Jeremy Chang

Identifying protein-protein interactions with statistical coupling analysis

Abstract: We used an algorithm known as statistical coupling analysis (SCA)[1] to create a set of features for building a biologically relevant and evolutionarily stable protein interaction network. The heart of the algorithm is to use protein sequence information from different organisms to determine which residues of a protein coevolve: the idea is that coevolving proteins probably also interact in a way that is important for an organism's survival. We have concatenated the sequences of three proteins (two of which are known to interact with each other; the third was a negative control) and, treating them as a single protein sequence, used SCA to see which residues coevolve. We found that the SCA analysis on our sequence data produced noisy results and we were not able to identify significant interactions among the proteins. We conclude with suggestions on how to improve this method.

Introduction: One of the central challenges of biology is creating a map of all biologically relevant protein-protein interactions. High-throughput screens may be useful in giving a set of possible interactions, but it is unclear which interactions are actually important parts of an organism's survival.

One method, known as statistical coupling analysis, has been shown to be successful in singling out the most important interactions within a protein[2,3]. The method is based on a multiple sequence alignment of orthologs of a given protein across many species. The statistical coupling between two sites is given essentially by how much the probability for finding a certain amino acid site A changes upon fixing the amino acid at site B to a given amino acid. In order to extend this technique proteome-wide, we will concatenate the sequences of many proteins and treat them as a single superprotein, and then perform SCA on this superprotein. What follows is a brief introduction to SCA; for further explanation, please consult Ranganathan (2003).

Consider the frequency distribution of amino acids at a given position (say, position *i*) in our superprotein (for example, position *i* could be valine 30% of the time and cystein 70% of the time in our multiple sequence alignment). Let's say another site in the protein (site *j*) also has a given distribution of amino acids. If we now examine only the sequences in which the amino acid at position *i* is cystein, the distribution at position *j* may or may not change. If it changes dramatically, this would imply that the two sites are statistically coupled; if not, they are not coupled.

We can quantify the amount of coupling with the following formula:

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

Here, $ddG^{stat}_{i,j}$ is the statistical coupling energy between positions *i* and *j* in the multiple sequence alignment. $P^{x}_{i|dj}$ is the probability of observing amino acid *x* at position *i*, given a perturbation *dj* (i.e. selecting only the sequences where a certain position is fixed to a certain amino acid). $P^{x}_{MSA|dj}$ is the probability of observing amino acid *x* across the total multiple sequence alignment. P^{x}_{i} and P^{x}_{MSA} are the equivalent probabilities without the perturbation. kT^{*} is an arbitrary energy value, which we have used for normalization.

We applied SCA to an MSA of three proteins: MAP kinase 1 (MAPK1), MAP kinase kinase 1 (MAP2K1), and p38 MAP kinase (p38) in order to determine which residues were statistically coupled.

Methods and results: We downloaded sequences for our three proteins of interest from PSI-BLAST by searching for the mouse versions. We removed duplicate homologs from a given species simply by randomly selecting one to keep, and removing the remaining sequences. We then did a multiple sequence alignment on each of the proteins separately, and then concatenated the resulting alignments by matching up the sequences by species. This resulted in an alignment with 110 orthologs (from 110 species), with a total length of 3739 amino acids.

We then proceeded with the SCA by identifying 39 perturbations. The criteria for selecting perturbations were that there must be at least 20 sequences that contain a certain amino acid at a certain site and, after the perturbation, there must be at least 20 sequences remaining. We performed SCA with these perturbations, and finally produced the graphs displaying *ddG* for each position given a perturbation.

In the MSA, MAPK1 is located at amino acids #1-1155; MAP2K1 is located at #1156-2568, and p38 is located at #2568-3739. We found that the average *ddG* values did not change significantly across these three proteins, indicated that we did not identify any significant interactions. Each plot in Figure 1 shows the *ddG* values across all positions for a given perturbation.

We had hoped to be able to (1) cluster our results to identify amino acids that coevolve across all perturbations and (2) use supervised learning techniques to train a classifier on what properties the matrix of *ddG* values of two interacting proteins has, but because the *ddG* values for all proteins had similar values, we decided not to proceed.

Conclusions: We believe that the main reasons we were unable to recover significant interactions were due to the original quality of the multiple sequence alignment. To improve it, we would need (1) a larger number of sequences and (2) cleaner sequences, in the sense that PSI-BLAST returns sequences for proteins that are not necessarily orthologous to the query sequence, and therefore some of our MSA is in fact composed of unrelated, junk proteins that happen to share sequence homology with our query sequence.

The alignment process itself can also be improved. The SCA algorithm assumes that each position in the alignment refers to the same 3D position in all of the amino acids. Without adjusting the alignments so that they correspond to the (most likely) 3D structure, we can't be confident in the results from SCA.

Finally, it may be worth trying to use mutual information as our coevolution metric, rather than SCA. Both methods have given similar results, and it is unclear which one gives better results.

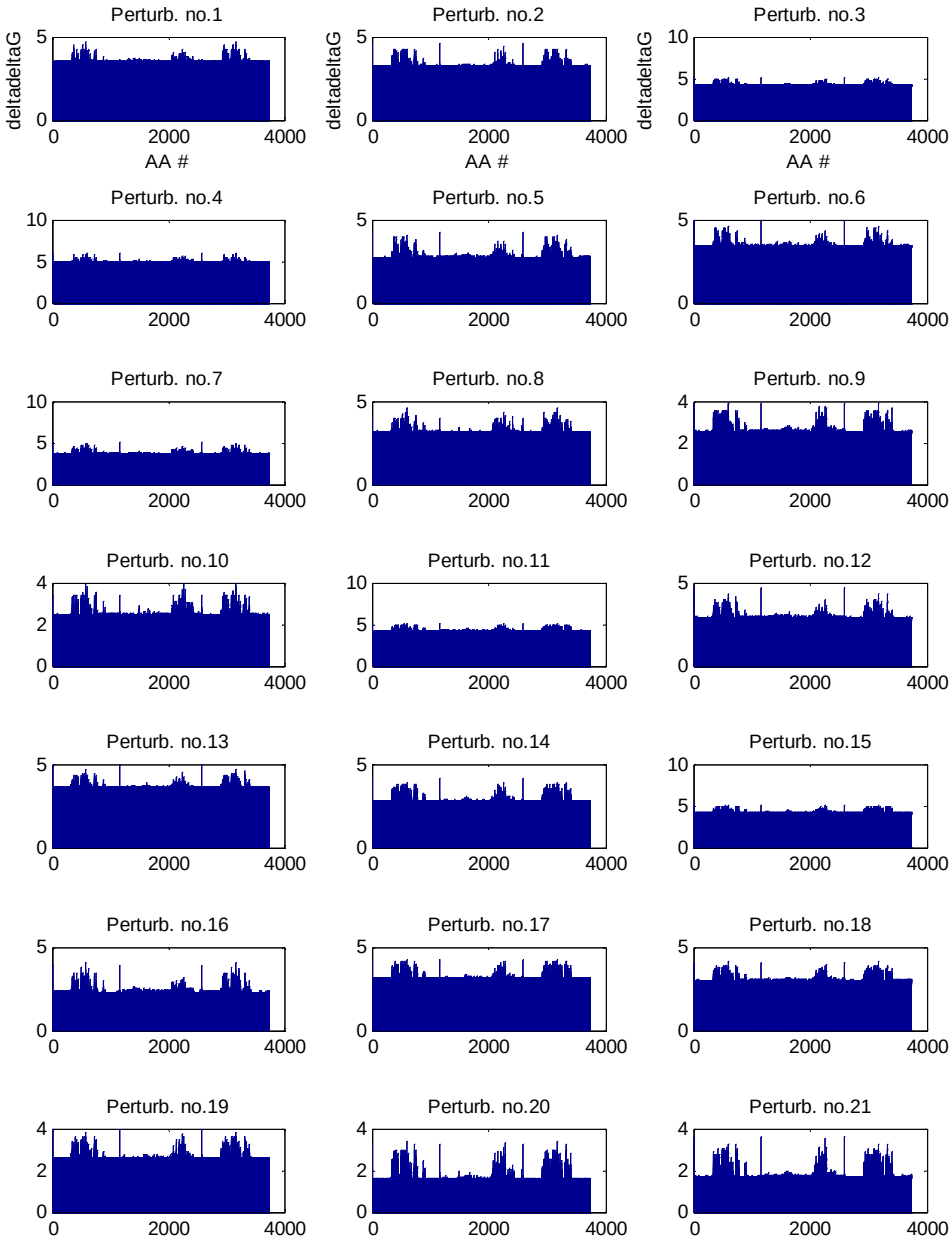
Please email Jeremy Chang (jbchang@stanford.edu) for the multiple sequence alignment and/or MATLAB code used in this project.

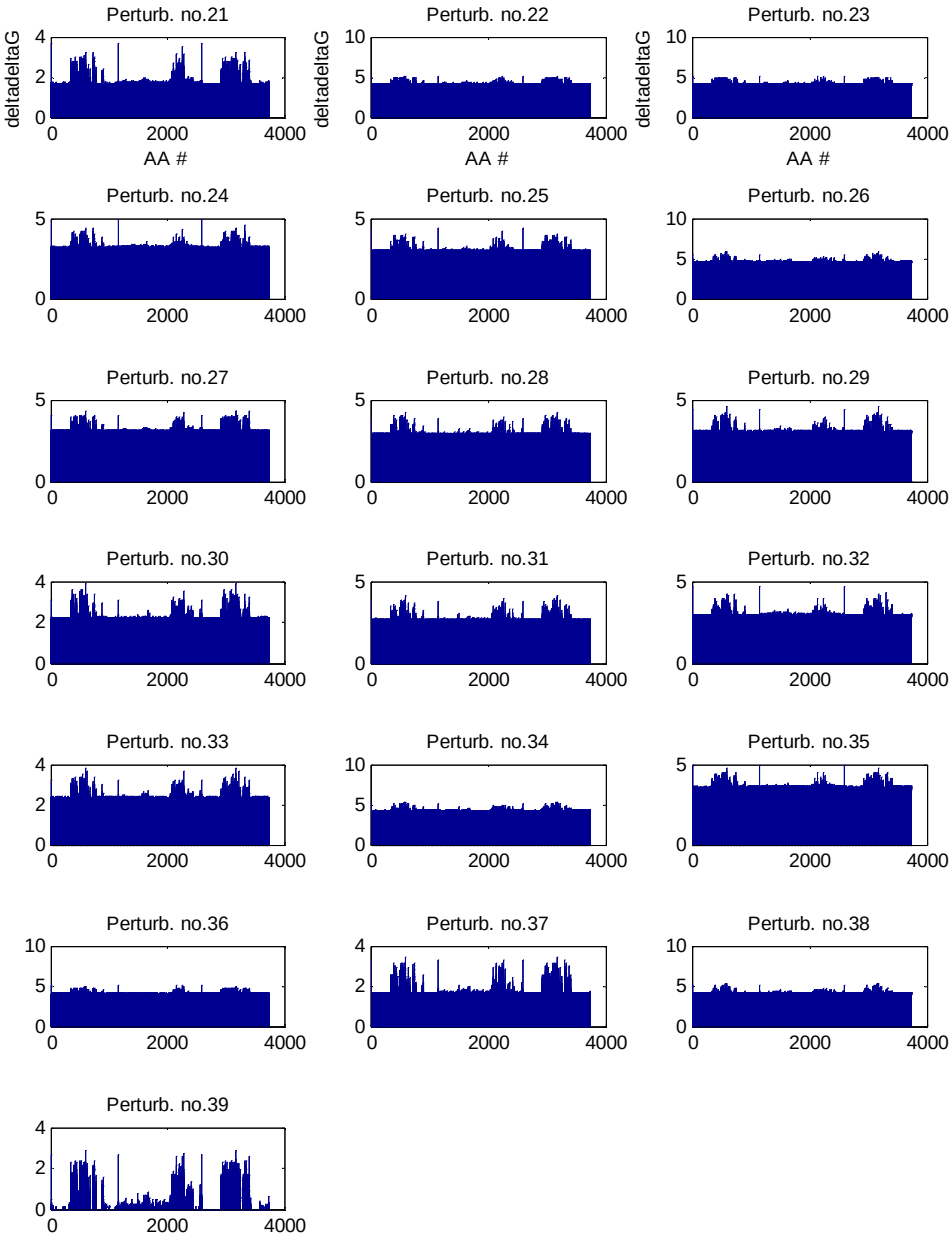
References

1. Lockless, S.W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-9(1999).

2. Socolich, M. et al. Evolutionary information for specifying a protein fold. *Nature* **437**, 512-8(2005).
3. Russ, W.P. et al. Natural-like function in artificial WW domains. *Nature* **437**, 579-83(2005).
4. Zhang, X. et al. An Allosteric Mechanism for Activation of the Kinase Domain of Epidermal Growth Factor Receptor. *Cell* **125**, 1137-1149(2006).
5. Ben-Hur, A. et al. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol* **4**, e1000173(2008).

Figure 1 (two pages)




Perturb. no.21
Perturb. no.22
Perturb. no.23
Perturb. no.24
Perturb. no.25
Perturb. no.26
Perturb. no.27
Perturb. no.28
Perturb. no.29
Perturb. no.30
Perturb. no.31
Perturb. no.32
Perturb. no.33
Perturb. no.34
Perturb. no.35
Perturb. no.36
Perturb. no.37
Perturb. no.38
Perturb. no.39
deltadeltaG
AA #
0
2000
4000