Molecular Structure and Evolution of Genomes

In: Evolutionary Genomics and Systems Biology (Section II Evolution of Molecular Repertoires)

Todd A. Castoe[1], A. P. Jason de Koning[1], and David D. Pollock[1*]

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045 USA

*Corresponding author: David D. Pollock, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, Aurora, CO, 80045 USA. *Email:* David.Pollock@UCDenver.edu *phone:* 303-724-3234 *fax:* 303-724-3215

Abstract

Prior to the availability of multiple eukaryotic genomes, it was expected that innovation and divergence at the phenotypic level would be readily explained by molecular innovation and divergence in protein-coding genes. Thus far, however, evidence for adaptation in proteins as a causative explanation of organismal diversity is rare, particularly in the vertebrates. While it may be unreasonable to expect to explain the origins of all phenotypic diversity through adaptation of proteins, it is only reasonable to assume that we have missed an extremely large number of such cases. Given the tremendous acceleration of genome biology enabled by next-generation sequencing, we must revisit this question and ask ourselves what we may intuitively expect and how we can reasonably search for it. This chapter represents our perspective on how this may be achieved.

Introduction

The field of evolutionary genomics is extremely young, with multiple complete (or nearly complete) eukaryotic genomes having been sequenced only recently. It is an exciting but challenging time for evolutionary genomics. Some of the biggest hurdles have been mechanical, as methods that worked well on a few genes from 10-20 species have had to be redesigned to rapidly handle tens of thousands of genes. The fundamentally interesting challenges, though, have been to understand the mechanistic factors that have shaped the evolution of genomes and the genes they contain. This requires an integrated understanding of mutation, population genetics, and the functional, structural, and thermodynamic bases of selection. Ultimately, we would like to know how functional molecules fold and interact, how structure is dynamically altered, how novelty is created, and to reconstruct how these factors have effected evolutionary change. In short, we seek to understand the interface of sequence and structure within the genomic context.

As more genomic resources accumulate, the field of genomics and molecular biology are rapidly morphing into the new fields of evolutionary (comparative) genomics and systems biology. This transformation largely represents a more realistic, comparative, and holistic reassessment of previous research aimed at appreciating the true complexities of the evolutionary process and the biological dynamics and interrelationships in nature. Molecular evolutionary biology, in contrast, has largely maintained a reductionist perspective that focuses on analysis of how single molecules change through time. The new fields are primed to deliver novel understanding on how sequence evolution shapes the evolutionary complexity of biological diversity.

Before going further, it is important to note that the following is a highly personal account of one laboratory's perspective on the topic of deciphering molecular structure and the evolution of genomes. It is not intended to be a review. This serves as an apology in advance to the authors of the many fundamentally interesting and relevant papers that will not be covered here, but also to the thoroughly relevant topics that could have been discussed, but aren't.

RNA, for example, is an extremely important functional molecule, but we will focus more on proteins because we have done more work on them.

Overview of Considerations in Studying Protein Evolution

There are some general points about the interface of sequence and structure that need to be clarified. Probably the most important of these is to emphasize how little we know about the mechanistic details of how proteins fold, undergo dynamic movements, and function. This is not to complain about the rate of progress in the field of structural biology, but simply to note that our ability to predict structures and the structural and functional effects of amino acid replacements in a protein, or even worse, a series of replacements, is limited. If the crystal structure of a protein is available, we usually must assume that the structure is about the same in proteins from the various species in a dataset we are using. Predictions of things such as interactions between amino acid replacements must thus be tempered by a healthy skepticism as to their accuracy. In statistical language, we must put a fairly large prior on the possibility that precise inferences from structural data are wrong. We also must rely heavily on information that is more likely to remain accurate over evolutionary time, such as the approximate distance between residues, orientation of the vector from the Cα and Cβ atoms, and distance from the surface or from active sites.

Another related point is that we are even further from understanding the complete thermodynamic explanations for why proteins evolve the way they do. Although we do not have space here to go into work on recreating protein-like complex systems in a thermodynamic setting (Goldstein and Pollock 2006; Williams et al. 2001; Williams et al. 2006b; Xu et al. 2005), such work is important, and in the current climate needs to be justified because it is not always seen as relevant because the models used are not reflective of real proteins (see previous paragraph). Oftentimes, for computational reasons, simulations are grossly simplified far below even the limits of our best knowledge of structural biophysics. The rationale for this is that evolving complex systems in a thermodynamic setting can often produce dramatically counterintuitive results. While such results do not constitute "proof" of anything, they do provide clues to what to look for in when observing the products of protein evolution, and how to interpret it, and also provide null expectations. Important examples include the ideas that proteins may evolve to lessen the deleterious effects of mutations, that different protein structures may have different degrees of "designability" and thus different freedom to vary and still form the same structure, and that proteins at evolutionary equilibrium will tend to be only marginally stable (Taverna and Goldstein 2002a; Taverna and Goldstein 2002b; Williams et al. 2006a). This last example means at a minimum that there is no inherent requirement to invoke stabilizing selection to explain the marginal stability of most proteins, and increases the burden of proof to demonstrate that stabilizing selection actually exists in some cases.

A particularly interesting (but often ignored) component of complex systems is ploidy, since it strongly alters the effect of mutations on fitness. By providing redundancy by default in the genetic system, it strongly interacts with redundancy and divergence in terms of duplicate gene copies. Another key question is how the form of the relationships between binding, expression, and fitness affect the outcome, and whether analysis of the outcome provides enough data to discriminate information about the form of these relationships. For example, our default will be to assume that expression levels are proportional to percent binding, that expression levels are additive across binding sites (unless they are overlapping or interacting), and that fitness levels due to the function of an expressed gene linearly increase with expression level up to a maximum, where they are constant, but that this is balanced by a constant energy/fitness cost for each protein synthesized.

A useful feature of inferences in evolutionary genetics, then, is that it can be pursued independently of knowledge of structure and function, and that is the route that has mostly been pursued in the past. Another way of putting this is that evolutionary genetics inference can be made based on models of the process, rather than using mechanistic or fitness-based models. We hope that this will change, however, because mechanistic models are bound to be more realistic and therefore presumably more effective for making inference, but also because inclusion of mechanistic models will allow an independent (from basic biophysics) method of understanding mechanisms. With even a passing knowledge of thermodynamics, for example, it is difficult to believe that individual residue positions evolve independently from one another. Much of this chapter will focus on how this integration can be achieved.

If we are to achieve such integration, however, there are still a few more introductory points to consider. When studying proteins and protein structure, it is easy to forget sometimes that all proteins are coded by DNA before they are transcribed to RNA and translated to protein. Thus, a complete approach should consider mutation processes in the DNA, and translation processes as well as possibly selectable structural and functional properties of the DNA and RNA. All of these are affected by the local genomic context, and thus the study of protein evolution may ultimately be inseparable from the study of evolutionary genomics. In vertebrate mitochondria, for example, the mutation process changes over the entire genome (Faith and Pollock 2003; Krishnan et al. 2004a; Krishnan et al. 2004b), while in mammalian (and probably more divergent animal) mitochondrial genomes, there are strong differences in mutation process along the genome that bias the equilibrium G+C concentration (Gu and Zhang).

Another point that is often underappreciated is the need for large amounts of taxon sampling to obtain accurate site-specific models. This is important for understanding coevolutionary interactions and other forms of context-dependent evolution. At the genomic level, this is important even for predicting such basic features as whether a region is under functional constraint. Such reasoning forms the basic rationale for sequencing more (at least 26) mammal genomes more thoroughly simply to predict the existence of transcription factor binding sites (Amemiya et al.). For understanding the evolutionary properties of functional molecules (proteins, as well as transcription factor binding sites), the benefits continually increase as more and more taxa are sampled, particularly if they are sampled to break up long branches (rather than adding more and more deeply-branching taxa). For that reason, we have been using the mitochondrial genome as a sort of pilot for predicting the benefits of expanded sequencing of complete genomes. The number of vertebrate mitochondrial genomes has increase from 67 in 2000 (Pollock et al.) to over 1000 today.

Finally, it is key to consider the role of selection and adaptation, and how they might interact with protein structure and function. Our biggest evolutionary predictor of such things is the observation of changes in nucleotide substitution (or amino acid replacement) rates, but we should be careful not to employ circular reasoning in our inferences: conservation predicts function but is not the same thing, and changes in conservation predict changes in function (functional divergence), but functional divergence is a generally historical causative inference and should not be defined as changes in evolutionary rate. Adaptation is also a concept that is difficult to pin down; here, we will use the term in the sense of evolving to improve the optimization of a trait with regard to its average functional role in maintaining the relative fitness of an organism in its usual environment. An adaptive event will result in long-term alteration of the physical characteristics that describe an interaction, and will usually involve multiple amino acid replacements.

“Convergence” is another slippery term, and it is necessary to distinguish between random convergence due to neutral processes and convergence due to natural selection or adaptation. We summarize below some recent work in vertebrate mitochondria, particularly snake mitochondria, to illustrate how these processes, as well as coevolution between residues, can be detected. “Innovation” is an adaptive event that creates a new functional role for a protein or regulatory element, and will often involve gene duplication and/or a change in relative evolutionary rates of amino acids residues or nucleotides. “Function” is itself a somewhat ill-defined concept, but much of it can be defined it as the degree that two molecules or regions of molecules bind together (including binding as a step in catalysis). Note that the relationship between the degree of binding and its average effect on organismal fitness is not necessarily direct, and may be described by a variety of parameterized distributions. The “fitness” of a genetic element (*e.g.*, a protein or DNA segment, haplotype, or genotype) will refer to the expected relative fitness of an organism bearing the element, averaged over all genotypes and environmental variables not being directly considered. Finally, the “speciation” question will be addressed by limiting our interest in it to the narrow question of how well regulatory interactions are maintained when their constituent elements (proteins and DNA segments) are brought back together after having diverged via multiple substitutions during independent evolution in separate species or sub-populations.

Function and Evolutionary Genomics

Deciphering Complexities of Protein Evolution

As proteins accumulate change and diverge over time, they must continue to satisfy the structural and energetic constraints that enable them to function properly. Because of this, the diversity of protein sequences available from living organisms represents a wealth of data on the relationships between protein sequence, structure, and function. Extracting insights from these data, however, remains a challenge. In principle, the optimal approach to decoding functional information in protein sequence biodiversity is to use parametric statistical inference with realistic phylogeny-based models. Historically, this approach has been limited by the amount of sequence data available, and by the difficulty and prohibitive computational complexity of the probability calculations needed.

The evolution of proteins is complex primarily because it is directed by a large number of underlying stochastic processes that are bounded by structural and functional constraints (Bloom et al. 2006; Drummond et al. 2006; Golding and Dean 1998; Julenius and Pedersen 2006; Lemos et al. 2005; Lopez et al. 2002; Robinson et al. 2003; Rodrigue et al. 2005; Rodrigue et al. 2006; Thorne 2007). Structural and functional features of proteins, and how they interact with the evolutionary process, are not easily or accurately predictable based on first principles (Hayashi et al. 2006; Wood and Pearson 1999; Xu et al. 2005). Nevertheless, the potential rewards of incorporating these factors into an accurate yet feasible framework for modeling protein evolution are truly immense. An understanding of protein functional evolution is essential for identifying mutations of functional significance that may lead to disease, using multiple sequences in structure and function prediction, ancestral reconstruction of sequence and function (Williams et al. 2006a), identifying sites involved in protein-protein interactions, and identifying changes in function (see (Philippe et al. 2003; Wang and Pollock 2005) for perspectives). Identifying and understanding the principles that dictate evolutionary diversification would also help to improve strategies for protein design (Glasner et al. 2007; Tobin et al. 2000).

The principal difficulty in modeling protein evolution is that it is highly context-dependent, meaning that the probability of amino acid residue replacement during evolution is expected to vary across positions and over time (Andreeva and Murzin 2006; Buchner 1999; Koshi and Goldstein 1995; Kozak 1999; Midelfort and Wittrup 2006; Pollock and Goldstein 2002; Templeton et al. 2004; Xu et al. 2005). Particular positions in a protein will have different structural and functional environments and thus different evolutionary constraints, and will therefore have distinctive patterns of substitution at the corresponding codons. Context-dependent changes over time may occur when the structural or functional environment around that position changes via replacement of interacting residues, or in general when any intrinsic or extrinsic factors (*e.g.,* protein function, physiological role, or expression pattern) change, altering selective constraints. When the context changes, the process of evolutionary change at each position may also change (Fig 1). When speaking of evolutionary interactions among protein residue positions, this is sometimes called molecular coevolution, and positions that interact are said to coevolve (Pollock 2002; Pollock and Taylor 1997; Pollock et al. 1999; Wang and Pollock 2005; Wang and Pollock 2007; Wang and Pollock 2009).

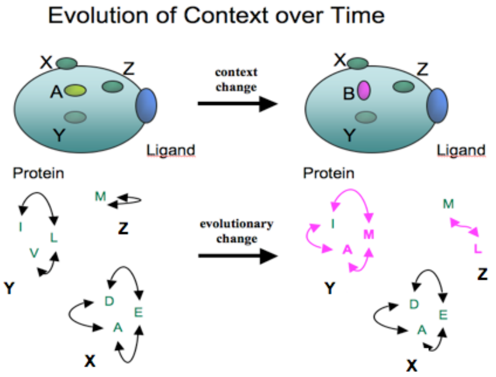

Figure 1: Functional or structural context changes. Replacement of amino acid A by amino acid B, may lead to changes in selective pressure that alter substitution processes at different sites depending on the constraints at those sites.

Despite its critical importance, traditional approaches to phylogenetic analysis of protein evolution have not substantially taken context dependence into account. This is mostly for computational and conceptual reasons, but also because of data limitations. It is this context-dependence, however, that is central to proper understanding of evolutionary genomics and its relation to structure and function. If the replacement of an amino acid at one residue position alters the functional effect (e.g., pathogenicity) of mutations at an adjacent position, the first replacement has changed the evolutionary context at the adjacent position. If another replacement leads to loss of binding affinity for a ligand, such functional divergence may alter the functional context of many residue positions, and the resultant alteration of evolutionary flexibility should be detectable in the descendants. Subtle changes in flexibility, packing of side chains, or distribution of charge on the surface may yield correspondingly subtle, yet functionally significant changes in contextual effects on the evolutionary process.

Our ability to predict and interpret the details of evolutionary shifts caused by changes in structure and function requires that we develop and observe the outcomes of models and datasets that are capable of accurately reflecting the details of these context dependent effects. The unrealistically simple models used in the past cannot detect biochemical realities that they are not designed to reflect. They thus fail to reveal subtle yet critical details of the true process and reduce the accuracy of model-dependent inferences of functional innovation or ancestral reconstruction of sequence and function (Williams et al. 2006a).

The future of modeling protein evolution: merging realism with tractability

One of the main constraints hampering the development and analysis of more realistically complex models has been computational limitations, although progress by our group and others have dramatically decreased these limitations (Hwang and Green 2004; Krishnan et al. 2004c; Nielsen 2002; Rodrigue et al. 2006). Key innovations that enable more computation are mostly due to truncations or simplifications of the calculations made in computing probabilistic models. Essentially, these approaches function by sacrificing a small degree of computational accuracy for massive gains in computational efficiency. Given the computational ability to efficiently incorporate complex models of the evolutionary process in a tractable framework, we can consider many of the potential complexities of both underlying mutational processes and also complex patterns of protein evolution that may elucidate changes in function that may represent adaptation.

Our group and others have made substantial progress developing methods that are capable of efficiently evaluating complex evolutionary models. We developed a fast likelihood-based “conditional pathway” approach that scales extremely well with the number and complexity of context-dependent models used in phylogeny-based analyses. This approach removes many computational barriers that have previously limited the types of model-building experiments that were feasible. The conditional pathway approach allows exploration of fundamentally novel levels of model complexity, and thus provides new potential to reconstruct and understand sequence, structural, and functional changes that have occurred through evolution.

The goal of modeling protein evolution is not to develop complex substitution models for their own sake, but to develop models that reflect the true complexity of protein evolution. This will lead to novel insight into how features of protein structure and function affect substitution processes, and will enable novel and fundamental insight into the relationship between sequence, structure, and function. A central need is to develop flexible context-dependent models to assess how substitution processes can differ in different parts of proteins and change through time. Such models should accommodate uncertain knowledge of which features are important. By allowing this flexibility in the substitution models, we will in the future be able to accurately infer the biologically relevant features that determine the evolutionary process. Previously, assessing context-dependent process in proteins would have been confounded by the incorrect assumption that the substitution process does not change over time, or was the same across large numbers of protein residues.

The accurate assessment of protein evolutionary processes is essential in determining the surrounding structural and functional features that constrain how amino acid residues evolve. Although *post hoc* correlation of substitution states to structure and function is an important means of understanding the causal basis of differences among sites and changes over time, it is ultimately best to integrate structural information directly into the models. Some basic structural information is essentially static or requires little additional computation, and can be incorporated into models directly; examples are the amino acid composition in residues adjacent to a site, the secondary structure, accessible surface, and the local side chain or charge densities. Incorporation of each particular category of information should be well justified based upon the amount of data available and its demonstrable effect on substitution processes. This avoids superfluous model complexity yet ensures inclusion of all potentially contributing information. The impact of structural information on key functional information, such as proximity to ligand binding sites, proton channels, and other functional features, should also be considered, and both continuous functions and discrete effects should be considered.

Other structural information (e.g., from molecular modeling) can require computationally expensive energetic calculations. It is essential to incorporate such information carefully to control the computational burden and maintain the tractability of calculations. Levels of structural integration should incorporate progressively more and more detailed energy potentials, beginning with simple pseudo-energy contact potentials, moving up to more detailed and realistic physical models including all-atom rotamer-based potentials and pseudo water potentials, and later incorporating flexible backbones and simulated water molecules. It is also important to move towards free energy approximations by considering distributions of alternative (competing) structures and local folding and unfolding processes. As the calculations become more complex, it will be feasible to make energy calculations on only a small proportion of the possible ancestral replacements.

The effect of increasing taxon sampling and sequence biodiversity

Inferences about the evolutionary process and how it has changed through time are highly dependent on taxon sampling (Abbott et al.). Thus, dense evolutionary sampling of genomic biodiversity is necessary for the powerful detection of subtle changes in protein evolution. Probabilistic methods of evolutionary inference rely critically on information about patterns of change at each site in order to accurately portray and estimate the evolutionary process. Also, the redesign of probabilistic evolutionary models to incorporate further biological realism, to both avoid systematic error and be able to detect subtle departures from regimes of selection is vital. Particularly in analyses of a single protein through evolutionary time, the only way to increase the information used to estimate evolutionary models is through dense taxon sampling to provide many examples of site patterns to inform the probabilistic models.

Taxonomic sampling is particularly important for studies of site-specific evolution and coevolution in proteins. Such studies have had moderate success, but it is clear that studies of this kind are currently limited by the need to include dense taxonomic sampling so that there are multiple amino acid substitutions at each site over the entire tree, but not too many multiple substitutions along individual branches. Twenty to one hundred diverse taxa appears to be a minimum for successful analysis, and more would improve the power considerably. When considering coevolution between genes, both datasets must include the same specific taxa. Furthermore, the amount of evolution between any two points on the tree should not be so large that there have been changes in the structural context of individual sites. For most proteins, the only means of obtaining a large dataset is to include widely divergent bacteria and eukaryotes, but even conserved proteins will undergo considerable change since these taxa have diverged.

Another reason for good taxon sampling is to obtain accurate phylogenies for the study of gene duplication and evolutionary innovation. Estimates of the phylogenetic history of multi-gene families to understand the process of gene duplication is uniquely difficult because only the gene of interest (i.e., the duplicated gene) and related sequences are available to make this inference. This factor is often under appreciated, but poses a severe limitation because the phylogenetic signal available for making such inferences of phylogeny is limited to a single locus. Thus, particular attention and care is required in interpreting and inferring multi-gene family phylogenies, and major sources of inference error must be kept in mind and evaluated. Taxon sampling helps to increase certainty of the duplication placement, and also because it allows for better models that will then reconstruct phylogeny better (Pollock et al.).

Removing the mutational noise and context-dependent biases from protein evolution

To infer changes in the evolutionary process in protein-coding genes that may represent adaptation or functional change, it is critical that analyses of how proteins evolve begin by appreciating, and removing, the effects and dynamics of the mutational processes at the nucleotide level. Equally important, extremely large datasets are subject to extremely large systematic error when probabilistic evolutionary model assumptions are violated, and such is the case if underlying mutational biases or context-dependencies are ignored. In addition to mutational effects at a local or small scale, larger scale context effects, such as genomic contexts, may also require appreciation in the modeling process, although only a bit is known regarding the ways in which genome architecture might affect the various aspects of genome function and evolution (including replication, transcription, and function of proteins and RNAs). Nevertheless, patterns linking mitochondrial genome structure, function, and nucleotide evolution have begun to emerge (Krishnan et al. 2004b; Krishnan et al. 2004c; Raina et al. 2005a). Thus, at the core of any good model of protein evolution is a good model of how the DNA alone would evolve if it were not involved in determining protein structure and function. To meet this goal, the DNA models underlying the amino acid replacement process must be accurate and realistic to avoid confounding estimates of DNA evolution with influences from amino acid evolution.

At a fine spatial scale, the local nucleotide environment or context may affect nucleotide evolutionary dynamics. In this case, the nucleotide content at adjacent sites may have a notable context-dependent effect on the probability of nucleotide substitution. We have previously investigated and demonstrated this context-dependent effect by modeling the nucleotide evolutionary process in alignments of *SINE* elements in the opossum (*Monodelphis domestica*) genome (Gu et al. In Review). Based on analysis using a symmetric fully context-dependent dinucleotide model, it is clear that adjacent nucleotide content can have an important effect on adjacent site substitution rates, and that accounting for such context-dependent effects represents an important feature of modeling the underlying nucleotide evolutionary process.

On a broader spatial scale, the genomic location of a locus can also have a notable effect on nucleotide evolution. Molecular evolutionary analysis showed that the substitution process was different in different *SINE1* elements with different adjacent GC content in opossum genome (Gu et al. 2007). Also, different genomic regions may have different degrees of accelerated nucleotide substitution at CpG dinucleotides. Elevation of substitution rates at CpG dinucleotides are thought to be linked to relative degrees of cytosine methylation which leads to higher rates of stochastic mutations (especially elevated transition substitutions). Examples of such acceleration can be seen in the *SINE* elements of the opossum based on the context-dependent dinucleotide model described above; note that transition substitutions at CpGs are an order of magnitude higher than other transitions.

Another striking example of how genomic location may lead to different mutational contexts has been demonstrated in vertebrate mitochondrial genomes, primarily affecting transition substations (purine⇔purine or pyrimidine⇔pyrimidine). We previously showed that the mutation process is different at every position, but differences among sites are fairly predictable (Faith and Pollock 2003; Krishnan et al. 2004a; Krishnan et al. 2004b; Krishnan et al. 2004c), largely based on the asymmetrical replication of the mitochondrial genome. According to the "classic" model of mitochondrial replication, different positions in the mitochondrial genome spend different amounts of time in an asymmetric and mutagenic single-strand state during replication. This apparently leads to gradients of thymine (T) to cytosine (C) and adenine (A) to guanine (G) substitution caused by thymine to uracil and adenine to hypoxanthine deaminations on the displaced heavy strand (Faith and Pollock 2003). The response to the mutation gradient differs between these two substitution types, with T⇒C having a roughly asymptotic response and A⇒G a strikingly linear response (Faith and Pollock 2003; Krishnan et al. 2004c). To account for this, we developed a nucleotide model that allowed evolutionary patterns to vary at each site in the mitochondrial genome, and applied this model to four-fold and two-fold redundant 3rd codon positions (Krishnan et al.). In the case of vertebrate mitochondria, and also in other circular genomes such as plastids and bacterial genomes, different locations in the genome may experience very different background mutational processes due to mutational gradients that result from the process of genome replication.

In addition to changes in mutational processes on various spatial contexts, it is important to consider changes to the process through time, and the combination of spatial and temporal dynamics in the evolutionary process. For example, in primate mitochondrial genomes, genome-wide gradients of substitution bias have been show to evolve rapidly across lineages such that different primate species may have quite different mutational gradients (Krishnan et al. 2004b).

Where is protein evolution is going?

Once the mutation process and amino acid replacement process are modeled separately and integrated, there are two main routes to more biologically realistic models. The first route is to generalize process-based models, what we call non-stationary context-dependent (NSCD) models. In the broadest sense, these would allow context-dependent mixture model processes to vary across sites and over time (Fig 2), and could incorporate structural information to inform the mixture and mixture switching. Mixture models without structural information can be compared to structure *post hoc* to determine profitable structural guides for future modeling (Fig 3). We have found that the conditional pathway method is highly amenable to methods to reduce the number of rate parameters in arbitrary ways relative to the mixture rate matrices. These methods, under development, are called “rates across rates”, or RAR methods, and can allow for the rapid testing of otherwise difficult protein models (Fig 4). The second route is to incorporate fitness into the equation, basing it on some estimation of the thermodynamic stability of variants. We call these SEF (structure/energetic/fitness) models. Ultimately, of course, these two approaches should be integrated, and all components thoroughly tested on large datasets to determine the justification for their inclusion. Although these models are simple to describe, we anticipate that a huge amount of interesting biology will occur as we begin to understand what physical factors truly influence how proteins evolve and under what conditions.

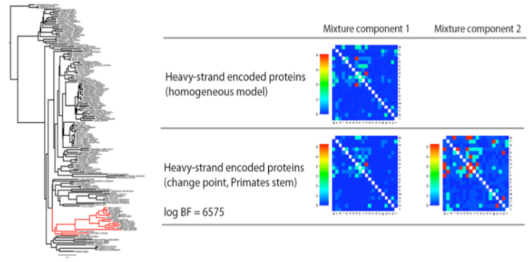

Figure 2: A simple example of a mixture of models changing over time. The mammalian phylogenetic tree is shown to the left, and the portion in red consists of the primates, where the mixture change point was set. Substitution rates in heavy-strand encoded proteins for the homogeneous model and the non-homogeneous model (two mixture components) are shown to the right. The two-class model had a log BF improvement of 6575, indicating strong support for a primate-specific alteration in evolutionary patterns/rates of some sites.

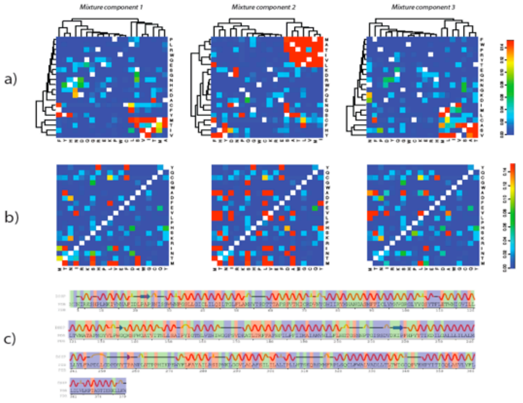

Figure 3: Posterior estimates for mammalian cytb under the RAR model with three independent non-reversible rate class components. a) Mean rate estimates for each mixture class component, with rows and columns clustered according to posterior average rates. b) The same as (a), but with rows and columns in the same order among mixture classes. (c) The most likely posterior model class overlaid (by color) onto a secondary structure diagram of cytb (alpha helices are squiggly lines, beta strands are arrows).

Detecting adaptation and functional innovation

One of the most important problems in protein evolution is that of detecting and understanding adaptation and functional innovation. Ideally, these phenomena will leave traces on the evolutionary record, possibly including accelerated evolution at some sites, bursts of substitution along particular branches, and changes in the models of protein evolution. Adaptive events may be detected as an excess

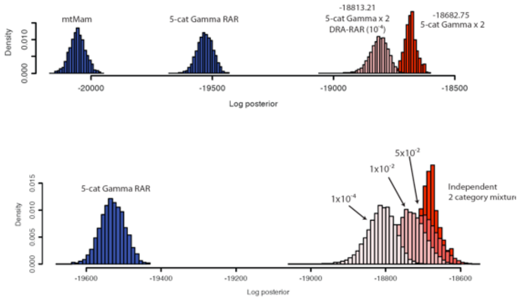

Figure 4: Posterior probability distributions for mammalian cytb using various amino-acid substitutions models. a) mtMam versus a single (5-cat Gamma RAR) unrestricted model and a mixture of two unrestricted (5-cat Gamma x 2) or two dependent rate assignment (DRA) RAR models. b) Comparison of DRA models with different priors. Marginal log likelihoods shown above the distributions were estimated using the harmonic mean over MCMC samples), and DRA priors are labeled.

number of substitutions in a particular set of sites, or excess numbers on a particular branch, compared to the expectation calculated from synonymous sites. We have used this approach in analyzing bursts of evolution in snake mitochondrial genomes, and been able to discriminate differential rate acceleration in different genes (Castoe et al. 2008c; Jiang et al. 2007). Also, based on our previous experience (Castoe et al. 2008c; Jiang et al. 2007; Wang and Pollock 2005; Wang and Pollock 2007; Wang and Pollock 2009), adaptation and coevolution often go together; many substitutions during an adaptive burst may be closely paired in the three-dimensional structure, and these same pairs tend to substitute together on different lineages. Patterns of coevolution also differ depending on whether the residues are involved in adaptive bursts. When such events are detected, they may be further dissected with non-stationary mixture models.

When (possibly adaptive) functional divergence is inferred, an important means of testing this inference is to reconstruct ancestors in the laboratory and examine their functional features through biochemical analysis. Unfortunately, ancestral reconstruction can be subject to a variety of errors and biases that lead to incorrect functional inferences (Krishnan et al. 2004c; Williams et al. 2006b). The improved biological realism of NSCD models should lead to improved accuracy in ancestral reconstruction, and we will test this through simulation studies. This will also provide the means to examine the relationship between model accuracy and phylogenetic structure (i.e., density and relationships of taxon sampling).

Another relevant point to consider, based on our previous experience (Castoe et al. 2008c; Jiang et al. 2007; Wang and Pollock 2005), is that adaptation and coevolution often go together; many substitutions during an adaptive burst may be closely paired in the three-dimensional structure, and these same pairs tend to substitute together on different lineages. Patterns of coevolution also differ depending on whether the residues are involved in adaptive bursts.

Integrating inferences to detect and interpret adaptation: An example with Snake metabolic proteins

Snake metabolic proteins – integration of inferences for adaptation

The best approach to identifying important functional change in proteins that may represent adaptation is through integration of multiple lines of evidence for functional change. Thus, because protein evolution is highly complex, and detecting changes in protein evolution that may represent adaptation and functional change may be confounded by so many factors, no single statistic is sufficient to convincingly demonstrate when adaptation and functional change happens in proteins.

Recently, we discovered that snakes are an excellent system for studying adaptive evolution and functional change in protein-coding genes, and this system demonstrates how multiple inferences of functional change in proteins can be integrated to provide a more holistic inference of adaptation and also of potential selective factors that may have led to functional change. The proteins involved in aerobic metabolism encoded in their mitochondrial genomes have undergone an extreme burst of adaptive evolution that appears to have led to functional innovation and reorganization of snake oxidative metabolism. To infer how and why this even may have occurred, we conducted extensive molecular evolutionary analyses of selection and coevolution in snake mitochondria and evaluated the results in the context of the structure and function of snake mitochondrial proteins.

Detection of accelerated non-synonymous change

The first indication of a burst of adaptive protein evolution in snake mitochondria was that snake proteins appear to have experienced greatly elevated rates of non-synonymous change compared to other tetrapods (Castoe et al. 2008b). Mitochondrial protein-coding genes are subject to strong purifying selection to conserve protein function (Reyes et al. 1998; Yang et al. 2000), normally leading to low rates of non-synonymous change compared to synonymous change (*dN*/*dS*). Consistent with this, the median *dN*/*dS* ratio (inferred from codon-based selection analyses) for the tetrapod mitochondrial dataset is 0.12, and for cytochrome C oxidase subunit 1 (COI), the most conserved mitochondrial protein, it is 0.02. In contrast, along the branch leading to snakes the *dN*/*dS* for all proteins combined is 25-fold higher (3.14), and is 40-fold higher for COI (0.81) (Castoe et al. 2008b). Ratios are also high along the COI branch leading to the alethinophidian snakes, and along these same two branches for the protein Cytochrome b (CytB). Furthermore, branch-site models (Yang et al.) indicate that a large number of sites across all 13 mitochondrial proteins experienced excess non-synonymous substitutions and positive selection. Paralleling the inferences based on standard *dN*/*dS*, the highest number of positively selected sites occur in COI and CytB.

Although *dN*/*dS*-based analyses of protein adaptation are a standard in the field, they are also very susceptible to error from a number of sources, mostly related to the high potential for inaccurate estimation of the *dS* component. Estimates of *dS* for both long branches and ancient (deep) branches, both of which were the case in our tetrapod mitochondrial dataset, are particularly susceptible to saturation and underestimation. Furthermore, in the mitochondrial genome a vast majority of synonymous substitutions are comprised of transition substitutions that evolve at a high rate and are thus likely to saturate. Mitochondrial transition substitution rates and substitution gradients across the genome may also evolve substantially across lineages (Raina et al. 2005b). Because transversion (TV; purine⇔pyrimidine) substitution dynamics in mtDNA are slower and far more consistent than transitions (Raina et al. 2005b), they are much less prone to saturation, the use of exclusively transversions for relative rate comparisons (e.g., *dN*/*dS*) can eliminate many potential errors (Raina et al. 2005a; Yang et al. 2000). Thus, the transversion component of *dN*/*dS* was estimated by averaging over all 3rd codon positions in the mtDNA with conserved four-fold redundancy (dS_{TV4X}), while the non-synonymous transversion rate was measured at first and second codon positions (dN_{TV12}) for each gene under consideration. It is notable that non-synonymous transversions at first and second codon positions result primarily in amino acid replacements with radical physico-chemical differences and major functional effects, thus dN_{TV12} may reflect more radical and functionally relevant amino acid replacements than standard measures of *dN*.

The dN_{TV12}/dS_{TV4X} ratios strongly supported the finding that mitochondrial proteins endured dramatic bursts of amino acid replacement early in snake evolution (Fig. 5). Notably, high ratios are not maintained in descendant snake lineages, indicating that strong purifying selection subsequently dominates snake mtDNA evolution (Fig. 5). These finding provide an excellent example of an apparent context-dependent change in protein evolution in snake mitochondrial genes, in which an episodic burst of selection disrupted the normally neutral equilibrium patterns of protein evolution.

Changes at conserved sites and coevolutionary signal

The impact of the most functionally relevant amino acid replacements in snake mitochondrial proteins was studied at “unique sites” that had replacements in snakes and were otherwise conserved across most tetrapods (Castoe et al. 2008b). COI and CytB have the greatest number of unique sites among mitochondrial proteins, and amino acid replacements at the 23 unique COI sites are concentrated in the earliest branches in the snake tree, with 25-31 estimated changes. Nine sites had reversions or multiple replacements, usually leading to parallel or convergent evolution, and about half of these sites underwent substantial changes in polarity or charge (Castoe et al. 2008b).

The 23 unique snake sites show an excessively high degree of coevolution with each other in this analysis: among all possible combinations of unique site pairs, 66% and 89% have significantly coevolved ($p<0.05$; 28% and 36% at $p<0.01$) according to polarity and volume, respectively (Castoe et al. 2008b). When these

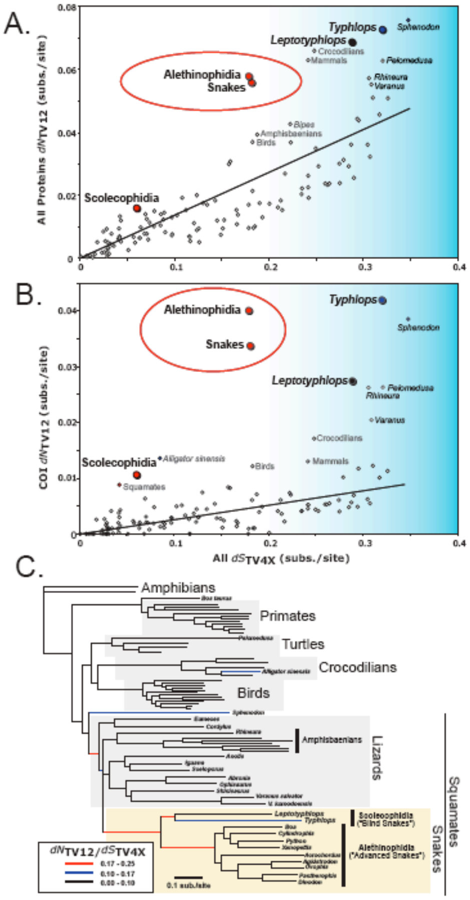

Figure 5: Mitochondrial proteins have had highly elevated rates of amino acid replacement early in the evolution of snakes. The conservative transversion-based approximations of the relative rates of non-synonymous to synonymous substitution (dN_{TV12} / dS_{TV4x}) rates are shown as open or colored circles for each branch of the phylogenetic tree; linear regression lines (excluding points in the red ellipse) are shown in black (A and B). The calculations shown are from (A) all mitochondrial proteins and (B) cytochrome C oxidase subunit 1 (COI). Blue-shaded areas of A and B indicate very long branches with high dS_{TV4x} values where the (dN_{TV12} / dS_{TV4x}) estimate may be inaccurate, possibly due to dS_{TV4x} saturation and underestimation. The phylogenetic tree of relationships among species in our comparative dataset is shown in (C). Branches with extremely high values of dN_{TV12} / dS_{TV4X} for COI are indicated with colored lines (black, blue, red) following the key in the bottom left. The circles for branches in (A) and (B) were colored according to the same legend for ratios of COI (dN_{TV12} / dS_{TV4x}).

23 unique sites are visualized on the structure of cow CO, seventeen of these 23 unique sites clearly form structurally clustered pairs or triplets, most of which appear to be in physical contact, and these clusters occur primarily in the core functional regions of the COI protein (Figure 6). To our knowledge, such a high proportion of physically close (or touching) clusters of replaced residues has not been previously observed in any protein, nor has this degree of concentrated coevolutionary change been previously reported for a protein. The physical clustering of unique sites strongly supports the hypothesis that these sites have coevolved, independent of the statistical coevolution analysis. Therefore, such tight physically paired coevolving residues at otherwise conserved (and therefore presumably functionally critical) sites are unlikely to have occurred without the influence of strong positive selection for evolutionary redesign. The general coevolutionary signal in snake COI at all sites (not just the unique ones) is also inordinately strong (Castoe et al. 2008b).

Integrating evolutionary inferences with structure and function information

The structural basis of CO function is complex. Oxidative phosphorylation is carried out by five complexes that generate a proton gradient and drive the synthesis of ATP. CO is the penultimate complex in this chain, where the reduction of oxygen is coupled to proton pumping (Tsukihara et al. 1995; Tsukihara et al. 1996). Of the 13 CO subunits, the three encoded by the mitochondrial genome (I, II, and III) are at the structural and functional core of the complex (Tsukihara et al. 1995; Tsukihara et al. 1996). A copper atom and two heme groups in COI are critical to the coordinated electron transport, oxygen reduction, and proton pumping function of CO (Tsukihara et al. 1995; Tsukihara et al. 1996). Protons transported or "pumped" along three putative channels (D, H, and K) from the mitochondrial matrix to the mitochondrial intermembrane space contribute to the proton gradient utilized by the ATP synthase complex to produce ATP, and also facilitate the reduction of

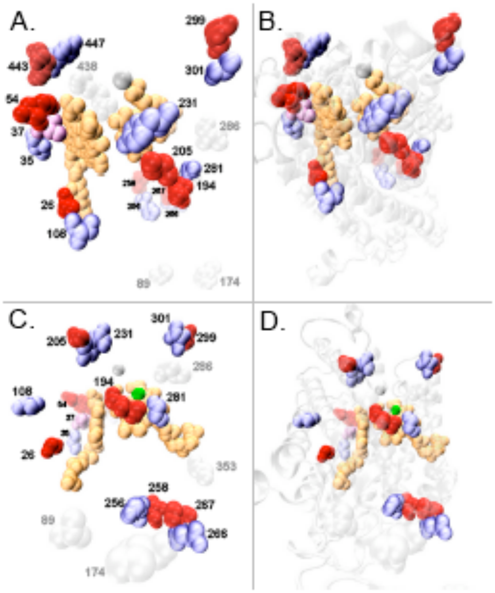

Figure 6: The twenty-three unique amino acid replacements in the cytochrome C oxidase subunit 1 (COI) protein of snakes form seven pairs and one triplet of spatially clustered amino acid replacements, concentrated at the core functional region of the COI protein. The seven spatially adjacent pairs of amino acid residues, strongly suggestive of coevolutionary adaptive change, are shown in blue/red paired spacefill combinations, and one triplet cluster is shown in a blue/purple/red combination. Unique sites that did not form clusters are shown in gray spacefill representations. The two heme groups are shown in gold spacefill shapes, the COI backbone in white, and the magnesium and copper atoms are shown as magenta and green balls, respectively. Two different perspectives are depicted, one in A and B, and a second in C and D; Figure sets A/B and C/D are the same views with B and D showing the ribbon structure of the COI backbone in transparent grey.

oxygen to water. The three core COI proton channels appear to have been extensively redesigned during the evolution of snakes. At least two unique site residues (unique residues) are located in or adjacent to each of three proposed channels, and most other unique residues are distributed around these channels.

Further evidence of adaptation from molecular convergence

Convergent molecular evolution is believed to be rare in nature, although few studies have explicitly searched for it. When it is observed, it is often taken as good evidence for directional selection for functional change, which has acted in parallel on independent lineages. There are an exceptionally large number of convergent changes between independent lineages of snakes, and between snakes and another group of legless squamate reptiles – amphisbaenians. In COI, convergent changes included some of the most conserved, structurally and functionally important sites in all three proton channels of COI (Castoe et al. 2008b).

Increased taxon sampling and a novel statistical approach for detection and analysis of convergent molecular evolution revealed evidence that a significant excess of convergent molecular evolution has occurred, at an unprecedented scale, between snake and agamid lizard mitochondrial genomes (Castoe et al. 2008a). There is a strong linear relationship between the number of divergent and convergent substitutions using both ML and Bayesian methods, and this allows for good statistical accounting of the effect of branch lengths on convergence expectations. Although previous analyses of molecular convergence utilized ML approaches (Zhang and Kumar 1997), there were big differences between ML and Bayesian results, probably due to error in ML approach, which ignores error in the unknown ancestral states; failure to integrate over unknown ancestral states can generally lead to misleading biological conclusions (Krishnan et al. 2004c; Williams et al. 2006a; Yang 2003). Likely convergent sites were concentrated in COX1 and ND1, but were present in other proteins as well (Castoe et al. 2008a).

A thorough analysis of alternative hypotheses to explain this convergence (e.g., nucleotide frequencies, heterogeneous models, and long branch attraction) eliminated all reasonable neutral explanation. Thus, the remaining obvious potential explanation for this case of excess convergent evolution is adaptation. Combined with other evidence for adaptive protein evolution in snakes (discussed above) the excess convergence levels observed here are consistent with the action of natural selection rather than random homoplasy. The evolutionary burst in snakes may have been driven by selection related to physiological adaptations for metabolic efficiency and to allow radical fluctuations in aerobic metabolic rate (Castoe et al. 2008b). The molecular convergence between snakes and agamid lizards may thus have resulted from shared adaptive pressures on metabolic function. Whatever the underlying cause, since the convergence extends across most regions of the mitochondrial genome, any common adaptive force must have been exceptionally strong and broad in scope.

Integrating inferences with possible causal factors

Adaptive evolution and coevolution in COI early in snake evolution appear to have redesigned core functions. In particular, the roles of the various amino acid residues and channels in proton transport, coupling of proton transport to oxygen reduction, and regulation of these processes appear to have been reorganized. Although the structural and functional evidence is best in COI, there is also compelling evidence for adaptive evolution in other mitochondrial proteins early in snake evolution (Castoe et al. 2008b). The distribution and number of unique amino acid replacements, the elevated dN/dS for the entire mitochondrial proteome, site-specific selection analyses, as well as nucleotide dynamics (Jiang et al. 2007) collectively suggest that most snake mitochondrial proteins have experienced extraordinary levels of functional adaptive change. Snake mitochondrial function and oxidative metabolism thus appear to be exceptional system-wide, implying that snakes are an excellent model system for further metabolic research.

Conclusion

The problem of genome evolution and molecular structure/function is of fundamental importance to a wide variety of scientific and health-related research. The better we understand the relationship between sequence, structure, and function, the better we will be able to predict structure and function, manipulate proteins to achieve our aims, and understand and predict protein failure through mutation that leads to disease. The evolutionary record provides a vast amount of information on the subject of how sequences change under the constraints of structure, function, and functional innovation; evolutionary genomics research should be designed to extract much more accurate and practically useful information about this process. Although evolutionary genomics is not designed necessarily to predict structure directly, the results obtained have obvious potential benefits for structural prediction. Such possible benefits include predicting mutational effects, predicting structural features in novel proteins, predicting protein-protein interactions, protein-substrate and protein-drug interactions, and guiding protein design. Every effort should be made to translate evolutionary genomics results into predictions that can be used in empirical research, or higher-level protein structure prediction, and to address any direct predictive utility that arises as an outcome of such research. In general, the next generation of evolutionary genomics should produce a more subtle and biologically realistic understanding of the significance of diversity and variation in proteins than is currently available.

REFERENCES

Abbott, C.L., M.C. Double, J.W.H. Trueman, A. Robinson, and A. Cockburn. 2005. An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in Thalassarche albatrosses. Molecular Ecology 14: 3605-3613.

Amemiya, C.T., J.M. Greally, R.L. Jirtle, E.S. Lander, K. Lindblad-Toh, R.D. Miller, D.D. Pollock, P.B. Samallow, M.S. Springer, and R.K. Wilson. 2003. Proposal for complete sequencing of the genome of a Marsupial, the gray, short-tailed opossum, Monodelphis domestica. NIHGRI White Paper.

Andreeva, A. and A.G. Murzin. 2006. Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16: 399-408.

Bloom, J.D., D.A. Drummond, F.H. Arnold, and C.O. Wilke. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol 23: 1751-1761.

Buchner, E. 1999. Molecular complexity at the synapse: new proteins and multiple isoforms detected in Drosophila. Ross Fiziol Zh Im I M Sechenova 85: 159-166.

Castoe, T.A., A.P. De Koning, H.-M. Kim, W. Gu, B.P. Noonan, Z.J. Jiang, C.L. Parkinson, and D.D. Pollock. 2008a. An ancient adaptive episode of convergent molecular evolution confounds phylogenetic inference. Nature Precedings 26 July 2008: http://hdl.handle.net/10101/npre.12008.12123.10101.

Castoe, T.A., Z.J. Jiang, W. Gu, Z.O. Wang, and D.D. Pollock. 2008b. Adaptive evolution and functional redesign of core metabolic proteins in snakes. PLoS ONE 3: e2201.

Castoe, T.A., Z.J. Jiang, Z.O. Wang, W. Gu, and D.D. Pollock. 2008c. Adaptive coevolution and functional redesign of oxidative phosphorylation proteins in snakes. PLoSONE In press.

Drummond, D.A., A. Raval, and C.O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327-337.

Faith, J.J. and D.D. Pollock. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics 165: 735-745.

Glasner, M.E., J.A. Gerlt, and P.C. Babbitt. 2007. Mechanisms of protein evolution and their application to protein engineering. Adv Enzymol Relat Areas Mol Biol 75: 193-239, xii-xiii.

Golding, G.B. and A.M. Dean. 1998. The structural basis of molecular adaptation. Mol Biol Evol 15: 355-369.

Goldstein, R.A. and D.D. Pollock. 2006. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. Mol Biol Evol 23: 1444-1449.

Gu, W., T.A. Castoe, A.P.J. de Koning, and P.D. D. In Review. Efficient Calculation of Complex Bayesian Models Using Conditional Pathway Analysis – A Dinucleotide Model Example. Gene #: ##-#.

Gu, W., D.A. Ray, J.A. Walker, E. Barnes, A.J. Gentles, P.B. Samallow, J. Jurka, M.A. Batzer, and D.D. Pollock. 2007. SINEs, evolution and genome structure in the opossum. Gene 396: 46-58.

Gu, X. and J. Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol 14: 1106-1113.

Hayashi, Y., T. Aita, H. Toyota, Y. Husimi, I. Urabe, and T. Yomo. 2006. Experimental rugged fitness landscape in protein sequence space. PLoS ONE 1: e96.

Hwang, D.G. and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101: 13994-14001.

Jiang, Z.J., T.A. Castoe, C.C. Austin, F.T. Burbrink, M.D. Herron, J.A. McGuire, C.L. Parkinson, and D.D. Pollock. 2007. Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. BMC Evol Biol 7.

Julenius, K. and A.G. Pedersen. 2006. Protein evolution is faster outside the cell. Mol Biol Evol 23: 2039-2048.

Koshi, J.M. and R.A. Goldstein. 1995. Context-dependent optimal substitution matrices. Protein Eng 8: 641-645.

Kozak, M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234: 187-208.

Krishnan, N.M., S.Z. Raina, and D.D. Pollock. 2004a. Analysis of among-site variation in substitution patterns. Biol Proced Online 6: 180-188.

Krishnan, N.M., H. Seligmann, S.Z. Raina, and D.D. Pollock. 2004b. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. DNA Cell Biol 23: 707-714.

Krishnan, N.M., H. Seligmann, C.B. Stewart, A.P. De Koning, and D.D. Pollock. 2004c. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. Mol Biol Evol 21: 1871-1883.

Lemos, B., B.R. Bettencourt, C.D. Meiklejohn, and D.L. Hartl. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22: 1345-1354.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19: 1-7.

Midelfort, K.S. and K.D. Wittrup. 2006. Context-dependent mutations predominate in an engineered high-affinity single chain antibody fragment. Protein Sci 15: 324-334.

Nielsen, R. 2002. Mapping mutations on phylogenies. Syst Biol 51: 729-739.

Philippe, H., D. Casane, S. Gribaldo, P. Lopez, and J. Meunier. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life 55: 257-265.

Pollock, D.D., J.A. Eisen, N.A. Doggett, and M.P. Cummings. 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol Biol Evol 17: 1776-1788.

Pollock, D.D. and R.A. Goldstein. 2002. Molecular evolution and phylogenetic analysis. Pac Symp Biocomput Tutorial.

Pollock, D.D., D.J. Zwickl, J.A. McGuire, and D.M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51: 664-671.

Raina, S.Z., J.J. Faith, T.R. Disotell, H. Seligmann, C.B. Stewart, and D.D. Pollock. 2005a. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res 15: 665-673.

Raina, S.Z., J.J. Faith, T.R. Disotell, H. Seligmann, C.B. Stewart, and D.D. Pollock. 2005b. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Research 15: 665-673.

Reyes, A., C. Gissi, G. Pesole, and C. Saccone. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Molecular Biology and Evolution 15: 957-966.

Robinson, D.M., D.T. Jones, H. Kishino, N. Goldman, and J.L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20: 1692-1704.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347: 207-217.

Rodrigue, N., H. Philippe, and N. Lartillot. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol 23: 1762-1775.

Taverna, D.M. and R.A. Goldstein. 2002a. Why are proteins marginally stable? Proteins 46: 105-109.

Taverna, D.M. and R.A. Goldstein. 2002b. Why are proteins so robust to site mutations? J Mol Biol 315:

479-484.

Templeton, A.R., R.A. Reichert, A.E. Weisstein, X.F. Yu, and R.B. Markham. 2004. Selection in context: patterns of natural selection in the glycoprotein 120 region of human immunodeficiency virus 1 within infected individuals. Genetics 167: 1547-1561.

Thorne, J.L. 2007. Protein evolution constraints and model-based techniques to study them. Curr Opin Struct Biol 17: 337-341.

Tobin, M.B., C. Gustafsson, and G.W. Huisman. 2000. Directed evolution: the 'rational' basis for 'irrational' design. Curr Opin Struct Biol 10: 421-427.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa. 1995. Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 A. Science 269: 1069-1074.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa. 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272: 1136-1144.

Wang, Z.O. and D.D. Pollock. 2005. Context dependence and coevolution among amino acid residues in proteins. Methods Enzymol 395: 779-790.

Wang, Z.O. and D.D. Pollock. 2007. Coevolutionary patterns in cytochrome c oxidase subunit I depend on domain structure and functional context. J Mol Evol 65: 485-495.

Wang, Z.O. and D.D. Pollock. 2009. Context dependent coevolution in protein complex cytochrome c oxidase detected by Bayes Factor analysis. in preparation.

Williams, P.D., D.D. Pollock, B.P. Blackburne, and R.A. Goldstein. 2006a. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol 2: e69.

Williams, P.D., D.D. Pollock, and R.A. Goldstein. 2001. Evolution of functionality in lattice proteins. J Mol Graph Model 19: 150-156.

Williams, P.D., D.D. Pollock, and R.A. Goldstein. 2006b. Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. Evolutionary Bioinformatics Online 2: 59-69.

Wood, T.C. and W.R. Pearson. 1999. Evolution of protein sequences and structures. J Mol Biol 291: 977-995.

Xu, Y.O., R.W. Hall, R.A. Goldstein, and D.D. Pollock. 2005. Divergence, recombination and retention of functionality during protein evolution. Hum Genomics 2: 158-167.

Yang, Z. 2003. Adaptive molecular evolution. In Handbook of Statistical Genetics (eds. D. Balding M. Bishop, and C. Cannings), pp. 229-254. Wiley, New York.

Yang, Z., R. Nielsen, N. Goldman, and A.M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

Zhang, J. and S. Kumar. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14: 527-536.