A Bayesian approach to systems biology: Measuring evolutionary constraints as protein properties reflecting underlying mechanisms

Andrew F. Neuwald

Cold Spring Harbor Laboratory

Though we view the properties of specific atoms and of the interactions between them as universal throughout the universe, biological phenomena are inherently stochastic in nature. Modeling DNA polymerase is fundamentally different from modeling the hydrogen atom, for example, because DNA polymerases vary both between species and within a population of the same species. Furthermore, the classical scientific method, in which one iteratively tests and refines a hypothesis until convergence (i.e., until widespread acceptance), is inadequate for understanding complex and highly correlated biological systems. There is simply not enough time to test the many feasible hypotheses that one can postulate for a complex system, and reductionism, in which the function of the whole is viewed as the sum functions of its parts, fails for highly correlated systems. The latter has played a key role in the rise of systems biology, which seeks to characterize the interplay between biological components.

To help address these issues, I outline a Bayesian approach to biological research where, instead of iteratively refining a single hypothesis or model, one iteratively applies Bayes' theorem to a space of hypothetical models (Neuwald and Liu, 2005). Suppose, for example, that we seek to understand the overall mechanism underlying DNA clamp loading by a protein complex called replication factor C. We may visualize the set of hypothetical mechanisms using a simple Venn diagram with each point representing a specific mechanism and with circles representing sets of mechanisms consistent with various sources of data, such as biochemical, genetic, sequence and structural data. The intersection of the circles then corresponds to those mechanisms most consistent with this data. Going from this Venn diagram to a Bayesian approach involves computing the joint probability distribution for each point using Bayes' theorem. Of course, the most interesting region in this distribution corresponds to mechanisms with high posterior probabilities. Typically one is unable to narrow the field down to a single, highly probable mechanism. Nevertheless, this approach will strongly or weakly favor certain mechanisms, disfavor others, and totally exclude others still. This can help guide further experimentation, leading in turn to recomputation of the posterior distribution and thus to further pruning of hypothetical mechanisms, such that our understanding of actual clamp loading mechanisms becomes increasingly clear over time. Given the nature of biology, we expect to converge not on a single mechanism, but on a set of similar mechanisms representing phenotypic variation among organisms. Incidentally, this approach provides a way to rigorously analyze the vast amounts of non-hypothesis driven experimental data that high throughput factories, such as the genome centers, are generating these days.

A Bayesian formulation of this kind is a long way off, of course, yet we have been moving in this direction beginning with models of protein function based on multiple sequence data. At this early stage it is not yet possible to model protein mechanisms directly, however, because far too little is known. Thus we first chose to model a surrogate property, one that in theory should be closely correlated with underlying mechanisms, namely the functional constraints

imposed on protein amino acid sequences during evolution. Using Bayesian and other statistical approaches, we optimally multiply align (Neuwald and Liu, 2004) and define the constraints acting upon a set of related protein sequences (Neuwald *et al.*, 2003). In an information theoretical sense optimizing over statistical models of evolutionary constraints in this way narrows the search for true mechanisms (Wilbur and Neuwald, 2000). For an expert biologist, measuring, categorizing and carefully examining inferred evolutionary constraints in the light of associated structural details and of other biochemical data is conceptually similar to what our idealized Bayesian approach aims to accomplish mathematically. Indeed, as illustrated by recent publications (Neuwald *et al.*, 2003; Neuwald, 2003; Kannan and Neuwald, 2004; Neuwald, 2005), such analyses do reveal important aspects of underlying mechanisms. In the future we aim to incorporate other types of biological data into our Bayesian statistical models and to refine these models toward closer representations of actual protein mechanisms. We are now developing such an approach for structural data. This will involve, of course, modeling the correlations between various kinds of data, which in turn will allow us to make statistical inferences regarding missing data.

References

Neuwald, A.F. and Liu, J.S. (2005). Edited by Jorde, L. B., Little, P. R. R., Dunn, M. J. and Subramaniam, S. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, Hoboken, NJ.

Neuwald, A.F. and Liu, J.S. (2004). Monte Carlo optimization and a hidden Markov model for gapped multiple alignment of protein sequence motifs. *BMC Bioinformatics*, **5**, 157.

Neuwald, A.F., Kannan, N., Poleksic, A., Hata, N. and Liu, J.S. (2003). Ran's C-terminal, basic patch and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras and Ran GTPases. *Genome Res.*, **13**, 673-692.

Wilbur, W.J. and Neuwald, A.F. (2000). A theory of information with special application to search problems. *Comput. Chem.*, **24**, 33-42.

Neuwald, A.F. (2005). Evolutionary clues to eukaryotic DNA clamp-loading mechanisms: analysis of the functional constraints imposed on replication factor C AAA+ ATPases. *Nucleic Acids Res.*, in press.

Neuwald, A.F. (2003). Evolutionary clues to DNA polymerase III beta clamp structural mechanisms. *Nucleic Acids Res.*, **31**, 4503-4516.

Kannan, N. and Neuwald, A.F. (2004). Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Sci.*, **13**, 2059-2077.