

Combined Statistical Methods for Describing Interacting Protein Networks

Jeremy Purvis

May 12, 2006

I recently came across a sixty-year-old paper published in *The American Naturalist* entitled, “A Statistical Evaluation of the Kinship of Protein Molecules.” In this short letter [1], the authors present an ‘index of correlation¹’ that describes the variation in amino acid composition among proteins:

$$I = \frac{\sigma_A^2}{\sigma_P^2 + \sigma_A^2}. \quad (1)$$

Using the modest sequence data available in 1955, the analysis included just twenty-four proteins. While this collection was enough to observe ‘a discernibly high degree of kinship’ among proteins, it is hardly enough to make any of the sweeping generalizations and projections that now typify the field of genome science.

In my studies, I am concerned with making predictions of this scale, involving large networks of interacting proteins [2]. In addition to knowing which proteins are expressed in a cell, their relative abundance, and in which reactions they participate—I am particularly interested in *where* these proteins are spending the most time, that is, how are they distributed spatially within a cell. One way to address this question is to ask which pairs of proteins have interacting domains, following the logic that molecules that ‘fit’ together are likely to co-localize. To this end, many studies have identified putative protein-protein or domain-domain or protein-domain interactions through analysis of whole genomes [3, 4]. Other studies have focused on finding specific signal sequences in proteins that target them for subcellular localization [5]. In each case, large data sets and recurring trends have

¹The index compares the variance in residue composition among all the proteins, σ_P^2 , to the variance in amino acids observed in proteins from the natural distribution of amino acids, σ_A^2

necessitated the use of statistical methods to make meaningful predictions about protein interactions. Here, I discuss a statistical method that identifies putative binding motifs within genomes. Then, I propose a hypothetical model² that uses this information to produce a spatially-sensitive model of protein-protein interactions.

Identifying Structural Interactions. Ideally, we would like to predict structural interactions from primary sequence. One approach is to look for ‘recognition motifs’ that match a statistical profile. The profile may consist of individual residue positions that are scored according to their binding properties. For this purpose, libraries of specific peptides or peptide-phage fusion proteins are systematically analyzed one position at a time to determine the sequence specificity of a particular signaling domain. These so-called oriented peptide libraries were first used by Songyang and colleagues to find cognate recognition partners for the Src homology 2 (SH2) domain [6]. The binding preference, or *selectivity*, of each amino acid was determined by sequencing high-affinity binding peptides eluted from a column containing the N-terminal SH2 domain of p85 (the method assumes that each residue binds independently of adjacent amino acids). Relative values of each residue are measured, and these preference values are used to generate a scoring matrix that reflects the relative binding strength of individual amino acids at each position in the motif. Results from a typical experiment are shown in Fig. 1.

Developing a Scoring Matrix. Using the method described above, raw selectivity values for each residue are calculated as the ratio:

$$\frac{\text{Mole \% high-affinity binding}}{\text{Mole \% expected}}. \quad (2)$$

Values greater than one represent preferred (high-affinity) residues; values close to zero represent residues that bind relatively poorly to the domain. Raw scores are rescaled³ into selectivity values χ_i , which are converted to bit scores $\ln \chi_i / \ln 2$ for each position i . Bit scores are used to construct a motif-specific position-dependent amino acid matrix, reminiscent of the scoring matrices used for BLAST analysis [8].

²Given the exploratory nature of this hypothesis, one should forgive any erroneous thinking or formulations in the proposed model.

³Like BLAST matrices, the selectivity scoring matrices are often manually curated to correct for numerical complications. This rescaling has been deemed acceptable so long as the absolute rank order is maintained (see supplementary material in [7]).

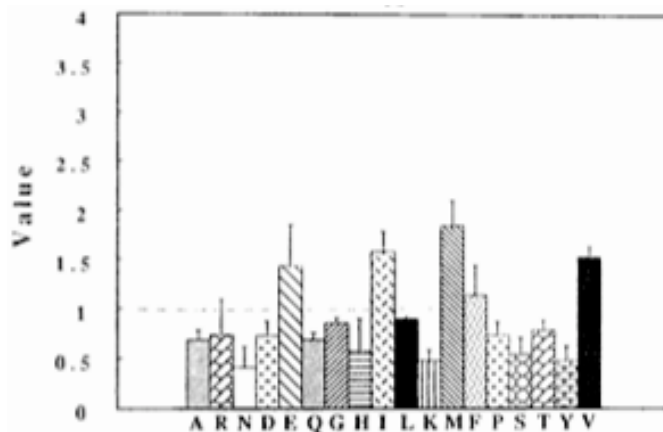


Figure 1: | **Phosphopeptides that bind the SH2 domain.** Degenerate phosphopeptides are run through a column containing the SH2 domain. Bound peptides are eluted and sequenced. Bars in the figure represent the ratio of each amino acid eluted from the SH2 column divided by that of a control column [6].

Putting these matrices to work, Yaffe *et al* devised a profile-scanning algorithm to find top-ranking signaling motifs [7]. The algorithm first looks for ‘critically invariant’ residues (e.g., serine or threonine kinase phosphorylation sites) and then scores the surrounding site using a 15-residue window centered about the invariant residue. A raw sequence score $S_{raw} = \sum(\ln \chi_i / \ln 2) / i$ is calculated for each putative site and compared to the optimal score S_{opt} to give the final sequence score $S_f = (S_{opt} - S_{raw}) / S_{opt}$.

We now have a numerical quantity that is an indicator of binding preference among proteins. Extending this thought, the scores may also be used to identify *groups* of proteins which share a common binding partner. In cells, such domains are often docking sites that tether one or more species together to form multimeric complexes. Associations of proteins are not only important for complex functions like signal integration, but also serve to increase the effective concentration of proteins which congregate near the docking domain, significantly altering the reaction kinetics [9].

Probabilistic Interaction Model. We now consider a collection of n interacting proteins $\mathbf{P} = \{P_1, P_2, \dots, P_n\}$ for which the aforementioned statistical rankings are available. Using these rankings, we may set an arbitrary

cutoff percentile to produce a subset of proteins that are likely to share a common binding partner. This procedure may be repeated for multiple recognition motifs k , so that each protein i in the collection is assigned a value for every other protein j , reflecting the extent to which the two species interact favorably:

$$H_{ij} = \sum_k S_i \ln \frac{S_i}{S_j} + \sum_k S_j \ln \frac{S_j}{S_i} \quad (3)$$

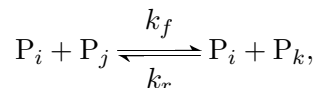
H is essentially a distance measure [8] and will produce a total of $n(n-1)/2$ graph edges for the entire protein collection. When two protein species interact, they spend a certain amount of time in close proximity. This time can range from zero (perfectly elastic collision) to infinity (irreversibly bound). The task is to relate this time of interaction to the structural complementarity of the reacting species described earlier.

Let us now assume, for the sake of simplicity, that the reacting protein species are initially distributed uniformly throughout the cell, where they may diffuse freely. In the model, we treat the number of interaction events between species i and j as a Poisson random variable X_{ij} that occurs at rate λ_{ij} . The probability of observing k interactions in one unit of time is $P(X_{ij} = k) = \lambda_{ij}^k e^{-\lambda_{ij}} / k!$. Moreover, the probability that the next interaction occurs within time τ is

$$P(W_{ij} \leq \tau) = 1 - e^{-\lambda_{ij}\tau} \quad (4)$$

where W_{ij} is the waiting time until a reaction between i and j occurs.

We now implement a kinetic Monte Carlo scheme in which the number of protein species in \mathbf{P} changes according to previously determined (or estimated) reaction kinetics. A typical equation for each reaction is



although alternate reaction schemes are also possible. The probability of moving from state \mathbf{P}_1 to state \mathbf{P}_2 is defined by a transition matrix \mathbf{T} constructed from rate constants k_f and k_r [10]. This method differs from traditional kinetic Monte Carlo, however, in that the probability of interaction between two species is sampled from the distribution of selectivity values described earlier, rather than from the concentration of the two species alone. In this manner, reactions between ‘nearby’ species have greater probability

density and are thereby more likely to occur.

The algorithm first chooses a discrete random value R distributed uniformly over $(1, n)$. R represents the next reaction that will be observed in the network model. A second random value τ is drawn from a continuous uniform distribution over $(0, \tau_{max})$ where τ_{max} is optimized so that half of the hypothetical reactions are accepted (explained below). The probability that reaction R occurs in time τ is

$$P(i \text{ and } j \text{ react}) = 1 - e^{-H_{ij}\lambda_{ij}\tau} \quad (5)$$

A uniformly distributed random value r_1 is then compared to P : if r_1 is less than P , the reaction is accepted and the collection vector \mathbf{P} is updated to reflect the new protein quantities; if r_1 is greater than P , the reaction is not accepted and the algorithm is repeated.

The statistical model described above may be useful for performing stochastic simulations of complex mixtures of interacting proteins. The method employs a kinetic Monte Carlo scheme, in which individual reactions are given preference according to genomic features of the protein species involved. This offers an improvement over many stochastic simulations in which the reaction species are assumed to be homogenous, or ‘well-mixed’, throughout the system.

References

- [1] S. W. FOX and P. G. HOMEYER, *The American Naturalist* **LXXXIX**, 163 (1955).
- [2] M. PELLEGRINI, D. HAYNOR, and J. M. JOHNSON, *Expert Review of Proteomics* **1**, 239 (2004).
- [3] Y. LIU, N. LIU, and H. ZHAO, *Bioinformatics* **21**, 3279 (2005).
- [4] M. DENG, S. MEHTA, F. SUN, and T. CHEM, *Genome Research* **12**, 1540 (2002).
- [5] G. SCHNEIDER and U. FECHNER, *Proteomics* **4**, 1571 (2004).
- [6] Z. SONGYANG, S. E. SHOELSON, M. CHAUDHURI, G. GISH, T. PAWSON, W. G. HASER, F. KING, T. ROBERTS, S. RATNOFSKY, R. J. LECHLEIDER, B. G. NEEL, R. B. BIRGE, J. E. FAJARDO, M. M. CHOU, H. HANAFUSA, B. SCHAFFHAUSEN, , and L. C. CANTLEY, *Cell* **72**, 767 (1993).

- [7] M. B. YAFFE, G. G. LEPARC, J. LAI, T. OBATA, S. VOLINIA, and L. C. CANTLEY, *Nature Biotechnology* **19**, 348 (2001).
- [8] W. J. EWENS and G. R. GANT, *Statistical Methods in Bioinformatics*, Springer, second edition, 2005.
- [9] M. DIVERSÉ-PIERLUISSI, *Sci. STKE* **2006**, tr4 (2006).
- [10] A. R. LEACH, *Molecular Modeling: Principles and Applications*, Prentice Hall, second edition, 2001.