Pacific Symposium on Biocomputing 6:164-166 (2001)

STRUCTURES, PHYLOGENIES, AND GENOMES: THE INTEGRATED STUDY OF PROTEIN EVOLUTION

R.A. GOLDSTEIN
Chemistry Department and Biophysics Research Division, University of Michigan,
Ann Arbor, MI, 48103, USA; richardg@umich.edu

D.D. POLLOCK
Department of Biological Sciences, Louisiana State University,
Baton Rouge, LA 70803, USA; daviddpollock@yahoo.com

J.L. THORNE
Department of Statistics, North Carolina State University,
Raleigh, NC, 27695, USA; thorne@stagen.ncsu.edu

1. Introduction

For the past decades, evolutionary biologists have tried to reconstruct evolutionary histories, to piece together phylogenetic trees, and to understand the network of hereditary relationships. Such approaches (whether it is admitted or not) are based on models of the evolutionary process. These tasks would be easier if the reality would better match the simplest models. Unfortunately for these scientists, evolution takes place in a complicated web of constraints, with changes in the DNA sometimes but not always translating to changes in amino acids which may or may not result in significant changes in the properties of these expressed proteins. All of this occurs in a complicated and interconnected fitness landscape, where different locations in the protein may be under radically different selective pressure. This situation has led a number of investigators to bring more of the biological and biochemical complexity into these evolutionary models, to develop approaches with a closer fidelity to the biological reality with the hope that more accurate pictures of biological history will result.

In the meantime, structural biologists and protein chemists have been trying to understand the properties of proteins and how these are determined by their constituent amino acids. Much effort has gone into making artificial mutants and relating similar sequences in an attempt to glean hints about structure and function. Some investigators have noticed the vast amount of information contained in the evolutionary record, and that the genomic data contains a record of how the forces of mutation and selection have sculpted modern protein sequences. They have embarked on projects to bring evolutionary thinking and phylogenetic modeling into studies of proteins. For these scientists, the complexities of the evolutionary process are a treasure trove of information, providing valuable hints of what is happening at the molecular level in these biological systems.

These approaches are obviously complementary. By bringing in a notion of the selective pressure acting on proteins and how this affects evolutionary change at the amino acid and DNA level, we are able to develop more accurate evolutionary models as well as understand how to decipher what the evolutionary record can tell us about the evolving proteins. One of the themes of this track is merging these approaches, using our knowledge of proteins to build better models of evolution and use our models of evolution to increase our understanding of proteins.

Two examples of this synthesis can be found in the papers of Z. Yang and J. Koshi. Both researchers develop models of the evolutionary process that explicitly consider evolution at the protein level and use these models to investigate the properties of the evolving proteins. Z. Yang has pioneered much of our thinking about the relationship between DNA and protein evolution, particularly the relationship between changes at the DNA and amino acid levels. In general, purifying selection will inhibit changes at the amino acid level, but will have a much reduced effect on mutations of the DNA that do not cause changes in the protein sequence, so-called "synonymous" or "silent" mutations. Conversely, positive adaptation may result in DNA mutations resulting in changes at the amino acid being accepted at a faster rate than silent mutations. The ratio of non-synonymous to synonymous substitutions provides a measure of whether the changes are made under purifying or adaptive pressure. Yang directs his attention to a specific protein, the envelope protein of HIV, locating positions in the protein sequence that seem to be under adaptive pressure.

An alternative direction of research is presented by J. Koshi. He and R. Goldstein present a model that represents the substitution rates in proteins directly in terms of the physical chemical properties of the constituent amino acids. This allows them to directly probe the nature of the selective pressures acting on proteins. As their model explicitly includes site heterogeneity, they are able to look at the variety of selective pressures that may be acting on supposedly similar locations in the protein. In contrast to Yang's paper, they adapt these models to a wide non-specific data set in order to derive general principles of conservation and variation.

Via innovations made by N. Goldman, R. Goldstein, Z. Yang and others, models of amino acid replacement and nucleotide substitution now routinely allow for variation of evolutionary rates among sites. The biological underpinnings of this rate heterogeneity are still poorly characterized. In their clear and concise manuscript, Tavaré and colleagues present a statistical framework for linking rate variation to covariates such as codon position, hydrophobicity, protein secondary structure, and degeneracy of the genetic code. In the future, this framework can be modified to include other covariates of interest. The appeal of this framework is that it directly assigns a biological meaning to parameters of evolutionary models.

Thereby, it allows statistical comparisons of competing models to provide biological insight in a straightforward fashion.

A. Rzhetsky and P. Morozov are also concerned with modeling rate heterogeneity in protein evolution. They continue progress on introducing wavelet models (wavelet decomposition of discrete functions) as an alternative to the gamma model for variation of evolutionary rates among sites. Since optimizing such models on a complete phylogenetic tree is computationally difficult, they use an approximation, or pseudolikelihood function, to speed calculation. Their heuristic is the pairwise likelihood function, with which they calculate estimates of a pseudoposterior distribution of parameter values using Markov Chain Monte Carlo. The importance of the method is shown in two applications. First, they look for differences in substitution rates between homologous subfamilies. Gu[1] has shown that this is a potential predictor of functional divergence between subfamilies, so it is very useful to see a more detailed model applied to this question. Second, they suggest that their method could be used to produce pairwise distances that best fit the pseudolikelihood function, which in conjunction with heuristic algorithms could then rapidly produce phylogenetic trees for large datasets.

Sequence alignment and phylogeny reconstruction are two of the central problems of computational biology. In his manuscript, J. Hein makes progress toward the simultaneous solution of both. He accomplishes this by developing algorithms that are based upon an explicitly evolutionary model of insertion and deletion and that allow all possible alignments to appropriately contribute toward the reconstruction of an evolutionary tree. Although the model of insertion and deletion is simplistic, Hein makes a substantial advance here and lays the groundwork for future major improvements to alignment and phylogeny inference methods.

Finally, J. S. Conery and M. Lynch have contributed an analysis that attempts to directly link new genomic data with past theoretical work on assessment of synonymous and nonsynonymous substitutions. They have developed an exploratory software system for entire genomes that finds duplicate genes, aligns their amino acid sequences, generates the alignment of the underlying nucleotides, and then assesses the substitutions. Their software is designed to shed light on processes of creation and evolution of duplicate genes, and they use their scan to estimate the age distribution of duplicate genes and the change in synonymous versus nonsynonymous substitution levels over time.

References

1. X. Gu, "Statistical methods for testing functional divergence after gene duplication" *Mol. Biol. Evol.* **16**:1664-1674 (1999).