Purifying Selection: Action on Silent Sites

R Nielsen, *Cornell University, Ithaca, New York, USA*
H Akashi, *Pennsylvania State University, University Park, Pennsylvania, USA*

Advanced article

Article contents

The role of selection on silent-site evolution in humans is controversial. Mounting evidence suggests that selection for optimal base composition, codon usage or both may be common.

Introduction

Synonymous or ‘silent’ deoxyribonucleic acid (DNA) mutations do not affect the primary structures of encoded proteins and have been considered to be strong candidates for molecular evolution in the absence of natural selection (King and Jukes, 1969; Kimura, 1977). Over the last 20 years, a combination of biochemical studies of protein synthesis, quantification of transfer ribonucleic acid (tRNA) pools and analyses of codon usage within and among species has established that natural selection discriminates among synonymous codons to enhance the efficiency of protein synthesis in organisms ranging from bacteria to unicellular eukaryotes, plants and invertebrates (reviewed in Ikemura, 1985; Andersson and Kurland, 1990; Sharp *et al*., 1995; Akashi, 2001). Vertebrates, and in particular mammals, appeared to be a notable exception to this widespread phenomenon (Eyre-Walker, 1991; Duret and Mouchiroud, 2000; Urrutia and Hurst, 2001). We briefly review silent-site selection in model organisms and discuss recent evidence that silent mutations also affect fitness in humans.

Major Codon Preference

In *Escherichia coli*, highly expressed genes tend to use a subset of the available synonymous codons, termed ‘major’ codons, to encode each amino acid (Ikemura, 1981). Among synonymous codons recognized by different aminoacyl-tRNAs, major codons tend to be recognized by abundant isoacceptors. Among codons recognized by the same tRNA through ‘wobble’ pairing, the codon that forms a perfect Watson–Crick pairing with the tRNA anticodon is generally favored in highly expressed genes. Biochemical studies have shown that major codons are translated roughly three times more quickly and 10 times more accurately than their synonymous counterparts (reviewed in Andersson and Kurland, 1990). Thus, the ‘major codon preference’ appears to reflect adaptation of both tRNA pools and synonymous codon usage to reduce investments of time and energy in protein synthesis. Population genetic studies in bacteria (Hartl *et al*., 1994) as well as in *Drosophila melanogaster* (Akashi, 1999) support the notion that the biased usage of synonymous codons reflects weak selection in favor of translationally superior codons. (***See*** Codon Usage.)

Synonymous codon usage can also have a substantial impact on gene expression in mammals. Studies of synthetic genes designed for DNA vaccines have shown that alterations in synonymous codon usage of both foreign (Andre *et al*., 1998; Zhou *et al*., 1999; Narum *et al*., 2001) and human (Kim *et al*., 1997) genes can increase protein production more than 10-fold in mice and in human cell lines. Thus, synonymous codon usage may have similar effects on protein synthesis in mammals, invertebrates, unicellular eukaryotes and bacteria. (***See*** Synonymous and Nonsynonymous Rates.)

Several factors could explain why evidence for major codon preference remains notably weaker in mammals than in other taxa. In natural populations of *E. coli*, yeast and *D. melanogaster*, the rapidity of protein synthesis may be directly related to organismal fitness, whereas this may not be the case in mammals. Alternatively, the effective population sizes of mammals may be considerably lower than that of microbes and invertebrates (Sharp *et al*., 1995). The relative contribution of genetic drift and natural selection to the fate of a DNA mutation is critically dependent on the parameter $N_e s$, the product of the effective population size of a species (N_e) and the fitness effect of the mutation (s) (Kimura, 1983). Major codon preference may not exist in mammals because $N_e s$ for silent mutations is not sufficiently large; the evolutionary dynamics of mutations with $N_e s << 1$ are governed primarily by genetic drift. (***See*** Evolution: Selectionist View; Evolution: Neutralist View.)

Finally, it is also possible that additional complications in the analysis of mammalian genomes have made the relevant evidence more difficult to access than for other taxa. Below, we consider studies that provide evidence for selection at silent sites in mammals.

Base Composition in Coding and Noncoding Regions of DNA

One factor that greatly complicates the study of human codon bias is that codon frequencies can be expected to be unequal simply because the four nucleotides do not occur at equal frequencies in the genome. Therefore, all possible mutations among nucleotides do not occur at the same rates – for example, transitions typically occur at a higher rate than transversions. In mammals, two additional confounding factors are drastic %GC variation among regions and at multiple scales (Bernardi, 1995; Eyre-Walker and Hurst, 2001; Lander *et al*., 2001) and dependence of nucleotide frequencies on neighboring bases (Karlin and Burge, 1995). In *D. melanogaster* and *Caenorhabditis elegans*, preferred codons almost all end in G and C. Apparent codon usage bias may, therefore, be caused by mutation biases and does not in itself provide evidence of selection for optimal codon usage. In many cases, the problem boils down to excluding the possibility that an excess of G and C nucleotides in silent sites could be caused by mutational biases. (***See*** Isochores.)

Several studies have attempted to control for regional base composition to determine whether codon usage departs from mutational equilibrium. One found higher GC content at silent sites within coding regions of highly expressed histone genes than in surrounding regions (DeBry and Marzluff, 1994). Another study of a much larger number (2396) of human genes demonstrated that codons do not occur at the frequencies expected from the overall nucleotide composition in nondegenerate sites (Urrutia and Hurst, 2001).

Gene Expression and Codon Usage in Mammals

Under translational selection, the fitness benefit conferred by the presence of a major codon should be strongly related to the translation rates of genes. Relationships between codon usage biases and gene expression levels have served as key lines of evidence for major codon preferences in *E. coli* (Grantham *et al*., 1981; Grosjean and Fiers, 1982; Gouy and Gautier, 1982) and yeast (Bennetzen and Hall, 1982; Coghlan and Wolfe, 2000) as well as *Arabidopsis thaliana*, *C. elegans* and *D. melanogaster* (Duret and Mouchiroud, 1999). In the human genome, highly expressed genes may cluster in regions of high GC content (Bernardi, 1995), and so regional base composition must be taken into account in assessing relationships between codon usage and expression levels.

One study examined the correlation between breadth of expression and the level of codon bias in human genes (Urrutia and Hurst, 2001). Breadth of expression was calculated by counting the number of tissues in which a gene was expressed, and was used as a proxy for gene expression levels. The authors found a strong correlation between breadth of expression and codon usage bias, but noted that this correlation could possibly be explained by an association between expression levels and gene length, because of a bias in the estimate of codon usage. After correction for this effect, there was little or no correlation between expression breadth and codon usage bias.

To control for regional compositional effects in comparisons of codon usage and gene expression, other researchers analyzed alternatively spliced human genes (Iida and Akashi, 2000). In all genes, they divided exons into those common to all known isoforms of the protein (constitutive exons) and those found in only a subset of isoforms (alternative exons). Constitutive exons are translated more often than alternative exons, and selection for translational efficiency should therefore be stronger on constitutive than alternative exons. The researchers concluded that GC-ending codons are more frequent in constitutive exons, consistent with selection to increase translational efficiency. (***See*** RNA Processing; Splicing of pre-mRNA.)

Finally, another study has shown correlations between gene expression levels, as determined using complementary DNA (cDNA) microarray and serial analysis of gene expression (SAGE) techniques, and the GC content of coding regions in rodents (Konu and Li, 2002). No correlations were seen for the 5′ and 3′ untranslated (but transcribed) regions of these genes, suggesting that regional mutational biases do not account for the greater GC content of highly expressed genes.

Within- and Between-species Sequence Comparisons

Population genetic theory provides testable predictions about codon usage bias maintained under weak selection. Under a null model of no selection at silent sites, rates of molecular evolution should be identical at silent sites within coding regions and within pseudogenes. Analysis of data from 121 processed pseudogenes in humans, mouse and rat using codon-based likelihood models showed that the rate of substitution in silent sites is approximately 70% of the rate of substitution in paralogous pseudogenes, when comparing pseudogenes and functional genes in regions with similar %GC (Bustamante *et al*., 2002). Assuming that the only positional genomic effect on the rate of substitution is the local %GC, this result is consistent with purifying selection acting at silent sites. An alternative explanation might be that associations between DNA repair and transcription result in differences in mutation rates between pseudogenes and their functional counterparts. One study, however, found no relationship between expression levels and rates of silent DNA divergence between human and rodent (Duret and Mouchiroud, 2000). (***See*** Homologous, Orthologous and Paralogous Genes.)

Within-gene correlations between silent and protein divergence (Alvarez-Valin *et al.*, 1998) are also consistent with selection pressures at silent sites that correlate with those affecting protein structure. However, intragenic variation in mutation rates could also give rise to such patterns.

Within- and between-species patterns of variation are highly sensitive to even very weak selection (Kimura, 1983). Comparisons of the allelic distribution of putatively preferred and nonpreferred mutations have demonstrated that natural selection is responsible for the nonrandom codon usage in *Drosophila simulans* (Akashi, 1999). Such analyses do not require homogeneity of mutational processes among genes but assume that mutational processes have remained constant over the time period examined.

A similar method was used to test a null model of mutational equilibrium at silent sites in the human genome (Eyre-Walker, 1999; Smith and Eyre-Walker, 2001). Assuming selective neutrality and an equilibrium process, the frequencies of GC→AT and AT→GC mutations should be equal. The researchers used DNA sequence data from the chimpanzee and the parsimony method for inferring the ancestral nucleotide state, and could thereby distinguish between GC→AT and AT→GC mutations. Contrary to the prediction of neutrality, they found that the numbers of GC→AT mutations segregating within human populations were higher than the numbers of AT→GC mutations. The patterns observed are consistent with either biased gene conversion or natural selection favoring GC at silent sites in humans. If biased gene conversion has a strong effect favoring gametes with high %GC, this could explain the difference in the frequency of GC→AT and AT→GC mutations. The authors argued that their results could not be explained by a sampling bias, biases in the use of the parsimony methods, effects of multiple mutations or recent changes in the mutational pattern (Smith and Eyre-Walker, 2001). Interestingly, they found similar patterns in noncoding regions and at silent sites within coding regions, consistent with selection based on regional base composition rather than major codon preference.

These studies appeared to provide compelling support for selection on silent sites in humans. However, recent evidence that the GC content is undergoing reduction in the human genome (Duret *et al.*, 2002) complicates the interpretation of these patterns, as the null model for these population genetic analyses assumes a base composition at mutational equilibrium (Eyre-Walker, 1999).

Other Causes of Selection at Silent Sites: Splicing Enhancers

Although most studies of selection at silent sites have focused on selection for optimization of protein synthesis, other selective forces may prove to be important in mammalian genes. A small region (roughly 100 bp) of the *BRCA1* locus has been shown to have a markedly reduced level of silent divergence among mammals (Hurst and Pal, 2001). Putative exonic splice enhancer elements (ESEs) have been identified within this region (Orban and Olah, 2001). Synonymous and nonsynonymous mutations that affect ESEs have been proposed to underlie a number of aberrant splicing events in human disease (reviewed in Liu *et al.*, 2001). Putative splicing signals are present in most exons examined but are relatively small. The extent to which selection for proper pre-mRNA splicing constrains intron, silent, and perhaps even protein evolution in humans remains relatively unexplored. (***See*** Exonic Splicing Enhancers.)

Future Directions

Together, recent findings suggest that selection may be acting on silent sites in humans to increase translational efficiency or to maintain proper pre-mRNA processing or optimal base composition. However, most studies are confounded by variation in %GC, biases of the statistical estimators, lack of appropriate data regarding expression levels and availability of tRNA pools, and uncertainty regarding local genomic effects. Future studies using new statistical techniques, more data and detailed knowledge regarding local genomic effects in humans may provide more reliable information.

See also

Codon Usage
Synonymous and Nonsynonymous Rates

References

Akashi (1999) Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene* **238**: 39–51.

Akashi H (2001) Gene expression and molecular evolution. *Current Opinion in Genetics and Development* **11**: 660–666.

Alvarez-Valin F, Jabbari K and Bernardi G (1998) Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *Journal of Molecular Evolution* **46**: 37–44.

Andersson SG and Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews* **54**: 198–210.

Andre S, Seed B, Eberle J, *et al*. (1998) Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. *Journal of Virology* **72**: 1497–1503.

Bennetzen JL and Hall BD (1982) Codon selection in yeast. *Journal of Biological Chemistry* **257**: 3026–3031.

Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Review of Genetics* **29**: 445–476.

Bustamante CD, Nielsen R and Hartl DL (2002) A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution* **19**: 110–117.

Coghlan A and Wolfe KH (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**: 1131–1145.

DeBry RW and Marzluff WF (1994) Selection on silent sites in the rodent H3 histone gene family. *Genetics* **138**: 191–202.

Duret L and Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 4482–4487.

Duret L and Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution* **17**: 68–74.

Duret L, Semon M, Piganeau G, Mouchiroud D and Galtier N (2002) Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162: 1837–1847.

Eyre-Walker AC (1991) An analysis of codon usage in mammals: selection or mutation bias? *Journal of Molecular Evolution* **33**: 442–449.

Eyre-Walker A (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**: 675–683.

Eyre-Walker A and Hurst LD (2001) The evolution of isochores. *Nature Reviews Genetics* **2**: 549–555.

Gouy M and Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* **10**: 7055–7074.

Grantham R, Gautier C, Gouy M, Jacobzone M and Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research* **9**: r43–r74.

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

Hartl DL, Moriyama EN and Sawyer SA (1994) Selection intensity for codon bias. *Genetics* **138**: 227–234.

Hurst LD and Pal C (2001) Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends in Genetics* **17**: 62–65.

Iida K and Akashi H (2000) A test of translational selection at ‘silent’ sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**: 93–105.

Ikemura T (1981) 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. *Journal of Molecular Biology* **151**: 389–409.

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**: 13–34.

Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11**: 283–290.

Kim CH, Oh Y and Lee TH (1997) Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* **199**: 293–301.

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275–276.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press.

King JL and Jukes TH (1969) Non-Darwinian evolution. *Science* **164**: 788–798.

Konu O and Li MD (2002) Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *Journal of Molecular Evolution* **54**: 35–41.

Lander ES, Linton LM, Birren B, *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Liu HX, Cartegni L, Zhang MQ and Krainer AR (2001) A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nature Genetics* **27**: 55–58.

Narum DL, Kumar S, Rogers WO, *et al*. (2001) Codon optimization of gene fragments encoding *Plasmodium falciparum* merozoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. *Infection and Immunity* **69**: 7250–7253.

Orban TI and Olah E (2001) Purifying selection on silent sites – a constraint from splicing regulation? *Trends in Genetics* **17**: 252–253.

Sharp PM, Averof M, Lloyd AT, Matassi G and Peden JF (1995) DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **349**: 241–247.

Smith NG and Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. *Molecular Biology and Evolution* **18**: 982–986.

Urrutia AO and Hurst LD (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**: 1191–1199.

Zhou J, Liu WJ, Peng SW, Sun XY and Frazer I (1999) Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *Journal of Virology* **73**: 4972–4982.

Further Reading

Akashi H (1999) Within- and between-species DNA sequence variation and the ‘footprint’ of natural selection. *Gene* **238**: 39–51.

Andersson SG and Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews* **54**: 198–210.

Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Review of Genetics* **29**: 445–476.

Bustamante CD, Nielsen R and Hartl DL (2002) A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution* **19**: 110–117.

Iida K and Akashi H (2000) A test of translational selection at ‘silent’ sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**: 93–105.

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**: 13–34.

Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11**: 283–290.

Orban TI and Olah E (2001) Purifying selection on silent sites – a constraint from splicing regulation? *Trends in Genetics* **17**: 252–253.

Sharp PM, Averof M, Lloyd AT, Matassi G and Peden JF (1995) DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **349**: 241–247.

Smith NG and Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. *Molecular Biology and Evolution* **18**: 982–986.