Our implementation of the SCA method

Introduction

The Statistical Coupling Analysis (SCA) method is one of the earliest and most popular methods for measuring the coevolution of pairs of sites. It was first described in Lockless and Ranganathan (1999). We based our implementation of the SCA method on the description in this article, as well as the description in Suel et al. (2003), and its web supplement at http://www.hhmi.swmed.edu/Labs/rr/SCA.html. In the following we will call them "the reference sources". We have also referenced the Matlab software of the original authors (SCA version 1.5) for some implementation details.

Since not all the algorithmic details are given in the reference sources, and we need to fit our implementation into the overall software framework, we have made a number of design choices. We have made our best effort in having the choices reasonable and close to the original definitions in the reference sources. Yet we have to stress that our implementation is not completely the same as the one of the original authors, and users of our system should be aware of the details of our implementation, which we describe below.

We have also referenced Dekker et al. (2004) since the authors also implemented the SCA method and made some design choices. However, our choices are not exactly the same as theirs.

We have discussed our design with some of the SCA inventors (Rama Ranganathan and William Russ). We follow their suggestion to produce a non-square, non-symmetric SCA matrix as was done in the original SCA papers (based on the details described below, which are close to, but not completely the same as their SCA software), with each row being a site and each column a perturbation. Then, to produce one single coevolution score for each pair of sites, we use a statistic to summarize the validated scores of the different perturbations. To distinguish this final statistic from the original perturbation scores, we call the former "Modified SCA" to emphasize its difference from the latter.

Design choices

- The constant kT^*: the exact value of the temperature parameter T^* is not described in

the three reference sources. Since kT^* is a constant in all calculations, and it is described as "an arbitrary energy unit" in both Lockless and Ranganathan (1999) and Suel et al. (2003), we fix its value to one. We remind users who want to compare the SCA scores of different MSAs to perform proper normalizations according to the number of sequences in the MSAs. (Note: in the following we assume some data preprocessing may have been performed. So for example when we talk about an MSA, we actually refer to the preprocessed MSA that may have some sequences and/or sites filtered.)

- Normalization by P_{MSA}^{x}: it is mentioned in the reference sources, and implemented in the original code used to produce the results of the reference sources. However, as mentioned in Dekker et al. (2004), and confirmed by the SCA inventors (personal communication), it is not implemented in the software package. Since this normalization is not important according to the SCA inventors (personal communication), it is not included in our implementation.
- Acceptance criteria: the three reference sources emphasize that the multiple sequence alignments (MSAs) used in the analysis should be large enough and diverse enough to be a statistical representative of the evolutionary constraints on the protein family. These concerns are partially handled by the various data preprocessing methods of our system. But in order to have our implementation of the SCA method as close to the original definitions as possible, we also implement the acceptance criteria described in the three reference sources.

  The most detailed description of the acceptance criteria is in the web supplement of Suel et al. (2003). It lists three acceptance criteria:
  - "The MSA should be so diverse that several sites display amino-acid distributions near to the mean in all natural proteins."
  - "The MSA should be so large that random elimination of sequences from the alignment does not change the amino-acid frequencies at sites much."
  - "Perturbations at sites in the MSA should produce sub-alignments that are also large and diverse, such that they remain a representative subset of the parent MSA and do not substantially alter the state of statistical equilibrium."

  We were unable to precisely locate the corresponding implementation details in the Matlab code in a form that we could re-implement. Therefore we designed our acceptance procedure according to the above criteria as follows:

1. We first calculate the statistical energy vectors $\overrightarrow{\Delta G_j^{stat}}$ [1] for all sites j at which at least 85% of the sequences are not a gap. The five sites with the smallest magnitudes (2-norms) of $\overrightarrow{\Delta G_j^{stat}}$ (i.e., the five most unconserved sites) are chosen to have the magnitudes averaged. If less than five sites pass the 85% requirement, the MSA is rejected as having too many gaps. If the average magnitude is more than a threshold α (default to 1.0), the MSA is rejected as being too conserved.
2. Using the five sites in step 1, we perform random sequence elimination. It is divided into iterations. At the i-th iteration, the number of sequences to be eliminated is $0.03 \times i \times n$, where n is the total number of sequences in the MSA. 100 random eliminations are performed per iteration, and the average of the averaged $\overrightarrow{\Delta G_j^{stat}}$ of the five sites are recorded. If the averaged $\overrightarrow{\Delta G_j^{stat}}$ reaches α before β% (default to 50) of the sequences in the MSA are eliminated, the MSA is rejected as being too small.
3. Using the five sites in step 1, we perform another random sequence elimination to determine the size threshold for a perturbation. The same number of sequences is eliminated in each iteration, but instead of recording $\overrightarrow{\Delta G_j^{stat}}$, we treat each random elimination as a perturbation, and record the average coupling energies $\Delta\Delta G_{ij}^{stat}$ over the 100 random eliminations of the iteration. When the average $\Delta\Delta G_{ij}^{stat}$ reaches γ (default to 0.07), the number of sequences not eliminated in the iteration is set as the size threshold. When calculating the perturbation scores, any perturbation that results in less remaining sequences than the threshold is rejected.

- Perturbation and symmetry: in the original SCA implementation, a perturbation is performed by retaining one of the 20 amino acids at a site. Therefore for a pair of sites, there are in total 2 x 20 = 40 possible perturbations. All perturbations passing the acceptance criteria are recorded. The final SCA matrix is a rectangular matrix

---

[1] As in SCA 1.5, both $\overrightarrow{\Delta G_j^{stat}}$ and $\Delta\Delta G_{ij}^{stat}$ are divided by 100 after the calculations described in the reference sources.

with each site as a row and each perturbation (a site number-residue pair) as a column.

Our system outputs this original SCA matrix whenever the SCA method is chosen. In order to fit in the overall framework with one single coevolution score for each site pair, we also use a statistic to summarize the (at most 40) non-rejected perturbation scores into one final coevolution score, which we call the “modified SCA score”.

Let SCA(i, j, a) be the original SCA score between sites i and j by retaining only residue a at site j, our modified SCA score is calculated as $\frac{1}{2n}\left[\sum_{a:n_{j,a}\geq t} n_{j,a} SCA(i,j,a) + \sum_{a:n_{i,a}\geq t} n_{i,a} SCA(j,i,a)\right]$, where n is the number of sequences in the MSA, $n_{i,a}$ is the number of sequences having residue a at site i, and t is the size threshold determined by the acceptance criterion. Basically, the score is a weighted sum of the SCA scores of the various perturbations, but taking into account only the non-rejected ones.

References

Dekker J. P. et al. (2004) A Perturbation-based Method for Calculating Explicit Likelihood of Evolutionary Co-variance in Multiple Sequence Alignments. *Bioinformatics* 20(10) 1565-1572.

Lockless S. W. and Ranganathan R. (1999) Evolutionarily Conserved pathways of Energetic Connectivity in Protein Families. *Science* 286 295-299.

Suel G. M. et al. (2003) Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nature Structural Biology* 10(1) 59-69.