Minireview

Tracking adaptive evolutionary events in genomic sequences

David A Liberles* and Marta L Wayne†

Addresses: *Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden. †Department of Zoology, University of Florida, Gainesville, FL 32611, USA.

Correspondence: David A Liberles. E-mail: liberles@sbc.su.se

Published: 29 May 2002

Genome ***Biology*** 2002, **3(6)**:reviews1018.1–1018.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/6/reviews/1018

Abstract

As more gene and genomic sequences from an increasing assortment of species become available, new pictures of evolution are emerging. Improved methods can pinpoint where positive and negative selection act in individual codons in specific genes on specific branches of phylogenetic trees. Positive selection appears to be important in the interaction between genotype, protein structure, function, and organismal phenotype.

Today's genomes have been shaped both by random evolutionary processes and by selection. Positive selective pressure, in which particular changes give an organism an advantage over other organisms, can quickly alter the sequences found at selected sites within a genome. At the same time, a slower dynamic process called genetic drift, in which mutations spread through a whole population (are 'fixed') by chance rather than because of selection, dictates the changes at other sites. The combination of positive selection and genetic drift, plus the purifying selection that keeps advantageous sequences from being changed, can be analyzed to assess the effects of selection on genomes and to detect where specific nucleotide (or codon) positions have been under different pressures at different points in evolutionary history. There are many mechanisms by which function can evolve adaptively in genomes, including changes in gene expression, splice-site usage, gene duplication, and many other processes, but we will focus in this article on coding-sequence evolution.

There are several methods that can be used to evaluate selective pressures on protein-coding genes. The most straightforward approach is to determine the ratio of the rate of substitutions that change the amino acid (nonsynonymous substitutions, K_a) to those that do not (synonymous substitutions, K_s); the fraction K_a/K_s is also known as dN/dS or ω and provides a quantitative measure of selection [1]. This approach has been expanded to include consideration of how different the chemical structure of a substituted amino acid is to the original one [2], by comparing more sophisticated measures of protein and DNA distance (point accepted mutations and neutral evolutionary distance) [3], and by examining sub-regions of a gene that either are close together along the primary sequence, are close in the three-dimensional structure, or are picked out as variable or invariable positions from a multiple sequence alignment [4,5]. More sophisticated tests, like the Tajima D statistic and McDonald-Kreitman test, consider K_a/K_s but also take advantage of single-nucleotide polymorphism data [6,7]. For example, in the McDonald-Kreitman test, the difference between the ratio of nonsynonymous to synonymous divergence and the ratio of nonsynonymous to synonymous polymorphism is used to measure the fraction of sites under positive selective pressure.

A long debate has raged in the molecular evolution literature over what fraction of mutations are fixed through drift and what fraction are fixed through positive selection. In the 1960s, Kimura [8] proposed that molecular change is selectively neutral, a stance known as 'neutralism'. More recent studies show clear examples of positive selection of molecular sequences, leading some researchers to disagree with Kimura and revert to the previously prevalent view, selectionism. Using the K_a/K_s approach, an early study [9] systematically compared orthologous genes in rat and mouse and identified only one, the interleukin-3 gene, that was

clearly under positive selective pressure. A later, similar study found only two chordate genes under positive selective pressure - prostatic steroid binding protein and snake neurotoxin [4]. Subsequently, an innovative approach was introduced that combined phylogeny with K_a/K_s analysis, allowing the identification of specific branches of phylogenetic trees that have been under positive selective pressure [10,11]. Systematic analysis using this approach has identified 643 branches of gene family trees in chordates and 228 branches in land plants (embryophytes) that appear to be under positive selective pressure (when K_a/K_s was calculated averaging over the whole gene length) [12]. Further, the functions of the genes we found were not random, but appeared to be linked to organismal functions for which selective pressures are thought to play a major role, including the immune and reproductive systems.

Such genome-wide rate-testing approaches, along with individually accumulated examples [1] and a systematic analysis of divergence after gene duplication in completed genomes [13], have begun to point to the importance of positive selective pressures in shaping the protein-coding content of genomes. Three recent studies [14-16] have taken advantage of single nucleotide polymorphism data in primates and *Drosophila* and have provided additional evidence for the importance of positive selection in coding-sequence evolution. All three conclude that there is evidence for far more adaptation than had previously been suspected from various statistical measures.

Smith and Eyre-Walker [14] started by collecting a sample of genes for which there were data on sequence polymorphisms between *Drosophila simulans* and *D. yakuba* orthologs. They excluded some genes that were thought to be under selection *a priori* and others because they contained no polymorphisms and thus were uninformative; this left 35 genes. They then worked from the assumption that all polymorphic amino-acid variation within a species is selectively neutral, whereas amino-acid differences between species (substitutions) could be neutral or strongly advantageous. A certain number of substitutions would be expected from the neutral polymorphism rate; the authors reasoned that if there were more substitutions than this, the excess must indicate adaptive evolution. From the original sample of 35 genes, they estimated using this method that 24% of amino-acid substitutions between the species were adaptive; this was not statistically significant, however. Smith and Eyre-Walker [14] then excluded the genes that were contributing most of the variance and that were viewed as outliers, as they had five or fewer synonymous polymorphic sites. This left a set of 30 genes and an estimate of adaptive substitutions of 43%. Finally, they also excluded three genes that individually showed evidence of adaptive evolution, leaving a set of 27 genes and a statistically significant estimated adaptive substitution rate of 35%.

Fay, Wyckoff, and Wu [15] took the comparison of ratios of nonsynonymous to synonymous changes (K_a/K_s) within and between species to a new level of sophistication. They built on an initial result that the ratio is twice as large for divergence as for polymorphism, and tested a sophisticated hypothesis: that the difference between polymorphism and divergence can be attributed to only a few genes, rather than being a whole-genome phenomenon. This distinction is important because genome-wide phenomena could have multiple alternative explanations, such as changes in selective constraint, changes in population size or a number of other possibilities in addition to selection. In contrast, any phenomenon that varies between the genes in a genome is more likely to be due to selection. The authors [15] considered substitutions between *D. melanogaster* and its sibling species, *D. simulans*. They began with an initial set of 45 genes for which there were data on both polymorphism and divergence. No genes were excluded for *a priori* reasons, as the point of this test was to detect variation in patterns of selection among genes, precisely the sort of variation that was excluded by Smith and Eyre-Walker [14]. Fay *et al.* [15] concluded that there are two classes of gene: neutral genes that are evolving predominantly by drift, and rapidly evolving genes. The latter comprised approximately 25% of the genes surveyed and included genes involved in the *Drosophila* immune system and reproduction, among others (similar to the results of a previous study [12]).

Fay *et al.* [15] are less whole-heartedly selectionist in their conclusions than Smith and Eyre-Walker [14], in part because they find evidence that some substitutions that Smith and Eyre-Walker consider adaptive to actually be mildly deleterious. In a different study, Fay *et al.* [16] found evidence for mildly deleterious segregating variation in the human genome as well. Their evidence is that a significant excess of rare amino-acid polymorphism was found in autosomal but not X-linked genes, suggesting that these rare polymorphisms are mildly deleterious and are being eliminated more efficiently from the X chromosome because they are partially recessive.

Methods have recently been developed to detect shifts over time in selective pressures on particular amino acids (sites) within a protein sequence [17-20]. One study on the mitochondrial cytochrome b, ATP synthase A chain, NADH dehydrogenase subunit 3, and cytochrome oxidase subunit 2 gene families went so far as to indicate that such shifts in selective pressure are the rule rather than the exception, also in line with a selectionist model of evolution [21]. Whether over a short or a long evolutionary time, such adaptive evolution must be ultimately governed by selective pressures acting on three-dimensional protein structures. Different sites within a protein will be subject to different selective pressures and modeled with a different substitution matrix. This concept has led to the development of context-dependent substitution matrices [22] and subsequently site-class-dependent

substitution matrices [23] to explicitly model structure through evolution, and evolution from structure, respectively. Interestingly, the evolutionary site-dependent substitution matrices resulting from these studies have no clear correlation with biophysical properties of the amino acids. The ultimate general connection between secondary structure, specific mutations between amino acids with different chemical functionalities, and changes in binding properties and enzymatic activities remains an active area of research.

Approaches linking sequence evolution, structural evolution, and organismal selective pressures have become increasingly common. For example, Naylor and Gerstein [24] examined patterns of variation in the globin genes in different mammalian families using both sequence alignments and secondary and tertiary structural considerations. They found that regions of the myoglobin sequence differed in their variability between whales and primates, possibly reflecting the differences in selective pressure for oxygen binding between the muscles of the aquatic whales and terrestrial primates. A preliminary systematic analysis has identified 749 sequences for which automated homology model construction has produced protein structural context data for mutations occurring on branches with high K_a/K_s ratios; these structures now need an automated method to generate and test ecological hypotheses to explain specific mutations in specific proteins ([12], D.A.L. and A. Elofsson, unpublished observations). Systematic, genome-wide analyses combining sequence, structural, and organismal evolution will undoubtedly reveal much of interest about both the mechanisms and results of selection as we progress further into the genome sequencing era.

Scientific theories can go in cycles, even with the steady progression of data. Before Kimura presented his controversial mathematical analysis on the limited newly available protein sequence data from that time [8], selectionist arguments dominated molecular evolution. As genome sequencing data accumulate, the tide seems to be shifting back to a selectionist view. The reason the neutral theory had such a tremendous impact on evolutionary biology is not just that it explained much of the data well, but that it provided a testable null hypothesis [25]. Selection at specific times on specific positions within specific genes in each species may account for much of the functionally significant change in proteins that correlates with species diversification, particularly in genes subject to evolutionary 'arms races' in the fight for survival and reproduction. Molecular evolution shows a backdrop of negatively selected and slowly drifting sites combined with positive selected sites evolving rapidly to produce new protein-binding specificities, binding kinetics, and enzymatic catalysis. A core of residues in each protein might be conserved to retain fold structure and stability, while other residues may be freer to vary (as seen, for example, in leptin [5] and the globins [24]). As mutations are fixed during divergence, different core residues may have to be selected in the new genetic background (sequence landscape) to retain the desired fold and stability. Ultimately, this results in distantly related proteins with similar folds performing chemically similar yet biologically divergent functions [26,27].

The 45 genes analyzed by Fay *et al*. [15] are a tiny fraction of the estimated 13,601 proteins in *D. melanogaster*. We must wait for truly genome-wide estimates of molecular polymorphism and divergence if we are to estimate rates and patterns of adaptive substitutions with confidence. The datasets that could scarcely be imagined ten years ago will soon be at our fingertips. As coding-sequence analyses are carried out on more and more genes, we will see whether the current support for selection is merely another pendulum swing in the selectionist/neutralist debate or a true change in our understanding of evolution. Moreover, future genome-sequencing efforts will enable us to define better the molecular evolutionary processes shaping the genomes themselves, ultimately yielding a better understanding of functions in genomes at all levels of resolution.

References

1. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *TREE* 2000, **15:**496-502.
2. Zhang J: **Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes.** *J Mol Evol* 2000, **50:**56-68.
3. Liberles DA: **Evaluation of methods for determination of a reconstructed history of gene sequence evolution.** *Mol Biol Evol* 2001, **18:**2040-2047.
4. Endo T, Ikeo K, Gojobori T: **Large-scale search for genes on which positive selection may operate.** *Mol Biol Evol* 1996, **13:**685-690.
5. Siltberg J, Liberles DA: **A simple covarion-based approach to analyse nucleotide substitution rates.** *J Evol Biol*, in press.
6. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
7. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in *Drosophila*.** *Nature* 1991, **351:**652-654.
8. Kimura M: *Molecular Evolution Protein Polymorphism and the Neutral Theory.* Berlin: Springer; 1982.
9. Wolfe KH, Sharp PM: **Mammalian gene evolution: nucleotide sequence divergence between mouse and rat.** *J Mol Evol* 1993, **37:**441-456.
10. Messier W, Stewart CB: **Episodic adaptive evolution of primate lysozymes.** *Nature* 1997, **385:**151-154.
11. Benner SA, Trabesinger N, Schreiber D: **Post-genomic science: Converting primary structure into physiological function.** *Advan Enzyme Regul* 1998, **38:**155-180.
12. Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA: **The adaptive evolution database (TAED).** *Genome Biol* 2001, **2:**research0028.1-0028.6.
13. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of duplicate genes.** *Genome Biol* 2002, **3:**research0008.1-0008.9.
14. Smith NGC, Eyre-Walker A: **Adaptive protein evolution in *Drosophila*.** *Nature* 2002, **415:**1022-1024.
15. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome.** *Genetics* 2001, **158:**1227-1234.
16. Fay JC, Wyckoff GJ, Wu CI: **Testing the neutral theory of molecular evolution with genomic data from *Drosophila*.** *Nature* 2002, **415:**1024-1026.
17. Gu X: **Statistical methods for testing functional divergence after gene duplication.** *Mol Biol Evol* 1999, **16:**1664-1674.
18. Gu X: **Maximum-likelihood approach for gene family evolution under functional divergence.** *Mol Biol Evol* 2001, **18:**453-464.

19. Gaucher EA, Miyamoto MM, Benner SA: **Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors.** *Proc Natl Acad Sci USA* 2001, **98:**548-552.
20. Knudsen B, Miyamoto MM: **A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins.** *Proc Natl Acad Sci USA* 2001, **98:**14512-14517.
21. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution.** *Mol Biol Evol* 2002, **19:**1-7.
22. Koshi JM, Goldstein RA: **Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees.** *Prot Eng* 1995, **8:**641-645.
23. Dimmic MW, Mindell DP, Goldstein RA: **Modeling evolution at the protein level using an adjustable amino acid fitness model.** *Pac Symp Biocomp* 2000.
24. Naylor GJP, Gerstein M: **Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins.** *J Mol Evol* 2000, **51:**223-233.
25. Ohta T, Gillespie, JH: **Development of neutral and nearly neutral theories.** *Theor Popul Biol* 1996, **49:**128-142.
26. Hegyi H, Gerstein M: **The relationship between protein structure and function: a comprehensive survey with application.** *J Mol Biol* 1999, **288:**147-164.
27. Penny D, McComish BJ, Charleston MA, Hendy MD: **Mathematical elegance with biochemical realism: the covarion model of molecular evolution.** *J Mol Evol* 2001, **53:**711-723.