

Part III. Understanding biases in synonymous codon usage

The results of the neutrality test in Part II are consistent with accumulating evidence that synonymous sites are under selection (reviewed in Chapter 7). In Part III, I explore several different mechanisms by which this might occur. In one model, selection affects synonymous codon usage to prevent premature degradation of the mRNA (Chapter 4). The other model posits that there is codon choice to ensure that introns are efficiently spliced-out of the initial transcript (Chapters 5 and 6).

In principle, the C preference I observed at four-fold synonymous sites (Chapter 3) could occur as a consequence of Comeron's (2004) proposed set of preferred codons, of which three-quarters are C-ending. Given, however, that there is still disagreement over whether selection does (Urrutia & Hurst 2003; Comeron 2004; Lavner & Kotlar 2005) or does not (Kanaya et al. 2001; Duret 2002; dos Reis, Savva & Wernisch 2004) maximise the efficiency of protein synthesis in mammals, I decided to test an alternative model.

Both theoretical (Seffens & Digby 1999; Cohen & Skiena 2003) and empirical (Duan et al. 2003; Capon et al. 2004) data have shown that synonymous codon usage can be important for mRNA stability. In Chapter 4, I test the hypothesis (Fitch 1974; Klambt 1975) that selection might act upon synonymous mutations to optimise the thermodynamic stability of mRNA secondary structure. I provide several lines of evidence that supports this idea. Most importantly, the C preference at third (usually four-fold) sites can potentially be explained by selection to favour strong G:C pairs, which increase mRNA stability. This effect may have arisen by virtue of exploiting a tendency for amino acids to use G at the first two sites within codons. Indeed, the stability of wild-type mRNAs relative to artificial transcripts is highest when there is a strong third site skew towards C, and mRNAs are less stable when Gs and Cs are interchanged. Through a novel simulation, I show that, had the synonymous mutations observed in the mouse lineage occurred elsewhere, transcripts would have been less stable. Interestingly, consistent with their proteins being under strong purifying selection, I find that the transcripts of housekeeping genes are also under the greatest pressure to maintain stability.

Is it likely that selection on mRNA stability is the only form of selection on synonymous mutations? Increasing evidence suggests that selection associated with splicing is also important. Two reports have described biases in codon usage that increase as one approaches intron-exon junctions, reflecting selection on codon choice at the pre-mRNA level. Although Willie and Majewski (2004) claimed that the findings

of Eskesen et al. (2004) were compatible with their own, this model of selection is actually two subtly different models that can predict similar effects. The ‘cryptic splice site avoidance model’ (Eskesen, Eskesen & Ruvinsky 2004) predicts that the gradients in bias are caused by the avoidance of particular codons that might be inappropriately recognised by the intron excision machinery as splice sites. By contrast, the ‘enhancer model’ hypothesises that specific codons are preferred near intron-exon junctions because they are found in exonic splicing enhancers (ESEs). Curiously, even though the Majewski group had previously described a generalised A+T enrichment at exon ends (Louie, Ott & Majewski 2003; Willie & Majewski 2004), neither they nor Eskesen et al. (2004) attempted to control for this effect. In Chapter 5, I confirm that a generalised A+T enrichment exists, then clarify and test the predictions of the models after controlling for the generalised effect. Overall, I find that there is good support for the enhancer model, but little evidence that codon usage is biased to avoid potential cryptic splice sites.

It is unclear whether the generalised nucleotide bias I observe can easily be explained by either of the models that I described in Chapter 5. The bias may even have nothing to do with selection for splicing efficiency. In Chapter 6, I return to a more direct test for a specific function, namely the well-supported splicing enhancer model. An early study demonstrated that SNP density is lower near intron-exon junctions (Majewski & Ott 2002), which seems to be explained by the presence of ESEs (Fairbrother et al. 2004; Carlini & Genut 2005). Chapter 6 shows that ESEs are under purifying selection, as I find that the synonymous sites in putative ESEs evolve more slowly than the remaining exonic sequence. I also observe that substitutions at four-fold synonymous sites become increasingly less frequent as one approaches the ends of exons, consistent with the trends seen in SNPs. Given the relative abundance of ESEs and the reduced rates of evolution, it appears that the effect of purifying selection on ESEs only leads to around a 10% underestimate the genomic mutation rate estimated from synonymous substitutions.

Chapter 7 features a review of the evidence for selection at synonymous sites and the implications of this finding. In contrast to the discussion from most chapters, which focus on the impact of these results on molecular evolution and underestimating the true point mutation rate, Chapter 7 also describes how selection at synonymous sites will improve our understanding transgenic gene expression and of disease etiology. Lastly, in Chapter 8, I briefly summarise my findings and discuss some of the future perspectives in this field.

References

- Capon, F., Allen, M. H., Ameen, M., Burden, A. D., Tillman, D., Barker, J. N. & Trembath, R. C. (2004) A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum. Mol. Genet.* **13**: 2361-2368.
- Carlini, D. B., & Genut, J. E. (2005) Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **In press**.
- Cohen, B., & Skiena, S. (2003) Natural selection and algorithmic design of mRNA. *J. Comp. Biol.* **10**: 419-432.
- Comeron, J. M. (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. *Genetics* **167**: 1293-1304.
- dos Reis, M., Savva, R. & Wernisch, L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**: 5036-5044.
- Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J. & Gejman, P. V. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **12**: 205-216.
- Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**: 640-649.
- Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. (2004) Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics* **167**: 543-550.
- Fairbrother, W. G., Holste, D., Burge, C. B. & Sharp, P. A. (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **2**: e268.
- Fitch, W. M. (1974) The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA. *J. Mol. Evol.* **3**: 279-291.
- Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**: 290-298.
- Klamt, D. (1975) A model for messenger RNA sequences maximizing secondary structure due to code degeneracy. *J. Theor. Biol.* **52**: 57-65.
- Lavner, Y., & Kotlar, D. (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**: 127-138.

- Louie, E., Ott, J. & Majewski, J. (2003) Nucleotide frequency variation across human genes. *Genome Res.* **13**: 2594-2601.
- Majewski, J., & Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**: 1827-1836.
- Seffens, W., & Digby, D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27**: 1578-1584.
- Urrutia, A. O., & Hurst, L. D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* **13**: 2260-2264.
- Willie, E., & Majewski, J. (2004) Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**: 534-538.