

Coding sequence evolution

Martin Kreitman* and Josep M Comeron†

Dramatic progress has been made in the past ten years in the development of statistical and experimental techniques for investigating features of molecular evolution. Applied to coding regions, these techniques have produced remarkable advances in our understanding of selection for codon usage but, ironically, have had little impact on our understanding of protein evolution. That may be about to change.

Addresses

Department of Ecology and Evolution, University of Chicago,
1101 East 57th Street, Chicago, Illinois 60637, USA

*e-mail: mkre@midway.uchicago.edu

†e-mail: jcomeron@midway.uchicago.edu.

Current Opinion in Genetics & Development 1999, **9**:637–641

0959-437X/99/\$ – see front matter © 1999 Elsevier Science Ltd.

All rights reserved.

Abbreviations

cSNP	SNP in a coding region
IDH	isocitrate dehydrogenase
IMDH	isopropylmalate dehydrogenase
MSD	Mutation-Selection-Drift
NAD	nicotinamide adenine dinucleotide
NADP	NAD phosphate
SNP	single-nucleotide polymorphism

Introduction

The preponderance of the available data for analyzing DNA sequence evolution is from coding regions, and so with the exponential increase in data acquisition it should not be unreasonable to expect a corresponding expansion in our knowledge about the evolutionary forces governing the entities they encode — proteins. Unfortunately, this expectation is far from the truth. Comparative analysis shows that proteins are generally conserved entities, and therefore that selective constraint on protein function must be the dominant mode of natural selection acting on amino acid replacement mutations. But beyond that, progress remains slow in understanding adaptive evolution at the molecular level, that is positive natural selection, and how it contributes to evolutionary change in proteins. That said, several research thrusts investigate the role of selection in driving amino acid replacement change and there are hints at more dramatic progress to come.

Coding regions consist of DNA triplets or codons, which each represent a specific amino acid. Degeneracy in the triplet code allows for synonymous substitutions, which do not change the protein. Despite this lack of effect on protein sequence, codon usage is, in some species, highly nonrandom and is governed by natural selection. Substantial progress has been made in the past few years in understanding the nature of this selection.

This article describes several statistical and empirical methodologies that we believe are most useful ones for

exploring evolutionary forces governing the evolution of coding regions. On the statistical side, we make no attempt to offer an exhaustive compendium of all available approaches, but rather focus on ones that consistently suggest interesting non-neural patterns of evolution. These techniques are highlighted in recent work on the adaptive basis of protein evolution and on the nature of selection governing biased codon usage.

Selection on proteins

Two relatively simple-minded statistical techniques are available for asking whether adaptive evolution has driven amino acid replacements in phyletic evolution. The first is one was promulgated by Kimura [1] soon after the advent of DNA sequencing: positive selection for replacement changes must be operating when the number of amino acid replacements per site (K_a) exceed the number of synonymous changes (K_s) in the same protein. This is an extremely stringent criterion for inferring the action of positive selection, as most proteins are slow-evolving (relative to the synonymous substitution rate) despite the fact that many may be evolving entirely by positive selection. But with large amounts of published data there are now numerous reports of proteins that conform to this criterion for inferring selection.

A very interesting case of rapid protein evolution comes from the work of Wu and colleagues [2*,3,4] on proteins related to male reproduction. Wu studies the genetics of male sterility in *Drosophila* species hybrids — a common phenomenon known as Haldane's Rule. They have mapped many genes contributing to this sterility and have even cloned one of them ([2*]; more about this 'speciation' gene later). The rapid evolution of male sterility (in hybrids) and the large number of genes involved, led them to hypothesize that the proteins encoding these genes should also be evolving rapidly, possibly driven by sexual selection. This apparently is the case, not only in the fruit fly [3–5] but also for the mammalian sex determining protein SRY [6] and many other male reproductive proteins in the human lineage [7].

Another useful application of the $K_a > K_s$ test is in the study of newly evolving genes. It has long been known that there can be a speed-up in the rate of amino-acid substitution following the origination of a novel gene through gene duplication [8,9]. For novel genes that arose sufficiently recently, this acceleration can be dramatic and can be shown to be significant by this test [10]. *Odysseus*, a *Drosophila* 'speciation' gene, has, upon cloning and sequencing, turned out to be a newly evolved homeobox gene with an extraordinary rate of amino acid substitution — even in the homeobox portion of the protein [2*].

The second technique for detecting accelerated protein evolution is the McDonald-Kreitman test, which compares ratios of replacement : synonymous polymorphisms within species to that between species [11]. This test is potentially more powerful than the $K_a > K_s$ test but is not without its pitfalls (for example, difficulty in interpreting a significant departure from the null hypothesis — neutrality). This simple test has been widely applied and there is a large list of genes whose polymorphism and divergence depart from neutrality in the direction expected under an alternative model of adaptive evolution. One biologically interesting example is the *G6pd* gene in *Drosophila melanogaster* and *D. simulans*, which shows a 10-fold relative excess of replacement changes between species according to this test, and which is likely to be an example of rapid adaptive evolution of a gene in two species that have undergone dramatic range expansions [12].

This test has also revealed many instances of significantly non-neutral patterns of polymorphism and divergence that are the inverse of that expected for adaptive evolution. In these departures from neutrality, there is a relative excess of amino acid polymorphism compared to divergence. This phenomenon has been most extensively studied by Nachman and by Rand, who have shown that this pattern is common in the mitochondrial genes of rodents and primates [13–15], as well as many other species [16••,17••]. They each propose that a substantial number of deleterious amino acid mutations in mitochondrial proteins contribute to polymorphism within a species but rarely become fixed and contribute to divergence.

There is recent evidence in support of mutation-selection balance for nuclear genes as well, and the evidence has come from one of the first reports of human coding region polymorphism based on large-scale single-nucleotide polymorphism (SNP) hunts [18••]. In a survey of 106 coding regions and averaging 114 independent alleles per gene, these authors identified 392 cSNPs (SNPs in a coding region). Most interestingly, amino acid replacement polymorphisms were skewed to lower frequencies than synonymous polymorphisms, a strong indication that these replacement polymorphisms are, on average, deleterious. Perhaps this is a harbinger of things to come, as SNP surveys become commonplace. More refined statistical methods for detecting departures of observed nucleotide frequencies from neutral expectations are possible and these methods can have considerable statistical power to detect even very weak selection (i.e. nearly neutral mutations) [19]. But because some SNP surveys have strong (and undocumented) ascertainment biases (for example, a study that identifies polymorphisms only if they exceed a certain minimum frequency in a sample) as well as sampling biases, it may not be possible to use them in these kinds of analyses. Careful attention needs to be paid to the ascertainment bias problem if these SNP surveys are to be used to address issues in population genetics.

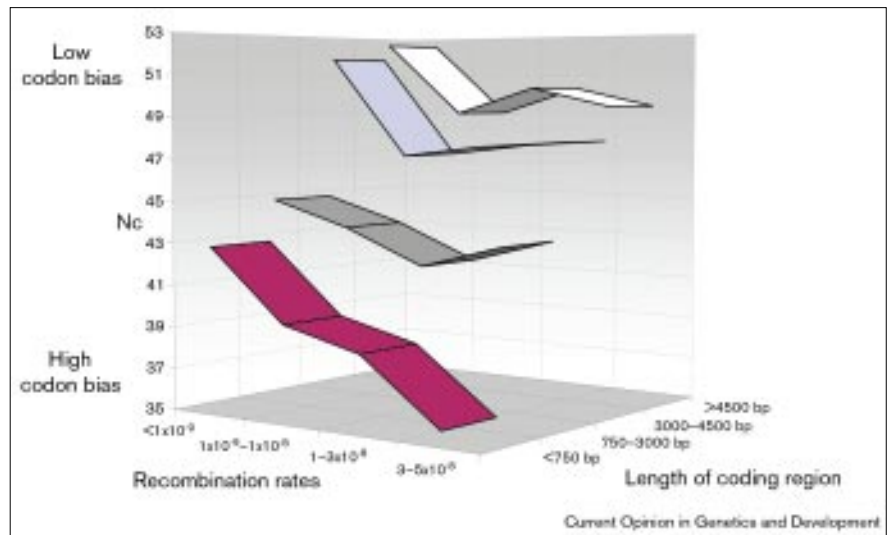
The third thrust to detect selection compares rates of evolution in different lineages. Ohta and Kimura [20] were the first to show that the variance in the rate of protein evolution (for β hemoglobin and cytochrome *c*) exceeded that predicted by the strictly neutral model, and this phenomenon of an ‘overdispersed’ molecular clock has been further documented for mammalian genes by Gillespie [21] and Ohta [22] (see also [23]). One potential limitation of this research thrust, however, is that it has been confined entirely to mammalian gene evolution. The generality of the overdispersed clock has been challenged by a recent study of twenty-four *Drosophila* genes, which found no evidence for an overdispersed molecular clock for replacement changes [24•]. But all of the published estimates of the variance in rates of protein evolution are based on relatively small numbers of genes. The advent of whole-genome sequence comparisons of related species will undoubtedly provide a fresh opportunity to investigate this vexing issue. Further work on theoretical models of molecular evolutionary processes will be needed as well [25].

Finally, we would like to call attention to an entirely different approach to understanding protein adaptation that we believe holds great promise: the integration of phylogenetics, site-directed mutagenesis, and knowledge of protein structure/function [26••]. As an example of this emerging approach, Dean and colleagues [27,28] investigated an ancient evolutionary change in isocitrate dehydrogenase (IDH) that shifted the enzyme from being NAD-dependent to being NADP-dependent. This evolutionary event, it is argued, allowed an ancestor of modern eubacteria to utilize acetate as a carbon source, thus opening up an immense new ecological niche for bacterial exploitation. By applying their knowledge of high-resolution X-ray structures of the binary complexes of *E. coli* IDH with NADP and *Thermus thermophilus* isopropylmalate dehydrogenase (IMDH) with NAD, Chen *et al.* [27] identified six amino acids that together could cause a shift in preference of IDH from NADP to NAD. Two additional changes allowed the modified *E. coli* enzyme to function at comparable levels to the eukaryotic NAD-dependent enzyme. The coenzyme specificity of *T. thermophilus* IMDH was also inverted by an even more spectacular feat of molecular engineering [28]. In this manner these authors were able to infer the sequence of ancestral IDH and to elucidate likely changes that occurred over three billion years ago to bring about this major adaptation!

Missing from Dean’s functional substitution analysis is any assessment of fitness differentials of the substitutions made to these proteins. Such estimates would be highly environment-dependent but will nevertheless be needed to completely flesh out this research approach. Molecular technology has come to the rescue to allow recent progress on this front. The use of classical genetic techniques to produce stocks for fitness comparison make it very difficult to be assured of having lines that differ only at the locus of interest and not anywhere else in the genome. The

Figure 1

A graph showing the relationship between the estimated recombination rate, and the codon bias (N_c , an inverse measure of synonymous codon usage bias) for four different coding region length ranges in *D. melanogaster* (from [55**]).



time-honored approach to deal with this problem has been either to randomize genetic backgrounds or to introduce the alleles of interest into a common genetic background by repeated backcrossing. Neither approach has been entirely satisfying. Genetic engineering, of course, now allows transgenes to be introduced into a standardized genetic background and, in some species, these introductions can be location-specific. In *Drosophila* this is not practical in most situations but to overcome this limitation Siegal and Hartl [29] have developed a method of transgene coplacement, which allows comparisons of transgenes in the same position of the genome.

Fitness analysis between alleles can now be made in truly identical genetic backgrounds, but one largely insurmountable problem remains: natural selection in the wild is at least several orders of magnitude more sensitive than the most sensitive experimental measurements of fitness in laboratory populations (even *E. coli* chemostats). So, this approach will only be useful for investigating mutations with large fitness effects and it will not allow investigation of small selection differentials, which at the end of the day may be the dominant form of selection governing molecular evolution.

Selection on codon usage

Synonymous mutations were initially assumed to be effectively neutral [30,31] but it was apparent upon compilation of the first coding region sequences that the different synonymous codons were not used equally. Codon bias shares several features across taxa: genes that are expressed at high levels tend to use a more restricted set of synonymous codons (preferred or major codons), although this set can differ between organisms [32]; preferred codons correspond to the most abundant tRNA for each amino acid (relative tRNA abundance) [33]; the local chromosomal base composition (variable mutational biases across the genome) can influence codon bias, but the magnitude of

its influence varies considerably; and there is a negative relationship between the extent of codon bias (taking into account mutational tendencies) and the rate of synonymous substitutions [34,35].

The relationships between gene expression, relative tRNA abundance, and degree of codon bias is generally explained by selection acting at the level of translation efficiency, both for translational speed and/or accuracy [36–38]. It has also been attributed to conflicting selection pressures associated with ribosome assembly and translation initiation [39], to mRNA stability and structure [40,41**], and to context dependent nucleotide preferences [42,43]. Natural selection on synonymous mutations has been invoked for a multitude of nuclear genomes, including bacteria, yeast, *Drosophila*, nematode, plants, and also for the chloroplast genome [44]. In mammals, however, evidence for codon selection remains equivocal [45].

Kimura [46] argued that nonrandom codon usage supported the neutral theory by invoking purifying (or negative) rather than positive selection. Li [47] and Bulmer [37] proposed a Mutation-Selection-Drift (MSD) model to better account for the data, which features selection coefficients on synonymous mutations (s) that are very small and close to the reciprocal of the effective population size (N_e ; $N_e s \approx 1$), deleterious and advantageous effects for preferred and unpreferred codons, and variable mutational biases.

Akashi [48] applied a modification of the McDonald-Kreitman test to prove that unpreferred synonymous mutations (those changing a preferred codon to a synonymous unpreferred codon) in *Drosophila* are under different evolutionary constraints than preferred mutations [11]. He showed that unpreferred mutations exhibit a higher ratio of polymorphism to divergence than preferred mutations, and segregate in the population at a lower frequency

[48,49]. The application of population genetics theory under the MSD model [50] has allowed selection coefficients against unpreferred codons to be estimated, expressed as multiples of the effective population size (N_e). Based on the ratio of polymorphism to divergence in *Drosophila*, Akashi [48] estimated $N_e s \approx -2.3$ in *D. simulans* and $N_e s < -1$ for *D. melanogaster*. Using polymorphism data, Hartl *et al.* [51] estimated in *E. coli* $N_e s \approx -0.82$ (based on only two genes), a value close to that estimated when the proportion of preferred codons present in a gene is taken into account ($N_e s \approx -0.54$).

Population genetics models of selection that take into account the non-independence of linked mutations predict a positive correlation between the effectiveness of selection on a mutation and the recombination rate [52,53]. Taking advantage of the fact that recombination rates in *D. melanogaster* genome vary by at least two orders of magnitude, Kliman and Hey [54] were able to show that codon bias in genes located in regions of very low recombination rates is lower than those in other regions. Comeron *et al.* [55**] (see Figure 1) further refined this result by showing that there is a correlation between recombination and codon bias across the whole range of recombination rates, but that this correlation is only detected when genes with different length are analyzed separately. Thus, both the length of the coding region and the recombination rate modulate codon bias in *Drosophila*. Indeed, in *Drosophila* there is negative correlation between gene length and codon bias [55**,56**], as well as a positive relationship between gene length and the rate of synonymous substitution [57].

These observations in *Drosophila* are congruent with stronger selection on synonymous mutation in short genes than long genes, and two possible causes have been proposed [55**]: the functional effect of individual unpreferred mutations decreases with the number of codons of a gene, and interference among selected mutations increases with gene length, making it harder for codon bias to evolve in long genes.

Similar effects of gene length on codon bias have been also detected in *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae* [56**,58]. A different pattern is seen in *E. coli*, however, in which gene length might be optimized in highly expressed genes and a strong component of selection on synonymous mutations acting at the level of translational accuracy has been proposed [56**,59].

Conclusions

The genome-sequence era will undoubtedly spark a fresh attack on classical problems in evolutionary genetics, and the issue of whether protein variation is mutation-selection driven is at the top of that list. One of the most satisfying aspects of the recent work on codon usage is the realization that statistical techniques developed to analyze sequence variation and evolution are indeed powerful enough to reveal the most subtle selection possible — nearly neutral

mutations. Codon preference, however, applies to every gene in a genome, whereas the evolution of each amino acid and each protein is highly context-dependent. The new millennium will tell whether massive amounts of evolutionary data will overcome inherent limitations in the evolutionary analysis of proteins.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.** *Nature* 1977, **267**:275-276.
 2. Ting CT, Tsaur SC, Wu ML, Wu CI: **A rapidly evolving homeobox at the site of a hybrid sterility gene.** *Science* 1998, **282**:1501-1504. The positional cloning of the *Odysseus* gene was a *tour de force*. That it turns out to be a very rapidly evolving homeobox gene came as a great surprise. The selective force(s) driving the rapid evolution is unknown.
 3. Tsaur SC, Wu CI: **Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*.** *Mol Bio Evol* 1997, **14**:544-549.
 4. Tsaur SC, Ting CT, Wu CI: **Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*: II divergence vs. polymorphism.** *Mol Bio Evol* 1998, **15**:1040-1046.
 5. Civetta A, Singh R: **Sex-related genes, directional sexual selection and speciation.** *Mol Bio Evol* 1998, **15**:901-909.
 6. Whitfield LS, Lovell-Badge R, Goodfellow PN: **Rapid sequence evolution of the mammalian sex-determining gene *SRY*.** *Nature* 1993, **364**:713-715.
 7. Wyckoff G, Wang W, Wu GI: **Rapid evolution of male reproductive genes in the descent of man.** *Nature* 1999, in press.
 8. Ohta T: **Multigene families and the evolution of complexity.** *J Mol Evol* 1991, **33**:34-41.
 9. Ohta T: **Further examples of evolution by gene duplication revealed through DNA sequence comparisons.** *Genetics* 1994, **138**:1331-1337.
 10. Long M, Langley CH: **Natural selection and the origin of *jingwei*, a chimaeric processed functional gene in *Drosophila*.** *Science* 1993, **260**:91-95.
 11. McDonald JH, Kreitman M: **Adaptive protein evolution at the *Adh* locus in *Drosophila*.** *Nature* 1991, **351**:652-654.
 12. Eanes WF, Kirchner F, Yoon J: **Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages.** *Proc Natl Acad Sci USA* 1993, **90**:7475-7479.
 13. Nachman ME, Brown WM, Stoneking M, Aquadro CF: **Nonneutral mitochondrial DNA variation in humans and chimpanzees.** *Genetics* 1996, **142**:953-963.
 14. Nachman ME, Boyer SN, Aquadro CF: **Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice.** *Proc Natl Acad Sci USA* 1994, **91**:6364-6368.
 15. Kennedy P, Nachman ME: **Deleterious mutations at the mitochondrial ND3 gene in south american marsh rats (*Holochilus*).** *Genetics* 1998, **150**:359-368.
 16. Nachman ME: **Deleterious mutations in animal mitochondrial DNA.** •• *Genetica* 1998, **102-103**:61-69. This paper and the following one [17**] are independent compilations of mitochondrial polymorphism and divergence data, and both apply the McDonald-Kreitman test [9] to show that a substantial number of data sets depart from neutrality in the direction predicted for deleterious mutations.
 17. Rand DM, Kann LM: **Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA.** *Genetica* 1998, **102-103**:393-407. See annotation [16**].

18. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, ●● Lim EP, Kalayanaraman N, Nemesh J *et al.*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22**:231-238.
This paper is the first of its kind reporting the results of a large-scale screen for human nucleotide polymorphisms.
19. Akashi H: **Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination.** *Genetics* 1999, **151**:221-238.
20. Ohta T, Kimura M: **On the constancy of the evolutionary rate of cistrons.** *J Molec Evol* 1971, **1**:18-25.
21. Gillespie JH: *The Causes of Molecular Evolution.* Oxford: Oxford University Press; 1991.
22. Ohta T: **Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory.** *Mol Evol* 1995, **40**:56-63.
23. Nielsen R: **Robustness of the estimator of the index of dispersion for DNA sequences.** *Mol Phylogenet Evol* 1997, **7**:346-351.
24. Zeng LW, Comeron JM, Chen B, Kreitman M: **The molecular clock • revisited: the rate of synonymous vs. replacement change in *Drosophila*.** *Genetica* 1998, **102-103**:369-382.
The analysis of the variance to mean ratio – or Index of Dispersion – of substitutions in three *Drosophila* lineages reveals an overdispersion of synonymous substitutions but not nonsynonymous substitutions, the opposite of results for mammals.
25. Ohta T, Gillespie JH: **Development of neutral and nearly neutral theories.** *Theor Popul Biol* 1996, **49**:128-142.
26. Golding GB, Dean AM: **The structural basis of molecular •• adaptation.** *Mol Biol Evol* 1998, **15**:355-369.
This is a provocative essay intended to push evolutionary geneticists towards functional and empirical approaches to the study of molecular adaptation. Six examples are taken from the literature to illustrate the power of this approach.
27. Chen R, Greer A, Dean AM: **A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity.** *Proc Natl Acad Sci USA* 1995, **92**:11666-11670.
28. Chen R, Greer A, Dean AM: **Redesigning secondary structure to invert coenzyme specificity in isopropylmalate dehydrogenase.** *Proc Natl Acad Sci USA* 1996, **93**:12171-12176.
29. Siegal ML, Hartl DL: **An experimental test for lineage-specific position effects on alcohol dehydrogenase (*Adh*) genes in *Drosophila*.** *Proc Natl Acad Sci USA* 1998, **95**:15513-15518.
30. Akashi H, Eyre-Walker A: **Translational selection and molecular evolution.** *Curr Opin Genet Dev* 1998, **8**:688-693.
31. Kimura M: **Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles.** *Genet Res* 1968, **11**:247-269.
32. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R: **Codon catalog usage is a genome strategy modulated for gene expressivity.** *Nucleic Acids Res* 1981, **9**:R43-R74.
33. Ikemura T: **Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system.** *J Mol Biol* 1981, **151**:389-409.
34. Sharp PM, Li W-H: **The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias.** *Mol Biol Evol* 1987, **4**:222-230.
35. Sharp PM, Li W-H: **On the rate of DNA sequence evolution in *Drosophila*.** *J Mol Evol* 1989, **28**:398-402.
36. Kurland CG: **Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes.** *Trends Biochem Sci* 1987, **12**:169-171.
37. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129**:897-907.
38. Akashi H: **Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy.** *Genetics* 1994, **136**:927-935.
39. Eyre-Walker A, Bulmer M: **Reduced synonymous substitution rate at the start of enterobacterial genes.** *Nucleic Acids Res* 1993, **21**:4599-4603.
40. Parsh J, Tanda S, Stephan W: **Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 1997, **94**:928-933.
41. Shen LX, Basilion JP, Vincent P, Stanton P Jr: **Single-nucleotide •• polymorphisms can cause different structural folds of mRNA.** *Proc Natl Acad Sci USA* 1999, **96**:7871-7876.
This study shows the effects on mRNA secondary structure of two SNPs in the coding region of two human genes.
42. Karlin S, Mrazek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262**:459-472.
43. Antezana M, Kreitman M: **The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences.** *J Mol Evol* 1999, **49**:36-43.
44. Morton BR: **Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages.** *J Mol Evol* 1998, **46**:449-459.
45. Mouchiroud D, Gautier C, Bernardi G: **Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions.** *J Mol Evol* 1995, **40**:107-113.
46. Kimura M: *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press; 1983.
47. Li W-H: **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons.** *J Mol Evol* 1987, **24**:337-345.
48. Akashi H: **Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA.** *Genetics* 1995, **139**:1067-1076.
49. Akashi H, Schaeffer SW: **Natural selection and the frequency distributions of 'silent' DNA polymorphism in *Drosophila*.** *Genetics* 1997, **146**:295-307.
50. Sawyer SA, Hartl DL: **Population genetics of polymorphism and divergence.** *Genetics* 1992, **132**:1161-1176.
51. Hartl DL, Moriyama EN, Sawyer SA: **Selection intensity for codon bias.** *Genetics* 1994, **138**:227-234.
52. Hill WG, Robertson A: **The effect of linkage on the limits to artificial selection.** *Genet Res* 1966, **8**:269-294.
53. Felsenstein J: **The evolutionary advantage of recombination.** *Genetics* 1974, **78**:737-756.
54. Kliman RM, Hey J: **Reduced natural selection associated with low recombination in *Drosophila melanogaster*.** *Mol Biol Evol* 1993, **10**:1239-1258.
55. Comeron J, Kreitman M, Aguadé M: **Natural selection on •• synonymous sites is correlated with gene length and recombination rate in *Drosophila*.** *Genetics* 1999, **151**:239-249.
See annotation [56**].
56. Moriyama EN, Powell JR: **Gene length and codon usage bias in •• *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*.** *Nucleic Acids Res* 1998, **26**:3188-3193.
This paper and [55**] establish the relationship between gene length and codon bias in several organisms, and offer hypotheses to account for the phenomenon.
57. Comeron JM, Aguadé M: **Synonymous substitutions in the *Xdh* gene of *Drosophila*: heterogeneous distribution along the coding region.** *Genetics* 1996, **144**:1053-1062.
58. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*.** *Proc Natl Acad Sci USA* 1999, **96**:4482-4487.
59. Eyre-Walker A: **Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy?** *Mol Biol Evol* 1996, **13**:864-872.