Gene expression and molecular evolution

Hiroshi Akashi

The combination of complete genome sequence information and estimates of mRNA abundances have begun to reveal causes of both silent and protein sequence evolution. Translational selection appears to explain patterns of synonymous codon usage in many prokaryotes as well as a number of eukaryotic model organisms (with the notable exception of vertebrates). Relationships between gene length and codon usage bias, however, remain unexplained. Intriguing correlations between expression patterns and protein divergence suggest some general mechanisms underlying protein evolution.

Addresses
Institute of Molecular Evolutionary Genetics and Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 06138, USA; e-mail: akashi@psu.edu

Current Opinion in Genetics & Development 2001, **11**:660–666

0959-437X/01/$ – see front matter

Abbreviation

EST	expressed sequence tag

Introduction

Sequence data from complete genomes and global estimates of gene expression provide rich sources of new information with which to determine mechanisms of both codon-usage bias and protein evolution. Synonymous changes do not alter the amino acid sequence encoded in DNA and were thought to be the among the best candidates for canonically neutrally evolving sites [1,2]. However, selectionist responses immediately following proposals of neutral molecular evolution by Kimura [3] and King and Jukes [1] included speculation that translational selection occurs at silent sites.

Both Clarke [4] and Richmond [5] noted that, because tRNAs for a given amino acid are found in unequal concentrations, synonymous codon choice might affect fitness through their effect on the rate of protein synthesis. Small, but evolutionarily significant, differences in fitness among synonymous alternatives are now supported by a combination of laboratory studies and DNA sequence analyses in a wide range of taxa (reviewed in [6–8]).

Relationships between gene expression and levels of synonymous codon usage provide an important line of evidence for translation selection. Recently, genome-wide estimates of expression levels have become available for model organisms and have helped both to establish major codon preferences, especially among multicellular eukaryotes, and to determine the phenotypic bases of selection at silent sites. Although a simple evolutionary model of mutation–selection drift appears to account for many features of silent DNA variation within and between genomes, some patterns, such as negative associations between gene length and codon-usage bias, remain unexplained.

Perhaps more surprising than the relationships between codon-usage bias and gene expression are associations between the evolutionary rates of proteins and the tissue-specificity and breadth of their expression. Such patterns may shed light on factors that underlie ‘functional constraint’, such as the biochemical complexity of protein interactions. In addition, expression estimates may help to determine whether the primary structures of proteins reflect natural selection to enhance translational efficiency as well as the specific functions of polypeptides. This review will focus on recent studies that explore how knowledge of gene expression levels can shed light on the causes of both DNA and protein evolution.

Preference for major codons at silent sites in DNA

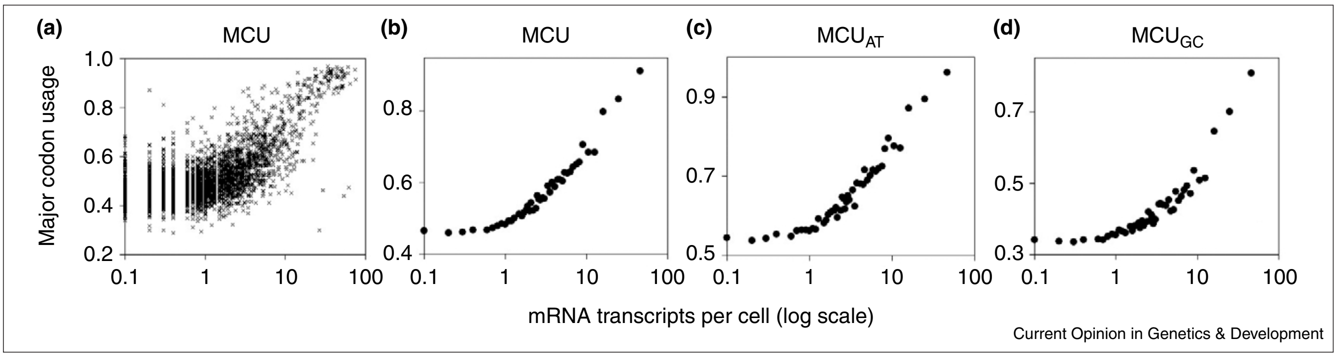
In most genomes, synonymous codon usage shows an overall ‘bias’, or departure from random usage, towards what have been termed ‘major codons’. Major codons are, almost without exception, recognized by cognate tRNAs that are relatively abundant and/or have perfect Watson–Crick pairing [9–11]. Laboratory experiments in *Escherichia coli* have shown that major codons are recognized and translated more quickly, and with fewer errors, than their less abundant counterparts (reviewed in [6]).

Faster rates of ribosomal elongation allow more efficient use of the protein synthesis machinery in the cell (i.e. more incorporations per ribosome per unit of time). In addition, major codons may reduce the energetic costs of proofreading (i.e. rejecting non-cognate tRNAs) during protein synthesis and may decrease the costs of synthesizing dysfunctional peptides, by reducing the probabilities of both misincorporations and processivity errors (ribosomal frameshifting and drop-off; reviewed in [12]). Under such a scheme, the fitness benefit to encoding a major codon is predicted to be a function of the number of aminoacyl-tRNA selections at a given codon; therefore, major codons should be more beneficial in highly expressed genes.

Levels of gene expression and codon usage

Early studies in *E. coli* [13–15] and the budding yeast *Saccharomyces cerevisiae* [16] revealed strong bias in synonymous codon usage for genes encoding abundant proteins such as ribosomal proteins and elongation factors, RNA polymerase subunits and glycolytic enzymes. In contrast, the codon usage of presumably low-expression transcription factors and other regulatory genes is more uniform. Quantitative data for mRNA abundances [16] and for protein abundances measured by 2D gel electrophoresis [17] have established strong correlations between the bias

Figure 1



Preference for major codons in *S. cerevisiae*: mRNA abundance and codon-usage bias. **(a)** Relationships between major codon usage, MCU = major/(major + minor) codons, and mRNA abundance [20] among all genes with detected transcripts. **(b)** The same relationships among expression-level classes. Genes were ranked by mRNA abundance and binned into categories containing at least 10,000 codons per category. Average MCU and mRNA levels among genes are plotted. **(c)** Relationships among expression categories for major codon usage among synonymous families with all A- and T-ending major codons (MCU_{AT}). **(d)** Relationships among expression categories for major codon usage among synonymous families with all G- and C-ending major codons (MCU_{GC}). Codon preferences are taken from [11].

in synonymous codon usage and estimates of translation rates for a small number of genes (eight in both studies).

More recent data from serial analysis of gene expression (SAGE), high-density oligonucleotide arrays (GeneChips) and expressed sequence tag (EST) libraries (reviewed in [18]) have increased the scope of expression analyses to a scale of thousands of genes. The highest quality whole-genome mRNA abundance data are found in studies of *S. cerevisiae* [19,20]. Although the numbers of identified protein spots on 2D gels are small, the abundances of mRNA and their corresponding proteins show fairly strong associations (n = 148, Spearman rank correlation, r_s = 0.8 in [21]; and n = 156, r_s = 0.59 in [22]).

But comparisons between different studies of mRNA quantification reveal unexpected inconsistencies. Coghlan and Wolfe [23••] found that estimates of mRNA levels from different laboratories show a surprisingly weak association (r_s = 0.68), even though the studies used GeneChips from the same manufacturer with similar strains of yeast grown in similar conditions [19,20]. These differences in expression patterns may be due to both a lack of precision in estimating mRNA levels and actual differences in the mRNA populations of the cells examined. Although the laboratory conditions were similar in the experiments, the microenvironment and cell density may alter the patterns of expression of several genes [24,25]. In addition, it is unclear how well mRNA abundances in a single laboratory environment reflect expression patterns within cells exposed to a cross-section of natural environments.

Despite these caveats, measures of mRNA abundance seem to be informative predictors of codon-usage bias. Although the positive correlation between major codon usage and mRNA abundance among *S. cerevisiae* genes shows a great deal of scatter (Figure 1a), the *average* codon-usage bias increases steadily among expression classes of genes (Figure 1b).

Duret and Mouchiroud [26] have used sequence matches to EST libraries to show similar relationships between codon-usage bias and mRNA abundances in the multicellular eukaryotes *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. These data may be more error-prone than oligo-chip estimates of mRNA abundances because of biases in the tissues sampled, biases in cloning mRNAs, and the 'normalization' of cDNA libraries (adjustment toward uniform concentrations of cDNAs from different genes) prior to DNA sequencing. Gene prediction may also be less accurate than in unicellular organisms where introns (and alternative transcripts) are less abundant. However, broad-scale associations between mRNA abundance and codon-usage bias add support to strong anecdotal evidence for relationships between levels of gene expression and codon-usage bias [27–29], as well as to both associations between codon-usage bias and tRNA abundances [30,31•], and population genetic tests of weak selection at silent sites ([32–37]; but see also [38••,39••]) in model eukaryotes.

Transcription, mutation and major codon usage

Although models describing the evolution of codon-usage bias generally assume that mutational processes occur uniformly among genes ([40–42]; but see [43•]), relationships between transcription levels and mutation patterns have been established in *E. coli* and *S. cerevisiae*. In *E. coli*, C to T mutations occur at relatively high frequencies on the non-coding strand of DNA [44], presumably through a mechanism involving deamination of cytosines in single-stranded DNA during transcription [45•,46]. Such a process will alter both the rate and the spectrum of mutations as a function of the level of gene expression; that is, cytosine deamination should lead to an excess of thymines in highly expressed genes. However, the magnitude of transcription-coupled mutational processes, and its taxonomic breadth, remain to be determined. In yeast, frameshift mutations are strongly dependent on the rate of gene transcription [47,48]; however, the dependence of nucleotide mutation on transcription has not been established.

Associations between mutational processes and transcription do not seem to explain correlations between gene expression and codon-usage bias in *S. cerevisiae*, *D. melanogaster* or *C. elegans*. Increases in major codon usage with expression level for both AT-ending and GC-ending major codons argues against a mutational explanation for codon-usage bias in *S. cerevisiae* (Figure 1c, d). In *D. melanogaster* and *C. elegans*, almost all major codons encode G or C in the third position. In *D. melanogaster*, the G+C content is uniformly higher at silent sites in coding regions than in putatively neutrally evolving introns [49], and the base composition of introns shows no correlation with levels of gene expression [26]. In addition, within alternatively spliced genes, constitutively translated exons show higher major codon usage than alternatively spliced exons that are transcribed at the same rate but translated at lower levels [50•]. *C. elegans* shows a weak, but statistically significant, negative correlation between G+C content and gene expression levels [26]. In the absence of evidence for strong transcription-dependent mutational biases toward major codons, genome-wide estimates of expression levels strongly suggest that selection coefficients at silent sites depend on rates of translation.

Gene length and codon-usage bias

Relationships among gene expression, gene length and codon-usage bias may help to elucidate the phenotypic bases of selection at silent sites. Correlations between gene length and codon-usage bias are positive in *E. coli* and negative in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *A. thaliana* [26,51,52]. Jansen and Gerstein [53••] have noted that, among yeast genes, gene length seems to set an upper limit on mRNA abundance: although the correlation between gene length and mRNA abundance is weak, the maximum size of proteins decreases steadily as a function of mRNA levels.

Relationships between levels of gene expression and the function, or subcellular localization, of proteins might explain this pattern. For example, integral membrane proteins show markedly reduced transcript levels relative to cytosolic proteins [54] and are generally longer. Alternatively, selection for metabolic efficiency might affect both protein size and synonymous codon usage [55]. In terms of translational costs, reductions in protein size should have advantages similar to those of encoding major codons, with an added benefit of reductions in the costs of amino acid biosynthesis and/or transport.

Estimates of gene expression levels also allow us to test whether natural selection discriminates among synonymous codons to enhance translational fidelity. Selection to minimize translational misincorporations predicts stronger selection, and higher codon-usage bias, at codons where amino acid misincorporations result in the costly synthesis of dysfunctional peptides [33]. Such patterns have been observed within genes in both *D. melanogaster* [33] and *C. elegans* [56•]. Among-gene patterns can also be used to test for translational accuracy. Eyre-Walker [57] proposed that, for proteins translated at the same rates, selection to reduce translational misincorporations should be higher in longer genes because the cost of producing dysfunctional peptides will be proportional to their length.

Relationships between codon usage and gene length have been observed among ribosomal protein genes in both *E. coli* [57] and *S. cerevisiae* [51]. Coghlan and Wolfe [23••] extended this analysis to the entire yeast genome by employing partial correlations between gene length and codon-usage bias, excluding effects of expression level, and found positive correlations between gene length and codon-usage bias. Thus, gene expression data suggest that translational selection may both reduce the sizes of highly expressed genes and enhance the fidelity of protein synthesis (reduce the rate of misincorporations and/or processivity errors).

Mechanisms underlying relationships between gene length and codon-usage bias in multicellular eukaryotes are less clear. In contrast to patterns found in *S. cerevisiae*, Duret and Mouchiroud [26] found no relationship between transcript levels and gene length in *A. thaliana* and *D. melanogaster*, and in fact an increase in gene size with transcript levels in *C. elegans*. There is no evidence to support selection for reduced protein size.

In the absence of expression differences among protein size categories, negative correlations between gene length and codon-usage bias could arise from interference in the evolutionary dynamics of selected sites that are genetically linked [40,52,58]. All else being equal, sites within longer genes will be closely linked to greater numbers of segregating non-neutral mutations and thus have lower expected levels of codon-usage bias [40,52]. In *C. elegans*, however, this explanation does not appear to be satisfactory; highly expressed genes with highly expressed neighbors do not show lower codon-usage bias than those without highly expressed neighbors [26]. The causes of associations between gene size and codon usage in multicellular eukaryotes remain undetermined.

Gene expression and protein evolution

Studies of expression patterns at a genomic scale have, for the most part, confirmed existing models of silent DNA evolution. At the protein level, however, measures of mRNA abundance have revealed many surprising patterns for which the causes remain speculative.

Duret and Mouchiroud [59••] have established that both the tissue specificity and the breadth of expression have an effect on rates of mammalian protein evolution. They compared 2,400 genes between human and rodent (mostly mouse) and quantified expression levels by counting gene sequence matches in tissue-specific EST libraries. 'Ubiquitously' expressed proteins, whose mRNAs were detected in 16 out of 19 tissues, evolved at average rates that were threefold lower than those of tissue-specific genes. Among the genes whose transcripts were found in single (or few) tissues, rates were twofold lower for genes expressed in brain, muscle, retina and neurons, relative to those expressed in lymphocyte, lung and liver. Silent divergence does not show such patterns, suggesting that variation in mutation rates does not cause these differences in rates of protein evolution.

Duret and Mouchiroud [59••] attributed these patterns to greater constraints among proteins functioning in more complex biochemical environments than those that are active in a narrower range of complexity (including pH and the numbers of interacting proteins) [60,61]. They also noted that mutations affecting proteins expressed in a larger number of tissues may be more likely to affect an organism's fitness than those whose expression is tissue-specific. Their explanation is consistent with patterns of protein evolution within alternatively spliced genes; regions of proteins encoded by constitutive exons expressed in a larger number of tissues, and at higher levels, show slower rates of evolution than those encoded by alternatively spliced exons [50•].

Differences in expression patterns among proteins that fall into different functional categories, or that contain different structural elements, might underlie correlations between patterns of mRNA abundance and rates of protein evolution. Solvent-exposed amino acids evolve at roughly twice the rate of 'buried' (interior) sites in globular proteins [62] and are more likely to be polymorphic within species [63••]. Within solvent-exposed regions, putative sites of interaction among proteins may result in patches of conserved amino acids [64].

Transmembrane regions evolve more slowly than solvent-exposed regions [65,66] within membrane-associated peptides, and Drawid *et al.* [54] have shown that yeast membrane-associated proteins are generally expressed less than cytosolic proteins. Low rates of mammalian brain protein evolution might reflect both the greater number of physical interactions among proteins (slower evolution among solvent-exposed sites) and the expression of a relatively large number of membrane proteins (i.e. receptors). Closer examination of rate variation within and among functional classes and structural elements should help to delineate the causes of tissue-specific expression and rates of protein evolution.

Recently, Pál *et al.* [67••] have established relationships between expression patterns and protein evolution in *S. cerevisiae* that may not be explained by the factors discussed above. They compared 185 gene pairs related by a whole-genome duplication event and found a negative relationship between estimates of mRNA levels and protein divergence. Because yeast is unicellular, such a pattern suggests that the level of expression, rather than the breadth of expression among tissues, may be an important variable associated with evolutionary rates. To detect EST 'hits' in many tissue-specific libraries, the abundance of mRNA must be relatively high across tissues; 'broadly' expressed genes in Duret and Mouchiroud's study [59••] are necessarily also 'highly' expressed. The negative correlation found by Pál *et al.* [67••] between evolutionary rates and expression levels holds for proteins within the same functional category and thus cannot be explained by differences in rates among structural elements.

Direct or indirect relationships between translational selection and protein evolution offer potential explanations for the findings of Pál *et al.* [67••]. In *S. cerevisiae*, roughly half of the major codons end in AT and half in GC. A mutation between amino acids can result in a change in the translational preference status of a codon. For example, a first codon position mutation from AAG to CAG would convert the major codon for lysine to a minor codon for glutamine. Such a change could be neutral at the protein level but translationally unpreferred. Thus, major codon preference may limit the spectrum of neutral non-synonymous mutations. This would not seem to explain the observations of Duret and Mouchiroud [59••], however, because translational selection has not been established in mammals.

The argument posed above assumes that major codons for different amino acids have similar translational effects on fitness. It is possible, however, that the major codon for one amino acid is translationally superior or inferior to a major codon for a different amino acid (and similarly among minor codons). Correspondences between amino acid usage and total cognate tRNA concentrations, or gene copy numbers, for each amino acid have been found in *E. coli*, *Mycoplasma capricolum* [10] and yeast [68], as well as in *C. elegans* [31•]. Yamao *et al.* [10] argued that such correlations reflect selection on tRNA concentrations (mostly through selection on gene copy number) to adjust to the amino acid requirements of highly expressed proteins.

Percudani *et al.* [68] and Lobry and Gautier [69] favor selection both on tRNA pools and amino acid usage to explain the same pattern (see also [55,70]). The crucial difference between these schemes is whether translational selection affects amino acid usage and rates of protein evolution. Current evidence does not distinguish between unidirectional adjustment of tRNA pools to protein requirements and co-adaptation of tRNA abundances and amino acid composition.

Conclusions

Associations between gene sequence evolution and expression patterns on a scale of thousands of genes may help to determine mechanisms of molecular evolution.

Selection among synonymous codons demonstrates that even minute phenotypic effects may be subject to natural selection; genome-wide codon-usage bias seems to result from benefits of increasing the efficiency of resource utilization and allowing more rapid growth.

Gene expression patterns may reveal cause(s) of the >100-fold variation in rates of protein evolution [71]. By integrating gene expression information, the functional and structural categorization of proteins, the physical location and recombination rates experienced by genes, and cellular concentrations of tRNAs, we may shed light on this central issue in molecular evolution.

Acknowledgements

K Mineta, WY Ko, CF Lin and R David provided many suggestions that have helped to improve this manuscript. This work was supported by the Alfred P Sloan Foundation.

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. King JL, Jukes TH: **Non-Darwinian evolution.** *Science* 1969, **164**:788-797.

2. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.** *Nature* 1977, **267**:275-276.

3. Kimura M: **Evolutionary rate at the molecular level.** *Nature* 1968, **217**:624-626.

4. Clarke B: **Darwinian evolution of proteins.** *Science* 1970, **168**:1009-1011.

5. Richmond RC: **Non-Darwinian evolution: a critique.** *Nature* 1970, **225**:1025-1028.

6. Andersson SGE, Kurland CG: **Codon preferences in free-living microorganisms.** *Microbiol Rev* 1990, **54**:198-210.

7. Sharp PM, Stenico M, Peden JF, Lloyd AT: **Codon usage: mutational bias, translational selection, or both?** *Biochem Soc Trans* 1993, **21**:835-841.

8. Akashi H, Eyre-Walker AC: **Translational selection and molecular evolution.** *Curr Opin Genet Dev* 1998, **8**:688-693

9. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2**:13-34.

10. Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S: **Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins.** *Nucleic Acids Res* 1991, **19**:6119-6122.

11. Kanaya S, Yamada Y, Kudo Y, Ikemura T: **Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238**:143-155.

12. Kurland CG: **Translational accuracy and the fitness of bacteria.** *Annu Rev Genet* 1992, **26**:29-50.

13. 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-r73.

14. Grosjean H, Fiers W: **Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes.** *Gene* 1982, **18**:199-209.

15. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.** *Nucleic Acid Res* 1982, **10**:7055-7074.

16. Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3036-3031.

17. 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* translation system.** *J Mol Biol* 1981, **151**:389-409.

18. Gerstein M, Jansen R: **The current excitement in bioinformatics — analysis of whole-genome expression data: how does it relate to protein structure and function?** *Curr Opin Struct Biol* 2000, **10**:574-584.

19. Cho RJ, Campbell MJ, Winzeler EA, Stenmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.

20. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CH, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.

21. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**:7357-7368.

22. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.

23. •• Coghlan A, Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, **16**:1131-1145.

Relatively strong correlations exist between measures of codon-usage bias and gene expression estimates, although the correlations among expression studies from different labs are surprisingly weak. mRNA abundance shows a negative partial correlation with protein length (controlling for codon-usage bias that may influence transcription rates). Partial correlations that eliminate the contribution of mRNA levels show positive relationships between codon-usage bias and gene length.

24. Lander ES: **Array of hope.** *Nat Genet* 1999, **21**:3-4.

25. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: **Genome-wide expression monitoring in *Saccharomyces cerevisiae*.** *Nat Biotechnol* 1997, **15**:1359-1367.

26. 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.

27. Chiapello H, Fisacek F, Caboche M, Henaut A: **Codon usage and gene function are related in sequences of *Arabidopsis thaliana*.** *Gene* 1998, **209**:GC1-GC38.

28. Shields DC, Sharp PM, Higgins DG, Wright F: **'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.

29. Stenico M, Lloyd AT, Sharp PM: **Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases.** *Nucleic Acids Res* 1994, **22**:2437-2446.

30. Moriyama EN, Powell J: **Codon usage bias and tRNA abundance in *Drosophila*.** *J Mol Evol* 1997, **45**:514-523

31. • Duret L: **tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes.** *Trends Genet* 2000, **16**:287-289.

The cellular concentrations of tRNAs are strongly correlated with their gene copy numbers in *E. coli*, *Bacillus subtilis*, *M. capricolum*, and *S. cerevisiae*. The author shows that tRNA gene copy numbers in *C. elegans* correlate strongly with both synonymous codon preferences and amino acid usage, suggesting that there is a similar correspondence in a multicellular eukaryote.

32. Kliman RM, Hey J: **Reduced natural selection associated with low recombination in *Drosophila melanogaster*.** *Mol Biol Evol* 1993, **10**:1239-1258.

33. Akashi H: **Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy.** *Genetics* 1994, **136**:927-935.

34. Akashi H: **Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA.** *Genetics* 1995, **139**:1067-1076.

35. Kliman RM: **Recent selection on synonymous codon usage in *Drosophila*.** *J Mol Evol* 1999, **49**:343-351.

36. Llopart A, Aguade M: **Nucleotide polymorphism at the RpII215 gene in *Drosophila subobscura*: weak selection on synonymous mutations.** *Genetics* 2000, **155**:1245-1252.

37. Begun DJ: **The frequency distribution of nucleotide variation in *Drosophila simulans*.** *Mol Biol Evol*, **18**:1343-1352.

38. Marias G, Mouchiroud D, Duret L: **Does recombination improve**
•• **selection on codon usage? Lessons from nematode and fly complete genomes.** *Proc Natl Acad Sci USA* 2001, **98**:5688-5692.

The authors cast doubt on the relationship between genetic crossovers and the efficacy of selection at silent sites [32,52] by demonstrating a relationship between mutational processes and local recombination rates. Intron G+C content increases as a function of recombination rates in both *D. melanogaster* and *C. elegans*. The increase of codon-usage bias with recombination rates reflects that almost all preferred codons end in G or C in both genomes. The authors suggest that recombination itself might be mutagenic and, when corrected for mutational spectra differences, find no relationship between recombination and codon-usage bias. Future studies will need to incorporate the relationships among gene expression, recombination, and mutation rates and biases in testing weak selection at silent sites.

39. Dunn KA, Bielawski JP, Yang Z: **Substitution rates in *Drosophila***
•• **nuclear genes: implications for translational selection.** *Genetics* 2001, **157**:295-305.

Maximum likelihood models provide a rigorous statistical framework for estimating silent DNA divergence. The authors find strong evidence for a relationship between silent and replacement rates of substitution among genes but no evidence for a negative correlation between codon usage and silent divergence. Previous studies relied on methods that overestimate the numbers of silent 'sites' and underestimate levels of divergence. Negative correlations between codon-usage bias and divergence can result as an artifact of this bias because the underestimation of divergence is a function of codon-usage bias. This result challenges what has been held as an important piece of evidence supporting major codon preference.

40. Li WH: **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons.** *J Mol Evol* 1987, **24**:337-345.

41. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129**:897-907.

42. McVean GAT, Charlesworth B: **A population genetic model for the evolution of synonymous codon usage: patterns and predictions.** *Genet Res* 1999, **74**:145-158.

43. McVean GAT, Vieira J: **Inferring parameters of mutation, selection,**
• **and demography from patterns of synonymous site evolution in *Drosophila*.** *Genetics* 2001, **157**:245-257.

In this study, likelihood models are applied to estimate parameters and test models that simultaneously consider codon-usage bias and levels of divergence among species. By adding parameters within nested models, they find evidence for variation in selection intensity both among amino acids and among lineages. However, the overall fit of the model to the data is weak for a surprisingly large number of genes.

44. Francino MP, Chao L, Riley MA, Ochman H: **Asymmetries generated by transcription-coupled repair in enterobacterial genes.** *Science* 1996, **272**:107-109.

45. Francino MP, Ochman H: **Deamination as the basis of**
• **strand-asymmetric evolution in transcribed *Escherichia coli* sequences.** *Mol Biol Evol* 2001, **18**:1147-1150.

This study establishes a direct relationship between gene transcription and particular nucleotide mutations. The authors compare sequence divergence in transcribed but untranslated sequences, as well as in non-transcribed sequences, and find a pronounced C to T bias only in transcribed regions (regardless of whether they are also translated). Determining such relationships between transcription and mutation will be crucial to interpreting associations between gene expression and both silent DNA and protein evolution. Laboratory estimates and polymorphism data are, however, less likely to be affected by selection.

46. Beletskii A, Bhagwat AS: **Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*.** *Proc Natl Acad Sci USA* 1996, **93**:13919-13924.

47. Datta A, Jinks-Robertson S: **Association of increased spontaneous mutation rates with high levels of transcription in yeast.** *Science* 1995, **268**:1616-1619.

48. Morey NJ, Greene CN, Jinks-Robertson S: **Genetic analysis of transcription-associated mutation in *Saccharomyces cerevisiae*.** *Genetics* 2000, **154**:109-120.

49. Kliman RM, Hey J: **The effects of mutation and natural selection on codon bias in the genes of *Drosophila*.** *Genetics* 1994, **1**:1049-1056.

50. Iida K, Akashi H: **A test of translational selection at 'silent' sites in**
• **the human genome: base composition comparisons in alternatively spliced genes.** *Gene* 2000, **261**:93-105.

Gene sequence evolution can be related to gene expression by comparing codon-usage bias and levels of divergence among exons within alternatively spliced genes. Constitutively expressed exons show higher major codon usage and higher G+C content in *D. melanogaster* and human, respectively. However, silent divergence among mammals is greater in the more highly expressed exons, whereas protein evolution is reduced (the latter association is only marginally statistically significant). The method controls for both transcription rates and regional differences in base composition but requires careful identification of mature mRNA isoforms.

51. 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.

52. Comeron JM, Kreitman M, Aguade M: **Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*.** *Genetics* 1999, **151**:239-249.

53. Jansen R, Gerstein M: **Analysis of the yeast transcriptome with**
•• **structural and functional categories: characterizing highly expressed proteins.** *Nucleic Acids Res* 2000, **28**:1481-1488.

The authors integrate sequence data from the complete yeast genome with estimates of mRNA abundance and functional categories of proteins. Membrane proteins, as a class, are expressed less than soluble proteins. In addition, the upper limit of protein sizes decreases as a function of mRNA abundance. Interestingly, amino acid composition also varies among protein expression classes.

54. Drawid A, Jansen R, Gerstein M: **Genome-wide analysis relating expression level with protein subcellular localization.** *Trends Genet* 2000, **16**:426-430.

55. Akashi H: **Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*.** *Genetics* 1996, **144**:1297-1307.

56. Marias G, Duret L: **Synonymous codon usage, accuracy of**
• **translation, and gene length in *Caenorhabditis elegans*.** *J Mol Evol* 2001, **52**:275-280.

The authors elaborate the causes of codon-usage bias in *C. elegans*. Comparisons of 548 orthologous protein sequences between *C. elegans* and human identify functionally constrained positions, and higher codon-usage bias in such regions suggest that codon-usage bias reflects, at least in part, selection to minimize the misincorporation rate during protein synthesis. Average functional constraint also appears to be reduced in longer genes, but conserved positions show reduced codon-usage bias relative to those in shorter genes (gene expression was not controlled for in the comparison). Thus, selection for translational accuracy is supported in *C. elegans* but does not explain the lower codon-usage bias in longer genes.

57. 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.

58. McVean GAT, Charlesworth B: **The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation.** *Genetics* 2000, **155**:929-944.

59. Duret L, Mouchiroud D: **Determinants of substitution rates in**
•• **mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17**:68-74.

The authors establish associations between expression patterns and rates of mammalian protein evolution. Both the tissue-specificity and breadth of expression are related to rates of non-synonymous, but not synonymous, DNA evolution. The authors attribute these patterns to differences in the biochemical environments in which tissue-specific proteins function and to a higher probability of affecting fitness for mutations in broadly expressed proteins.

60. Hastings KEM: **Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families.** *J Mol Evol* 1996, **42**:631-640.

61. Kuma K, Iwabe N, Miyata T: **Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families.** *Mol Biol Evol* 1995, **12**:123-130.

62. Goldman N, Thorne JL, Jones DT: **Assessing the impact of secondary structure and solvent accessibility on protein evolution.** *Genetics* 1998, **149**:445-458.

63. Bustamante CG, Townsend JP, Hartl DL: **Solvent accessibility and**
•• **purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*.** *Mol Biol Evol* 2000, **17**:301-308.
Logistic regression models are used to establish predictors of amino acid polymorphism in *E. coli* and *Salmonella enterica*. Among five genes, solvent accessibility is a stronger predictor of polymorphism at a given amino acid position than the size of the encoded amino acid, its physicochemical properties or its location in the secondary structure of the protein. The probability of amino acid polymorphism, a measure of functional constraint, increases monotonically with solvent accessibility.

64. Kisters-Woike B, Vangierdegom C, Müller-Hill B: **On the conservation of protein sequences in evolution.** *Trends Biochem Sci* 2000, **25**:419-421.

65. Jones DT, Taylor WR, Thornton JM: **A mutation data matrix for transmembrane proteins.** *FEBS Lett* 1994, **339**:269-275.

66. Tourasse NJ, Li WH: **Selection constraints, amino acid composition, and the rate of protein evolution.** *Mol Biol Evol* 2000, **17**:656-664.

67. Pál C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve**
•• **slowly.** *Genetics* 2001, **158**:927-931.
The authors compare protein sequence divergence among genes related by an ancient genome duplication in *S. cerevisiae*. They use transcription rate estimates from GeneChip data and find a surprisingly strong negative correlation between mRNA abundance and protein distance. This relationship exists among gene pairs in the same functional category, suggesting that gene expression is directly related to yeast protein evolution.

68. Percudani R, Pavesi A, Ottonello S: **Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*.** *J Mol Biol* 1997, **268**:322-330.

69. Lobry JR, Gautier C: **Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes.** *Nucleic Acids Res* 1994, **22**:3174-3180.

70. Morton BR, So BG: **Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content.** *J Mol Evol* 2000, **50**:184-193.

71. Li WH: *Molecular Evolution*. Sunderland, Massachusetts: Sinauer Associates; 1997:179-182.