}^{\mathrm{stat},b,a_i}$ represents the change in the conservation of amino acid b at position j due to the perturbation introduced at position i, a measure of their correlation (the term was renamed to $\Delta\Delta E$ in subsequent publications and is ignored entirely now to avoid confusion with Gibbs energies). The first term on the right hand side, $-\frac{1}{M}\ln\left(P_M\left[f_j^{(b)}\right]\right)$, corresponds to $D_j^{(b)}$. A basic tenet of the original SCA approach was that perturbations lead to sub-alignments that are representative of the full alignment, a condition satisfied typically by only the most frequent amino acid at a subset of positions. Under this assumption, $f_{j|i}^{(b)|a_i} \approx f_j^{(b)}$ for most amino acids b at positions j. We may therefore expand the second term, $-\frac{1}{M}\ln\left(P_M\left[f_{j|i}^{(b)|a_i}\right]\right)$, by writing

$$f_{j|i}^{(b)|a_i} = \frac{f_{ij}^{(a_i b)}}{f_i^{(a_i)}} = f_j^{(b)} + \frac{f_{ij}^{(a_i b)} - f_i^{(a_i)}f_j^{(b)}}{f_i^{(a_i)}} = f_j^{(b)} + \frac{C_{ij}^{(a_i b)}}{f_i^{(a_i)}} \tag{30}$$

with $C_{ij}^{(a_i b)}$ defined as in Eq. (7), so that

$$-\frac{1}{M}\ln\left(P_M\left[f_{j|i}^{(b)|a_i}\right]\right) \approx D_j^{(b)} + \frac{C_{ij}^{(a_i b)}}{f_i^{(a_i)}}\frac{\partial D_j^{(b)}}{\partial f_j^{(b)}}. \tag{31}$$

This leads to

$$\Delta\Delta G_{j,i}^{\mathrm{stat},b,a_i} \approx -\frac{1}{f_i^{(a_i)}}\frac{\partial D_j^{(b)}}{\partial f_j^{(b)}}C_{ij}^{(a_i b)}, \tag{32}$$

which shows that the perturbation procedure also represents a weighted procedure for correlations that is fully consistent with the general principle of SCA outlined in equation 9.

C. Dimensional reduction

In previous implementations of the SCA method, a reduced matrix $\bar{C}_{ij}$ was defined from $\tilde{C}_{ij}^{(ab)}$ by

$$\bar{C}_{ij} = \sqrt{\left(\sum_{a,b}\left(\tilde{C}_{ij}^{(ab)}\right)^2\right)}. \tag{33}$$

This is known as the Frobenius norm of the 20×20 matrix $\tilde{C}_{ij}^{(ab)}$ and can be expressed in terms of the singular values λ_{ij}^c of this matrix as $\bar{C}_{ij} = (\sum_c(\lambda_{ij}^c)^2)^{1/2}$. Since $\lambda_{ij}^1 \gg \lambda_{ij}^c$ for $c \neq 1$, the Frobenius norm of $\tilde{C}_{ij}^{(ab)}$ is well-approximated by the spectral norm of the matrix, $\bar{C}_{ij} \simeq C_{ij}^{\mathrm{sp}} = \lambda_{ij}^1$.

D. Binary approximation

In Ref. (9), we make use of a so-called "binary approximation" of the full alignment in which we 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 contains the most frequent amino acid at position i, and 0 otherwise (i.e., $x_{i,s} = x_{i,s}^{(a_i)}$). As a consequence of the non-linear dependence of $D_i^{(a)}$, $\bar{D}_i^{(a)}$, and D_i with respect to $f_i^{(a)}$, the overall conservation D_i is well approximated by the conservation of the prevalent amino acid ($D_i^{(a_i)}$ or $\bar{D}_i^{(a_i)}$), a result that justifies the use of the binary approximation.

The $L \times M$ binary array $x_{i,s}$ is formally obtained from the $L \times M \times 20$ array $x_{i,s}^{(a)}$ by application of a "projector" B_i^a defined by $B_i^a = 1$ if $a = a_i$, and 0 otherwise: $x_{i,s} = \sum B_i^a x_{i,s}^{(a)}$. From the observation that $\phi_i^{(a_i)} B_i^a$ is an approximation of $\phi_i^{(a)} \bar{P}_i^a$, where $\bar{P}_i^a$ is defined in Eq. (**??**), it results that

$$\tilde{C}_{ij}^{(\mathrm{bin})} = \phi_i^{(a_i)} \phi_j^{(a_j)} |C_{ij}^{(\mathrm{bin})}|, \tag{34}$$

where $C_{ij}^{(\mathrm{bin})} = \langle x_{i,s} x_{j,s} \rangle_s - \langle x_{i,s} \rangle_s \langle x_{j,s} \rangle_s = f_{ij}^{(a_i a_j)} - f_i^{(a_i)} f_j^{(a_j)}$. The matrix $\tilde{C}_{ij}^{(\mathrm{bin})}$ corresponds to the matrix of positional correlations considered in Ref. (9).

E. Cleaning of the first mode

In Ref. (9), we "cleaned" the first mode of $\tilde{C}_{ij}^{(\mathrm{bin})}$ and identified the sectors based on the subsequent eigenvectors of this matrix. This procedure rests on the observation that statistical correlations between positions can arise from a combination of phylogenetic bias and functional constraints acting independently on the positions, but that such statistical correlations would directly depend on the degree of conservation of the positions. From this point of view, if four equally conserved positions i, j, k, ℓ satisfy $\tilde{C}_{ij}, \tilde{C}_{k\ell} \gg \tilde{C}_{ik}, \tilde{C}_{i\ell}, \tilde{C}_{jk}, \tilde{C}_{j\ell}$, the absence of strong inter-correlation between the pairs (i, j) and (k, ℓ), contrasted with the presence of strong intra-correlations within them, can be attributed to correlations of functional origin.

Implementing this principle requires estimating C_{ij}^*, the (weighted) correlation expected from phylogeny for a pair of positions with same degree of conservation as i, j. Given C_{ij}^*, functionally significant correlations can then be deduced as those for which $\tilde{C}_{ij} > C_{ij}^*$. An approximate estimation of C_{ij}^* follows from the premise that most correlations are actually not functionally significant. By averaging $\tilde{C}_{ij}$ over the positions j we thus get an estimation of the degree of correlation to be expected from a position with degree of conservation such as i. This suggests to estimating C_{ij}^* by[2]

$$\hat{C}_{ij}^* = \frac{\langle \tilde{C}_{i\ell} \rangle_\ell \langle \tilde{C}_{kj} \rangle_k}{\langle \tilde{C}_{k\ell} \rangle_{k\ell}}. \tag{35}$$

This procedure admits an equivalent spectral formulation when the first eigenvalue of $\tilde{C}_{ij}$ is separated by a gap from the rest of the spectrum. Standard perturbation theory (see SI of Ref. (9)) indeed indicates that in such a case[3]

$$\hat{C}_{ij}^* \simeq \tilde{\lambda}_1 \langle i | \tilde{V}_1 \rangle \langle \tilde{V}_1 | j \rangle. \tag{36}$$

In other words, subtracting $\hat{C}^*$ from $\tilde{C}$ is equivalent to subtracting the first mode from $\tilde{C}$.

[2] This estimate $\tilde{C}_{ij}^*$ should be only slightly above C_{ij}^* (we assume here that the absolute value has been taken to define a non-negative $\tilde{C}_{ij}$); in particular, the largest deviations are expected for the correlations involving some of the most correlated positions, leading to an overestimation of C_{ij}^*, i.e. $\hat{C}_{ij}^* > C_{ij}^*$ for the i, j that are the most functionally correlated. As a result, the matrix $\tilde{C}_{ij} - \hat{C}_{ij}^*$ may display negative correlations that are only reflective of this over-estimation of C_{ij}^*. A way to correct for this artifact that we used in Ref. (9), is to consider $\max(0, \tilde{C}_{ij} - \hat{C}_{ij}^*)$, which effectively sets to zero any negative correlation resulting from the subtraction of $\hat{C}_{ij}^*$.

[3] Perturbation theory also indicates that the first eigenvector of $\tilde{C}_{ij}$ satisfies $\langle i | \tilde{V}_1 \rangle \propto \langle \tilde{C}_{ij} \rangle_j$. Given the positivity of $\tilde{C}_{ij}$ when the absolute value is taken, this mode thus corresponds to a coherent mode where each position contributes in proportion to its general contribution to the correlations (Perron-Frobenius theorem in fact indicates that the first mode is coherent even outside the perturbative regime).

This simple procedure however rests on the assumption that the sequences in the alignment are subject to essentially the same functional constraints, which does not represent the general case. The procedure presented in SCAv5.0 (Sec. III), which does not involve cleaning the first mode, is a more general solution for alignments with strong heterogeneities in the distribution of sequences.

F. Projection matrix

In section III.D we described the empirical finding that for each position i, the top singular vector of amino acid correlations with all other positions j is essentially independent of j. This led to the key concept of a projection matrix $\bar{P}_i^a$ (based on averaging P_{ij}^{1a} over all j) which can be used to reduce the weighted alignment tensor $\tilde{x}_{i,s}^a$ to a weighted alignment matrix $\tilde{x}_{i,s}$. But, there is one technical complication to discuss that leads to the specific form of the projection matrix described in Eq. (18). To make a proper matrix of projection vectors $\bar{P}_i^a$ by averaging P_{ij}^{1a} over all j, it is essential that the signs of the singular vectors in P_{ij}^{1a} all be the same. However, since $\tilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b (f_{ij}^{(ab)} - f_i^a f_j^b)$ can have both positive and negative values, the signs of the singular vectors are arbitrarily fixed in computing the singular value decomposition and the average over these vectors is not appropriate. To avoid this, we can force the SCA correlation tensor to have only positive values by computing $\hat{C}_{ij}^{(ab)} = \phi_i^a \phi_j^b f_{ij}^{(ab)}$, a SCA correlation tensor computed without subtracting the randomly expected correlations (the random expectation for correlations are subtracted at a later step in defining the $\tilde{C}^P$ and $\tilde{C}^S$ matrices, Eqs. (19) and (20)). Since $\hat{C}_{ij}^{(ab)}$ has only non-negative entries, the elements of P_{ij}^{1a} have all same sign (Perron-Frobenius theorem) and we can compute $\bar{P}_i^a$ the average of P_{ij}^{1a} over j. The calculation in Eq. (18) is an excellent approximation to $\bar{P}_i^a$.

V. SCA TOOLBOXES

A. Distributions

Distributions are MATLAB Toolboxes and contain various accessory codes for data formatting, display, and analysis. Previous versions include:

(1) SCA Toolbox 1.5: The original SCA method as specified in Lockless and Ranganathan (4) with one modification that was used in all subsequent papers: the division of binomial probabilities by the mean probability of amino acids in the alignment is removed. This version is longer in active use.

(2) SCA Toolbox 2.5: The bootstrap-based approach for SCA. Position-specific conservation calculated as in Eq. (4) and correlations calculated as in Eq. (9). Matrix reduction per Eq. (33).

(3) SCA Toolbox 3.0: The analytical calculation of correlations weighted by gradients of relative entropy. Position-specific conservation calculated as in Eq. (4) and correlations calculated as in Eq. (9)-(34). For non-binarized alignments, matrix reduction is per Eq. (33).

(4) SCA Toolbox 4.0: Analytical calculations as in Toolbox, but now including sector identification methods as described in Ref. (9). Includes two tutorial with two sample alignments.

(4) SCA Toolbox 5.0: See below.

B. SCA Toolbox 5.0

what's new...

References

[1] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley-Interscience, New-York, 1991.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind source separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[3] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in neural information processing systems*, volume 8, pages 757–763, Cambridge MA, 1996. MIT Press.

[4] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–9, 1999.

[5] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59–69, 2003.

[6] M. E. Hatley, S. W. Lockless, S. K. Gibson, A. G. Gilman, and R. Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proc Natl Acad Sci USA*, 100(24):14445–50, 2003.

[7] A. I. Shulman, C. Larson, D. J. Mangelsdorf, and R. Ranganathan. Structural determinants of allosteric ligand activation in rxr heterodimers. *Cell*, 116(3):417–29, 2004.

[8] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–8, 2005.

[9] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–86, 2009.