Identification of residue contacts in protein–protein interaction from multi-species data

Weigt M[1,2], White RA[2,3], Szurmant H[3], Hoch JA[3], Hwa T[2]

[1] Institute for Scientific Interchange, Torino
[2] Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA
[3] Division of Cellular Biology, The Scripps Research Institute, La Jolla, CA

Motivation

The large majority of cellular functions are executed and controlled by interacting proteins. With up to several thousand types of proteins expressed in a typical bacterial cell at a given time, their concerted specific interactions regulate the interplay of biochemical processes that are the essence of life. Many protein interactions are transient, allowing proteins to mate with several partners or travel in cellular space to perform their functions. Understanding these transient interactions is one of the outstanding challenges of systems biology. The characterization of the molecular details of the interface formed between known interacting proteins is a requirement for understanding the molecular determinants of protein-protein interaction, the knowledge of which may be important for a variety of applications including synthetic biology, e.g., designing new specific interaction between proteins, and pharmaceutics, e.g., protein interaction surfaces as drug targets. Experimental approaches to identify surfaces of interaction between proteins such as surface-scanning mutagenesis and cocrystal structure generation are arduous and/or serendipitous. Cocrystal structures provide the best molecular resolution but are particularly challenging to obtain for transient interaction partners. In addition, independent evidence is required to ensure that the structure reflects an accurate picture of the physiologically relevant interaction.

Methods

Given the challenges of these experimental approaches, it is clear that the comprehensive identification of interaction surfaces between a large number of cellular proteins may be significantly expedited by computational methods. The availability of large protein databases generated from sequences of hundreds of bacterial genomes enables various statistical approaches to this problem. In this context covariance-based methods have been used to identify correlations between amino acid positions in interacting proteins. Such methods rely on the premise that amino acid substitution patterns between interacting residues are constrained. To maintain protein function (in this case the protein-protein interaction), the acceptance of a deleterious substitution at one position must be compensated for by substitution(s) in the residue(s) interacting with it, leading to correlations between residues in contact at the interaction surface. However, the covariance approach has a number of shortcomings that may significantly affect its predictive power. One important problem stems from the fact that correlation in amino acid substitution may arise from direct as well as indirect interactions. Traditional covariance analysis is an inherently local one, and cannot distinguish between directly and indirectly correlated residues. We therefore developed a new method that combines covariance analysis with global statistical model inference, adopted from use in statistical physics. Even is the resulting problem is algorithmically hard, it has been solved approximately by using messagepassing techniques.

Results

Applied to a set of >6,900 representatives of the bacterial two-component signal transduction system, the combination of covariance with global inference successfully and robustly identified residue pairs that are proximal in space without resorting to ad hoc tuning parameters, both for heterointeractions between sensor kinase (SK) and response regulator (RR) proteins and for homointeractions between RR proteins. The infered contact pairs are used for de novo prediction of the (currently unknown) SK/RR complex. Further more, the infered statistical model sheds light on the nature of specificity of the SK/RR interaction: It is able to explain the outcome of specificity switching experiments, and therefore elucidates the molecular recognition code between bacterial signaling proteins. The spectacular success of our approach illustrates the effectiveness of the global inference approach in identifying mechanisms and structure of protein-protein interaction based on sequence information alone.

Contact : weigt@isi.it