Modeling Protein Evolution

David D. Pollock and Richard A. Goldstein

Department of Biological Sciences and Biological Computation and Visualization Center
Louisiana State University, Baton Rouge, LA 70803
dpollock@lsu.edu, 225-578-4597, 225-578-2597 (Fax)

Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
richard.goldstein@nimr.mrc.ac.uk +44(0)20 8816 2293, +44 (0)20 8816-2460 (Fax)

A "Short Specialist Review" contribution (g101311) to "Encyclopedia of Genomics, Proteomics, and Bioinformatics", M Dunn, L Jorde, P Little, and S Subramaniam, eds. w/ DB Goldstein, section editor

Key words: protein evolution, evolutionary dynamics, protein structure, maximum likelihood, posterior probability, Bayesian, coevolution

Abstract

Modeling protein evolution has been frustratingly simplistic in the past, but new methodologies and approaches have been rapidly changing this situtation. Increased computational power, improved phylogeny-based maximum likelihood and Bayesian statistics, larger data sets, and better protein structure prediction methods are jointly improving the outlook and allowing researchers to improve the biological realism of protein models. They are also allowing more detailed analysis of differences in processes among sequence positions over space and time, of selection and adaptation, coevolution, and functional divergence, and of ancestral changes in function. The future is expected to bring improved integration of models of protein evolution with protein structure prediction, with the potential to dramatically improve the accuracy and power of both.

Introduction and Discussion

Proteins are the biological macromolecular entities most closely and directly related to organismal function, and yet our ability to model evolutionary processes in proteins has lagged far behind models of DNA evolution. An obvious and important reason for this is that proteins are composed of 20 amino acids, while DNA is composed of only four nucleotides; it is both harder to make calculations with amino acids and harder to find data sets sufficiently large to accurately estimate substitution probabilities between so many states. These problems only become worse when modeling of codons (with 64 states) is attempted. A somewhat more subtle point is that DNA models have progressed partly because they usually assume, implicitly or explicitly, that substitutions are occurring in a random, or neutral fashion. In contrast, amino acid propensities in proteins are so obviously linked to structure and function that neutral models, which assume that all positions in a protein evolve in the same manner and rate and do not change over time, appear painfully inadequate. Furthermore, although experimental evidence (e.g., mutagenesis and functional analysis) has long shown that different positions in proteins can have wildly different tolerances for different amino acids and that non-additive or interactive functional relationships among positions abound, the ability to predict protein structure and function from primary sequences has not been sufficiently robust, rabid, and accurate enough for modeling anything but the shortest evolutionary steps.

A number of recent developments have rapidly been changing this situation, and modeling protein evolution is becoming a more accurate, diversified, and broadly useful pursuit. An increase in computational power is perhaps foremost among these developments, but not simply because the same old calculations can be performed faster. In addition to increasing speed, computational advances have spurred and made practical the development of novel and sophisticated statistical methodologies using complex models that were unthinkable when computers were slower. Chief among these methodologies are maximum likelihood and Bayesian, or posterior probability analyses. These fundamentally model-based approaches allow incorporation of biochemical knowledge and testing of evolutionary hypotheses in a flexible and statistically sound manner. They also allow the incorporation of hypotheses concerning the phylogenetic relationships among genes or species, a component that is essential for reducing noise and spurious correlations in evolutionary analyses. The simple but obviously incorrect assumption of treating sequences as though they are independent (unrelated) entities is no longer necessary or advisable.

Another important development, the same that spurred the creation of the genomics and bioinformatics fields, was the advent of rapid, cheap, and large-scale DNA sequencing capabilities. Although much of the sequencing efforts have been focused on obtaining large quantities of sequence from one or a few species, which except for large multi-gene families are relatively useless for studying evolutionary processes, there has still been considerable production of homologous sequences from divergent organisms. This sampling of many taxa, or “genomic biodiversity”, is essential to the development of sophisticated and realistic models of protein evolution, since discerning the evolutionary behavior at individual sites requires enough biodiversity data such that many substitutions would have occurred at each individual site over the evolutionary time frame during which the sequences diverged.

Finally, the newest and potentially most revolutionary developments have been in the field of protein structure prediction. Despite a tradition of unbridled optimism, protein folding and structure prediction has long been problematical and even less tractable than studying protein evolutionary dynamics. Proteins are only marginally stable, and folding takes time; for accurate folding prediction, the time steps may need to be subdivided many orders of magnitude, and the energy functions may need to be much more complex, detailed and accurate if the compounding of errors across thousands of atoms and bond angles are not to be catastrophically erroneous. Despite this, the inverse problem, that of finding sequences that conform to a particular fold, has had dramatic success recently by using a heuristically optimized combination of simple energy potentials and observed distributions of conformations for both amino acid side chains and main chain oligomers with particular sequences of amino acid residues. Calculations for these methods are relatively rapid, and are in the range of being useful for evolutionary studies and prediction. In addition, observations of adjacency statistics between pairs of residues in an enlarged database of three-dimensional protein structures (from x-ray crystallography or NMR), although obviously crude, have provided useful information for modifying amino acid substitution probabilities, and have provided semi-realistic complex models for simulating long-term evolutionary processes and discovering consequences of these processes that may be reflected in real proteins.

How then, have protein models been developing and incorporating biological realism? For a long time, substitution models were never inferred from an individual data set of interest, but were instead obtained from observations of differences between many closely related protein pairs or sets of sequences. Since these observations accumulated over many protein positions and many unrelated proteins, they necessarily assumed an unwarranted consistency in the protein evolutionary process. Relatively early modifications included development of specialty substitution matrices that focused on slowly evolving sites, particular genome types (e.g., mitochondria or viral genomes), or residues involved in particular structural features (e.g., alpha-helices, beta-sheets, or buried residues and residues exposed to solvent). Eventually, hidden Markov models were used to allow incorporation of these matrices into a phylogenetic likelihood analysis, and some success in predicting secondary structure was achieved, along with more probable reconstructions of protein phylogenies. Still, though, there was not great reason to believe that these specialty matrices, based on pre-conceived conceptions of what might be most important in determining evolutionary processes, did not represent unknown amalgams of heterogeneous processes that were yet to be deciphered. Although incorporation of gamma-distributed average rates, a technique used in DNA evolutionary analysis, was used to address some of this hidden variation, an important advance was the use of mixture models and substitution matrices derived as functions of physico-chemical features of amino acids. These mixture models allow the association of substitution matrices with positions in an alignment to arise freely during the course of analysis, and thus can obtain novel information and produce inferences that are not possible when the substitution classes are pre-defined. Models in which the evolutionary process changes over time are also being developed, although they are clearly in their infancy, and still limited by the size of current data sets.

One of the more important reasons for improving models of evolution in proteins is to understand the forces of selection that act on proteins, and to separate these forces from random drift. Although it cannot be said that these two processes of change have been cleanly separated, in many cases statistical features have been evaluated that strongly indicate selection. The most clear-cut cases are those of diversifying selection, in which a sort of molecular cat-and-mouse scenario emerges that drives amino acid substitution at an accelerated rate that is greater than neutral expectations. Such scenarios are special cases, however, and not observed in most proteins. It is much more common that a burst of amino acid substitution might occur on a particular branch, and this may be detected as a brief elevated rate of amino acid versus nucleotide substitution. It is also possible to detect coevolution between residues in proteins, in which case substitutions at one position alter substitution probabilities at other positions. When proteins have duplicated, it is possible to detect changes in rates or patterns of substitutions at individual sites, and thus to identify changes that may have been due to functional divergence. Finally, it is possible to use evolutionary analysis to predict ancestral sequences; these can then be resurrected and analyzed to infer patterns of functional change along the phylogeny (a process sometimes referred to as "paleobiochemistry").

Despite the successes of these types of analyses, it may still be argued that we are far from a complete and realistic model of protein evolution. Many of the techniques used simply detect extreme evolutionary processes or unexplained changes in the evolutionary process, while the causal mechanism (adaptive burst, functional divergence, coevolution) is more a matter of perspective and hope than of direct evidence, and many of these mechanisms may be related and interact in ways that have yet to be deciphered. Evolutionary analyses have also not yet made dramatic progress in prediction of protein structure and function, although preliminary results are promising. The advances in prediction of the structural "reverse engineering" problem are therefore quite exciting, in that they may provide an avenue for integrating models of evolutionary change with realistic biophysical models that predict sequence compatibility with specific structures. The simpler pairwise pseudo-energy potentials have already been used to model sequence evolution in structures from the protein data bank, and the full integration of structure prediction methods with models of protein sequence evolution may soon provide more accurate and useful methods for both purposes.