

Note 103: A summary of SCA calculations

Olivier Rivoire¹ and Rama Ranganathan²

¹Laboratory of Living Matter, and Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.

²The Green Center for Systems Biology, and Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.

(Dated: August 18, 2008)

I. INTRODUCTION

This document provides a short summary of the principles and implementations of the SCA method. A more thorough description of the assumptions, justifications and open questions related to the method will be given elsewhere in formal publication.

II. PRELIMINARIES - MULTIPLE SEQUENCE ALIGNMENT AND FREQUENCIES

A. 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 computed as the number of sequences in the alignment having amino acid a at position i , divided by the total number of sequences, including those with a gap at i ; it can also be written

$$f_i^{(a)} = \langle x_{i,s}^{(a)} \rangle_s, \quad (1)$$

where $x_{i,s}^{(a)}$ is averaged over all sequences s .

B. Binary approximation

In Halabi et al., (manuscript submitted), we also 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., $x_{i,s} = x_{i,s}^{(a_i)}$). This reduction is useful since, as calculated below (Sec. III), the positional conservation in the full alignment is well-approximated by the positional conservation in the binary alignment. Other approaches to binary approximation are possible and are currently under investigation for improved approximation of positional conservation and correlation.

C. Background frequencies

As described in section III, positional conservation 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,

$$q = (0.073, 0.025, 0.050, 0.061, 0.042, 0.072, 0.023, 0.053, 0.064, 0.089, \\ 0.023, 0.043, 0.052, 0.040, 0.052, 0.073, 0.056, 0.063, 0.013, 0.033),$$

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

D. Gaps

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; alternatively, one can truncate the alignment to the positions present in an atomic structure of a representative member of the protein family. This prevents trivial over-representation of gaps in a sequence alignment and ensures calculations are made only at largely non-gapped sequence positions. Other approaches to defining a background expectation for gaps are possible that can also be consistent with the general approach outlined here.

III. POSITION-SPECIFIC CONSERVATION - FIRST ORDER STATISTICS

A. Relative entropy

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 relative entropy (1) of $f_i^{(a)}$ given $q^{(a)}$. Its 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 $q^{(a)}$:

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

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

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

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

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.

B. Equivalence with previous definitions

$D_i^{(a)}$ 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. (2-5)), and $\Delta E_i^{\text{stat},a}$ (SCA Toolbox v1.5, as reported in Ref. (6)):

$$\Delta G_i^{\text{stat},a} = \Delta E_i^{\text{stat},a} = -\frac{1}{M} \ln P_M[f_i^{(a)}] \simeq D_i^{(a)} \quad (5)$$

The pre-factor $-\frac{1}{M}$ 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.

C. Appropriate alignment sizes

A more precise relation between the probability $P_M[f_i^{(a)}]$ and the relative entropy $D_i^{(a)}$ is

$$-\frac{1}{M} \ln P_M[f_i^{(a)}] = D_i^{(a)} + \frac{\ln M}{2M} + O\left(\frac{1}{M}\right). \quad (6)$$

The values of $D_i^{(a)}$ are typically of order 1-3 (the scale is given by $\ln 20 \approx 3$), as 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; provided one operates above this limit, the previous measurements of conservation are quantitatively close to $D_i^{(a)}$.

D. Overall positional conservation

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. II.D). Denoting $f_i^{(0)} = 1 - \sum_{a=1}^{20} f_i^{(a)}$ the fraction of gaps at position i , we can then write the probability of jointly observing the frequencies $(f_i^{(1)}, \dots, f_i^{(20)})$ of each of the 20 possible amino acids at position i as

$$P_M[f_i^{(1)}, \dots, f_i^{(20)}] = \frac{M!}{(Mf_i^{(0)})! \dots (Mf_i^{(20)})!} (\bar{q}^{(0)})^{Mf_i^{(0)}} \dots (\bar{q}^{(20)})^{Mf_i^{(20)}} \simeq e^{-MD_i} \quad (7)$$

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

E. Relation between overall and amino acid-specific positional conservations

If considering gaps, we also define $\bar{D}_i^{(a)} = f_i^{(a)} \ln \frac{f_i^{(a)}}{\bar{q}^{(a)}} + (1 - f_i^{(a)}) \ln \frac{1 - f_i^{(a)}}{1 - \bar{q}^{(a)}}$, the equivalent of $D_i^{(a)}$ (the positional conservation for amino acid a) corresponding to a background probability distribution that includes gaps. As a rule, $\bar{D}_i^{(a)} \leq D_i$ and in practice, $\bar{D}_i^{(a)}$ is found to be maximal for $a = a_i$, the most frequent amino acid at position i .

F. Equivalence of various definitions in the binary approximation limit

Note that $D_i^{(a)}$, $\bar{D}_i^{(a)}$, and D_i are non-linear functions of $f_i^{(a)}$ that rise more and more steeply as $f_i^{(a)}$ approaches one. A consequence is that for all but the least conserved positions, the overall conservation D_i is well approximated by the conservation of the most prevalent amino acid ($D_i^{(a_i)}$ or $\bar{D}_i^{(a_i)}$), a result that justifies the use of the binary approximation in Halabi et al. (submitted). The same argument indicates that the definition of overall positional conservation introduced in previous reports of the SCA method, namely

$$\Delta G_i^{\text{stat}} = \Delta E_i^{\text{stat}} = \frac{1}{M} \sqrt{\sum_{a=1}^{20} (\ln P_M[f_i^{(a)}])^2}, \quad (8)$$

behaves essentially as $D_i^{(a_i)}$.

IV. CORRELATED CONSERVATION - SECOND ORDER STATISTICS

The basic principle for defining a SCA correlation matrix is to weight correlations between pairs of positions by a function of the positional conservations. Different implementations of the SCA method correspond to different definitions of the weights, but are all based on this same principle (described below). 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. IV.E). The current implementation is based on a bootstrap (or, more precisely, jackknife) procedure, which amounts to using weights that are gradients of positional conservations (Sec. IV.D). With regard to distributions, SCA Toolbox 2.x computes the SCA correlation matrix using the bootstrap, and SCA Toolbox 3.0 using weights. In practice, versions 2.x and 3.0 report nearly identical values.

A. Unweighted covariance matrix

In general, a covariance matrix reporting pair-wise correlations between amino acids at positions in a multiple sequence alignment can be defined as

$$C_{ij}^{(ab)} = \langle x_{i,s}^{(a)} x_{j,s}^{(b)} \rangle_s - \langle x_{i,s}^{(a)} \rangle_s \langle x_{j,s}^{(b)} \rangle_s = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}. \quad (9)$$

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 . The corresponding expression in the binary approximation is

$$C_{ij} = \langle x_{i,s} x_{j,s} \rangle_s - \langle x_{i,s} \rangle_s \langle x_{j,s} \rangle_s = f_{ij}^{(a_i a_j)} - f_i^{(a_i)} f_j^{(a_j)}. \quad (10)$$

B. Weighted covariance matrix

As a general principle, SCA matrices can be obtained by weighting these covariance matrices by a functional ϕ of the positional conservations $D_i^{(a)}$ (or, more generally, a function of the frequencies $f_i^{(a)}$ and $q^{(a)}$ with properties similar to that of $D_i^{(a)}$)

$$\tilde{C}_{ij}^{(ab)} = \phi\left(D_i^{(a)}\right) \phi\left(D_j^{(b)}\right) C_{ij}^{(ab)}, \quad (11)$$

or, in the binary approximation

$$\tilde{C}_{ij} = \phi\left(D_i^{(a_i)}\right) \phi\left(D_j^{(a_j)}\right) |C_{ij}|. \quad (12)$$

Although not essential, the absolute value taken in the last formula eliminates negative correlations that originate from alternative choices of amino acids at a position. In uses of SCA for characterizing positional correlations, the sign of amino acid-specific correlations is not considered. Eqs. (11)-(12) represent the most general description of the SCA matrix - a weighted correlation matrix that measures the significance of amino acid correlations by the conservation of the residues involved. A possible choice of weights, implemented in SCA Toolbox 3.0 and discussed in section IV.D below, is

$$\phi\left(D_i^{(a)}\right) = \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} = \ln \left[\frac{f_i^{(a)}(1 - q^{(a)})}{(1 - f_i^{(a)})q^{(a)}} \right]. \quad (13)$$

These weights rise even more steeply than $D_i^{(a)}$ as the frequencies of amino acids $f_i^{(a)}$ approach one, a property that reduces correlations arising from weakly conserved amino acids (since the gradient of $D_i^{(a)}$ approaches zero as $f_i^{(a)} \rightarrow q^{(a)}$), and emphasizes conserved correlations. This property addresses a central issue in assessing functional correlations in sequence alignments - the need to minimize the contribution of purely historical correlations between positions that derive from many small clades of sequences with close phylogenetic relationships.

C. Reduced SCA matrix and binary approximation

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

$$\bar{C}_{ij} = \sqrt{\left(\sum_{a,b} \left(\tilde{C}_{ij}^{(ab)} \right)^2 \right)} \quad (14)$$

Within the range of validity of the binary approximation, this matrix corresponds to \tilde{C}_{ij} and therefore yields equivalent results. Other approaches to reduction of the four-dimensional tensor $\tilde{C}_{ij}^{(ab)}$ to the positional correlation matrix \tilde{C}_{ij} are possible but, as long as the binary approximation is justified, these are expected to give similar results.

D. Weights derived from the bootstrap procedure

If we introduce $D_{i,s}^{(a)}$, the positional conservation of amino acid a at position i for an alignment obtained by leaving out sequence s , the covariance matrix associated with this bootstrap procedure is:

$$\hat{C}_{ij}^{(ab)} = \langle D_{i,s}^{(a)} D_{j,s}^{(b)} \rangle_s - \langle D_{i,s}^{(a)} \rangle_s \langle D_{j,s}^{(b)} \rangle_s. \quad (15)$$

This approach is implemented in the MATLAB distribution SCA v2.0. The $\hat{C}_{ij}^{(ab)}$ matrix can also be analytically derived by rewriting it as a weighted correlation matrix (7). 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 sequences s is left out, the frequency $f_i^{(a)} = M_i^{(a)}/M$ becomes

$$f_{i,s}^{(a)} = \frac{M_i^{(a)} - x_{is}^{(a)}}{M - 1} = \left(1 + \frac{1}{M} \right) f_i^{(a)} - \frac{x_{is}^{(a)}}{M} + O\left(\frac{1}{M^2} \right), \quad (16)$$

where $x_{i,s}^{(a)} = 1$ 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

$$D_{i,s}^{(a)} \approx \hat{D}_i^{(a)} - \frac{x_{i,s}^{(a)}}{M} \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}}, \quad (17)$$

where $\hat{D}_i^{(a)}$ is the relative entropy $D_i^{(a)}$ with $f_i^{(a)}$ replaced by $(1 + 1/M)f_i^{(a)}$. It thus follows that, to first order in $1/M$,

$$\hat{C}_{ij}^{(ab)} \approx \frac{1}{M^2} \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} \frac{\partial D_j^{(b)}}{\partial f_j^{(b)}} \left(\langle x_{i,s}^{(a)} x_{j,s}^{(b)} \rangle_s - \langle x_{i,s}^{(a)} \rangle_s \langle x_{j,s}^{(b)} \rangle_s \right). \quad (18)$$

Or, per Eq. (9)

$$\hat{C}_{ij}^{(ab)} \approx \frac{1}{M^2} \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} \frac{\partial D_j^{(b)}}{\partial f_j^{(b)}} C_{ij}^{(ab)} \quad (19)$$

The bootstrap procedure thus corresponds to weighting the raw covariance matrix $C_{ij}^{(ab)}$ by gradients of positional conservation (compare Eq. (19) with Eqs. (11)-(13)). A scaling factor of $1/M^2$ relates the bootstrap and weighting approaches to the SCA correlation matrix:

$$\hat{C}_{ij}^{(ab)} \approx \frac{1}{M^2} \tilde{C}_{ij}^{(ab)}. \quad (20)$$

E. Weights derived from the original perturbation procedure

The implementation of the SCA method introduced originally in Lockless and Ranganathan 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 a highly 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:

$$\Delta\Delta G_{j,i}^{\text{stat},b,a_i} = \Delta\Delta E_{j,i}^{\text{stat},b,a_i} = -\frac{1}{M} \left[\ln \left(P_M \left[f_j^{(b)} \right] \right) - \ln \left(P_M \left[f_{j|i}^{(b)|a_i} \right] \right) \right], \quad (21)$$

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 . $\Delta\Delta G_{j,i}^{\text{stat},b,a_i}$ 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 first term on the right hand side, $-\frac{1}{M} \ln \left(P_M \left[f_j^{(b)} \right] \right)$, corresponds to $D_j^{(b)}$. Given the assumption 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), $f_{j|i}^{(b)|a_i} \approx f_j^{(b)}$ 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

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

with $C_{ij}^{(a_i,b)}$ defined as in Eq. (9), so that

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

This leads to

$$\Delta\Delta G_{j,i}^{\text{stat},b,a_i} \approx -\frac{1}{f_i^{(a_i)}} \frac{\partial D_j^{(b)}}{\partial f_j^{(b)}} C_{ij}^{(a_i,b)}, \quad (22)$$

which shows that the perturbation procedure also corresponds to a weighting procedure for correlations. Finally, within the range of validity of the binary approximation,

$$\Delta\Delta G_{j,i}^{\text{stat}} = \sqrt{\sum_{b=1}^{20} (\Delta\Delta G_{j,i}^{\text{stat},b,a_i})^2} \approx \left| \Delta\Delta G_{j,i}^{\text{stat},a_i,a_i} \right| \approx \frac{1}{f_i^{(a_i)}} \frac{\partial D_j^{(a_j)}}{\partial f_j^{(a_j)}} |C_{ij}|. \quad (23)$$

V. DISTRIBUTIONS OF SCA

(1) SCA v1.5: The original SCA method as specified in Lockless and Ranganathan (2) 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 v2.5: The bootstrap-based approach for SCA. Position-specific conservation calculated as in Eq. (4) and correlations calculated as in Eq. (11). Matrix reduction per Eq. (14).

(3) SCA v3.0: 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. (11)-(12). An update to this version is expected shortly that will also includes codes for new statistical methods for identifying groups of correlated amino acid positions (the "sectors" in Halabi et al., submitted) and for assessing the statistical independence of sectors. For non-binarized alignments, matrix reduction is per Eq. (14). Current version.

Distributions are MATLAB Toolboxes that include various accessory codes for data formatting, display, and analysis through hierarchical clustering. A tutorial with a sample alignment that illustrates the analytic process is also provided.

References

- [1] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New-York, 1991.
- [2] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–9, Oct 1999.
- [3] Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59–69, Jan 2003.
- [4] Mark E Hatley, Steve W Lockless, Scott K Gibson, Alfred G Gilman, and Rama Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proc Natl Acad Sci USA*, 100(24):14445–50, Nov 2003.
- [5] Andrew I Shulman, Christopher Larson, David J Mangelsdorf, and Rama Ranganathan. Structural determinants of allosteric ligand activation in rxr heterodimers. *Cell*, 116(3):417–29, Feb 2004.
- [6] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–8, Sep 2005.
- [7] B. Efron and R. J. Tishirani. *An introduction to the bootstrap*. Chapman and Hall, 1994.