STUDIES OF SYNONYMOUS CODON EVOLUTION IN MAMMALS

Adam C. Eyre-Walker

Thesis presented for degree of
Doctor of Philsophy
University of Edinburgh
1992

DECLARATION

I declare that this thesis is of my own composition and that any help received in its preparation has been acknowledged.

ACKNOWLEDGEMENTS

I am very grateful to Bill Hill for allowing me to pursue my own research interests; to Peter Keightley for his interest and encouragement; to Paul Sharp, Ken Wolfe and Michael Bulmer for helpful discussion; and to Ian Hastings, Sarah Knott, Awinder Sohal and those already mentioned for reading parts of this thesis. I would also like to thank the SERC for financial support, and Ken Wolfe for providing the sequences used in chapter 5.

Finally, I express my deepest gratitude to my parents and family for love and education, and to Awinder and my friends, running and non-running, for many good times.

CONTENTS

ABSTRACT

Although tremendous progress has been made in many other groups, the forces and factors which affect synonymous codon use in mammals remain something of a mystery. At least some of the differences in codon usage between mammalian genes can be summarised in terms of composition: within any one species some genes have very low G+C contents (<30%) and others very high G+C content (>90%), with the majority lying somewhere in between. The very simplicity of this trend and the fact that this composition is correlated to that of introns and isochores suggests that the differences in synonymous codon use may be the result of variation in the pattern of mutation across the genome.

This hypothesis is examined by considering the three most likely ways in which the mutation pattern might vary across the genome: (1) temporal changes in the performance of the replicative machinary; (2) variation in the efficiency of DNA repair; and (3) variation in the frequency of gene conversion across the genome. Evidence is found against all these hypotheses. Principally none of them predict the silent substitution rate to be related to G+C content in the manner which is observed. Furthermore the lack of any discernible difference between the silent site G+C contents of early and late replicating genes, and the very small parameter range over which DNA repair can generate large differences in synonymous codon use, support the conclusions that replication and repair, respectively, are not responsible for the codon use of mammalian genes.

It is therefore suggested that selection might act upon synonymous codon use. However an analysis of codon usage within genes suggests that selection of the type commonly found in other groups, selection upon tRNA interaction, is not operative in mammals. It is tentatively suggested that selection upon mRNA secondary structure might be the responsible agent.

Some of the results obtained also have implications for the maintainence of isochores. Since the G+C contents of isochores and silent sites are correlated, the lack of any distinction with respect to composition between early and late replicating genes suggests that the differences in isochore G+C content are not caused by DNA

replication. However it is hypothesised that variation in the frequency of recombination can provide a very elegant explanation of the differences in isochore G+C contents, and the relationship between gene density and isochore G+C content.

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION

With at least 61 codons to encode 20 amino acids the genetic code is degenerate as far as protein sequences are concerned. This degeneracy has led sites at which substitutions can be synonymous to be known as 'silent'. However it is clear that synonymous codons are not neccessarily equivalent in all respects; some may bind tRNA's with greater efficiency or speed, make or disrupt mRNA and DNA secondary structure, or form motifs essential for gene expression (see Clarke 1970). Of course the potential for silent sites to encode information in no way guarantees that selection will act upon it, or that the effects of selection will manifest themselves; in particular it has been argued that mammalian population sizes are too small for selection to be effective (Sharp 1989). However there is evidence accumulating that selection does act on silent sites with observable effect.

1.1.1 tRNA concentrations

The best evidence for selection on synonymous codon usage comes from studies in *Escherichia coli* and *Saccharomyces cerevisiae* in which it has been possible to show that there is selection on silent sites to use the codons which bind the commonest tRNA's, or which have optimal codon/anti-codon binding strengths. Ikemura (1981,1982) has shown that codon use in these two species is different but biased towards the use of codons which match the commonest tRNA's. Furthermore the level of this bias is correlated to the level of gene expression; highly expressed genes show almost exclusive use of the optimal codons, whereas the weakly expressed genes show more even codon usage (Guoy and Gautier 1982, Ikemura 1985, Sharp and Li 1986a,b, Bulmer 1988).

1.2.2 Codon/anti-codon binding

On the basis of *ad hoc* arguments and *in vitro* experiments a number of authors have also suggested that the strength with which the anti-codon binds to the codon may be important, especially when a tRNA binds to two or more codons by 'wobbling' in the third position. For instance Ikemura (1981, 1982) has suggested on the basis of *in vitro* experiments conducted by Nishimura (1978) and Weissenbach and Dirheimer (1978) that anti-codons with certain types of uridine in the wobble position lead to the use of A rather than G ending codons, since the affinity of uridine for the former is greater than for the latter. This is Ikemura's rule 2, rule 1 being the influence of tRNA concentration. Rules 3 and 4 were suggested on a more *ad hoc* basis: Ikemura (1981, 1982) and Benetzen and Hall (1982) have suggested that inosine, which can bind U, C and A, might prefer the former two to avoid unorthodox purine-purine interactions (rule 3); and Grosjean and Fiers (1982) and Guoy and Gautier (1982) have suggested that codons with A or U in the first two positions should favour C in the third position to generate a codon/anti-codon bond of reasonable strength (rule 4).

Although these four rules allow one to explain most of the codon usage patterns seen in *E.coli* and *S.cerevisiae* there is some doubt as to whether all these forms of selection are acting in *E.coli* since only rules 1 and 3 show a correlation with gene expression level (Bulmer 1988). It is possible that some of the codon usage patterns are mutational in origin, not selective.

1.1.3 Is gene expression regulated by codon usage?

Despite the undeniable evidence that selection acts upon synonymous codon usage in *E.coli* and *S.cerevisiae*, the precise nature of this selection remains somewhat elusive. In particular it is not clear why genes of medium and low levels of expression have lower levels of codon bias than highly expressed genes. Is gene expression actually regulated by codon usage, or is low codon bias simply a consequence of weaker selection? In essence this is an argument of whether selection is stabilising or directional in character. The crucial question to be answered is whether initiation is rate limiting in translation, for if it is, codon usage can have no direct effect upon the level of gene expression. Bulmer (1991a) has

presented a number of arguments and observations to suggest that initiation is indeed rate limiting: (1) Ribosomes are the most costly part of the translational apparatus to produce, and should therefore not saturate the system. (2) Studies suggest that there is a substantial gap between each ribosome in a polysome which would not be expected if elongation was rate limiting, one would expect them to be adjacent. And (3) models of translation suggest that initiation is rate-limiting under reasonable parameter values (Von Heijne, Blomberg and Lijenstrom 1987). Furthermore there are several other observations which are highly consistent with a view of low codon bias as a consequence of weak directional selection: (1) There appears to be no tendency to increase the frequency of very sub-optimal codons in weakly expressed genes (Sharp and Li 1986b). (2) The rate of silent substitution increases as codon bias decreases as one might expect if selection is weaker in genes of low codon bias (Sharp and Li 1987). And (3) the effect of neighbouring bases on the pattern of mutation appears to become apparent in weakly expressed genes (Shields and Sharp 1987, Bulmer 1990). It is however worth mentioning that the last three observations can be explained under a model of stabilising selection acting on codon usage if the system is in a mutation-selection balance.

Intriguingly one is left with something of a paradox if initiation is rate-limiting: why does selection act upon elongation when it has no direct effect on gene expression, and therefore phenotype? Bulmer (1991a) has proposed a rather elegant solution to this puzzle: selection acts upon elongation rate to reduce the number of ribosomes bound to each mRNA molecule thus increasing the pool of free ribosomes and increasing the total rate of protein production. This very neatly explains why codon bias is related to gene expression. However it should be mentioned that a model of this selection predicts the codon bias to be far greater than it is in *E.coli*, for reasons which are none to clear (Bulmer 1991a).

1.1.4 Trends in codon usage in other species

Besides the very detailed work mentioned so far there is good evidence in a number of other species that selection acts upon synonymous codon usage. In *Bacillus subtilis* (Shields and Sharp 1987), *Salmonella typhimurium* (Sharp and Li 1987) and *Drosophila*

melanogaster (Shields *et al.* 1988) there are correlations between codon bias and gene expression; and in *Bombyx mori* it has been shown that the tRNA populations are adapted to the amino acid composition of fibroin and seracin, the two components of silk (Garel 1974).

1.1.5 mRNA structure

Selection might act upon mRNA structure for a variety of reasons: to increase elongation rate to optimise translation, to decrease the rate of elongation to help protein folding (Zama 1990), to increase or decrease stability, or to control gene expression. The effect of mRNA secondary structure upon elongation rate has been neatly demonstrated by Zama (1990) who showed that areas of local stability in the chicken collagen α2(I) mRNA correspond to sites at which translation seems to pause. However there is no evidence that the pauses are in fact of any consequence.

The general dangers of ascribing function to secondary structure are well illustrated by the case of the bacteriophage MS2. The secondary structure of this RNA virus is well characterised and known to be involved in gene expression in some way, which led Hasegawa *et al.* (1979) to interpret the excess of G+C in the silent sites involved in stem structures as evidence that the whole secondary structure was under selection. However Bulmer (1989) has demonstrated that this interpretation is incorrect, by showing that non-silent sites involved in the stems also show an excess of G+C. Thus the excess is more consistent with the simple idea that RNA's always fold up in such a way as to maximise the number of G:C base pairs. As further evidence it would be interesting to compare the rates of evolution in paired and unpaired sections of the secondary structure.

Mita *et al.* (1988) have suggested on the basis of two observations that the secondary structure of the silk fibroin gene of *Bombyx mori* has been subject to selection. Firstly the codon usage of different areas of the gene differs markedly, and secondly the potential secondary structure changes when alterations are made to the codon usage. The second point is really hardly worth criticising, and since silk fibroin is made up of many repeated units, the differences in codon usage can be understood in terms of concerted evolution. Whether this is the case of course remains to be

ascertained.

Rather more abstract evidence of selection upon mRNA secondary structure has come from work by Huynen *et al.* (1992) on histones. They noted that the amount of secondary structure should be optimised when the ratio of G to (G+C) is 50%. By looking at this ratio separately for silent and non-silent sites they were able to show that the silent sites appear to compensate for shortcomings in the non-silent sites. So for example human histone H4 has rather too much G in the first two codon positions but a large deficiency of G at the third position, thus bringing the overall G/(G+C) ratio for the gene to around 50%. In human histone H3 the excess and deficiency are the other way round. Futhermore they were also able to show that the minimum free energies of the mRNA's are slightly less than they are if the third positions are randomly reassorted.

Although these observations are very interesting there are some problems, both with this particular study and the general idea behind the test. With reference to this particular study the main problem is their treatment of the same histone from different species and different loci as independent. This they clearly are not. In fact one could not choose a worse set of genes for this, since histones evolve very slowly at the amino acid level (e.g see Li *et al.* 1987). More generally it is not clear that their central assertion is always going to be correct. A G/(G+C) ratio of 50% will only optimise the secondary structure when the opportunity to increase the frequency of C is the same as for G. For instance if a gene has many more two-fold codons which end in C/T than those which end in A/G it seems likely that the optimal G/(G+C) ratio will not be 50%. Furthermore some caution must be taken over amino acid composition; for instance as the frequency of cysteine (TGT and TGC) increases so the G/(G+C) ratio increases at the first two positions but decreases at the third position. These last two problems present something of a challenge to development of the G/(G+C) ratio as a general test of selection upon mRNA structure. In fact the second line of attack Huynen *et al.* (1992) employed, comparing a mRNA secondary structure to similar sequences with scrambled codon usage may be a more profitable approach. In chapter 8 I will discuss just such a method.

Therefore the evidence for selection upon mRNA secondary structure is not strong, and even if we accept the evidence of Zama (1990),

Mita *et al.* (1988) and Huynen *et al.* (1992) it still remains unclear whether the cases of selection that they observed are exceptions or rules.

1.1.6 DNA structure

If evidence for selection upon mRNA structure is weak, evidence for selection upon DNA structure is practically non-existent. Wada and Suyama (1985) have noted that there is a negative correlation between the G+C content of the first two positions and the third position when a 21 codon window is slid along a gene, an observation which they interpreted as evidence for selection upon local DNA stability; and Aota and Ikemura (1986) have noted that there is periodicity in G+C content along the vertebrate genome which corresponds in scale to the amount of DNA involved in one solenoid of six nucleosomes. Both are intriguing, but unconvincing observations. However, even if hard evidence is shortcoming there is considerable speculation about the role of selection upon DNA structure in the evolution of genomic G+C content, either between species or within a genome (see Bernardi and Bernardi 1986, Bernardi 1989).

1.1.7 The role of mutation

Although selection has dominated much of the discussion on codon usage in unicellular organisms, the role of mutation has not been ignored. Not only does it appear that codon usage in lowly expressed genes is largely determined by mutation (Shields and Sharp 1987, Bulmer 1990), but there is now some evidence that the mutation rate, and therefore possibly pattern, might vary along the *E.coli* genome (Sharp *et al.* 1989). It has also been suggested that the differences in genomic G+C content between bacterial species, which are most sharply shown at silent sites, are the result of different mutation biases (Suoeka 1962, 1988, 1992).

1.2 CODON USAGE IN MAMMALS

The synonymous codon usage of mammals is dominated by variation in G+C content unlike anything seen in other species. For instance in humans there are genes with silent site G+C contents of less than 30%

(e.g acyl coA dehydrogenase) and some with G+C contents of greater than 90% (e.g zeta haemaglobin) (Mouchiroud *et al.* 1987, 1988, Bernardi *et al.* 1988, Shields *et al.* 1988, Ikemura and Wada 1991). In sharp contrast *E.coli* genes rarely exceed the bounds of 40% and 70% (see Suoeka 1988). As one might expect the large variation in G+C content is seen in all other mammalian groups for which sufficient data are available (Bernardi *et al.* 1988), although in rodents the range of G+C contents is somewhat narrower than in other groups (Mouchiroud *et al.* 1988). This contraction of the G+C content range in rodents has been termed the 'minor shift' (Mouchiroud *et al.* 1988).

A clue to the possible cause of the synonymous codon usage in mammals is afforded by the observation that some of the variation in silent substitution rate can be explained by silent site G+C content (Filipski 1988, Wolfe *et al.* 1989, Ticher and Graur 1989, Bulmer *et al.* 1991). The various studies differ in the relationships they observed: Filipski (1988) and Ticher and Graur (1989) found that the silent substitution rate declined with increasing G+C content, whereas Wolfe *et al.* (1989) and Bulmer *et al.* (1991) found that sequences of intermediate G+C content had the highest rates of silent substitution. The differences would appear to be due to the data sets, and techniques used; in particular Ticher and Graur looked at the divergence between rats and human sequences thus complicating the issue with the 'minor shift', and Filipski's data set was very small. Furthermore, in both of these analyses correction for multiple hits was not made using a formula which could take into account the G+C content of the sequence. The analyses of Wolfe *et al.* (1989) and Bulmer *et al.* (1991) used very similar techniques which did not suffer from any obvious shortcomings. The quadratic models fitted by Bulmer *et al.* (1991) are shown in figure 1.1. They found that primate and artiodactyl genes behaved in a similar fashion, with rodents being somewhat different, possibly because they are undergoing a shift in G+C content.

Much of this thesis is concerned with the relationship between silent substitution rate and G+C content since it seems a good, and possibly the only way to test various ways in which mutation might account for the synonymous codon usage in mammals. In chapters 2, 4 and 6 models of various mutation processes are developed to

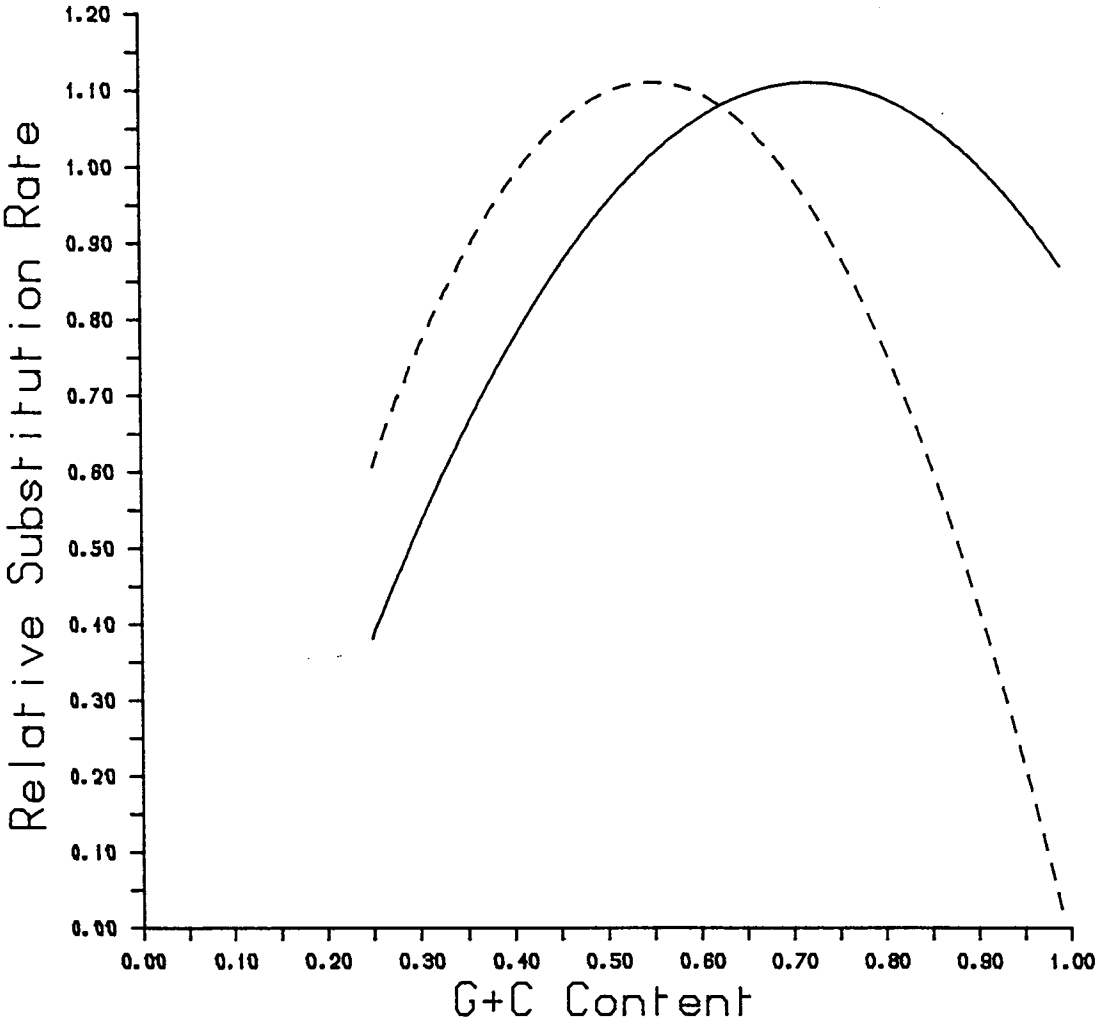


Figure 1.1 The quadratic models relating the silent substitution rate to the silent site G+C content fitted by Bulmer et al (1991). The substitution rates are relative to the mean substitution rate for the respective group (primates/artiodactyls or rodents). The solid line is for primates and artiodactyls combined, the dashed line for rodents.

investigate the possible relationships between the mutation rate and G+C content. However the assumptions made in these models mean that even the analyses of Wolfe *et al.* (1989) and Bulmer *et al.* (1991) are inappropiate. In chapter 5 the silent site substitution rate/G+C content relationship is re-analysed within the constraints of the models.

1.2.1 Mutation biases

One of the most intriguing and puzzling observations in mammals is the correlation between silent site, intron (Bulmer 1986, Shields *et al.* 1988, Aissani *et al.* 1991) and intergenic DNA G+C contents (Bernardi *et al.* 1985, Aota and Ikemura 1986, Ikemura and Aota 1988, Aissani *et al.* 1991). Superficially these correlations are highly consistent with the idea that the pattern of mutation varies across the genome, and that mammalian codon use is a simple consequence of mutation bias (Shields *et al.* 1988).

However there is a complication; the range of G+C contents seen at silent sites is far greater than that seen at introns, which is in turn slightly larger than the range of G+C contents seen in intergenic DNA. For instance in humans silent site G+C contents go from about 30% to 90%, whereas intron G+C contents range from about 35% to 65% (Bulmer 1986, Shields *et al.* 1988, Aissani *et al.* 1991, D'Onofrio *et al.* 1991), and intergenic DNA from about 39% to 53% (Bernardi 1989). Silent site G+C contents also tend to greater than intron G+C contents. One can explain these differences under a mutation bias model by noting that introns and intergenic DNA can fix the products of processes like replication slippage and transposition, which coding sequences generally cannot due to the constraints upon the amino acid sequence. Furthermore these constraints upon the coding sequence could affect the pattern of point mutation at silent sites in two other ways. (1) Approximately half the silent sites in a gene are only two-fold degenerate. Since such sites can only fix transitions generally they have the potential to exhibit different G+C contents to four-fold sites, introns and intergenic DNA. However this does not appear to be the case since four-fold and two-fold sites have remarkably similar G+C contents (chapter 5). (2) There is evidence that neighbouring bases affect the pattern of mutation at a site (Bulmer 1986, Bulmer 1990, Eyre-Walker

1991, Blake *et al.* 1992). For instance it is well known that the cytosine in CpG dinucleotides is often methylated in mammals causing it to become hypermutable (Bird 1986). Such neighbouring base effects will affect non-coding DNA in a different way to exons because they have little effect upon the composition of the non-silent sites. Thus it is, in principle at least, possible to explain the difference in the G+C contents of exons and introns/intergenic DNA without invoking selection. However the difference in intron and intergenic DNA G+C contents requires some further explanation if it really exists.

1.2.2 DNA replication

The type of mutations most likely to be fixed during the evolution of silent sites are point mutations, the formation of which involves two processes, mismatch formation and DNA repair. Let us first consider mismatch formation, which can come about in a number of ways: misincorporation during DNA replication, spontaneous chemical change, heteroduplex formation during recombination and possibly other pathways. It seems likely that DNA replication is a major source of mismatches (Topal and Fresco 1976, Friedberg 1985, Miyata *et al.* 1990) and there is clearly the potential for such a complex process to vary in accuracy across the genome by the temporal or spatial variation in the conditions under which replication occurs. Wolfe *et al.* (1989) and Wolfe (1991) have proposed a very neat hypothesis along these lines. They noted three observations: (1) that the pattern and rate of mutation is dependent upon the free nucleotide concentrations (Phear and Meuth 1989a,b, Meuth 1989, Kohalmi *et al.* 1991); (2) that the relative concentrations of the free nucleotides vary during the cell cycle (McCormick *et al.* 1983, Leeds *et al.* 1985); and (3) that different sections of the genome are replicated at different times during the cell cycle in most cell types. It should however be pointed out that the last two observations have never been made in germ-line cells. From these three observations one comes to the conclusion that sequences replicated at different times should have different G+C contents.

A theoretical analysis of this process by Wolfe (1991) suggested that the mutation rate should also vary with sequence G+C content, although not in a way reminiscent of the silent site substitution rate/G+C content relationship reported by him previously (Wolfe *et*

al. 1989). Furthermore the results of his model defied both his and my intuition. In chapter 2 Wolfe's model is extended and re-analysed.

However, quite by accident I came across some evidence that both G+C rich and G+C poor components of the genome are replicated during both halves of S phase, a result which suggests that generally DNA replication cannot be responsible for the variance in silent site G+C content found in mammalian genes. This is the subject matter of chapter 3.

1.2.3 Spontaneous chemical change

The two best characterised spontaneous chemical changes that occur in DNA are the deamination of methylated cytosine to thymine, and the conversion of cytosine to uracil (Friedberg 1985). Since the level of methylation is known to vary across the genome (Lewis and Bird 1991), and to increase the frequency of C→T transitions substantially (Coulondre *et al.* 1978, Bulmer 1986, Sved and Bird 1990, Blake *et al.* 1992) the deamination of methyl-cytosine could be responsible for the variation in silent site G+C content. However as I show in chapter 5 this does not appear to be the case since there is substantial G+C content variation at silent sites which are not preceded by C or followed by G. Ignoring such sites removes most of the effects of methylation because almost all methyl-cytosines are in CpG dinucleotides (Bird 1986).

Less is known about the conversion of cytosine to uracil in mammals. There does appear to be an efficient repair system to handle U:G mismatches (Brown and Brown-Leudi 1989) suggesting that such mismatches are quite common. However there is no *a priori* reason why the frequency should vary across the genome and therefore generate G+C content variance.

1.2.4 Recombination

Until very recently I had not really considered gene conversion as a likely determinant of synonymous codon usage in mammals. That was until I read Ikemura and Wada (1991) in which they showed (1) that A+T rich genes (at silent sites) occur much less frequently in chiasmata dense chromosome bands than G+C rich genes, and (2) that the frequency of chiasmata formation on a chromosome is proportional to the average silent site G+C content. These observations fit very

neatly in with the G+C bias in the repair of base mismatches reported by Brown and Jiricny (1988). The repair of base mismatches during recombination is considered in chapter 6 along with some thoughts on the relationship between the rate of evolution and the level of constraint.

1.2.5 DNA repair

DNA repair is an attractive hypothesis, for not only can the repair of base mismatches be biased (Brown and Jircny 1988, 1989), but the level of some types of repair appears to vary across the genome (Bohr *et al.* 1987). We might therefore expect DNA repair to generate G+C content variance. Furthermore the G:C bias in the repair of mismatches fits very well with the observation that silent sites generally have a higher G+C content than intergenic DNA. Not only might one expect this, since exons should be the most efficiently repaired areas of the genome, but there is evidence that repair is more efficient in coding sequences (Bohr *et al.* 1986, Mellon *et al.* 1986). However it should be emphasised that possible variations in the repair of base mismatches has never been studied. In chapter 4 a model of DNA repair is developed to study the relationship between the mutation rate and G+C content, and in chapter 5 the predictions of this model are tested.

1.2.6 Selection

The evidence for selection upon mammalian codon usage is really rather weak, especially now that it has become apparent that the mutation rate and pattern have the potential to vary across the genome. For instance the observations of Ikemura (1985), Newgard *et al.* (1986) and Wells *et al.* (1986) of relationships between gene function and G+C content can be explained by mutation by supposing that certain groups of genes have particular chromatin conformations, replication times or chromosomal positions. Furthermore the value of comparing pseuodgene and silent site rates of evolution is clearly limited, although such comparisons have been used (Miyata and Hayashida 1981, Wolfe *et al.* 1989).

Strangely the best evidence that selection acts upon mammalian synonymous codon use is offered by the intriguing effect non-silent substitution has on the rate of silent substitution. Fitch (1980),

and Lipman and Wilbur (1985) have both noted that codons which have undergone non-silent substitution have higher rates of silent substitution, an observation the latter authors interpreted as evidence for selection. They argued that since in species like *E.coli* different amino acids have different optimal codons, changing the amino acid will lead to certain silent mutations becoming advantageous, thus increasing the probability of fixation and the rate of silent substitution. There are two problems with this explanation. Firstly it is not clear that this knock-on effect will occur for all types of selection; it will work for selection via tRNA interactions but probably not for selection upon nucleic acid structure. And secondly it will only work if there is a substantial number of non-silent mutations which are advantageous with respect to protein function; the reason being that mutations neutral with respect to protein function will tend to be deleterious because of their effect on synonymous codon usage.

However the clustering of silent and non-silent substitutions has another more general explanation. Whatever selection acts upon silent sites, acts similarly upon non-silent sites. Hence those codons or regions of the gene which are under strong purifying selection for tRNA interactions, nucleic acid structure and controlling sequences will have low rates of both silent and non-silent substitution which will therefore lead to the apparent clustering.

Of course there is also a non-selective explanation. Michael Averoff and Paul Sharp (Trinity College Dublin, pers comm) have noted that switches between the two-fold and four-fold degenerate codons of serine occur rather more often than one would expect from the rate at which single nucleotides mutate, suggesting that the simultaneous mutation of two adjacent nucleotides is reasonably common. In keeping with this hypothesis Lipman and Wilbur (1985) found that the rate of silent substitution in codons 5' to sites which had undergone non-silent substitution was also elevated, whereas the rate in codons 3' was not. However these patterns are also consistent with selection, on mRNA for example.

Generally the investigation of selection at mammalian silent sites seems to have been hampered by a lack of suitable techniques. For instance it is not obvious how one might detect selection upon

tRNA interaction since little is known about gene expression and tRNA levels when and where genes are under the strongest selection. In chapter 7 I follow one route around these problems, and in chapter 8 I discuss ways in which one might set about detecting selection upon mRNA structure.

1.3 Other topics

Almost the whole of this thesis is concerned with molecular evolution. However in the appendix I offer a short critique of an idea about the evolutionary advantage of sex suggested by Kirkpatrick and Jenkins (1989). I will leave the introduction to this topic to the appendix.

CHAPTER 2

DNA REPLICATION AND THE MUTATION RATE

2.1 INTRODUCTION

It seems to be generally accepted that the replication of DNA is a major source of base mismatches (Topal and Fresco 1976, Friedberg 1985), although as far as I am aware, there is no direct evidence of this in mammals. Some rather neat indirect evidence comes from recent analyses of silent substitution rates in X-linked and autosomal mammalian genes. Miyata *et al.* (1990) found that the rate of silent substitution on the X chromosome was about 58% that on the autosomes, a finding highly consistent with replication as the major source of mutations since the male germ-line goes through many more divisions than the female, and X chromosomes on average spend rather less of their time in males than do autosomes. In fact if all mutations occured in males we would expect the rate of neutral evolution on the X to be 66% that on the autosomes, which is not significantly different from 58% ($p>0.10$). Of course there are other explanations but none of them predict the ratio of X to autosome to be around 66%.

As such DNA replication is a good candidate for the generation of G+C content variance, espiecally when one considers the three observations noted by Wolfe *et al.* (1989) and Wolfe (1991): (1) that the pattern of mutation is dependent upon the concentrations of the free nucleotides (Phear and Meuth 1989a,b, Meuth 1989, Kohalmi *et al.* 1991); (2) that the free nucleotide concentrations vary both relatively and absolutely during the cell cycle (McCormick *et al.* 1983, Leeds *et al.* 1985); and (3) that DNA replication is not synchronous for the whole genome in most tissues. Given these three observations, variation in the pattern of mutation across the genome appears to be almost inevitable. The catch is of course that none of these observations has been made in the germ-line.

Wolfe *et al.* (1989) supported their hypothesis with the observation that the silent substitution rate was related to G+C

content. Their argument, though never explicitly stated, seems to have been the following: if mutation patterns vary across the genome and cause G+C content variance, so we might expect the mutation rate to vary as well. A belief which seemed to be confirmed by theoretical analysis (Wolfe 1991). However there were problems with Wolfe's model which confused us both. For instance one might expect sequences of intermediate G+C content, replicated in free nucleotide pools containing the four nucleotides in equal quantitities, to have a higher mutation rate than a sequence of very high G+C content replicated in a free nucleotide pool composed entirely of G and C. The reason being that the DNA polymerase has to try more nucleotides before finding the correct one in the case where all the free nucleotides are equally common. His model did not predict this. The failure appears to come from one incorrect assumption, that the equilibrium sequence G+C content is equal to the free nucleotide G+C content. As I will show this is not the case.

The object of this chapter is to develop a slightly more detailed model of DNA replication than that given by Wolfe, and then to investigate the effect of free nucleotide pool composition and concentration on sequence G+C content and mutation rate. The object is not so much to test the model against experimental data, the model is far too simple for that, but to ascertain whether DNA replication is likely to be source of G+C content related mutation rate variance: i.e is it likely to be able to explain why the silent substitution rate is related to the G+C content?

2.2 THE MODEL

Let us consider a simple model of DNA replication, in which a free nucleotide collides with the DNA polymerase and is then either incorporated into the growing DNA sequence or rejected. The probability that the base, z (where z can be A,T,C or G), collides with the polymerase is P_z, the relative concentration of the free tri-phosphate nucleotide dZTP. Given that a collision has occurred let the probability that the nucleotide is subsequently incorporated be j if the nucleotide is correct and k_a if it is incorrect (where a is i for a type I mismatch and ii for a type II mismatch). Type I mismatches are those which if left unproofread and unrepaired give

G+C content altering mutations: i.e C↔T and A↔G transitions, and C↔A and G↔T transversions. If type II mismatches are left unaltered they give C↔G and A↔T mutations. The frequency with which y is misincorporated instead of z is then:

$$T_{zy} = k_i P_y / (jP_z + k_i P_m + k_i P_y + k_{ii} P_n) \quad \text{Type I mutation} \quad (2.1a)$$
$$T_{zy} = k_{ii} P_y / (jP_z + k_i P_m + k_i P_n + k_{ii} P_y) \quad \text{Type II mutation} \quad (2.1b)$$

where bases m and n also form mismatches. Under biologically realistic conditions where the probability of incorporating a mismatch is very small (i.e $k_a \ll j$) equations 2.1a and 2.1b reduce to the form used by Fersht (1979), Goodman (1988) and Wolfe (1991): $T_{zy} = \alpha P_y/P_z$ where α is a constant. It would of course be possible to further divide type I mutations into transitions and transversions, and to treate them separately. However transtions and type I transversions turn out to have the same dynamics in this model, and were thus treated together.

Once a mismatch is incorporated it may be removed by proofreading or mismatch repair. We will only consider elimination by proofreading in this model since the interaction between repair and replication is likely to be non-trivial. Let the probability of not proofreading a mismatch be N_a. Then the probability with which z will mutate to y per cycle of replication will be

$$U_{zy} = T_{zy} N_a \quad \text{where a is i or ii as appropiate} \quad (2.2)$$

2.2.1 Equilibrium G+C content

Now consider the change in the frequency of a nucleotide in a sequence each time the sequence is replicated. For instance the change in the frequency of G:

$$\Delta f_G = -f_G (U_{GC} + U_{GA} + U_{GT}) + f_C U_{CG} + f_A U_{AG} + f_T U_{TG} \quad (2.3)$$

where f_z is the frequency of nucleotide z in the sequence being replicated.

Let us make the simplifying assumption that the concentrations of dCTP and dGTP are the same (i.e $P_C = P_G$) and that $P_T = P_A$. Quite clearly

at equilibrium $f_C=f_G$ and $f_T=f_A$, so $P_A=(1-2P_G)/2$ and $f_A=(1-2f_G)/2$. If we let $p=P_G$ and $f=f_G$ then (2.3) simplifies to

$$\Delta f = 2k_i N_i \left[\frac{f(2p-1)}{2pB+2k_i} + \frac{p(2f-1)}{2pB-j-k_{ii}} \right] \qquad (2.4)$$

where $B=j-2k_i+k_{ii}$. Solving $\Delta f=0$ to get the equilibrium frequency of G (or C) in a sequence, gives under biologically realistic conditions (i.e $k_a<<j$)

$$\bar{f} = 2p^2 / (8p^2-4p+1) \qquad (2.5)$$

So the equilibrium frequency of G (or C) in the sequence is independent of proofreading and the probabilities of incorporation; essentially because of the symmetry in the model. As expected, when the free nucleotide pool is either all A+T, all G+C or half and half (p=0, 1/2 and 1/4) the equilibrium sequence G+C content is equal to the pool G+C content. Figure 2.1 shows the equilibrium frequency of G+C plotted against the free nucleotide concentration of G+C. The relationship between the two variables is sigmoidal, so that at intermediate G+C contents small changes in the free nucleotide concentrations have large effects on the equilibrium G+C content of the sequence (e.g a sequence of 80% G+C is replicated in a pool of only 70% G+C). This non-linearity arises because the probability of misincorporating a nucleotide is dependent on the probability of a nucleotide being incorrect per collision, and the number of collisions that occur, both of which are dependent upon the pool composition.

2.2.2 Proofreading

The probability that a mismatch will be proofread appears to depend on how long it takes to replicate the next position in the sequence, since once replication of the distal base has occured, the mismatch cannot be proofread, and must instead be corrected by other mechanisms which we shall not consider here. The main evidence for this model comes from studies in which the composition of the free nucleotide pools have been altered; not only does the nucleotide in

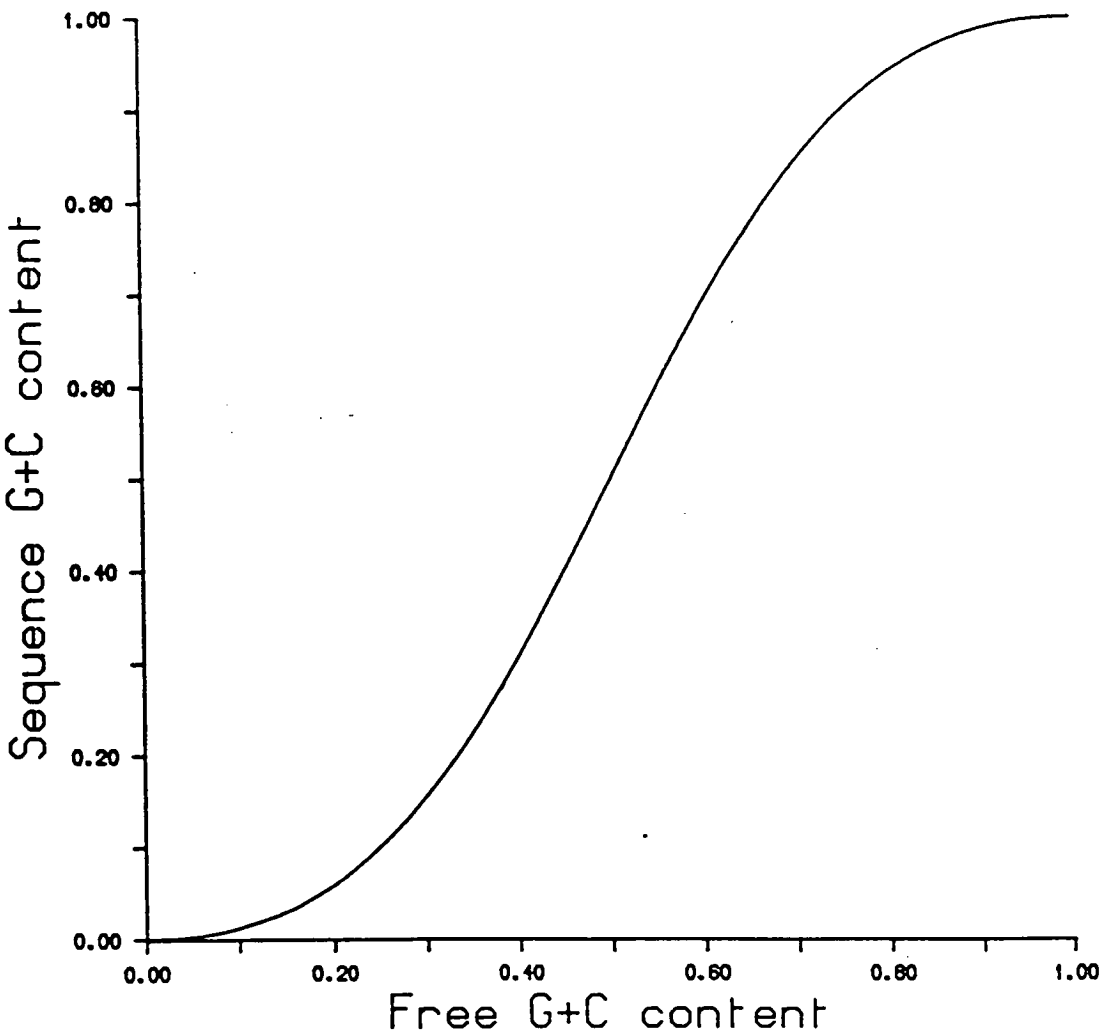


Figure 2.1 The equilibrium sequence G+C content.

excess tend be misincorporated, but such mutations tend to occur at sites immediately followed by the excess nucleotide (Phear and Meuth 1989a,b). So for instance when hamster cells are exposed to excess thymidine 73% of the C→T transitions are followed by T (Phear and Meuth 1989b).

Let the average probability of proofreading a mismatch between collisions be V_a (where a is i or ii for proofreading type I and type II mutations respectively); and let us imagine that the polymerase is waiting to replicate nucleotide z distal to a mismatch. Assuming that the probability of incorporation is not dependent upon whether the preceding site is a mismatch, and that proofreading cannot occur until the polymerase is ready to incorporate the next nucleotide, the probability that neither replication nor proofreading has occured after t collisions is $(1-V_a)^{t+1}(1-jP_z)^t$. So the probability that proofreading never occurs when a mismatch is followed by z is

$$jP_z(1-V_a)\sum_{t=0}^{\infty}((1-V_a)(1-jP_z))^t = \frac{P_z}{E_a + P_z} \qquad (2.6)$$

where $E_a=V_a/(j(1-V_a))$. Therefore the average probability of not repairing a mismatch is

$$N_a = \sum \frac{f_z P_z}{E_a + P_z} \qquad (2.7)$$

as given by Bernardi and Ninio (1978), Fersht (1979) and used by Wolfe (1991). In a sequence at equilibrium this becomes

$$N_a = \frac{p(2p-1)(8p^2-4p+1) + E_a(12p^2+6p-1)}{(8p^2-4p+1)\ (p(2p-1)+E_a(2E_a-1))} \qquad (2.8)$$

E_a is a measure of the proofreading stringency. When there is no proofreading $E_a=0$, and when proofreading is stringent $E_a\to\infty$. It is worth noting here that as proofreading becomes very stringent (i.e

$E_a \to \infty$)

$$(1+4E_a)N_a \to (24p^2-12p+2)/(8p^2-4p+1) \qquad (2.9)$$

2.2.3 Mutation rate

The average mutation rate per nucleotide per replication of a sequence is

$$U = f_G \sum_{z \neq G} U_{Gz} + f_C \sum_{z \neq C} U_{Cz} + f_A \sum_{z \neq A} U_{Az} + f_T \sum_{z \neq T} U_{Tz} \qquad (2.10)$$

which for a sequence at equilibrium simplifies under biologically realistic conditions (i.e $k_a \ll j$), to

$$U = \underbrace{\frac{8p(1-2p)(k_i N_i)}{j(8p^2-4p+1)}}_{\text{Type I mutations}} + \underbrace{\frac{k_{ii}N_{ii}}{j}}_{\text{Type II mutations}} \qquad (2.11)$$

Since we are interested in the relative, rather than absolute, mutation rate, let us divide U by the rate of mutation in an equilibrium sequence of 50% G+C content: i.e $2k_i/(1+4E_i) + k_{ii}/(1+4E_{ii})$. The relative rate of mutation is

$$R = w_i \frac{(1+4E_i)4p(1-2p)N_i}{8p^2-4p+1} + w_{ii}\,(1+4E_{ii})N_{ii} \qquad (2.12)$$

$$\text{where } w_i = \frac{2k_i/(1+4E_i)}{2k_i/(1+4E_i)+k_{ii}/(1+4E_{ii})}$$

$$\text{and } w_{ii} = \frac{k_{ii}/(1+4E_i)}{2k_i/(1+4E_i)+k_{ii}/(1+4E_{ii})}$$

w_i and w_{ii} are the proportion of mutations which are of type I and type II in a sequence of 50% G+C content.

2.3 RESULTS

2.3.1 Type II mutations

Let us consider the frequency of type II mutations alone. By setting $k_i=0$ in equation 2.12 we obtain

$$R_{ii} = N_{ii} (1+4E_{ii}) \qquad (2.13)$$

an expression which is solely dependent upon p, the concentration of dGTP or dCTP, and E_{ii} a measure of the proofreading stringency. R_{ii} is plotted against the equilibrium G+C content of a sequence in figure 2.2 (dashed line) for various levels of proofreading. Remember that as proofreading becomes stringent ($E_a \to \infty$) the expression $(1+4E_a)N_a$ becomes independent of E_a (see equation 2.9).

When there is no proofreading sequences of all G+C contents have the same rate of type II transversion mutations. However as the stringency of proofreading increases so sequences of extreme G+C content have higher mutation rates than sequences of intermediate G+C content. The reason: at extreme G+C contents the polymerase only has to try on average two nucleotides before the correct one is found, compared to the four that must be tested at intermediate G+C contents. Therefore the probability of not proofreading (when proofreading is stringent) at extreme G+C contents is twice that at intermediate G+C contents. This means the mutation rate is twice as great at extreme G+C contents.

2.3.2 Type I mutations

Now consider type I mutations alone. Setting $k_{ii}=0$ in equation 2.12 we obtain

$$R_i = \frac{N_i(1+4E_i)\ 4p(1-2p)}{(8p^2-4p+1)} \qquad (2.14)$$

an expression dependent upon p and E_i. Figure 2.2 (solid lines) shows R_i plotted against the equilibrium sequence G+C content. For all

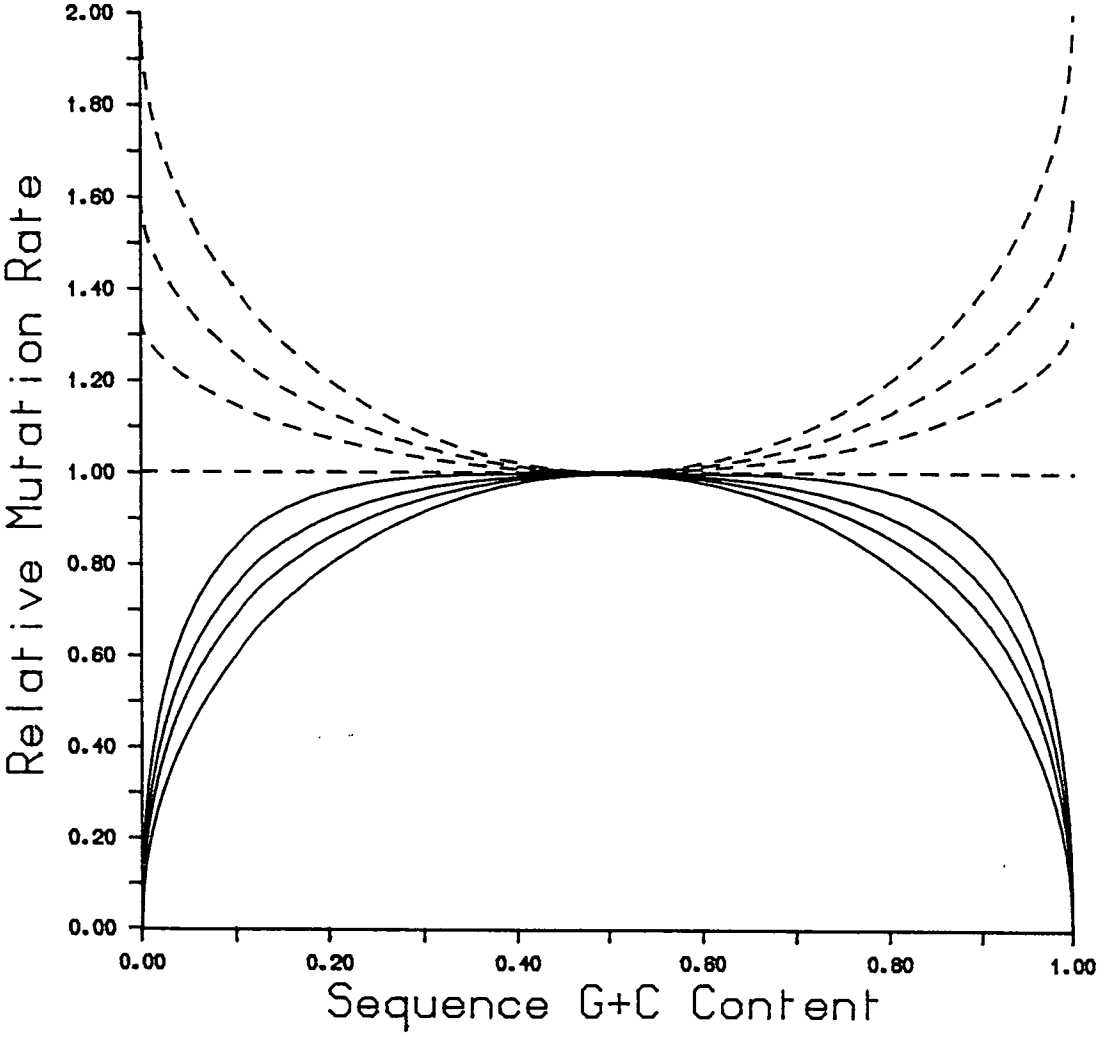


Figure 2.2 The effect of proofreading on the relative rates of type I and type II mutations. Figure shows the relative mutation rates for type I (solid lines) and type II (dashed lines) mutations for various levels of proofreading. In each case from bottom to top E_a, the strength of proofreading, is 0, 0.25, 0.75 and ∞. Under these values of E_a the probability of proofreading a mismatch in a sequence of 50% G+C content is 0, 0.5, 0.75 and $\rightarrow 1$ respectively.

levels of proofreading sequences of intermediate G+C content always have higher rates of transition (and type I transversion) mutations than sequences of extreme G+C content. The reason is that sequences of extreme G+C content are replicated in nucleotide pools which are deficient in the free nucleotides required to make transition mismatches. As the stringency of proofreading increases the curves become much flatter so that eventually the difference in mutation rates of sequences at 50% and 70% (or 30%) G+C content are negligible. Flattening occurs because proofreading elevates the mutation rate of sequences at extreme G+C content compared to sequences at intermediate G+C content (see figure 2.2 dashed lines).

2.3.3 The overall mutations rate

In graphical terms the overall mutation rate relative to that in a sequence of 50% G+C content (equation 2.12), is simply the average of the curves shown in figure 2.2. For instance if at 50% G+C content there are two unproofread type I mutations to every proofread type II mutation, then two of the bottom curves in figure 2.2 should be added to the top curve and the result divided by three. Thus the maximum variation in the mutation rate is achieved when all mutations arise via unproofread type I mismatches or stringently proofread type II mismatches.

Let us consider sequences in which both type I and II mutations can occur starting with the cases when type II mutations are not proofread. Since the rate of unproofread type II mutations is independent of the sequence G+C content, sequences of intermediate G+C content will always have higher mutation rates than sequences of extreme G+C content whether or not the type I mismatches are proofread. Qualitatively the curves will be similar to those given in figure 2.2 (solid lines) only flatter. Quantitatively the gradient at each point will be λ times the original gradient and the curve will bisect the abscissa at λ, where λ is the fraction of mutations which are type I mutations in an equilibrium sequence of 50% G+C content.

Consider now the case when type II mutations are stringently proofread and type I mutations are not proofread at all. Under these conditions equation 2.12 reduces to

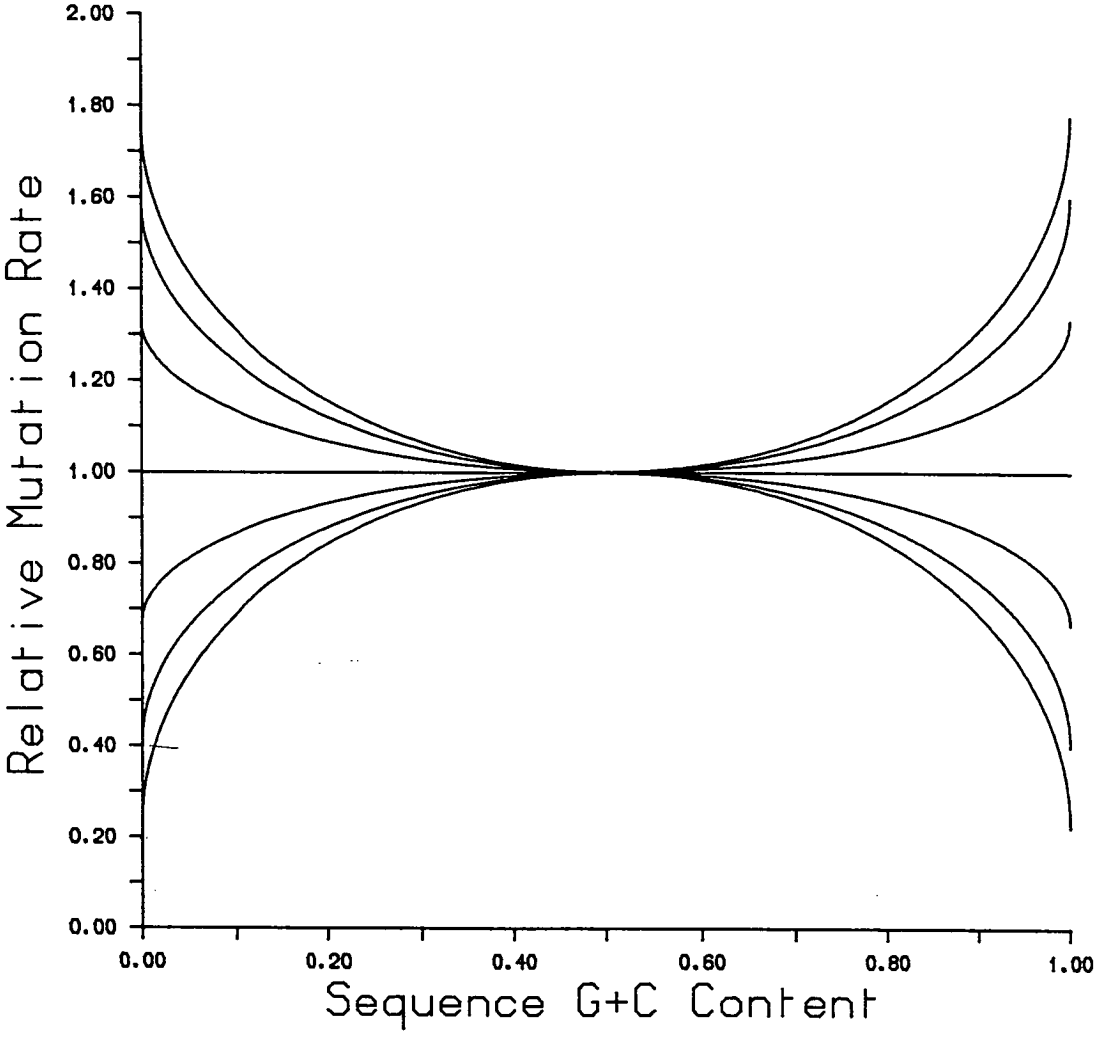


Figure 2.3 The relative mutation rate for a sequence undergoing unproofread type I mutation and stringently proofread type II mutation. The curves represent different ratios of type I and type II mutations. The ratio of unproofread type I mutations to proofread type II mutations in a sequence of 50% G+C content is, from top to bottom, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4 and 1:8.

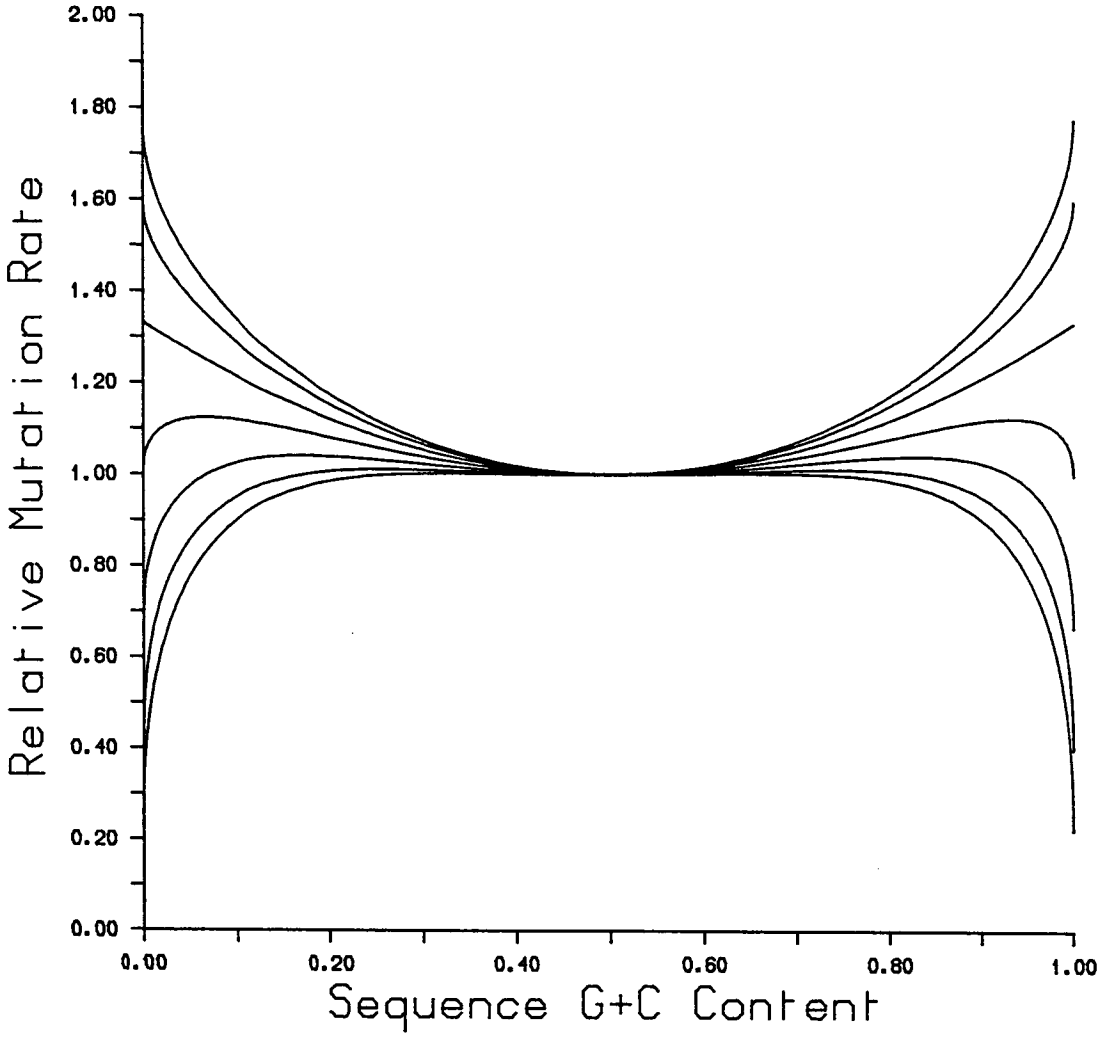


Figure 2.4 The relative mutation rate for a sequence undergoing stringently proofread type I and type II mutation. Each curve represents a different ratio of type I and type II mutations. The ratio of proofread type I mutations to proofread type II mutations in a sequence of 50% G+C content is, from top to bottom, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4 and 1:8.

$$R = \frac{8w_i p(1-2p) + 2w_{ii}(12p^2-6p+1)}{8p^2-4p+1} \qquad (2.15)$$

which can be shown to have a local maximum when more than half ($w_i>1/2$), and a local minimum when less than half ($w_{ii}<1/2$), of the mutations in a sequence of 50% G+C content are type I mutations. Equation 2.14 is plotted against the equilibrium sequence G+C content in figure 2.3. The rate of mutation is only weakly dependent upon sequence G+C content unless type I mutations are much more common than type II mutations, or vice versa. This is even more the case when all mismatches are stringently proofread (figure 2.4).

2.3.4 Changing the overall concentration of the free nucleotides

Up till now we have only considered the effect of varying the relative concentrations of the free nucleotides. However variations in the overall concentration of free nucleotides will affect the probability of proofreading, under this model, if the polymerase is not saturated by the free nucleotides.

Let the rate at which correct nucleotides collide with the polymerase be ρ and the rate at which proofreading occurs be ϵ. The time t between collisions is distributed exponentially and the probability of proofreading during a particular time interval is a Poisson process with mean ϵt. Therefore the probability of not proofreading a mismatch between collisions is

$$1-V = \int_0^\infty e^{-\epsilon t}\ \rho e^{-\rho t}\ dt = \frac{\rho}{\rho + \epsilon} \qquad (2.16)$$

and $E = \epsilon/(j\rho)$. ρ is proportional to the overall concentration of the nucleotides, so doubling the concentration leads to a doubling of ρ. When proofreading is slow (ϵ and E small) changes in the overall concentration of free nucleotides have little effect on the probability of proofreading (see equation 2.7). However when proofreading is stringent (ϵ and E large) the probability of proofreading becomes a linear function of the free nucleotide concentration: i.e $N \approx \rho j(12c^2-6c+1)/(\epsilon(16c^2-8c+2))$. Hence it is possible for fluctuations in the overall concentration of nucleotides to cause large variations in the probability of proofreading, and the rate of mutation.

2.4 DISCUSSION

The analysis above suggests that compositional changes alone are unlikely to cause much mutation rate variance, but that changes in the the total concentration can alter the mutation rate under certain conditions. Although the model is extremely simple, and the findings therefore subject to some uncertainty, it would seem likely that the antagonism between misincorporation and proofreading for type I mutations, and the antagonism between type I and type II mutations, are general phenomena. One might therefore expect the findings to be robust.

2.4.1 Changes in composition

It is not clear whether Wolfe *et al.* (1989) and Wolfe (1991) proposed their theory as an explanation of the G+C content variance at silent sites, or the smaller G+C content variance one finds between different blocks of intergenic DNA, the so called 'isochores' (Bernardi 1989). Which ever, it seems unlikely that changes in the nucleotide pool composition during DNA replication can explain the covariance between the rate of silent substitution and the G+C content since the model shows very little G+C content related mutation rate variance; far too little to explain the two-fold variation that is accounted for by G+C content in the data set of Bulmer *et al.* (1991)(see figure 1.1) unless type I mutations completely dominate the system. It is possibly worth noting here that in such a situation the mutation rate/G+C content relationship does at least take a shape similar to that of the silent substitution rate/G+C content relationship reported by Wolfe *et al.* (1989) and Bulmer *et al.* (1991). Unfortunately little is known about the mismatch formation pattern in mammals so one cannot elaborate on this possibility. Analysis of substitution patterns in pseudogenes suggests that transitions are about twice as common as transversions (Gojobori *et al.* 1982, Li *et al.* 1984), but since the probability and pattern of repair appear to differ between mismatches (Brown and Jiricny 1988) mutation patterns do not correspond to patterns of mismatch formation.

2.4.2 Changes in concentration

However even if compositional changes are unable to produce G+C content related mutation rate variance, there may be correlated changes in the overall concentration of the free nucleotides which can. One might expect composition and concentration to be correlated for one simple mechanistic reason: an increase in one free nucleotide, say C, will lead to an increase in the relative concentration of C and an increase in the total nucleotide concentration unless compensatory changes are made in the other free nucleotides. Since there is at least some proofreading in mammalian replication (Kunkel *et al.* 1987, Meuth 1989) and the polymerases appear to be unsaturated by free nucleotides (Matthews and Slabaugh 1986) such changes in the concentration should lead to variation in the mutation rate. However it should be appreciated that although correlated changes in the overall concentration of the free nucleotides can generate G+C content related mutation rate variation, such correlations will have to be quite complex to explain the quadratic relationship observed between the substitution rate and the G+C content at silent sites (see figure 1.1). Changes in the composition and concentration of the free nucleotide pools during replication would therefore appear not to be responsible for the differences in synonymous codon usage between mammalian genes.

CHAPTER 3

REPLICATION TIME AND G+C CONTENT

3.1 INTRODUCTION

For the most part this chapter deals with a paradox and its implications which are not an integral part of this thesis. The paradox revolves around the replication time of what are known as isochores: very large blocks of DNA (>200kb) which have very homogeneous G+C content (Bernardi 1989). In warm blooded vertebrates the genome appears to be a mosaic of isochores with different G+C contents (~39% to ~53% in humans) which are thought to replicate at different times. It is the connection between replication time and G+C content which is the major subject matter of this chapter, although the results have implications both for how isochores are maintained, and whether DNA replication is responsible for the G+C content variance seen at mammalian silent sites.

The paradox is as follows. It is generally believed that G+C rich isochores and housekeeping genes replicate early in the cell cycle, with G+C poor isochores and some tissue specific genes replicating late (Comings 1978, Goldman 1988, Bernardi 1989, Holmquist 1989). Since the G+C content of a gene is correlated to the isochore in which it lies (Bernardi *et al.* 1985, Aissani *et al.* 1991) housekeeping genes should be G+C rich compared to tissue specific sequences. However there appears to be no difference in the G+C contents of housekeeping and tissue specific genes (Mouchiroud *et al.* 1987).

The weak link in this paradox upon which I wish to focus, is the early replication of G+C rich isochores, since the evidence for it can be interpreted in two ways. Evidence for the early replication of G+C rich DNA comes from the 3-5% difference in G+C content that has been measured between the early and late replicating fractions of the genome (Tobia *et al.* 1970, Flamm *et al.* 1971, Comings 1971, Hutchison and Gartler 1973, Holmquist *et al.* 1982) and the coincidence of

chromosome bands produced by G+C content sensitive methods, such as quinacrine staining, and replication time bands (Comings 1978, Bernardi 1989). The simplest and most popular interpretation of these observations is that most, if not all of the isochores replicated early in the cell cycle have a higher G+C content than those replicated late in S phase (Holmquist 1989, Bernardi 1989, Wolfe *et al.* 1989, Wolfe 1991). However the observations are equally consistent with the replication of all fractions of the genome both early and late in the cell cycle, with the early replicating DNA only being on average slightly more G+C rich. The difference is very important since it has implications for our understanding of chromosome structure and evolution; in particular how isochores are maintained. It also seems inappropiate to talk about early replicating DNA being more G+C rich if in fact most of the variation in G+C content is within the early and late replicating fractions, not between them.

In order to distinguish between these alternatives a set of genes for which there is replication time data and sequence information was compiled. If we assume that genes replicate at the same time as their isochore, as they appear to do (Hatton *et al.* 1988) then gene G+C contents can give us a handle on the G+C contents of isochores replicated during the two halves of S phase, and thus allow us to test whether both G+C rich and G+C poor fractions of the genome replicate during the two halves of S phase.

3.2 MATERIALS AND METHODS

3.2.1 Replication time data

Data on the replication time of specific genes was taken from Holmquist (1989) with minor modifications (see below). His list is a compilation of data from Goldman *et al.* (1984), Calza *et al.* (1984), Iqbal *et al.* (1984, 1987), Hatton *et al.* (1988) and Goldman (1988). These studies suggest that all genes expressed in a tissue are replicated early, but that unexpressed genes may replicate at any time during S phase. If a gene does replicate late in one cell type it does so in most other cell lines in which it is not expressed. Therefore genes which were found to replicate late in most cell

types, in which they were not expressed, were classified as late replicating. All other genes were classed as early replicating. The classification was the same as that given by Holmquist (1989) except for albumin, which appears to have been misclassified, and complement C4 which was not included in the compilation. It is worth pointing out that the replication time of most genes was coincident over several cell types from several species. In particular there were no genes with different replication times between the two species groups (rodents and primates).

3.2.2 Sequence Information

Sequence information was taken from the Genbank (Release 68) and Embl (Release 27) databases using the GCG sequence analysis package (Devereux *et al.* 1984). Human and mouse sequences were extracted for all genes for which replication time data were available. If a mouse gene was not available the rat sequence was used instead. The G+C contents of mouse and rat genes are very similar (Mouchiroud *et al.* 1988, Bernardi *et al.* 1988) so mixing rat and mouse genes should not lead to substantial bias. If human sequences were not available other primate sequences were used. The classification as to housekeeping or tissue specific expression was taken from Holmquist (1989) who gave no details as to how the classification was performed.

Our primary interest in the G+C contents of early and late replicating genes is not the compositions of the sequences themselves, but what they tell us about the isochore in which they reside; i.e we are interested in whether G+C rich and G+C poor isochores replicate both early and late. It is therefore important to ensure that a particular isochore is only represented once in the data set, by including only one gene from a set of linked, or recently diverged, genes. Since isochores are thought to be at least 300kb in length (Bernardi 1989) genes within this distance of one another were regarded as representing the same isochore. The average G+C content over a set of linked genes was not used because it is possible for a linkage group to traverse two or more isochores of different compositions. Instead the longest sequence was used. Small scale physical distance information (<300kb) was taken from Iqbal *et al.* (1984) and Hatton *et al.* (1988). Large scale linkage was also checked using HGM10.5 (McAlpine *et al.* 1990, Davisson *et al.* 1990)

and a mouse map (Hillyard *et al.* 1991). No genes were excluded on the basis of this information because of the scale and the lack of accuracy involved. However suffice it to say that only six genes were found to be 'linked' (within a centimorgan or in the same chromosome band). Beta-globin and c-Ha-ras map to the same human chromosome band, 11p15.5, but are some 18 centimorgans apart in mice; immunoglobulin kappa constant and variable map to the same chromosome band, 2p12, in humans and the same centimorgan in mice, 6.32; and arginine succinate synthetase and c-abl are the same distance along chromosome 2 in mice.

If several sequences from a dispersed multigene family were available with replication time information (e.g beta and gamma actin), only one sequence was used since any recently diverged members will tend to correlate with the G+C content of the 'parental' isochore, not that of their present location. Such sequences will therefore tend to contribute information about the same parental isochore.

One further problem with multigene families is identifying which member the replication time is actually known for, espiecally if some members of the family are quite disimilar to each other. For instance the rodent and primate placental lactogen genes have very different G+C contents which suggests that they are paralogous, and such paralogous genes could have different replication times. However this source of error is only relevent if the paralogous genes have different G+C contents, of which there is little evidence in the data set (table 3.1) and most of the sequences used are probably single copy genes. Therefore any errors should be small.

3.3.3 Testing the data

Differences in the distribution of G+C contents, say of early and late replicating genes, were tested with a Mann-Whitney test. This tests whether two sets of data could have come from the same distribution, and fails if the medians are different, or if the medians are the same but the shapes of the distributions are asymmetrical and different.

Such tests ask whether two sets of data could have come from the same distribution, whereas we want to ultimately ask a slightly more subtle question: are the gene G+C contents we observe consistent with

all the early replicating isochores being more G+C rich than the late replicating isochores? In order to do this we need to take into account the less than perfect correlation between gene and isochore G+C contents; i.e it is possible for all early replicating isochores to be more G+C rich than late replicating isochores and yet for there still to be some overlap in the G+C contents of early and late replicating genes. The approach taken was as follows: isochore G+C contents were randomly generated from gene G+C contents in a way consistent with the available data for each gene. The isochore G+C contents so produced were then compared to see if any overlap existed in the range of early and late replicating fractions.

Of all the relationships between gene and isochore G+C content that have been published the best, in terms of sample size and correlation coefficient, is that given by Aissani *et al.* (1991) for human third position G+C contents. Aissani *et al.* chose to leave out two genes from their regression analysis because one of the genes had very biased amino acid composition, and the other had a very low G+C content. Since the second of these reasons appears to be arbitary and the first is not relevent to the third position G+C content both genes were included in this study. The relationship between isochore and gene G+C content obtained by least squares linear regression is:

(1) Isochore G+C = 31.3 + 0.229 x Third Position G+C

Since the error terms (residuals) appear to be normally distributed, and unrelated in magnitude or sign to the third position G+C contents, the predicted isochore G+C content for a gene of G+C content X_0 is t-distributed with N-2 degrees of freedom, a mean of Y_0, the isochore G+C content given by the regression line (1), and a standard deviation of

$$(2) \qquad S \left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2} \right]^{\frac{1}{2}}$$

where S is the standard deviation of the residuals and N the sample size of the data used in the regression. Thus by sampling at random from a t-distribution with the appropiate parameters it is possible

to convert gene G+C contents into isochore G+C contents in a way consistent with the data of Aissani *et al.*. The isochore G+C contents so produced can then be examined to see if early and late isochores overlap in G+C content. By repeating this procedure many times it is possible to assess how much overlap there must be between the G+C contents of early and late replicating isochores. For instance if we found that only 0.5% of a very large number of randomly produced isochore sets showed no overlap between early and late replicating fractions, then we would be able to reject the null hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at the 0.5% level. This test was only applied to human genes because there are far fewer rodent genes for which isochore location is known. The test would therefore be much less powerful.

TABLE 3.1

Gene	Rep Time[a]	Exp[b]	Time known[c]	Primate 3[d]	Time known[c]	Rodent 3[d]
HPRT	E	H	✓	39.6	✓	41.5
APRT	E	H	✓	81.6	✓	74.3
CAD	E	H	✓	71.5	✓	NA
DHFR	E	H	✓	42.5	✓	47.8
Argininesuccinate synthetase	E	H	✓	74.7	✓	67.9
Glucose-6-phosphate dehydrogenase	E	H	✓	85.1		62.5
β-tubulins	E	H	✓	81.8	✓	71.8
Phosphogylcerate kinase 1	E	H	✓	55.8		54.3
Tyrosine aminotransferase	E	H	✓	NA	✓	62.7
β-actin	E	H	✓	84.5		73.0
Metallothionein I	E	H	✓	80.0	✓	88.3
c-myc	E	H	✓	76.7	✓	75.8
c-Ha-ras	E	H	✓	81.4	✓	NA
c-ki-ras	E	H	✓	32.6	✓	43.2
c-fos	E	H		71.5	✓	68.1
c-raf	E	H	✓	37.3	✓	58.0
Histone H2A.1	E	H	✓	67.4	✓	94.6
α-globin	E	T	✓	88.8	✓	67.9
c-sis	E	T	✓	77.9		NA
c-myb	E	T		45.5	✓	55.6
c-fes/fps	E	T		80.3	✓	NA
c-rel	E	T		26.7	✓	NA
c-mos	E	T	✓	74.2	✓	67.3
c-fms	E	T		74.8	✓	67.5
Apolipoprotein AI	E	T	✓	85.0		71.3
Thy-1	E	T		79.4	✓	76.4
Placental lactogen	E	T	✓	74.5		43.9
Complement C4	E	T		70.8	✓	65.1

TABLE 3.1 cont/d

Immunoglobulin						
Kappa constant	E	T		NA	✓	59.4
Albumin	E	T		39.0	✓	57.0
N-ras	E	?	✓	45.7		55.7
c-abl	E	?		71.5	✓	68.4
Skeletal						
muscle actin	L	T	✓	89.1		77.1
β-globin	L	T	✓	66.4	✓	66.9
α1-antitrypsin	L	T	✓	68.8	✓	64.2
β-casein	L	T		53.0	✓	49.6
Phenylalanine						
hydroxylase	L	T	✓	52.5		56.0
Factor IX	L	T	✓	35.4		31.8
Fibronectin	L	T	✓	50.2		53.4
Myosin heavy						
α-cardiac	L	T		85.8	✓	80.7
N-myc	L	T	✓	79.8	✓	75.3
α-amylase 1	L	T		34.1	✓	40.0
Major urinary						
proteins	L	T		NA	✓	41.3
Immunoglobulin						
kappa variable	L	T		56.6	✓	45.2

The expression status, replication time and G+C content of a set of primate and rodent genes.

a The replication time of the gene: E-early and L-late.
b The expression status of the gene: H-housekeeping, T-tissue specific and ?-unknown.
c Replication time known in primate/rodent cell line.
d Third postion G+C content.

3.3 RESULTS

The G+C content, replication time and expression status of the 44 genes in the data set are given in table 3.1, and represented graphically in figures 3.1. Only third position G+C contents are given since the correlation between third position and isochore G+C contents is much better than for other positions (Bernardi *et al.* 1985, Aissani *et al.* 1991). Since there is no evidence that replication times differ between rodents and primates (table 3.1), and no evidence of differences in the G+C contents of genes whose replication time is known and those whose replication time is inferred from another group (table 3.2a), it was assumed in all subsequent analyses that replication times are identical in primates and rodents. The results are not qualitatively affected by this assumption.

Confirming the result of Mouchiroud *et al.* (1987) figures 3.1 and 3.2 show that there is no difference in the distribution of G+C contents of housekeeping and tissue-specific genes. Mann–Whitney tests confirm this (table 3.2b). More importantly there is also little difference in the distributions of early and late replicating genes. The early replicating genes appear to be slightly more G+C rich than the late replicating genes but this difference is not significant (table 3.2b).

To illustrate how inconsistent these results are with the replication of only G+C rich isochores early, and G+C poor isochores late in S phase, isochore G+C content is plotted against third position G+C content for a set of 21 human genes (Aissani *et al.* 1991) in figure 3.2. There is no horizontal line which would split the data so they look like the patterns in figures 3.1 and 3.2. For instance let us imagine that all isochores above 43% replicate early in S phase with the rest replicating late: there is little overlap between the G+C contents of early and late replicating genes.

It is possible to make this argument more quantitative by converting gene G+C contents to isochore G+C contents in a way consistent with the data of Aissani *et al.* (1991,figure 3.2), as detailed in the materials and methods. In 10000 simulated sets of isochores generated from the human early and late sets of genes, there was not a single case when the most G+C rich late replicating

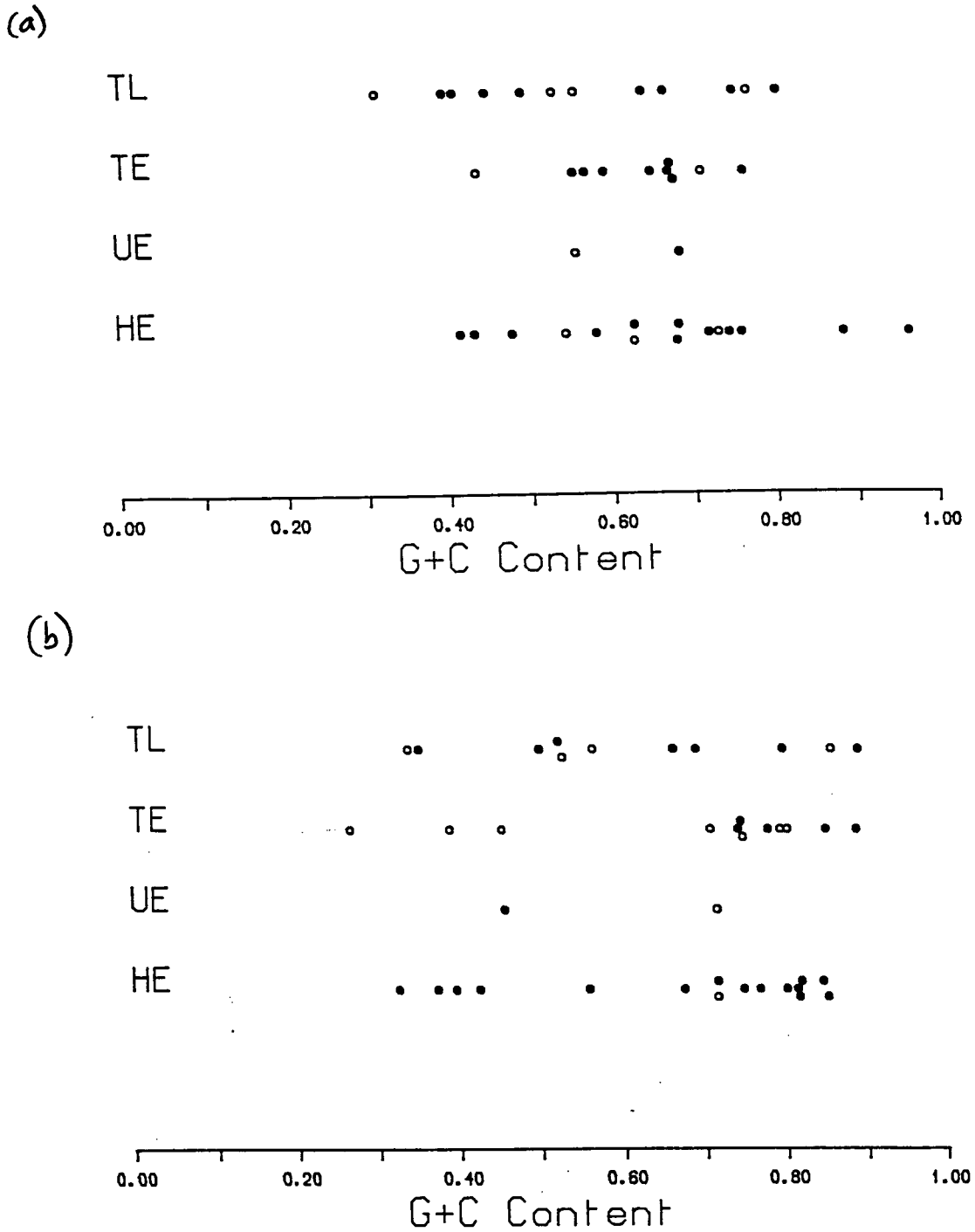


Figure 3.1 The third postion G+C contents of human (a) and rodent (b) housekeeping, tissue-specific, early and late replicating genes. **HE** - early replicating housekeeping genes. **UE** - early replicating genes of unknown expression. **TE** - Early replicating tissue specific genes. **TL** - Late replicating tissue specific genes. Filled circles are those genes whose replication time is known in a human (a) and rodent (b) cell line.

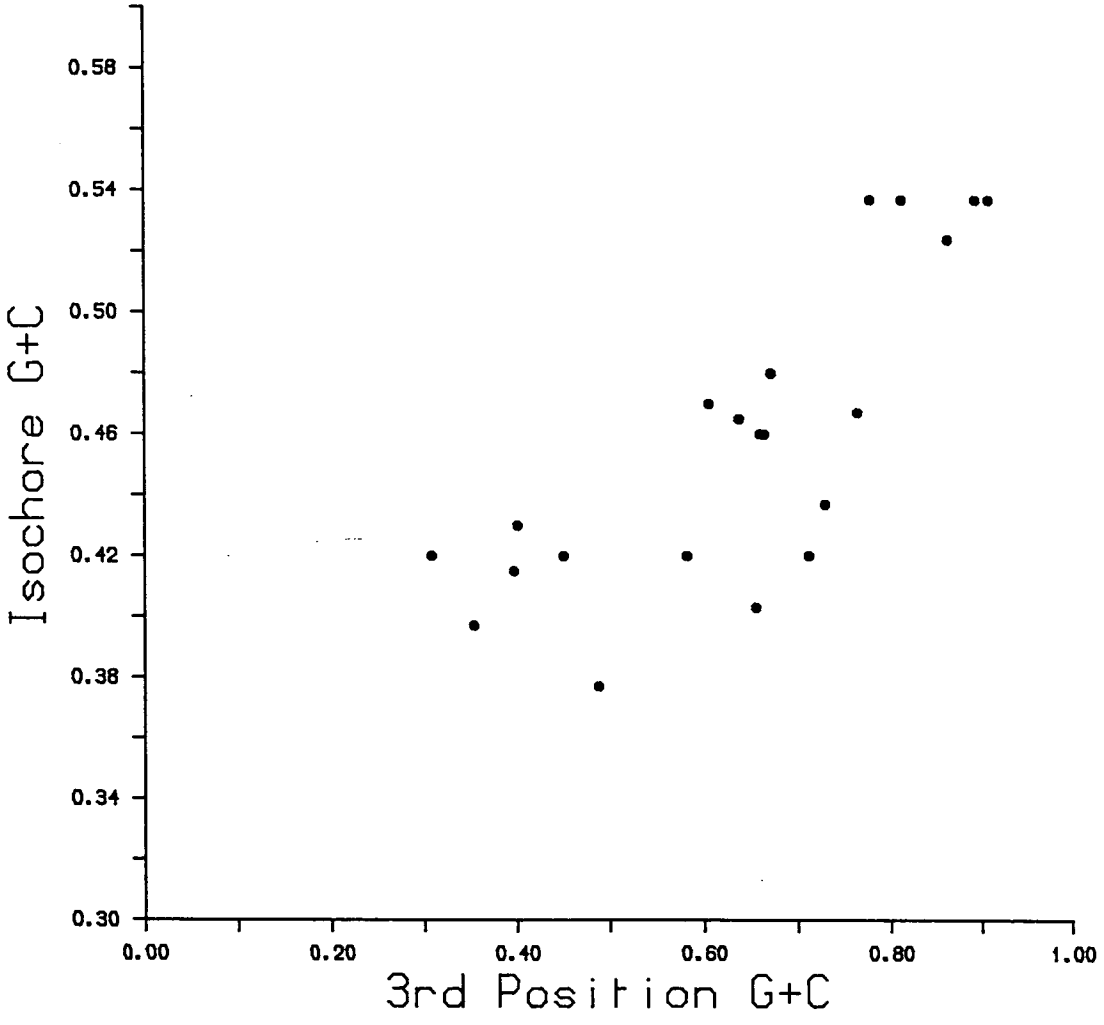


Figure 3.2 Isochore G+C content against third position G+C content for 21 human genes. Data from Aissani *et al.* (1991).

TABLE 3.2

Data set	Primates	Rodents
(a)		
H	——	0.72
TE	0.14	0.90
TL	0.78	0.93
E	0.21	0.37
T	0.25	0.53
(b)		
H v TE	0.91	0.62
TE v TL	0.41	0.28
H v T	0.72	0.28
E v L	0.41	0.17

Testing for differences in G+C content. Figures in the body of the table show the probability of the two data sets being more dissimilar than they are by chance alone in a Mann-Whitney test. In part (a) the G+C contents of genes whose replication time is known in a group (e.g primates) are compared against those whose replication is inferred from another group (e.g rodents). In part (b) genes with different characteristics are compared. H - housekeeping, T - tissue specific, E - early and L - late replicating genes. The test cannot be performed for primate housekeeping genes due to insufficient sample size.

isochore was less G+C rich than the least G+C rich early replicating isochore; i.e there was always some overlap between the early and late replicating isochores. We can therefore reject the hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at 0.05% significance or lower. Furthermore in 9999 cases the upper quartile (the value above which 25% of the observations lie) of the late replicating genes was greater than the lower quartile (the value below which 25% of the observations lie) of the early replicating genes. In other words the overlap was always

substantial. When the test was repeated on just early and late replicating tissue specific genes there was always an overlap of G+C contents, and in all but 22 cases the early replicating lower quartile was less than the late replicating upper quartile.

3.4 DISCUSSION

These results demonstrate that there is considerable heterogeneity in the G+C content of isochores replicated early and late in S phase. It is unclear however whether the replication of G+C rich and poor DNA is temporally separated, but over a much shorter time scale than the length of S phase, or whether sequences of different G+C content are simultaneously replicated. Several groups have looked at the G+C content of DNA being replicated at hourly intervals during S phase. Comings (1971) found in a hamster cell line that the average G+C content of replicating DNA changed continuously from relatively A+T rich, to G+C rich before decreasing again to G+C poor. Thus there was an overlap in the G+C content of sequences replicated early and late in S phase. In contrast Tobia *et al.* (1970) and Flamm *et al.* (1971) found that in mouse cell lines the G+C content of replicating DNA decreased monotonically during S phase. However in all analyses the range of average G+C contents replicated at different times (~5%) was insufficient to cover the range of isochore G+C contents (~9-15%). It therefore seems likely that sequences of very different G+C content are replicated simultaneously during S phase.

3.4.1 Isochore Replication Time

Further evidence that G+C rich and poor isochores replicate in both halves of S phase is provided by a few genes for which isochore location and replication time are known (table 3.3). In mice there appears to be no relationship between replication time and isochore class, and in humans early replicating genes are located in all isochore classes except the very G+C poorest.

3.4.2 The Maintainence Of Isochores

The fact that sequences of very different G+C contents may be replicated simultaneously has some implications for the maintainence of isochores. The mechanism by which isochores are maintained is the

TABLE 3.3

Gene	Isochore G+C	Replication time
Human		
Factor IX	39.7	L
β-globin	40.3	L
HPRT	41.5	E
c-mos	43.7	E
c-myc	46.7	E
Glucose-6-dehydrogenase	52.4	E
c-Ha-ras	53.7	E
α-globin	53.7	E
c-sis	53.7	E
Mouse		
IgK Variable	40.5	L
IgK Constant	42.0	E
β-globin	42.0	L
α-globin	49.1	E
Skeletal actin	49.1	L
c-abl	49.1	E

Gene replication times and isochore locations. Replication time data comes from references cited in the methods section. Isochore location data comes from Aissani *et al.* (1991) for humans, and from Bernardi *et al.* (1985) for mice.

subject of considerable debate (Wolfe *et al.* 1989, Bernardi *et al.* 1988, Filipski 1987). The simplest and most cogent hypothesis has been put forward by Wolfe and colleagues (Wolfe *et al.* 1989, Wolfe 1991). They proposed that different replicons are replicated in free

nucleotide pools of different compositions which biases the pattern of mutation, thus producing replicons/isochores of different G+C contents. This very neatly explains the relationship between replication time and G+C content that was originally thought to exist, since it has been shown that the free nucleotide pool composition changes through the cell cycle (McCormick *et al.* 1983, Leeds *et al.* 1985). The fact that isochores of different G+C contents appear to replicate simultaneously poses something of a problem for this hypothesis, unless the free nucleotide pools are spatially heterogeneous. Paradoxically one is loath to drop the Wolfe/Li/Sharp hypothesis because it provides a very elegant explanation of the correlation between gene and isochore G+C contents, one of the observations which led to the original paradox. The correlation arises under this hypothesis, because although selection and DNA repair may vary across a replicon, all sequences in a replicon have the same pattern of misincorporation which is different to other replicons replicated under different conditions. Therefore sequences within a replicon are expected to have correlated compositions.

However it should be appreciated that the conclusions reached via table 3.1 are only strictly applicable to the cell lines in which the gene replication times were studied. The conclusions do not neccessarily extend to the germ-line, which is the relevent tissue when discussing the origins and maintainence of isochores. It is possible that the pattern of replication is quite different in germ and somatic cell lines. However it is clear from the present work that in certain cell lines both G+C rich and G+C poor isochores replicate early and late in the cell cycle.

3.4.3 Implications For Silent Site G+C Content Variance

The easiest and possibly the only way in which the pattern of misincorporation might come to vary during DNA replication is through some temporal change in the conditions under which DNA replication is carried out. Conditions might vary spatially but this theory lacks the simplicity of the temporal change, and is completely unsupported by any evidence. Hence the lack of any major difference between the G+C contents of early and late replicating genes and isochores suggests that DNA replication cannot generally be responsible for the silent site G+C content variance. It might be responsible for some

fraction of the G+C content variance since the G+C content of replicated DNA does appear to change through the cell cycle (Tobia *et al.* 1970, Flamm *et al.* 1971, Comings 1971), but the changes in G+C observed are very small.

CHAPTER 4

DNA REPAIR - THEORY

4.1 INTRODUCTION

If the repair of base mismatches affects the pattern of mutation, and the probability of repair varies across the genome then we would expect different genes to have different codon usage patterns. Both these criteria appear to be met in mammals. Evidence for variation in the efficiency of repair comes from a number of studies of pyrimidine dimer and bulky adduct removal (Bohr *et al.* 1987). For instance pyrimidine dimers in the DHFR gene are much more efficiently repaired than those in the flanking DNA (Bohr *et al.* 1986, Mellon *et al.* 1986). The variation in repair is thought to be caused by differences in chromatin structure altering the accessibility of the DNA to the repair enzymes; an idea supported by the work cited above and the observation that pyrimidine dimers are more efficiently repaired when the histones are removed from DNA than when they are not (Wilkins and Hart 1974).

Evidence for the effect of repair upon the pattern of mutation comes from the work of Brown and Jiricny (1988) who showed that different mismatches introduced into SV40 were repaired with different efficiencies and biases. For instance they found that the mismatch G:T was repaired in 92% of the cases to G:C and in 4% to A:T, the other 4% being left unrepaired. In sharp contrast the other transition mismatch, A:C, gave values of 41%, 37% and 22%. Since repair can only leave the pattern of mutation unaffected if it repairs as many mismatches to G:C as to A:T these biases in repair will alter the pattern of mutation.

Like all other mutation hypotheses DNA repair struggles to explain why silent site G+C contents are greater in mean and range than intron G+C contents. However variation in repair can explain why silent site and intron G+C contents are different to isochore G+C contents; coding and non-coding DNA differ in their accessibility to the repair enzymes. In fact the idea that repair efficiency is linked

to chromatin structure, the work on the DHFR gene mentioned above and the G+C bias in DNA repair provide a very reasonable explanation of why silent sites and introns have higher G+C contents than isochores.

The aim of this chapter is to develop a model of DNA repair, and to use it to investigate the G+C content/mutation rate relationship with a view to developing a test of the DNA repair hypothesis. Unfortunately several assumptions have to be made in formulating the model which means that the results cannot be directly compared to the silent substitution rate/G+C content relationships previously reported (Wolfe *et al.* 1989, Bulmer *et al.* 1991). In the next chapter I reanalyse the data of Bulmer *et al.* within the constraints of the model.

It will be assumed throughout the following sections that DNA repair is G+C biased. Not only is this supported by the experimental work of Brown and Jiricny (1988) but it seems sensible; if repair was A+T biased we would be forced to conclude that most protein coding sequences were less efficiently repaired than the DNA flanking them since intergenic DNA is usually less G+C rich than silent site DNA (Bernardi *et al.* 1985, Aissani *et al.* 1991).

4.2 THE MODEL

Let us initially consider only the repair of heteromispairs (i.e C:T, C:A, G:T and G:A) since mutations via homomispairs (i.e C:C, A:A, T:T or G:G mismatches) do not change the G+C content of a sequence, have quite complex dynamics, and turn out to be best left out of the data analysis since it is better to estimate the number of substitutions without them.

An A:T base pair can be converted to a G:C (or C:G) base pair by any one of four different mismatches: A:T can mutate to C:G via A:G or C:T mismatches, and A:T can mutate to G:C via A:C or G:T. Let us label the mismatches 1 to 4. Assuming that the pattern of mismatch formation is the same on the two strands of the DNA duplex, an assumption supported by the work of Bulmer (1991b) let M_{bm} be the frequency with which base pair b becomes mismatch m. In all that follows the subscript b can either be C, standing for a C:G (or G:C)

base pair, or A for an A:T (or T:A). Once a mismatch is formed it may or may not be repaired. Given that the mismatch was formed from base pair b, let the probability of repairing mismatch m be α_{bm}, and the frequency with which the repair is to G:C (or C:G) be β_{bm}. When the repair machinery is confronted by a mismatch it may or may not have some way of detecting where it came from. For instance after DNA replication in *E.coli* the lack of methylation on the newly synthesised strand can be used to direct repair in favour of the information held in the older strand (Friedberg 1985, Radman and Wagner 1986). Under such conditions the probability of repairing a mismatch, and the direction in which the repair is made, depend upon where the mismatch came from (C:G or A:T): i.e $\alpha_{Am} \neq \alpha_{Cm}$ and $\beta_{Am} \neq \beta_{Cm}$. This is known as error-proof repair. However the repair machinery may have no epigenetic clues as to which base pair gave rise to the mismatch. This is the case in *E.coli* once the newly synthesised strand is methylated. Under such circumstances a guess must be made and $\alpha_{Am}=\alpha_{Cm}$ and $\beta_{Am}=\beta_{Cm}$; this is termed error-prone repair.

Consider the mutation of a G:C base pair to an A:T base pair via a particular mismatch, say G:T, which we will label '1'. There are two paths by which a G:C base pair can become an A:T. The mismatch can be formed and left unrepaired. The probability that this occurs is $M_{C1}(1-\alpha_{C1})$ and the frequency of G:C base pairs is reduced by $M_{C1}(1-\alpha_{C1})/2$. If the mismatch is repaired to A:T the frequency of G:C is reduced by $M_{C1}\alpha_{C1}(1-\beta_{C1})$. The change in the frequency of G:C base pairs is

$$\Delta f = -fM_{C1}\,((1-\alpha_{C1})/2 + \alpha_{C1}(1-\beta_{C1})) + (1-f)M_{A1}((1-\alpha_{A1})/2 + \alpha_1\beta_{A1})) \qquad (4.1)$$

where f is the frequency of G:C (or C:G) base pairs in a sequence. Hence the change in the frequency of G:C base pairs by all four mismatches is

$$\Delta f = -f \sum_{m=1,2,3,4} M_{Cm} \left\{ \frac{1-\alpha_{Cm}}{2} + \alpha_{Cm}(1-\beta_{Cm}) \right\} + (1-f) \sum_{m=1,2,3,4} M_{Am} \left\{ \frac{1-\alpha_{Am}}{2} + \alpha_{Am}\beta_{Am} \right\} \quad (4.2)$$

This can be simplified to

$$\Delta f = -fM_C \left\{ \frac{1-\alpha_C}{2} + \alpha_C(1-\beta_C) \right\} + (1-f)M_A \left\{ \frac{1-\alpha_A}{2} + \alpha_A\beta_A \right\} \quad (3)$$

where

$$M_b = \sum M_{bm}$$

$$\alpha_b = \frac{\sum M_{bm}\alpha_{bm}}{M_b}$$

$$\beta_b = \frac{\sum M_{bm}\alpha_{bm}\beta_{bm}}{M_b\alpha_b}$$

M_b is the probability of base pair b becoming a mismatch. α_b and β_b are the weighted averages of repairing, and repairing to G:C mismatches that come from base pair b.

Solving $\Delta f = 0$ to obtain the equilibrium frequency of G:C base pairs in a sequence, f, gives

$$\bar{f} = \frac{M_A\ (1+\alpha_A(2\beta_A-1))}{M_A\ (1+\alpha_A(2\beta_A-1)) + M_C\ (1-\alpha_C(2\beta_C-1))} \quad (4.4)$$

4.2.1 Mutation rate via heteromispairs

Using similar reasoning to that above the mutation rate via heteromispairs, U, in a sequence of G+C content f can be written as:

$$U = fM_C \left\{ \frac{1-\alpha_C}{2} + \alpha_C(1-\beta_C) \right\} + (1-f)M_A \left\{ \frac{1-\alpha_A}{2} + \alpha_A\beta_A \right\} \quad (4.5)$$

which in a sequence at equilibrium, substituting equation 4.4 into 4.5, simplifies to

$$\bar{U} = 2\, M_C\, (1-\alpha_C(2\beta_C-1))\, \bar{f} \qquad (4.6)$$

Since we are interested in relative, rather than absolute mutation rates, the mutation rate in a sequence undergoing repair is divided by that in an equilibrium sequence subject to no repair, i.e $\bar{U}_0 = 2M_CM_A/(M_A+M_C)$. Without loss of generality we can now define $M_C+M_A=1$ and the ratio of $\bar{U}$ to $\bar{U}_0$ is

$$R = \frac{(1-\alpha_C(2\beta_C-1))\bar{f}}{M_A} \qquad (4.7)$$

4.3 ANALYSIS

The expressions so far derived for the equilibrium G+C content and mutation rate (equations 4.4 and 4.7) are a general description of the effect of DNA repair. However it is not possible to progress further in the analysis of this model without making some simplifications. If we assume that the probability of repair is the same for all mismatches ($\alpha=\alpha_{cm}=\alpha_{Am}$) equations 4.4 and 4.7 can be rearranged and combined to give:

$$R = \frac{\bar{f}(1-\bar{f})(1+\Lambda)}{(1-\bar{f})M_A\Lambda + M_C\bar{f}} \quad \text{where } \Lambda = \frac{(2\beta_A-1)}{(2\beta_C-1)} \qquad (4.8)$$

an expression which relates the relative rate of mutation, R, and to equilibrium sequence G+C content, $\bar{f}$, as the probability of repair changes. Assuming that all mismatches will be repaired with equal efficiency is unrealistic. However this assumption turns out to be unimportant (see simulation section).

The parameter Λ summarises the direction and strength of repair; the numerator and denominator give the tendency for A:T and C:G (respectively) generated mismatches to be repaired to C:G (or G:C). When $2\beta_b-1$ is negative the repair is to A:T, when it is positive the repair is to C:G, and when it is zero there is effectively no repair. Thus when $\Lambda<0$ the repair is 'error-proof' in character: i.e X:Y

generated mismatches tend to be repaired back to X:Y. Whereas if $\Lambda>0$ repair is 'error-prone' since all mismatches, irrespective of their origins tend to be repaired to G:C.

When repair is G+C biased (i.e $df/d\alpha>0$) it can be shown that $\Lambda>-1$. Furthermore Λ must be less than 1.5 if DNA repair is to generate a wide range of G+C contents, say 50% to 90%; the reason being that $\Lambda>1$ implies that some G:C generated mismatches end up as A:T base pairs, which in turn implies that there will always be some A:T base pairs in a sequence.

4.3.1 Results

The relative mutation rate is plotted against the equilibrium G+C content under various mismatch, M_C, and repair, Λ, patterns in figure 4.1 according to equation 4.8. Generally the mutation rate decreases as the G+C content increases, unless Λ and M_C are quite large, when sequences of intermediate G+C content can have higher mutation rates than sequences of lower or higher G+C content. However it is important to note that these maxima are always at a G+C content of less than 50%, and are only ever large at very low G+C contents. Thus the mutation rate declines with increasing G+C content over most of the G+C content range. Furthermore the slope is quite steep over most of the G+C content range unless the repair of A:T generated mismatches is very biased back to A:T ($\Lambda \rightarrow -1$). However if the repair of A:T generated mismatches is generally biased to A:T the conditions under which G+C content variance can be generated become extremely restrictive. This is because the repair of C:G derived mismatches must be even more biased in the direction of C:G to compensate for the lack of A:T base pairs becoming C:G (see next section).

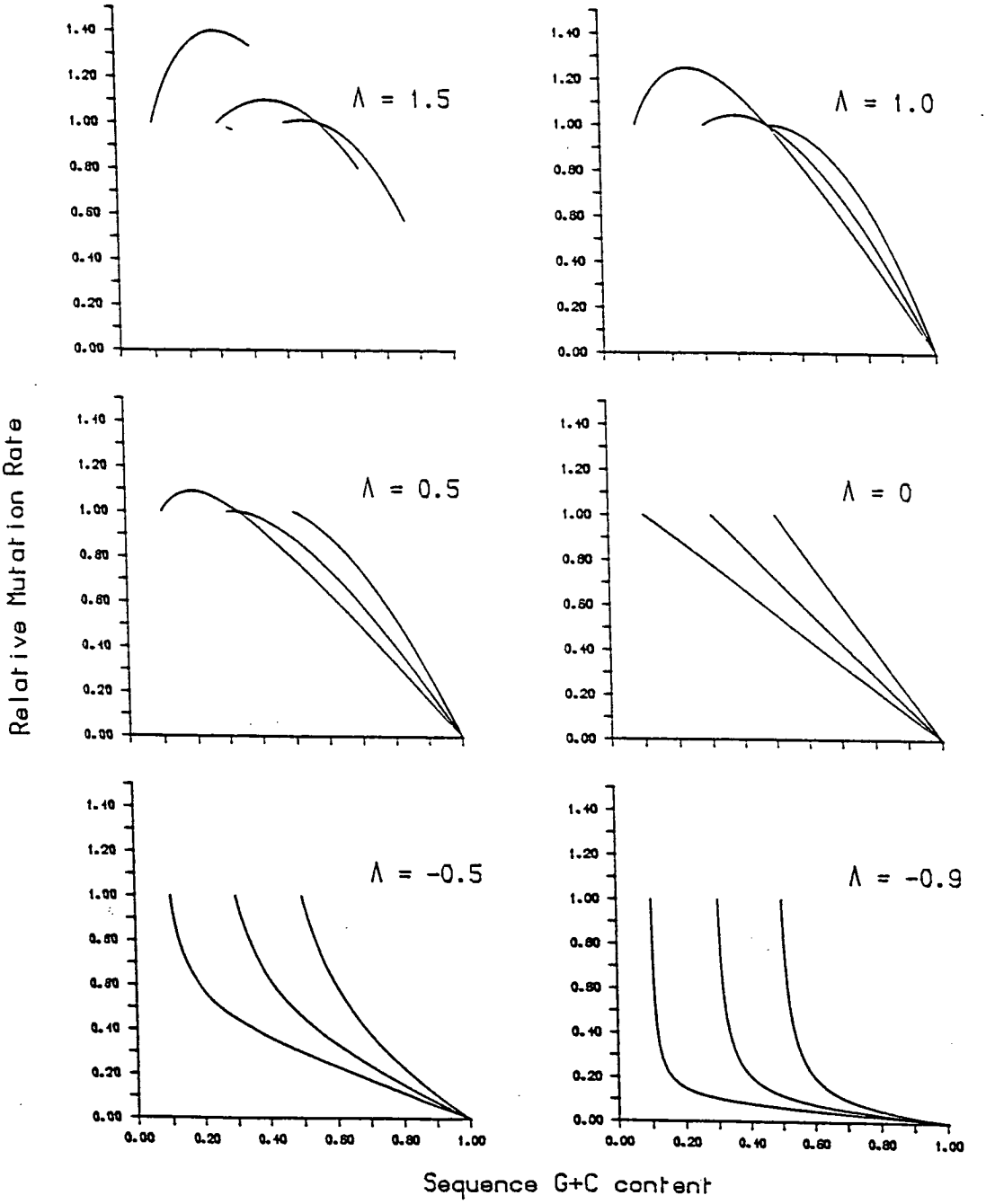


Figure 4.1 The relative mutation rate/G+C content relationship as predicted from the simplified model (equation 4.8) for various mismatch formation (M_C) and repair (Λ) patterns. In each graph the curves from left to right are for M_C = 0.9, 0.7 and 0.5 corresponding to sequences under no repair of 10%, 30% and 50% G+C.

4.4 SIMULATION

To assess how general the results from the simplified model were, a simulation was carried out. To model how the probability of repair varies with chromatin openness let us assume that the repair machinery repairs $r_{bm}t$ mismatches at random between each round of DNA replication, where t measures the chromatin openness and r_{bm} the rate at which mismatches m, formed from base pair b are repaired. For instance t might represent the time for which the DNA is accessible to the repair enzymes. The probability with which any one mismatch is repaired was taken to be a poisson process, and is thus

$$\alpha_{bm}=1-\exp(-r_{bm}t) \qquad (4.9)$$

This model is quite general, only assuming that the parameter t, the effect of chromatin openness, is the same for all mismatches. Therefore it does not matter what order the kinetics of the repair reaction is, or whether more than one repair system is operating. So long as different repair systems respond to the opening of chromatin in a similar fashion, the model will be applicable.

The simulation itself involved randomly sampling the parameters M_{bm}, $\alpha_{bm}(max)$ and β_{bm} from a uniform distribution, assessing whether they could generate a sufficiently large range of G+C contents and then calculating the relationship between the mutation rate and G+C content. $\alpha_{bm}(max)$ is the value of α_{bm} when the chromatin is fully open, from which r_{bm} can be calculated by rearrangement of equation 9 using the maximum value of t, t_{max}. The range of G+C contents produced by a particular set of parameter values was assessed by calculating the sequence G+C content under no repair (t=0), and under maximal repair ($t=t_{max}$) using equation 4. The relationship between the equilibrium G+C content and the mutation rate was calculated using equations 4.4 and 4.7.

Three distinct repair patterns were investigated. In the first M_{bm}, $\alpha_{bm}(max)$ and β_{bm} were independently sampled, whereas in the second error-proof repair was modelled by sampling β_{Am} from between 0 and 0.5, and β_{Cm} from between 0.5 and 1; i.e so the repair of an X:Y generated mismatch was always biased towards X:Y. In the third simulation error-prone repair was modelled by setting

$\alpha_{Cm}(max)=\alpha_{Am}(max)$ and $\beta_{Cm}=\beta_{Am}$. These three schemes are referred to as 'anything', 'error-proof' and 'error-prone' in table 4.1.

The number of mismatches in the model was also varied by setting $M_{Cm}=M_{Am}=0$ as appropiate; i.e if there was only one mismatch in the model only M_{C1} and M_{A1} were non-zero. The number of mismatches was varied for two reasons. Firstly it seems sensible to analyse two-fold (2 mismatches) and four-fold (4 mismatches) degenerate silent sites separately. And secondly it is a rapid way in which to investigate systems in which the dynamics of less than four mismatches dominate the system. For instance if the mismatch C:A forms a hundred times more frequently than any other mismatch, the dynamics would be those of a one mismatch system. Note however that there is no need to do simulations on one mismatch systems since they are gorverned by equation 4.8.

4.4.1 Results

In many cases it proved impossible to find parameter values which would give the desired range of G+C contents (table 4.1), illustrating a point not previously made; that the production of large G+C content variances is far from inevitable under a model of DNA repair. Even under the most favourable conditions, when one mismatch is more frequently formed than the others and repair is error-prone, only 0.2% of the parameter space allows a G+C content range of 0.5 to 0.9. This is because the generation of large G+C content variances depends ultimately upon the repair of all G:C generated mismatches being very efficient and biased to G+C: i.e from equation 4, $f\rightarrow 1 \Rightarrow 1-\alpha_C(2\beta_C-1)\rightarrow 0 \Rightarrow \alpha_C\rightarrow 1$ and $\beta_C\rightarrow 1 \Rightarrow \alpha_{Cm}\rightarrow 1$ and $\beta_{Cm}\rightarrow 1$ unless $M_{Cm}\approx 0$. As expected it becomes more difficult to find parameter values the wider the desired G+C content range gets, and the more mismatches there are. Note also that it is easiest to find parameter values when repair is error-prone because under such a system you a more likely to get both A:T and G:C generated mismatches being repaired to G:C.

However in all cases for which a set of parameters could be found, the relationship between the relative mutation rate and G+C content was consistent with the relationships depicted in figure 4.1. The consistency was such that even the error-prone and error-proof simulations gave relationships which looked like $\Lambda>0$ and $\Lambda<0$ graphs

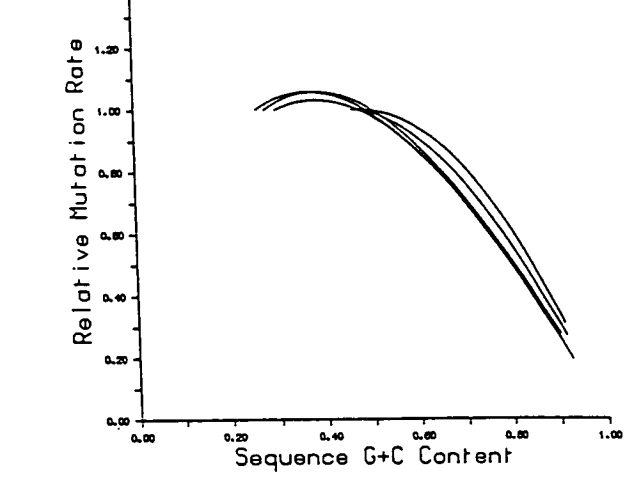
respectively (figure 4.2). As expected the graphs in each group tended to look like $\Lambda \to 1$ and $\Lambda \to 0$ figures since it is easiest to produce large G+C content ranges when A:T→C:G mutations are common. It therefore seems that equation 4.8 is quite general and that the mutation rate should decline with increasing G+C content with a fairly steep gradient, at least for all sequences with a G+C content greater than 50%.

The simple model (equation 4.8) would seem to give general results for two reasons. Firstly to produce sequences of high G+C content requires that the repair of all common G:C generated mismatches is efficient and very biased to G:C (i.e $\alpha_{Cm}(max) \to 1$ and $\beta_{Cm} \to 1$ unless $M_{Cm} \approx 0$), which means that the mismatches which dominate the system must have very similar values of r_{Cm} and β_{Cm}, and thus similar dynamics. And secondly there is some redundancy in the parameters α_{bm} and β_{bm} which allows one to model different efficiencies of repair. The redundancy arises because a mismatch repaired with high efficiency and low bias is very similar in dynamics to a poorly repaired mismatch with high bias. In the extreme, repairing a mismatch with no bias is exactly equivalent to not repairing a mismatch at all.

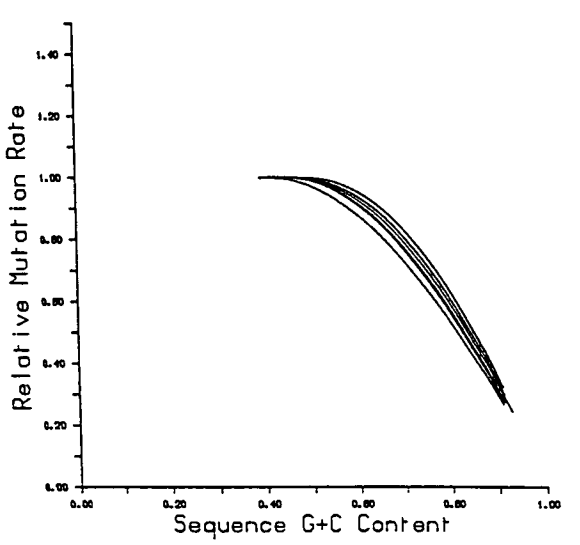
TABLE 4.1

No of mismatches	G+C Content Range 0.5→0.9	0.3→0.9	0.1→0.9
Anything			
1	636	40	0
2	33	0	0
3	0	0	0
Error-Prone			
1	1859	124	3
2	119	0	0
3	4	0	0
Error-Proof			
1	626	46	0
2	23	0	0
3	0	0	0

Figures give the number of times in 1,000,000 randomly generated parameter sets that the given G+C content range was attained. Three simulations were carried out, 'anything', 'error-prone' and 'error-proof' (see text for details). Note that no parameter sets with four mismatches were found which could generate the G+C content ranges.



(b)



(c)

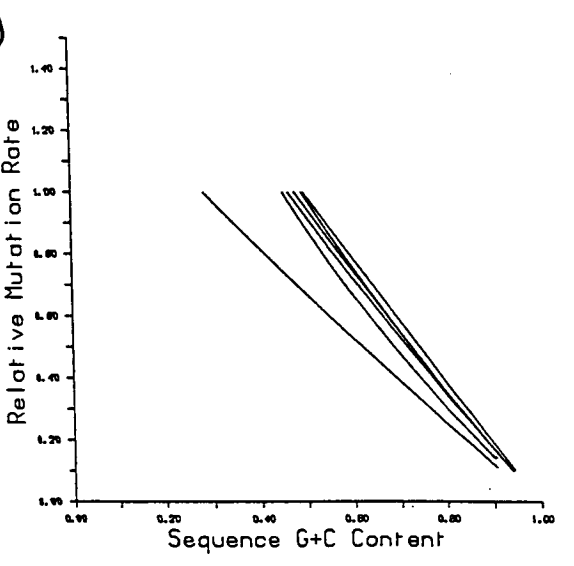


Figure 4.2 Examples of the relative mutation rate/G+C content relationship for randomly produced parameter sets for (a) error-prone repair with two mismatches, (b) error-proof repair with three mismatches, and (c) error-prone repair with 2 mismatches. Six examples are given in each case. Equations 4.3 and 4.7 were used to generate these curves.

4.5 DISCUSSION

Two points come out of analysing the model. The first is the difficulty DNA repair has in producing large G+C content ranges. This is well illustrated by the results of the simulation study which shows that only a small fraction of the total parameter space can produce large G+C content ranges (table 4.1). Of course such random production of parameter values fails to take account of evolution; we might expect the commonest mismatches to be repaired with the greatest efficiency and bias, which would make DNA repair more likely to produce G+C content variance.

The second, and more important observation is that the mutation rate should decline with increasing G+C content for all sequences of G+C content greater than 50%. In fact we would generally expect the mutation rate to decline over most of the G+C content range with quite a steep gradient. This prediction appears to be quite robust, being independent of the mutation pattern and the type of repair. Therefore if variation in the efficiency of DNA repair is solely responsible for the differences in codon usage between mammalian genes we would expect the rate of silent substitution to decline with increasing G+C content. This will be true even if other mutation processes vary or selection acts. So long as the other mutation processes and selection do not vary with G+C content and the rate of evolution is proportional to the mutation rate the silent substitution rate should decline with G+C content under the repair hypothesis.

4.5.1 Assumptions

Two crucial assumptions were made in formulating the model which need to be either justified or tested.

(1) It was assumed that every site in the sequence was independent; i.e that the adjacent bases do not influence either the mismatch formation pattern at a site or the probability of repair. Although it is well documented that neighbouring bases do affect the mutation pattern (Bulmer 1986, Hanai and Wada 1988, Eyre-Walker 1991, Blake *et al.* 1992), these effects should only contribute noise to the system. The saving grace in this respect is the relationship between silent and non-silent site G+C contents. Not only is the G+C content

range of non-silent sites much narrower (~40% to ~60%) than that of silent sites, but the correlation between the two is quite poor (r^2 = 0.28 from Shields *et al.* 1988, see also Bernardi *et al.* 1988). If the correlation had been good and the range of non-silent site G+C contents similar to silent sites, then the probability of mismatch formation and repair would have become G+C content dependent greatly complicating the dynamics of the system. The non-silent sites effectively act as a buffer; so long as the neighbouring base effects do not extend beyond two bases, the silent sites will be effectively independent. However neighbouring base effects will generate noise. Imagine that the probability of repair is different if a site is surrounded by two thymines but the same for all other sites. The sites surrounded by thymines will be governed by a different curve in figure 4.1 to other sites, so at any given G+C content different sites will have different mutation rates which will appear as random variation around a central trend.

(2) The model assumes that the sequences are at equilibrium. There are two possible reasons why this condition may not be met. Firstly the sequences may not have had time to reach an equilibrium; and secondly non-silent substitution between two-fold and four-fold degenerate sites may lead to deviations from equilibrium if the G+C contents of two-fold and four-fold sites, with their different patterns of mutation, have different equilibrium G+C contents. For instance two-fold sites may have a lower equilibrium G+C content compared to 4-fold sites, say 50% to 90%. If non-silent substitutions lead to the conversion of two-fold sites to four-fold sites, and vice versa, the G+C content of both types of silent site may approach 70%. Both of these possibilities can be tested.

4.5.2 Summary

It has been demonstrated that the mutation rate will decline with increasing G+C content if DNA repair is responsible for the silent site G+C content variance, for all sequences of greater than 50% G+C. In fact the mutation rate is generally expected to decline with increasing G+C content over most of the G+C content range with a fairly steep gradient. A number of assumptions were made in reaching this prediction, all of which are delt with in the analysis of data which follows in the next chapter.

CHAPTER 5

DNA REPAIR - ANALYSIS OF DATA

5.1 INTRODUCTION

In the preceding chapter I have shown that the mutation rate should decline with increasing G+C content if variation in the efficiency of DNA repair is responsible for the variation in G+C content seen at mammalian silent sites. This prediction is true for all sequences of over 50% G+C, and is expected over most of the G+C content range.

On certain evidence DNA repair appears a good candidate for the generation of G+C content variance. Not only does the repair of base mismatches appear to be G+C biased (Brown and Jiricny 1988, 1989) but certain types of repair appear to vary in efficiency across the genome (Bohr *et al.* 1987). Furthermore DNA repair can provide a reasonable explanation for why silent site G+C contents are usually greater than isochore G+C contents. However like all mutation hypotheses attempting to explain the silent site G+C content variance, it has problems explaining the difference between intron and silent site G+C contents (see chapter 1). The theoretical analysis of the preceding chapter also suggests that a broad range of silent site G+C contents can only be produced by variation in DNA repair under a very restricted set of parameters.

The object of this chapter is to test the hypothesis that variation in the efficiency of DNA repair is responsible for the variation in silent site G+C content by examining the relationship between silent site G+C content and substitution rate; the silent substitution rate being proportional to the mutation rate under this hypothesis. Some insight should also be gained into the cause of the silent substitution rate variance.

Although several groups have reported relationships between the rate of silent substitution and the G+C content of the sites involved (Filipski 1988, Wolfe *et al.* 1989, Ticher and Graur 1989, Bulmer *et al.* 1991), none of these studies can be used as a test because of the assumptions made in the model, and the techniques used in estimating

the rates and G+C contents. The model of the previous chapter which was used to predict that the mutation rate should decline with increasing G+C content, makes four assumptions, all of which can be tested, justified or controlled for: (1) the mutation patterns on the two strands of the DNA duplex are the same; (2) every site in the sequence is independent; (3) C↔G and A↔T mutations are not included in the analysis; and (4) the sequences are at equilibrium with respect to G+C content. Each of these assumptions will be addressed in the following section.

5.2 MATERIALS AND METHODS

5.2.1 The data set

Three mammalian groups are well represented in the DNA sequence databases: primates, rodents and artiodactyls. Of these primates and artiodactyls appear to be the most suitable for the present analysis since their silent sites seem to be at equilibrium with respect to G+C content. The evidence for this comes from comparing the third position G+C contents of primate and artiodactyl genes; they are very similar despite a reasonable amount of sequence divergence (Bernardi *et al.* 1988). In contrast rodent genes show a much narrower range of G+C contents compared to primate genes suggesting that they are undergoing what has been termed the 'minor-shift' (Mouchiroud *et al.* 1988, Bernardi *et al.* 1988).

The 58 genes used in this study were the same as those analysed by Bulmer *et al.* (1991), and were kindly provided by Dr. Ken Wolfe (Trinity College, Dublin). Details of the data set are shown in table 5.1. Only the primate and artiodactyl genes were used. It would of course be possible to use the rodent sequence data as well and obtain individual branch lengths for the primates and artiodactyls. However this approach has no advantages and there are two problems. First of all one can get non-sensical negative branch lengths; and secondly the methods for estimating the number of substitutions assume that the sequences are at equilibrium, which rodent sequences evidently are not.

5.2.2 Testing the equilibrium assumption

Simply comparing the silent site G+C contents of homologous

genes from different species is not a very powerful method by which to detect departures from equilibrium, since the correlation of G+C contents is very dependent upon the level of sequence divergence. A better method is to concentrate on those sites which differ between two sequences. If the sequences are at equilibrium we expect half the differing sites to be G or C, and half to be A or T, in each species (see section 5.6). Note that since the G+C contents of the sequences being compared must add upto one, it is only neccessary to test the sites of one sequence. The primate sequences were therefore used.

Such investigations test most departures from equilibrium. However non-silent substitution can cause departures from equilibrium by changing two-fold sites into four-fold sites, and vice-versa, if two-fold sites have a different G+C content to four-fold sites. The first test would not pick up this type of departure if the rate and pattern of non-silent substitution were similar in the two lineages, just as it cannot detect departures if primate and artiodactyl genes behave in a similar fashion. A possible solution is to plot the G+C content at two-fold sites against that at four-fold degenerate sites; deviations from a slope of one would suggest that two-fold and four-fold sites do indeed have different G+C contents, and that non-silent substitution must therefore move each type of site away from its equilibrium G+C content. However it is important to appreciate that the difference in two-fold and four-fold G+C contents must be considerable for non-silent substitution to have any effect. This is because the rate of non-silent substitution is generally much lower than that at silent sites (e.g Li *et al.* 1987) and switches between two-fold and four-fold sites can only occur by substitution at the slowest evolving codon position, the second (Kimura 1983).

5.2.3 Estimating the rate of silent substitution

It is neccessary to correct for multiple, back and parallel substitutions to obtain an accurate estimate of the number of substitutions which separate two sequences. Many ways have been suggested in which this should be done but the following method, based on the ideas of Bulmer (1991c) and Tajima and Nei (1984), turns out to be 'accurate' (see below and appendix 5.1), removing the unwanted C$\leftrightarrow$G and A$\leftrightarrow$T mutations which were ignored in the model. Following Wolfe *et al.* (1989), Bulmer *et al.* (1991) and Bulmer

(1991c) we will only consider those silent sites which are preceeded by two unchanged non-silent sites. Futhermore we will consider two-fold and four-fold sites separately. If p is the proportion of sites which differ between two sequences, counting C↔G and A↔T substitutions as unchanged sites, a suitable formula for the correction of multiple, back and parallel substitutions is

$$K = -b \, \mathrm{Ln} \, (1 - p/b) \qquad (5.1)$$

where

$$b = 1 - (C+G)^2 - (A+T)^2 \qquad (5.2)$$

C, *G*, *A* and *T* are the average proportions of sites in the sequences which are C, G, A and T respectively. Bulmer (1991c) proposed equation 5.2 from an intuitive argument; b is the expected proportion of sites which differ between two sequences after an infinite amount of divergence. In section 5.5 I show that this formula is in fact correct, in the sense that it gives the correct answer for a sequence of infinite length. The variance of K is given by Bulmer (1991c) as:

$$V(K) = \frac{p\,(1-p)}{n\,(1-p/b)^2} \qquad (5.3)$$

Throughout the following sections I will refer to K, the estimated number of substitutions per site between two sequences as the substitution rate.

5.2.4 G+C contents

G+C contents were always calculated on two-fold and four-fold sites separately, using only those sites involved in the calculation of the substitution rates; i.e only those silent sites preceded by two unchanged non-silent sites.

5.2.5 Neighbouring base effects

The model assumes that every site is independent, so that the pattern of mismatch formation, and the probability of repair are not

dependent upon the composition of the surrounding bases. I have argued that this assumption is likely to be irrelevent since the composition of the sites (non-silent sites) adjacent to a silent site do not vary from gene to gene very much, and the correlation between silent and non-silent G+C contents is weak (r^2 = 0.28 Shields *et al.* 1988, Bernardi *et al.* 1988). In other words there are two bases between each pair of silent sites which act as a buffer. However neighbouring base effects will add noise to the system which it may be useful to remove. In particular it seems worth removing the CpG effect since this is known to be quite large (Bulmer 1986, Hanai and Wada 1988, Sved and Bird 1990, Blake *et al.* 1992). The cytosine in a CpG dinulceotide can become hypermutable if it is methylated, so eliminating silent sites preceded by cytosine, or followed by guanine will largely control for this source of noise. Other neighbouring effects, of which there are many (Bulmer 1986, Hanai and Wada 1988, Eyre-Walker 1991, chapter 7), were deemed too small to be worth removing since there is always the penalty of a smaller sample size to be paid.

5.2.6 Other assumptions

One final assumption was made in formulating the model: the mutation patterns are the same on the two strands of the DNA duplex. Recent work on the primate beta-globin locus seems to confirm that this assumption is met (Bulmer 1991b).

5.3 RESULTS

5.3.1 Testing the equilibrium

Under the null hypothesis that both primate and artiodactyl sequences are at equilibrium 50% of the sites which have changed between primates and artiodactyls should be G or C. Overall this prediction is clearly not met (table 5.1). At four-fold sites 157 sites in humans are G+C, whereas 241 are A+T ($p<0.00001$ in a binomial test). At two-fold sites 282 are G+C compared to 377 which are A+T ($p<0.00001$). However there does appear to be some heterogeneity in the degree of departure from equilibrium. If we accumulate genes at random into groups of ten or eight to make the sample size reasonable

TABLE 5.1 The composition of sites which differ between primate and artiodactyl sequences.

Gene	4 fold		2 fold	
	G+C	A+T	G+C	A+T
Lactalbumin	1	2	4	1
Luitenising hormone β	1	1	1	0
SPARC	2	3	8	7
Oxytocin	1	0	2	1
Pdi disulphide isomerase	3	0	11	16
Protein kinase C β2	4	11	9	16
Protein kinase C γ	3	5	4	6
Phospholipase A2	1	0	5	5
POMC	5	3	9	0
Prolactin	5	3	5	4
SUBTOTAL	26	28	58	56
Parathyroid hormone	1	2	3	2
Terminal transferase	5	3	10	11
Thyrotropin β	1	1	6	6
Vasopressin/neurophysin II	0	1	1	1
Nicotinic acetylcholine receptor α	4	6	5	6
Atrial natriuretic factor	4	3	3	3
Mitochondrial ATPase β	7	9	8	9
G-α-i-1 protein	1	4	9	14
Endozepine	3	0	2	0
G-0-α protein	2	4	2	9
SUBTOTAL	28	33	49	61
β 1-4 galactosidase	3	4	2	7
Growth hormone	2	3	3	5
G-α-S protein	3	3	10	5
β globin	0	1	3	1

TABLE 5.1 cont/d

Interleukin 2	0	0	1	4
Enkephalin A	5	6	1	8
Glucagon	0	1	2	5
α glycoprotein	0	1	1	2
Interleukin 2 receptor	2	2	2	8
M2 muscarinic acetylcholine receptor	6	8	13	8
SUBTOTAL	21	29	38	53
Elastase	3	0	1	6
Myoglobin	0	0	4	0
Protein kinase R-i-α	11	4	6	7
Urokinase	0	7	11	6
Corticotropin	4	1	6	0
Fibroblast basic growth factor	0	0	1	2
Cholecystokinin	1	4	5	1
Enkephalin B	3	2	3	8
Relaxin	1	2	4	1
Cytochrome p450 α-17	6	8	6	9
SUBTOTAL	29	28	47	40
M1 muscarinic acetylcholine receptor	2	8	2	5
Alkaline phosphotase	2	6	8	6
G-1-α matrix protein	1	1	4	2
Albumin	2	11	15	24
Interleukin 1α	2	2	2	6
Interleukin 1β	2	4	2	4
Neuroleukin	5	10	4	6
Thrombomodulin	3	2	4	5
Nicotinic acetylcholine receptor γ	12	5	7	8
Transferrin	7	6	11	11

TABLE 5.1 cont/d

SUBTOTAL	38	55	59	77
Glutathione peroxidase	2	4	5	1
Opsin	1	5	1	5
LDL receptor	2	5	4	5
ATPase Na^+ K^+ $\alpha 1$	5	23	11	40
ATPase Na^+ K^+ β	2	6	4	8
Protein phosphatase 2A	1	5	2	4
Protein kinase C α	1	16	3	27
α globin	1	4	1	0
SUBTOTAL	15	68	31	90
GRAND TOTAL	157	241	282	377

Table 5.1 The number of differing sites which are G+C, or A+T, in primates. Subtotals were used to test whether there was heterogeneity in the degree of departure from equilibrium.

(table 5.1), we can perform a χ^2 independence test, which shows that there is heterogeneity in the degree of departure for both two-fold and four-fold sites (in both cases χ^2=22 with df=5, $p<0.005$). It is therefore possible that just one or two genes are out of equilibrium.

The bias from 50% G+C, scaled according to the expected standard error ($\surd(0.25/N)$ where N is the sample size), is plotted in figure 5.1. Those genes which lie outside plus or minus two standard errors are probably not at equilibrium, since the probability of them doing so by chance alone is less than 5%. There are nine genes which lie outside this boundary, four genes at four-fold sites and seven at two-fold sites (two genes are out of equilibrium at both types of site). However even if these are removed there is still some evidence of a departure from equilibrium; overall there are 149 four-fold

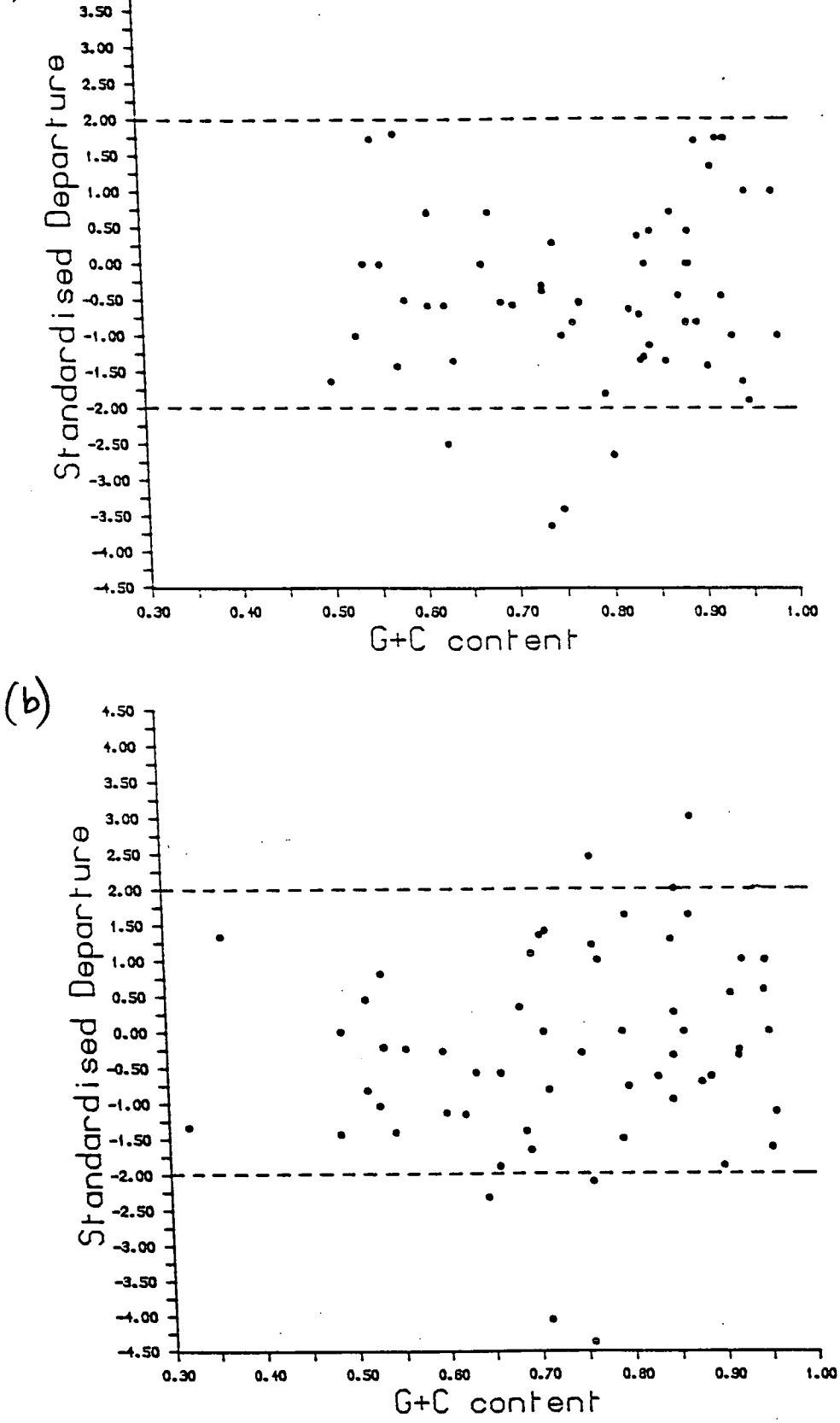
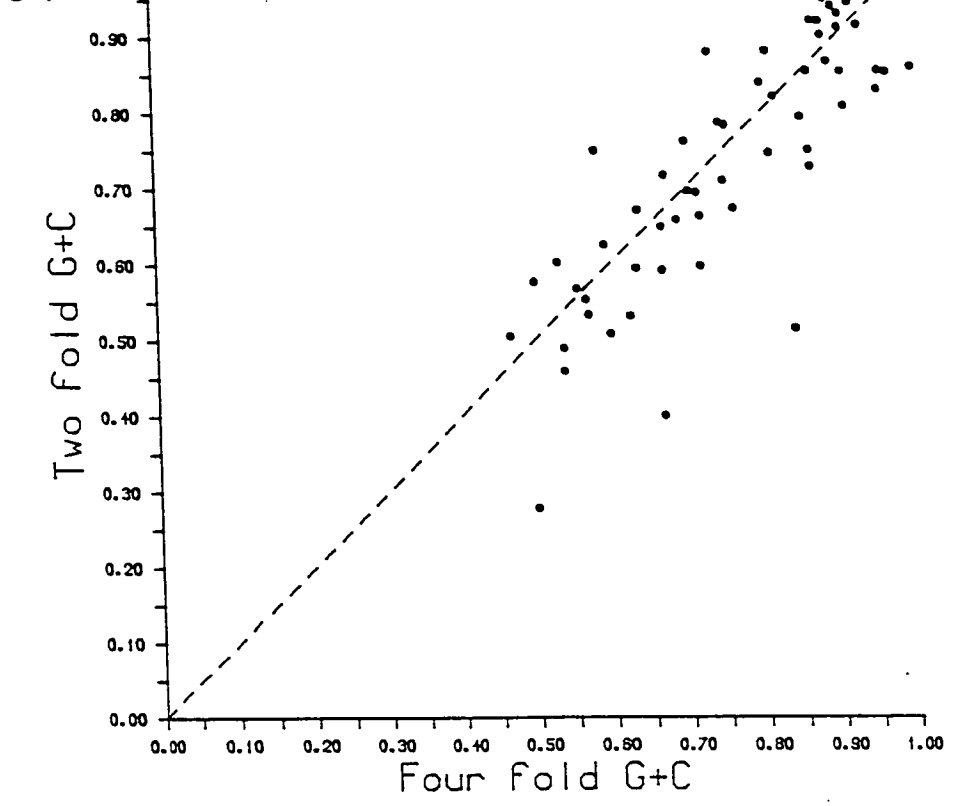


Figure 5.1 The standardised degree of departure from equilibrium plotted against the G+C content for four-fold sites (a) and two-fold sites (b). Dotted lines mark plus and minus two standard errors, outside which genes are unlikely to be at equilibrium.



(b)

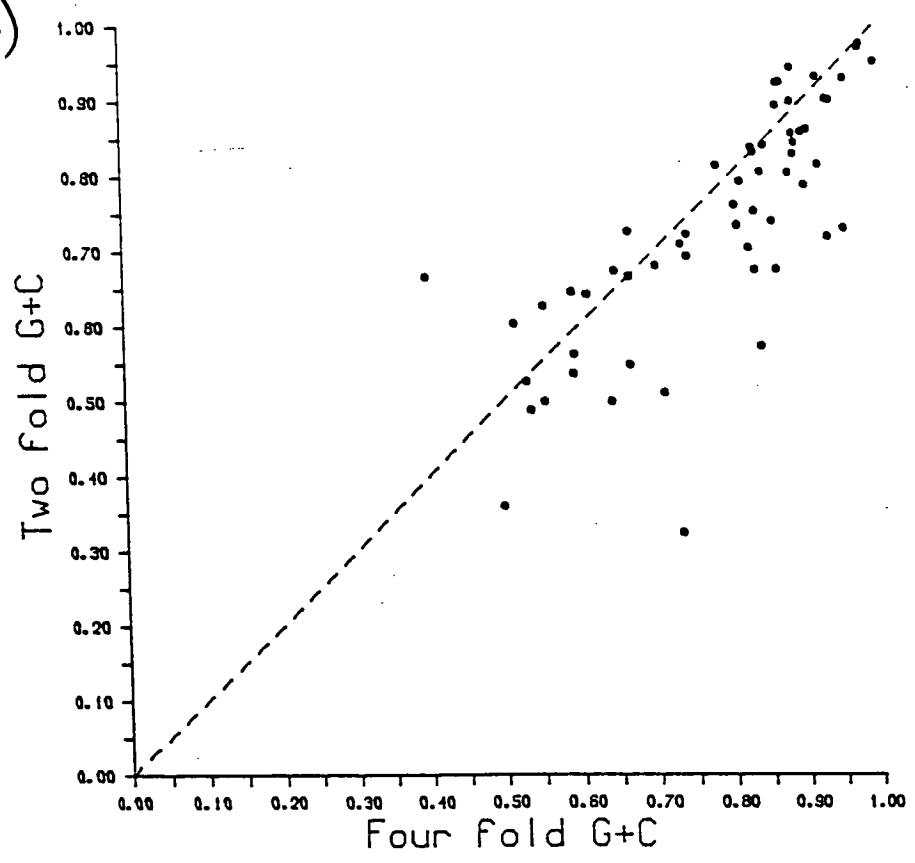


Figure 5.2 The four-fold site G+C content plotted against that at two-fold sites for primates (a) and artiodactyls (b). The dotted line passes through the origin and has a slope of one.

sites which are G+C out of a total of 333 four-fold sites which differ between primates and artiodactyls ($p<0.06$). At two-fold sites there are 246 out of 539 sites ($p<0.05$). Despite this the equilibrium condition seems to be approximately true. Ignoring the nine genes which are clearly out of equilibrium the proportion of sites which differ between the two species, which are G or C in humans is about 44-45%; not greatly different from 50%. It is important to also note that the degree of departure does not appear to be related to the G+C content of the sites involved (see figure 5.1), just as it is not related to the silent substitution rate (results not shown).

In figure 5.2 the G+C content at two-fold sites is plotted against that at four-fold sites for both primate and artiodactyl genes in order to assess the possible impact of non-silent substitution. In neither case does there appear to be any great departure from the slope of one suggesting that non-silent substitution will have little effect on the equilibrium of two-fold and four-fold sites.

5.3.2 The substitution rate/G+C content relationship

The two-fold and four-fold degenerate silent substitution rates are plotted against their respective G+C contents in figure 5.3. All C$\leftrightarrow$G and A$\leftrightarrow$T mutations have been ignored, as have all silent sites preceded by C or followed by G. Note that both two-fold and four-fold sites show a very broad range of G+C contents despite the removal of CpG effects; in fact a number of genes are not shown in figures 5.3 because their G+C content is so high that equation 5.1 becomes undefined with the smallest amount of divergence. For instance oxytocin (98%) and phospholipase A2 (95%) at four-fold sites, and oxytocin (95%) at two-fold sites are undefined.

There is clearly a relationship between the rate of silent substitution and the G+C content at four-fold degenerate sites. This is confirmed by both weighted and unweighted least squares regression. If the reciprocals of the substitution rate variances (equation 5.3) are used as weights the quadratic model

$$Y = 0.22 - 3.31\ (X - 0.70)^2 \qquad (5.4)$$

where Y is the substitution rate and X is the G+C content, explains

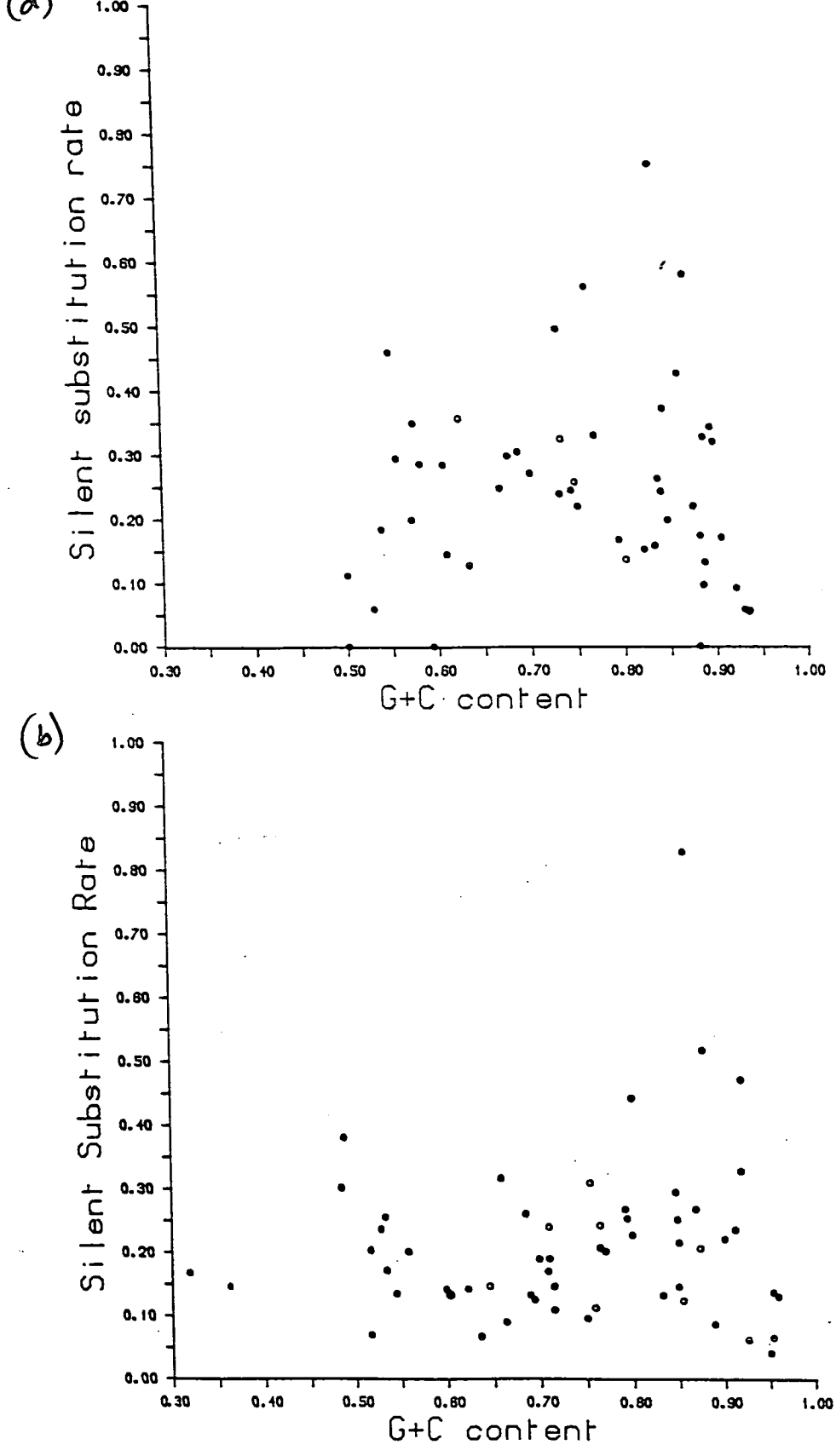


Figure 5.3 The silent substitution rate plotted against the G+C content for four-fold (a) and two-fold (b) sites. Open circles denote those genes identified as being out of equilibrium (see text).

45% of the variance in substitution rate. All the coefficients are significant, as is the model overall (all $p<0.001$). Note that the genes which show the greatest departure from equilibrium, marked as unfilled circles, do not fall as outliers, and that the maximum substitution rate is attained at a G+C content way in excess of 50%. Equation 5.4 is very similar to that obtained by Bulmer *et al.* (1991) by slightly different methods.

In contrast to the picture at four-fold sites there appears to be no relationship between the rate of silent substitution at two-fold sites and the G+C content of the sites concerned. No model, fitted by weighted or unweighted least squares regression explains a significant proportion of the variance in the substitution rate.

5.4 DISCUSSION

I have previously shown that the rate of silent substitution should decline with increasing G+C content if the variation in silent site G+C contents is due to different genes being subject to different levels of repair. Both two-fold and four-fold degenerate sites show very large G+C content ranges yet neither shows a decrease in silent substitution rate with increasing G+C content. The four-fold case is possibly more convincing since there is a highly significant quadratic relationship between the rate of silent substitution and the G+C content, whereas two-fold sites may just be subject to too much noise to show any relationship. However it seems difficult to suggest that two-fold and four-fold sites, which show such similar G+C content ranges, are subject to different processes. It therefore seems that DNA repair is not solely responsible for the variation in silent site G+C contents.

This conclusion is supported by two other observations. First, just like all other mutation hypotheses, DNA repair does not *a priori* predict that introns should have a lower mean and smaller range of G+C contents than silent sites. And second there is some evidence that the efficiency of repair is related to level of transcription (Okumoto and Bohr 1987), so we might expect genes expressed in the germ line to have higher G+C contents than those which are not. This does not appear to be the case since housekeeping genes are not more G+C rich at silent sites than tissue specific sequences (Eyre-Walker

1992a, chapter 3).

5.4.1 The equilibrium assumption

However one of the assumptions made in the model does not appear to be met; the sequences do not appear to be at equilibrium. For most genes the departure is quite small and importantly, not related to the G+C content. Futhermore two-fold and four-fold genes show very similar levels and patterns of departure yet seem to show very different substitution rate/G+C content relationships, and those genes which show the largest departures from equilibrium to not appear to be outliers. If the degree of departure from equilibrium had been related to the G+C content we might have had better reason to suspect that the state of non-equilibrium was responsible for the relationship at four-fold sites. For instance if four-fold sites with a G+C content of ~70% had shown the greatest departure we might have suspected that the maximum in the silent substituion rate/G+C content relationship (figure 5.3) was caused by departures from equilibrium. It therefore seems unlikely that departures from equilibrium are responsible for the shape of the silent substitution rate/G+C content relationship at four-fold sites, although such departures could be a source of noise which might be responsible for the lack of a relationship at two-fold sites.

5.4.2 Bias in the substitution rate measures

It seems unlikely that equation 5.1 and 5.2 give unbiased estimates of the substitution rate, so the quadratic nature of the relationship at four-fold sites might be an artifact. The bias would appear to potentially come from two sources. Firstly equations of the form $-bLn(1-p/b)$ become undefined if $p>b$, so since b (equation 5.2) is a function of the G+C content one might expect a G+C content related bias to be introduced into the calculation. This does appear to be the case since eight genes have undefined substitution rates in the four-fold site data set, all of which have G+C contents in excess of 90%, except cholecystokinin which has a G+C content of 83%. However this source of error will tend to bias estimates downwards, in favour of the DNA repair model; i.e it is the rise in substitution rate from 50% G+C to 70% which the DNA repair model cannot explain.

Secondly bias would appear to be introduced by measuring the rate

over a finite number of sites. To illustrate this let us imagine that we have a large number of sites and a reasonable level of divergence so that the proportion of sites which differ between two sequences is normally distributed. Furthermore let us assume, for simplicity, that the value of b is known with no error. If p is normally distributed so is p/b (although b scales the distribution) and so is 1-p/b. However taking the logarithm of a normal distribution skews it to the right biasing the estimate downwards. Since b scales the distribution of p the level of bias is dependent upon sequence G+C content. It should be possible to estimate the size of this bias, but for the moment it remains a weakness in the argument against the repair hypothesis.

5.4.3 Substitution rate variance

Although some of the silent substitution rate variance can be explained by G+C content, and therefore not by DNA repair, there is a possibility that repair is responsible for the remaining variance. On the other hand this residual variance might simply be due to error associated with measuring the substitution rates over a finite number of sites (i.e equation 5.3). To investigate this each residual was divided by the standard error associated with the relevent substitution rate. These standardised residuals are plotted in figure 5.4. If the substitution rate is normally distributed (see section 5.4.2) and G+C content the only factor upon which the substitution rate at four-fold sites depends, then we would expect the majority of the standardised residuals to lie between plus and minus two. This they do, suggesting that G+C content might explain all the variance in silent substitution rate there is to explain. It therefore seems unlikely that DNA repair contributes anything to the silent substitution rate variance at four-fold sites, although one must remember that the substitution rate is not strictly normally distributed and the variances are only approximate. From the analysis of four-fold site substitution rates we might also infer that it does not cause any variance in substitution rate at two-fold sites.

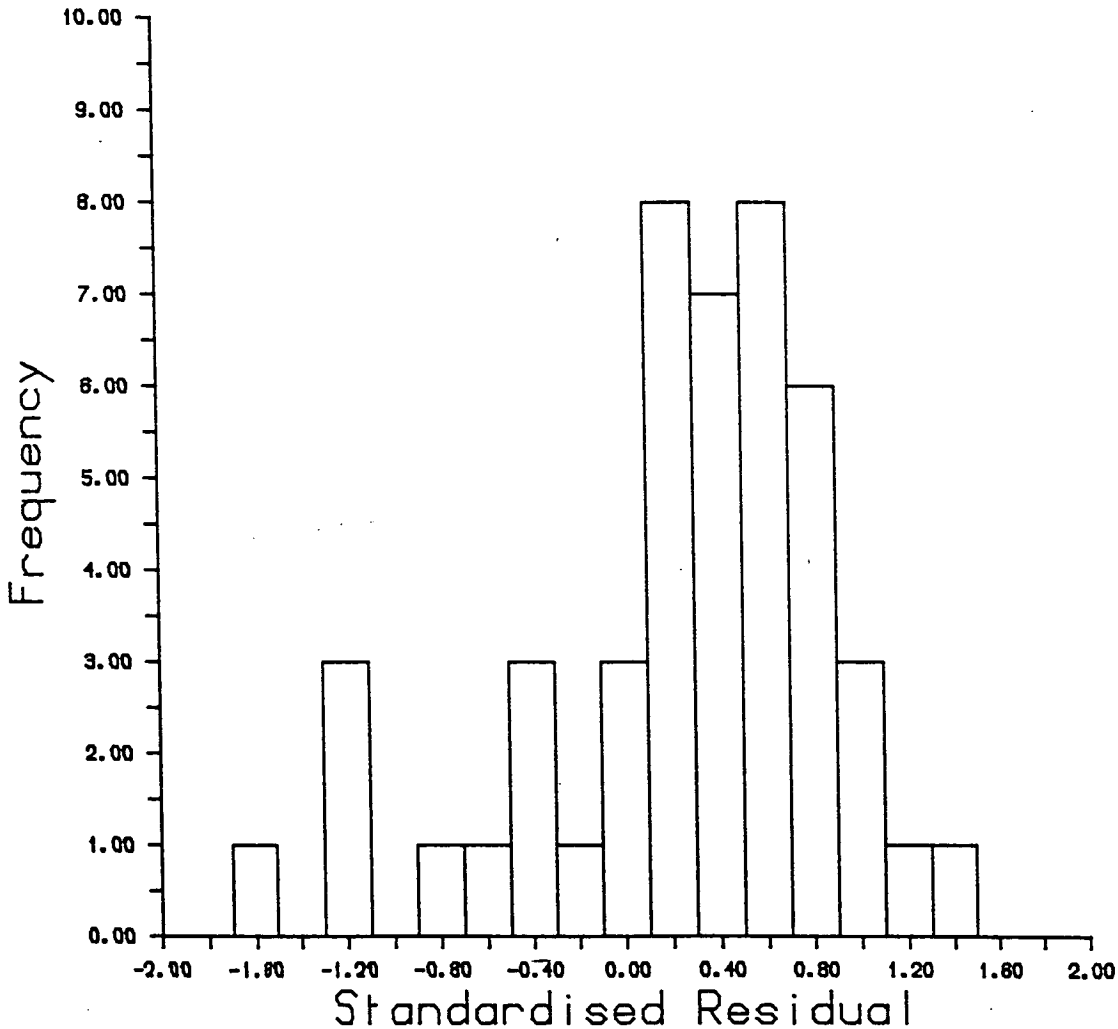


Figure 5.4 Histogram of the standardised residuals for four-fold sites; i.e the residuals from the regression analysis divided by the standard error associated with the substitution rate estimation.

5.4.4 Summary

In the previous chapter it was possible to show that the mutation rate should decline with increasing G+C content if variation in the efficiency of repair is responsible for the silent site G+C content variance between mammalian genes. At neither two-fold or four-fold sites is there any evidence that this prediction is met.

5.5 APPENDIX I

In this appendix equations 5.1 and 5.2 are derived. Let us assume that the pattern of substitution is the same on the two strands of the DNA duplex so at equilibrium the frequency of C equals the frequency of G (and A=T). The pattern of substitution can then be described with just two parameters if we ignore C$\leftrightarrow$G and A$\leftrightarrow$T changes: let the probability with which C:G base pairs change to A:T (or T:A) be γ and the probability of the reverse process be $\gamma\delta$. Let X_C be the probability that a C:G base pair at some time t=0, is C:G or G:C at time t, and let X_A be the probability that an A:T base pair at time t=0 is A:T or T:A at time t. Then

$$\Delta X_C = X_C\,(1-\gamma\delta) + (1-X_C)\gamma - X_C \qquad (5.A1)$$

which can be approximated by the continous function

$$\partial X_C/\partial t = \gamma\,(1-X_C(1+\delta)) \qquad (5.A2)$$

Integrating and noting that X_C=1 when t=0 we obtain

$$X_C = \frac{1 + \delta\,\exp(-\gamma(1+\delta)t)}{1+\delta} \qquad (5.A3)$$

By similar steps it can be shown that

$$X_A = \frac{\delta + \exp(-\gamma(1+\delta)t)}{1+\delta} \qquad (5.A4)$$

The proportion, P, of sites which differ between two sequences at equilibrium of G+C content f is then

$$P = 1 - f\,(X_C^2 + (1-X_C)^2) - (1-f)(X_A^2 + (1-X_A)^2) \quad (5.A5)$$

If we note that $f=1/(1+\delta)$ the above can be rearranged to give

$$2\gamma t = \frac{-1}{1+\delta} \mathrm{Ln}\left[1 - \frac{(1+\delta)^2}{2\delta} P\right] \quad (5.A6)$$

which can be used to give K, the average number of substitutions per site:

$$K = f\,2\gamma\delta t + (1-f)\,2\gamma t$$

$$= \frac{-2\delta}{(1+\delta)^2} \mathrm{Ln}\left[1 - \frac{(1+\delta)^2}{2\delta} P\right] \quad (5.A7)$$

which is -b Ln (1-p/b) with $b=1-f^2-(1-f)^2$; i.e equations 5.1 and 5.2.

5.6 APPENDIX II

To demonstrate that sites which differ between two sequences at equilibrium have a G+C content of 50%, irrespective of whether the two lineages evolve at different rates, consider the following. Let the rate at which lineage x evolves be γ_x and the probability that a C:G base pair is C:G (or G:C) at time t be X_{Cx}. Then the proportion of sites which are G:C in sequence 1, and which differ between two sequences is

$$U = f\,X_{C1}\,(1-X_{C2}) + (1-f)(1-X_{A1})X_{A2} \quad (5.A8)$$

and the total proportion of sites which differ is

$$V = f\,(X_{C1}(1-X_{C2}) + (1-X_{C1})X_{C2}) + (1-f)((1-X_{A1})X_{A2} + X_{A1}(1-X_{A2})) \qquad (5.A9)$$

If one substitutes in equations 5.A3 and 5.A4 for X_{Cx} and X_{Ax}, it can be shown that the G+C content of the sites which differ between two sequences at equilibrium is equal to one half; i.e U/V = 1/2.

CHAPTER 6

DNA RECOMBINATION
AND
THE RELATIONSHIP BETWEEN RATE AND CONSTRAINT IN NEUTRAL MODELS

6.1 RECOMBINATION

6.1.1 Introduction

The possibility that the frequency of recombination is a major determinant of synonymous codon use in mammalian genes is suggested by two observations made by Ikemura and Wada (1991): (1) G+C rich genes are found in chaismata dense chromosome bands more frequently than A+T rich genes; and (2) the average silent site G+C content of genes on a chromosome is positively correlated to the mean number of chiasmata.

The mechanism by which recombination can cause G+C content variance is quite simple. It is generally believed that heteroduplex DNA forms during recombination producing base mismatches at sites for which the individual is heterozygous. Such base mismatches might be subject to repair, which will probably be biased. So if the frequency of heteroduplex formation and the probability of repair are sufficient gene conversion will affect the probability of mutant fixation and therefore the composition of a sequence. Thus if there is variance in the frequency of recombination across the genome G+C content variance will be produced.

The work of Brown and Jiricny (1988) suggests that the repair of base mismatches is generally efficient and biased in the direction of G+C, thus tieing in neatly with the observations of Ikemura and Wada (1991). Furthermore there are good reasons why one might expect the frequency of recombination to vary across the genome. Recombination is generally thought to have advantages above and beyond those associated with its role in DNA repair, since it reduces interference between advantageous alleles (Fisher 1930, Muller 1932), generates genetic variation (Muller 1932), reduces the mutation load if there is synergistic epistasis (Kondrashov 1982, 1988, Charlesworth 1990)

and stops the accumulation of genetic damage (Muller 1964). We might therefore expect recombination to be favoured in gene dense regions of the genome, or in areas associated with particular sets of genes, like those involved in the immune system. Of course there may also be selection to reduce recombination in certain areas to preserve coadapted complexes of genes. It is important to appreciate that there does not have to be any genetic variation for chiasmata density; those individuals which survive may be those which have undergone recombination events in 'useful' places. In other words even if there is no variation in the frequency of recombination across the genome there will be variation in the probability with which recombinant events survive. Interestingly the potential association between gene density, recombination and G+C content may explain why G+C rich isochores have higher gene densities than A+T rich isochores (Bernardi *et al.* 1985).

Following the precedent of earlier chapters the relationship between the expected substitution rate and G+C content will be investigated. This turns out to be a simple problem for two-fold degenerate sites but rather more complex for four-fold sites.

6.1.2 The fixation probability for a mutation subject to gene conversion

The first step in assessing the role of recombination in the generation of G+C content variance is to assess the effect biased gene conversion has on the fixation probability of a new mutation. If the frequency of a mutant allele is x and the bias in conversion is ω, such that a proportion $(\omega+1)/2$ of the gametes from a heterozgote carry the mutant allele, then it is not difficult to show that the change in the frequency of the mutant allele is

$$\Delta x = \omega\, x(1-x) \qquad (6.1)$$

assuming non-overlapping generations. This is the same as for a semi-dominant allele with selective advantage ω. Since the variance in allele frequency, $x(1-x)/2N$ (where N is the effective population size), is also the same as for a semi-dominant allele, the fixation probability of a new mutation subject to biased gene conversion multiplied by 2N is

$$P = \frac{4N\omega}{1 - \exp(-4N\omega)} \qquad (6.2)$$

assuming that ω is small and ignoring subsequent mutation (Kimura 1957, Nagylaki 1983). The parameter ω can be split into two parts: τ, the probability of a site being involved in heteroduplex DNA, and ψ a parameter which measures the bias in repair such that a proportion $(\psi+1)/2$ of the strands from such a site are G or C ($\omega=\tau\psi$). If there is no repair, or repair is unbiased then $\psi=0$. If repair is totally efficient and biased to G+C then $\psi=1$, and if it is totally efficient and biased to A+T, $\psi=-1$. Note that I have, and will continue to refer to the recombination hypothesis as a mutation hypothesis despite the similarity between the processes of gene conversion and selection; biased gene conversion is essentially selection upon the sequence as a phenotype. I do this because I believe this is more in keeping with the generally accepted view of what selection and phenotypes are.

6.1.3 Two fold degenerate sites

Let us consider the equilibrium G+C content, and substitution rate at two-fold degenerate sites. If the mutation pattern is the same on the two strands of the DNA duplex and there is no tendency to include one strand more often in heteroduplex than the other, then a two-fold site can be represented as a two allele system.

Let the probability of a mutation from C:G to T:A be U_{CT} and reverse be U_{TC}; and let the probability of fixing (multiplied by 2N) a G:C mutation in a population of A:T be P_{TC}, the reverse being P_{CT}. If we assume that $NU_{ij}<<1$ the equilibrium frequency of G:C, $\bar{f}$, can be obtained by considering the flux between the two alleles (C:G and A:T), since the population will generally be monomorphic. Furthermore P_{ij} will be given by equation 6.2 such that $P_{TC} = 4N\omega/(1-\exp(-4N\omega))$ and $P_{CT} = -4N\omega/(1-\exp(4n\omega))$. The equilibrium frequency of an allele in this context is the time for which a site is fixed for an allele, or the frequency of the allele over an infinite number of sites.

By considering the change in the frequency of G:C base pairs

$$\Delta f = -f\,(U_{CT}P_{CT}) + (1-f)(U_{TC}P_{TC}) \qquad (6.3)$$

it is simple to show that the equilibrium G+C content is

$$\bar{f} = \frac{U_{TC}\, P_{TC}}{U_{TC}\, P_{TC} + U_{CT}\, P_{CT}} \qquad (6.4)$$

which simplifies to

$$\bar{f} = \frac{U_{TC} \exp(4N\omega)}{U_{TC} \exp(4N\omega) + U_{CT}} \qquad (6.5)$$

An equivalent formula has been given by Li (1987) for the haploid selection case.

The rate of substitution in a sequence at equilibrium is

$$S = \bar{f}\, U_{CT}\, P_{CT} + (1-\bar{f})\, U_{TC} P_{TC} \qquad (6.6)$$

If we substitute equation 6.4 into 6.6, divide by the rate in a sequence without gene conversion ($2U_{TC}U_{CT}/(U_{CT}+U_{TC})$) and note that we can define $U_{CT}+U_{TC}=1$, the rate of substitution relative to that in a sequence undergoing no gene conversion is

$$R = \frac{P_{TC}\, P_{CT}}{U_{TC}\, P_{TC} + U_{CT}\, P_{CT}} \qquad (6.7)$$

If we now rearrange equation 6.5 to get an expression for $N\omega$, equation 6.7 becomes

$$R = \frac{\bar{f}(1-\bar{f})}{\bar{f}-1+U_{CT}} \operatorname{Ln}\left[\frac{U_{CT}\, \bar{f}}{(1-U_{CT})(1-\bar{f})}\right] \qquad (6.8)$$

an expression which relates the relative substitution rate to the equilibrium G+C content as the value of $N\omega$ changes. This expression is plotted in figure 6.1 for the case of a G+C biased repair system

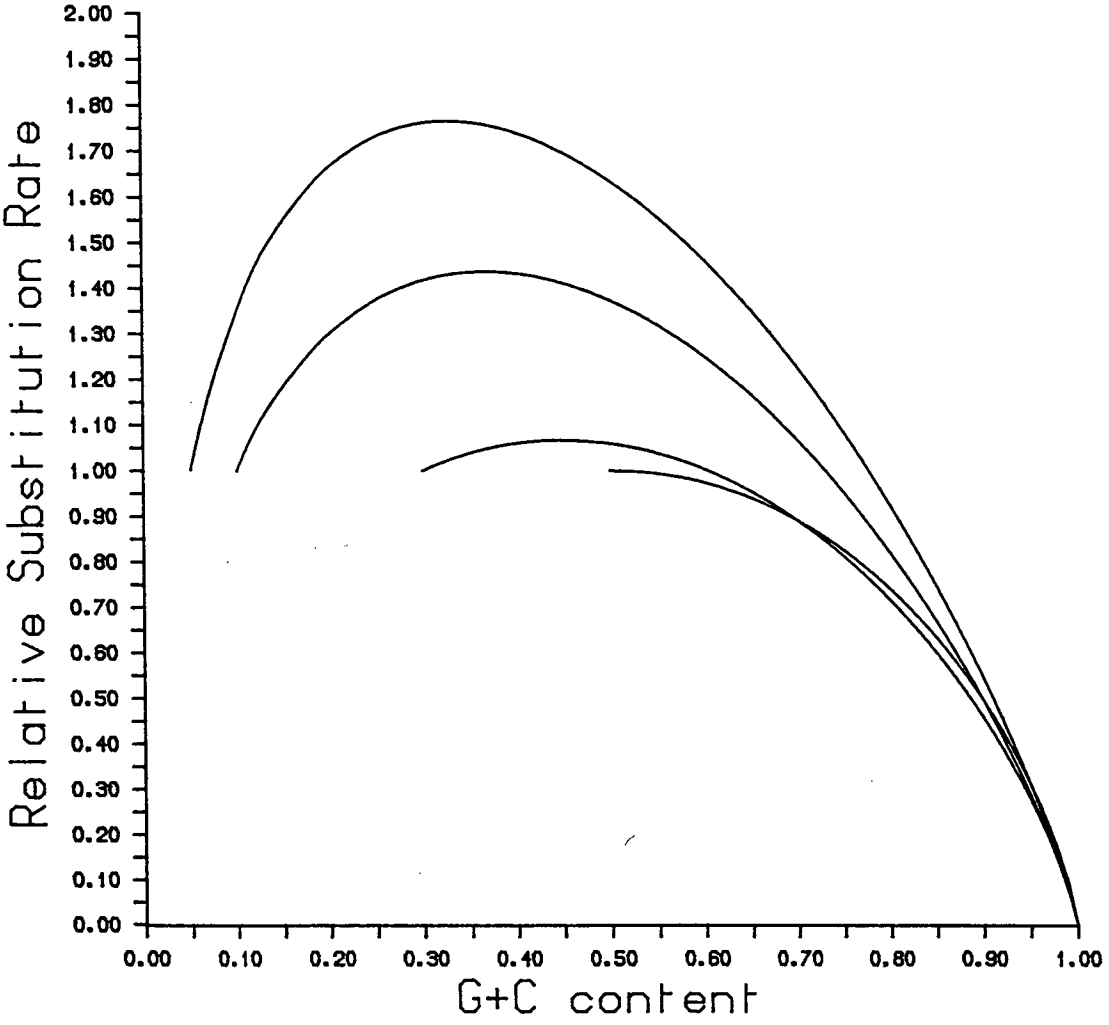


Figure 6.1 The relative substitution rate plotted against the equilibrium G+C content for two-fold degenerate sites subject to various levels of gene conversion. The curves beginning from left to right are for U_{CT} = 0.95, 0.90, 0.70, 0.50 corresponding to sequences with no gene conversion of 5%, 10%, 30% and 50% G+C.

($\omega>0$). Somewhat surprisingly there are conditions under which an increase in the level of gene conversion can lead to an increase in the rate of substitution. This is surprising because gene conversion is equivalent to selection in its effect, so we have here a system in which the imposition of selection (constraint) leads to an increase in the rate of evolution, which is counter to the often cited 'neutral' dictum that an increase in constraint leads to a decrease in the rate of evolution (Kimura and Ohta 1974, Kimura 1983). In section 6.2 I investigate this problem of rate and constraint in more depth.

The relationships shown in figure 6.1 do not square easily with the findings of chapter 5 in which we failed to find a relationship between the rate of silent substitution and the G+C content at two-fold sites. However since the lack of a relationship at two-fold sites might be due to excessive noise in the system it is not sensible to place too much emphasis on this result.

6.1.4 Four fold degenerate sites

The situation at four-fold degenerate sites is rather more complicated since we have to consider transitions and tranversions separately; although we can ignore C$\leftrightarrow$G and A$\leftrightarrow$T mutations since they were not included in the analysis in chapter 5 and are unlikely to be affected by gene conversion. Following methods similar to those above it is easy to derive expressions for the equlilibrium G+C content and relative substitution rate. If we let the subscripts CA and AC denote transversions the equilibrium G+C content can be written as

$$\bar{f} = \frac{U_{TC}\,P_{TC} + U_{AC}\,P_{AC}}{U_{TC}\,P_{TC} + U_{CT}\,P_{CT} + U_{AC}\,P_{AC} + U_{CA}\,P_{CA}} \qquad (6.9)$$

and the relative rate of substitution as

$$R = \frac{(U_{TC}P_{TC} + U_{AC}P_{AC})\,(U_{CT}P_{CT} + U_{CA}P_{CA})}{(U_{TC} + U_{AC})(U_{TC}P_{TC} + U_{CT}P_{CT} + U_{AC}P_{AC} + U_{CA}P_{CA})(U_{CT} + U_{CA})} \qquad (6.10)$$

These do not simplify further, and it is readily apparent from

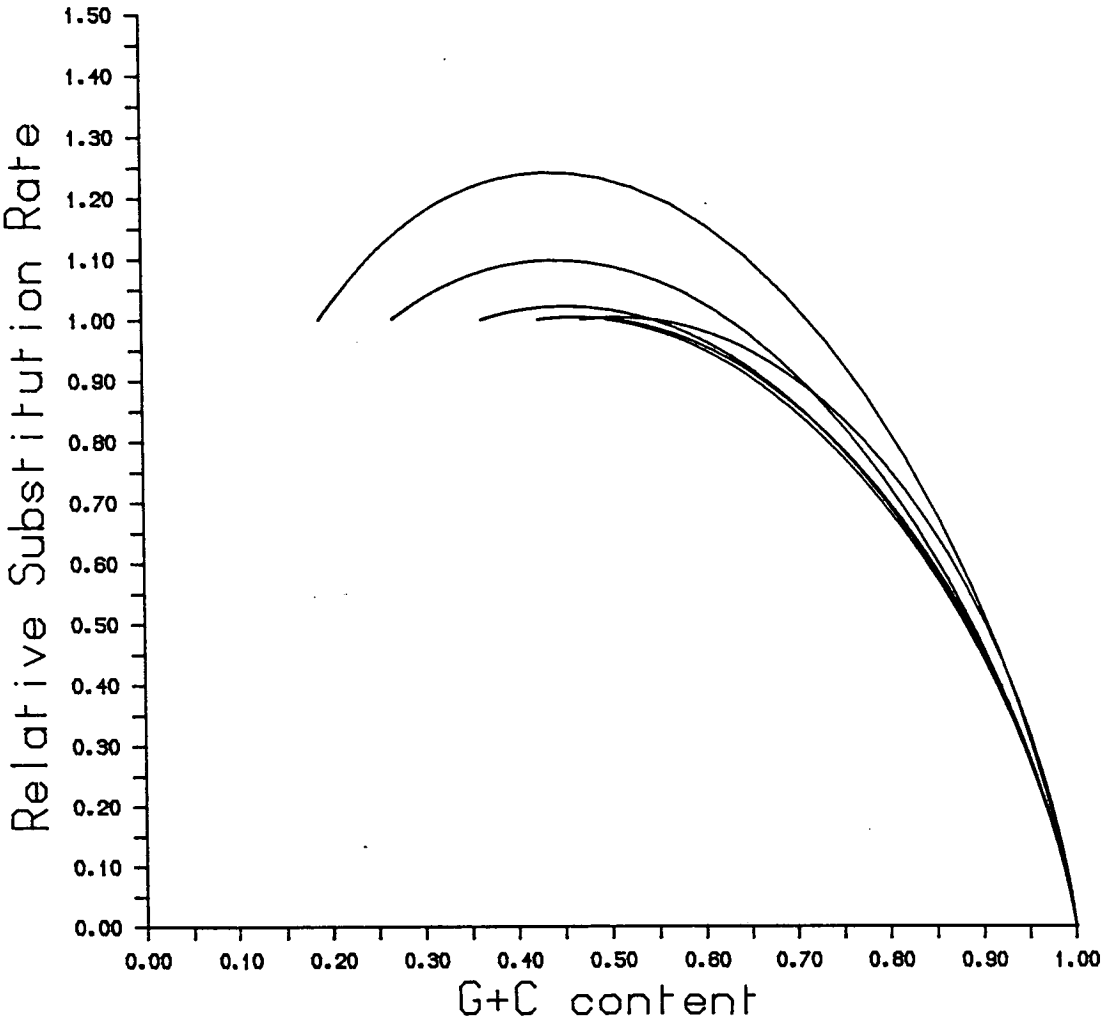


Figure 6.2 Examples of the the relative substitution rate/G+C content relationship at four degenerate sites when the mutation pattern is randomly generated, but the biases in repair are those calculated from the data of Brown and Jiricny (1988).

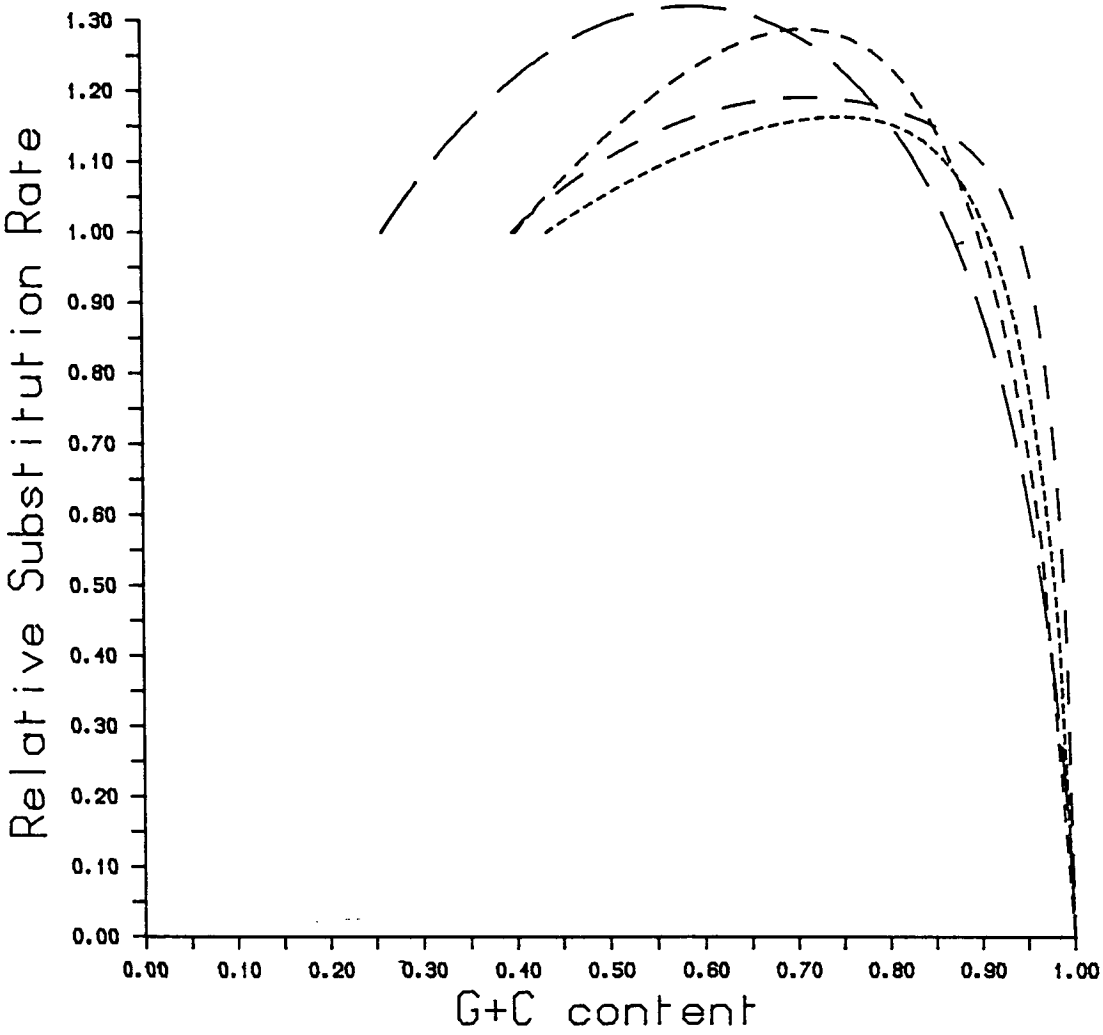


Figure 6.3 Examples of the relative substitution rate/G+C content relationship when both the mutation pattern and the repair bias is randomly generated, and the maximum substitution rate is attained at a G+C content of about 70%. In ascending order of dash length (shortest to longest) the curves had the following parameters.

ψ (Transitions)	ψ (Transversions)	U_{CT}	U_{TC}	U_{CA}	U_{AC}
0.064	0.728	0.44	0.21	0.12	0.22
0.065	0.531	0.58	0.23	0.02	0.17
0.897	0.060	0.23	0.37	0.37	0.03
0.852	0.194	0.30	0.25	0.44	0.00

evaluating these expressions that the relationship between the substitution rate and G+C content is complicated. The behaviour of the system depends very much on the relative values of ψ for transitions and transversions. For instance if the repair of transitions is G+C biased and the repair of transversions A+T biased, the rate of evolution eventually increases with the probability of heteroduplex formation; whereas it eventually declines if the repair of transitions and transversions is biased in the same direction.

Fortunately the work of Brown and Jiricny (1988) allows us to estimate the values of ψ for the two types of mutation. They estimated that 94% of the strands from G:T mismatches would end up either C or G, with 52% from A:C (G:T and A:C being the two transition mismatches). Under the assumption that there is no bias in heteroduplex formation with respect to strand these two mismatches will form equally frequently. The value of ψ for transitions is thus 0.46 (=(0.94+0.52)-1) and that for transversions 0.32. These values are reasonably similar and one might therefore expect the dynamics of the two types of mutation to be alike, and the overall substitution rate/G+C content relationship to be similar to that for the two allele case. This is confirmed by randomly generating mutation patterns and evaluating equations 6.9 and 6.10 over a wide range of heteroduplex formation probabilities ($N\tau$). Out of 1000 randomly generated mutation patterns none gave a maximum substitution rate at greater than 56% G+C. Some examples of relationships produced are given in figure 6.2.

However it is evident from randomly generating both mutation and bias (ψ) parameters that relationships very similar to the substitution rate/G+C content relationship found for four-fold sites can be produced (chapter 5). Four examples are given in figure 6.3 along with their parameter values. Although the sample size is clearly too small to draw firm conclusions, it does seem that one type of mutation (transitions or transversions) must have a low degree of repair bias compared to the other, and that this mutation should favour A+T base pairs.

6.1.5 Is the pattern of repair unique?

The critical question therefore is whether the values of bias observed by Brown and Jiricny (1988) are generally applicable to

mammalian systems, or whether they are unique to the monkey cell line in which they were measured. Afterall cell lines, which are under intense selection to proliferate, tend to assume their own biology. The uniqueness of the monkey cell line repair system can be partially tested since the repair of G:T mismatches has been assayed in seven other independent human cell lines (Brown and Jiricny 1988, 1989)(table 6.1). If we group 'unrepaired' and 'repaired to A:T' classes, and also group together the NHF and BS cell lines we can perform a chi-square test for heterogeneity; there is no evidence that the probability and pattern of repair varies between cell lines (χ^2 = 2.056 with 3 degrees of freedom, p>0.10). Thus we might expect the substitution rate/G+C content relationship at four-fold sites to look like that at two-fold sites and the curves given in figure 6.1.

TABLE 6.1

Cell line	Unrepaired	Repaired to G:C	Repaired to A:T
NHF 1187	1	36	2
NHF 1222	0	27	1
NHF 3229	0	8	0
XP(A) 1223	2	77	8
BS 916	0	25	0
BS 1492	1	10	2
BS 2548	0	27	3
Monkey	3	72	3

Table 6.1 The repair of G:T mismatches in seven human fibroblast cell lines and one African Green monkey kidney cell line. Data are from Brown and Jiricny (1988, 1989).

6.1.6 Conclusions

The two models presented above make very similar assumptions to those made in the DNA repair model; the sequences are at equilibrium, sites are independent and the pattern of mutation is the same on the two strands of the DNA duplex. These assumptions appear to be largely met in the data set of chapter 5 so the discordance between the relationships found there and those theoretically derived here suggest that variation in the frequency of recombination across the genome is not responsible for the differences in synonymous codon usage seen between mammalian genes. However variation in recombination frequency might be responsible for the differences in G+C content between isochores, a point discussed in more detail in chapter 8.

6.2 RATE AND CONSTRAINT IN NEUTRAL MODELS

6.2.1 Introduction

In the two allele model above we saw how an increase in the frequency of gene conversion/selection could lead to an increase in the rate of evolution. This seems to go against one of the central tenets of the neutral theory of molecular evolution, which states that the rate of evolution is inversely related to the level of constraint upon a sequence (Kimura and Ohta 1974, Kimura 1983). One might argue that the imposition of weak selection upon a system is not equivalent to constraint since both advantageous and deleterious alleles segregate as a consequence of the selection. However as I will now show the imposition of strong purifying selection can also elevate the rate of evolution despite decreasing the number of potential neutral alleles, i.e increasing the constraint. The approach taken is to consider the rate of evolution at single sites, comparing the rates at sites which have two, three or four neutral alleles (with all other alleles being very deleterious). Note that a site with one neutral allele does not undergo substitution.

6.2.2 Single sites

Following Wright (1969 chapter 3) let us consider a single site in a DNA molecule at which n alleles can segregate. Normally n will have

a maximum value of four corresponding to the four bases which can occupy a site. Let the frequency of the ith allele be f_i, the probability of a mutation from allele i to j be U_{ij} and the probability of fixing a newly arising j mutant in a population of i be P_{ij}. Let $K_{ij}=2NU_{ij}P_{ij}$ where N is the population size of a diploid organism. If we assume $NU_{ij}<<1$ then $K_{ij}=U_{ij}$ (Kimura 1968). Furthermore the population will generally be monomorphic for one of the alleles and the expected frequency of allele i ($\bar{f}_i$)(i.e the relative amount of time for which the population is fixed for i) can be obtained from a consideration of the flux between the n alleles: i.e by solving n-1 simultaneous equations of the form:

$$\Delta f_i = -f_i \sum_{j \neq i} K_{ij} + \sum_{j \neq i} f_j K_{ji} = 0 \qquad (6.11)$$

The average rate of substitution at the site is then

$$R_n = \sum_i \bar{f}_i \sum_{j \neq i} K_{ij} \qquad (6.12)$$

The equilibrium frequencies for two, three and four allele systems are given by Wright (1969 chapter 3). As one might expect the expressions for R_3 and R_4 are complex and do not yield readily to further analysis. However there is an informative simplification that can be made. If we consider the mutation pattern symmetrical about allele 4 such that $U_{12} = U_{21} = U_{13} = U_{31} = U_{23} = U_{32} = U$, $U_{14} = U_{24} = U_{34} = U_{*4}$ and $U_{41} = U_{42} = U_{43} = U_{4*}$ the rates of evolution in the following systems of two, three and four neutral alleles become:

$$R_4 = \frac{6\, U_{4*}\, (U + U_{*4})}{3\, U_{4*} + U_{*4}} \qquad (6.13a)$$

$$R_3(*,*,4) = \frac{2\, U_{4*}\, (U + 2U_{*4})}{2\, U_{4*} + U_{*4}} \qquad (6.13b)$$

$$R_3(*,*,*) = 2U \qquad (6.13c)$$

$$R_2(*,*) = U \qquad (6.13d)$$

where * refers to one of alleles 1,2 or 3. Using the expressions in (6.13) it is simple to show that

$$R_3(*,*,*) > R_4 \quad \text{when} \quad U > 3U_{4*} \qquad (6.14a)$$

$$R_2(*,*) > R_4 \quad \text{when} \quad U > \frac{6\,U_{4*}\,U_{*4}}{U_{*4} - 3\,U_{4*}} \quad \text{and} \quad U_{*4} - 3\,U_{4*} > 0 \qquad (6.14b)$$

$$R_2(*,*) > R_3(*,*,4) \quad \text{when} \quad U > 4U_{4*} \qquad (6.14c)$$

In other words there are mutation patterns under which a reduction in the number of potential neutral alleles at a site leads to an increase in the average rate of evolution. Note however how biased the mutation patterns must be for this to occur. Essentially the elimination of allele 4 in each case leads to an increase in the time for which the site is occupied by alleles 1, 2 and 3. So if the mutation rate between alleles 1, 2 and 3 is high, elimination of allele 4 increases the overall flux. As an extreme example consider a mutation pattern which is very biased so the site is almost permanently fixed for allele 4. Clearly there can be little substitution. Removal of allele 4 from the system allows the other alleles to occupy the site alternately thereby increasing the rate of substitution at the site. Of course eliminating an allele from a system also removes several mutation pathways (e.g those between alleles 1 and 4). This explains why the conditions under which three allele systems evolve faster than four allele systems are rather less stringent than those for other comparisons.

It turns out that some of the relationships can be applied to mutation patterns in general, not just those in which $U_{12}=U_{21}$..etc, by using average mutation rates. Thus if we define

$$U = \frac{(U_{12}+U_{21}+U_{23}+U_{32}+U_{13}+U_{31})}{6} \qquad (6.15a)$$

$$U_{*4} = \frac{U_{14}+U_{24}+U_{34}}{3} \qquad (6.15b)$$

$$U_{4*} = \frac{U_{41}+U_{42}+U_{43}}{3} \qquad (6.15c)$$

the relationships 6.14a and 6.14c almost always hold. This was demonstrated by sampling mutation rates at random from a uniform distribution. In less than 2% of the 50000 mutation patterns sampled was either $R_3(*,*,*)>R_4$ and $U<3U_{4*}$, or $R_2(*,*)>R_3(*,*,4)$ and $U<4U_{4*}$. However in 20-30% of the cases where the mutation pattern inequality held (i.e $U>3U_{4*}$ or $U>4U_{4*}$) the rate inequalility did not hold.

A general expression for the case when a two allele system evolves faster than a four allele system eluded this investigator. Suffice it to say that the conditions under which a two allele system will evolve faster than a four allele system are likely to be rather more stringent than the conditions under which a three allele system evolves faster than a four allele system. This is because in reducing a four allele site to a two allele site one loses five of the six possible mutation pathways, compared to the three that are lost in the reduction to a three allele site.

6.2.3 Multiple sites

So far the analysis has only delt with single sites. Although it is evident that a sequence with fewer potential alleles can evolve faster than one with more potential neutral alleles the conditions under which this occurs are more restricted for a number of reasons. Firstly a reduction in the number of potential alleles may be brought about by an increase in the number of one allele sites, not a reduction, say of four allele sites to three allele sites. Secondly the inequalities 6.14a and 6.14c are partially exclusive. For a three allele site to evolve faster than a four allele site requires that

allele 4 is removed to form the three allele site. However a three allele site must contain allele 4 if a two allele site is to evolve faster than it. Finally different alleles may be removed from different sites: e.g at one three allele site allele 4 may be removed whereas at another allele 3 may have been eliminated. However it is worth appreciating that a sequence of say three allele sites, with alleles removed at random, can evolve faster than a sequence of four allele sites. To see this consider the rate at four and three allele sites under the symmetrical mutation pattern used above, when $U_{4*}=0$. The rate of evolution at three and four allele sites containing allele 4 is zero (6.13a, 6.13b), whereas it is non-zero at the three allele site missing allele 4 (6.13c). Therefore the average rate of evolution over a set of three allele sites is greater than that at a four allele site. Of course the mutation patterns must be very biased for this to occur; although if such patterns did exist they would overcome the partial exclusivity of inequalilities 6.14a and 6.14c.

6.2.4 Summary

Hence in both sequences and sites, a reduction in the number of potential neutral alleles can increase the rate of evolution. However this does depend on there being extreme bias in the mutation pattern. Although it is unclear what mutation patterns one might expect to find in the natural world, the very limited data available suggest that the patterns are not very biased (Gojobori *et al.* 1982, Li *et al.* 1984), and that therefore the rate of neutral evolution will generally be inversely related to the level of constraint. However the present findings suggest some caution should be exercised in deducing the action of positive natural selection when the rate of substitution at non-silent sites exceeds that at silent sites, especially when few sites are involved or there is extreme codon bias, codon bias being an indicator of biased mutation.

CHAPTER 7

ANALYSIS OF CODON USAGE PATTERNS

7.1 INTRODUCTION

In the preceding chapters we have considered a number of ways in which mutation could be responsible for the pattern of codon usage seen in mammalian genes, in each case finding evidence against the mutation hypothesis. We now turn our attention to selection, and rather perversely find our first evidence that mutation is responsible for some of the patterns of codon usage seen in mammals.

Although it seems unlikely to be totally responsible for the G+C content variation seen at mammalian silent sites, testing for tRNA interactions seems a good place to start. This is not only because it is a testable hypothesis, but because it also seems to dominate the codon usage of many other species (see chapter 1). The main reason tRNA interactions seem unlikely to be solely responsible for the codon usage of mammals is because genes of very high G+C content (>90%) are not expected, since this would require that all optimal codons are G or C ending in roughly equal quantities, or that a fair number of genes have very extreme amino acid composition.

However there are problems with detecting selection upon tRNA interaction in multicellular organisms by the methods used in unicellular organisms since it is not generally possible to identify the levels of tRNA and gene expression at the relevent times and places when selection is likely to be strongest. There are of course exceptions; Garel (1974) has shown that the silkworm tRNA concentrations are adapted to the amino acid composition of the silk, and Shields *et al.* (1988) have been able to show that the degree of codon bias seems to be related to gene expression in *D.melanogaster*. However Shields *et al.* (1988) were able to substantially strengthen their argument by demonstrating that silent site G+C contents are not correlated to intron and non-silent site G+C contents, an argument which is not available to the those investigating mammalian codon usage.

A solution is to look at codon usage within a gene. If there is no selection acting we would expect all amino acids, with say four codons, to have the same synonymous codon usage (at the third codon position). We would not expect this to be true if selection was acting upon tRNA interaction, or upon motifs essential for nucleic acid structure.

7.2 MATERIALS AND METHODS

7.2.1 Codon bias tests

If there is no selection acting at silent sites the third position of a codon will simply reflect mutational processes. Codons of similar degeneracy should therefore show similar third position nucleotide frequencies. However Bulmer (1986) and Blake *et al.* (1992) have shown that the rate of base mutation is dependent upon the adjacent nucleotides, so we expect the third position nucleotide frequencies to be influenced by both the second base in the codon, and the first base of the distal codon. Correction for this source of error at the second position can be achieved by restricting comparisons to those amino acids which have the same nucleotide at the second position. It is thus possible to compare the third position nucleotide frequencies of the following amino acids using a chi-squared independence test.

(i) The C test : Alanine, threonine, proline and the four-fold degenerate codons of serine (Ser4) are compared. These amino acids all have C at the second position and any base at the third. Note that the four-fold degenerate codons of serine are not connected to the two-fold serine codons by a single base substitution, so substitution between them is very rare. Since these amino acids all have C at the second position and therefore very low G frequencies at the third it was necessary to combine data. The frequencies of G and C were combined since this will not remove heterogeneity due to selection for intermediate codon/anti-codon binding strength (Grosjean and Fiers 1982, Gouy and Gautier 1982) which appears to act in yeast (Bulmer 1988).

(ii) The A^{G}_{A} test : Glutamic acid, lysine and glutamine were compared; these are two-fold degenerate codons with A at the second position, and purines (A and G) at the third.

(iii) The A^{T}_{C} test : Tyrosine, histidine, aspartic acid and asparagine were compared. These are two-fold degenerate codons with A at the second position and pyrimidines (T and C) at the third.

(iv) The G^{T}_{C} test : Cysteine and the two-fold degenerate codons of serine (Ser2) were compared. These are two-fold degenerate codons with G at the second position and pyrimidines at the third.

As an example consider the codon usage tables of Cys and Ser2 (G^{T}_{C} test) for two genes (Table 7.1). If we calculate a standard 2x2 chi-square independence test the χ^2 value equals 2.33 for human dystrophin and 9.91 for human androgen. This suggests that Ser2 and Cys are subject to the same processes (mutation, selection) in dystrophin but not in androgen.

TABLE 7.1

	3rd posn		3rd posn	
	T	C	T	C
Ser-2	55	40	4	30
Cys	15	20	13	14
	(a)		(b)	

Table 7.1 The number of times the codons of cysteine and serine-2 are used by human dystrophin (a) and human androgen receptor (b). Cys and Ser-2 show different codon usage patterns in androgen (χ^2=9.91 $p<0.005$) but not in dystrophin (χ^2=2.33 $p>0.10$).

7.2.2 Distal base effects

The potential bias introduced by the first base of the distal codon (henceforth known as the fourth base) is not expected to be as strong as the bias introduced by the second base. However some bias can be produced if certain amino acids tend to be found together, as one might reasonably expect; for instance hydrophilic amino acids are likely to be found together. The bias caused by the fourth base can be eliminated by randomly drawing codons from the sequence without replacement in such a way that each amino acid involved in a comparison, say an A^{G}_{A} test, has the same fourth base frequencies. After resampling, if selection is not acting and mutation biases do not extend beyond the adjacent base, each amino acid in a test should have the same third position nucleotide frequencies. Resampling the data theoretically leaves the statistic calculated on the resultant contingency table χ^2 distributed. However the reduction in sample size leads to many cells having small expectations which leads to departures from a χ^2 distribution. So as an approximation the average codon usage after such a resampling event is calculated (see table 7.2 for a worked example, and the appendix section 7.5). Intuitively one might expect this approximation to be a slight underestimate of the true chi-square value since calculating the average codon usage reduces the sample size without increasing the variance of the frequencies. To understand this consider a random binomial variate with a mean of 0.5. Let us imagine that we take a sample of 1000 and get the unusual result of 531; the probability of getting such a deviant result is 5%. If we now keep that proportion (0.531) but reduce the sample size, the probability of getting such a proportion increases. For instance the probability of getting 53 or more (or 47 and less) out of 100 is about 55%. This reduction in sample size without any systematic change in the frequency is what happens in the fourth base pair correction.

TABLE 7.2

		4th base pair			
		A	G	A	G
Cys	TGT	4	4	3.5	4
	TGC	7	1	6.1	1
	Total	11	5	9.6	5
	Freq	0.69	0.31	0.66	0.34
Ser2	AGT	11	21	11	16.1
	AGC	20	0	20	0
	Total	31	21	31	16.1
	Freq	0.60	0.40	0.66	0.34
		Uncorrected		Corrected	

Table 7.2 An example of correcting for 4th base pair frequencies. First of all the frequency with which each amino acid is followed by each base is calculated; e.g Ser2 is followed by A 60% of the time. The minimum frequency with which each 4th base follows any of the amino acids is found. For A this is 0.60, for G it is 0.31. By multiplying the total number of each amino acid by these minimum frequencies one obtains the maximum sample size one is allowed to draw of that amino acid followed by that fourth base. So we can use 0.60x16=9.6 Cys codons followed by A, and 0.31x16=5.0 Cys codons followed by G. On average, when 9.6 Ser2 codons followed by G are drawn from the sequence there will be 3.5 TGT codons and 6.1 TGC codons. By summing these corrected totals over all fourth bases one gets the expected number of each codon after resampling. In the case of TGT this would be 3.5+4=7.5. It is these numbers which are used in the χ^2 tests. Note that once the correction has been performed each amino acid has the same fourth base pair frequencies.

7.2.3 The data set

The data set is 62 large (>1800 bp) human, rat and mouse genes which were extracted from the GenBank and Embl databases using the GCG (Devereux *et al.* 1984) and ACNUC (Gouy *et al.* 1985) packages. Humans and rodent genes were used since they are the best represented mammalian groups in the sequence databases, and only genes with more than 600 codons are included in the analysis to avoid the expectations in the χ^2 tests becoming too small through insufficient sample size.

7.3 RESULTS

7.3.1 Testing codon bias

Table 7.3 shows the results of the four codon bias tests performed on the 62 mammalian sequences. In the analysis which follows human and rodent sequences will be analysed separately so that different trends in the two groups may be revealed and the analysis is not complicated by covariances introduced by genes represented in both groups. These covariances arise because genes represented in both samples are not independent since they share a relatively recent common ancestor.

It is clear that there are significant amounts of heterogeneity in the A^G_A comparison. Not only are there several significant values but the overall level of heterogeneity is inconsistent with similar codon usage in Gln, Lys and Glu: the pooled χ^2 value for humans is 232.99 ($p<0.005$) and for rodents it is 172.30 ($p<0.005$).

There are also several significant values in the G^T_C test. However it is not immediately clear that this is not simply due to taking 62 samples since the overall pooled χ^2's are not significant in any species. However since many of the G^T_C tests have very small expected frequencies and the fourth base pair correction probably biases the estimates downwards, the chi-square values calculated are not strictly χ^2 distributed. (Note that one cannot simply eliminate the tests with very small expectations since this would constitute biased sampling). As a result the pattern of bias in the G^T_C test is investigated further.

The overall pooled χ^2 for the rat C tests is significant (χ^2=128.86

TABLE 7.3 The codon bias tests

Gene	C	A^{G}_{A}	A^{T}_{C}	G^{T}_{C}
$df^{X}=$	6	2	3	1
Human				
Dystrophin	3.41	16.11**	2.22	1.46
Androgen receptor	5.43	8.02*	2.72†	8.82**
Complement C5	7.96	10.86**	2.88	1.80
c abl	4.05	3.45	1.68	0.44
α-2 macroglobulin	3.85	5.82	0.41	0.62
Epidermal GF receptor	3.41	0.08	2.86	1.83
Glucocorticoid receptor	2.71	4.08	1.76†	1.14
Ca^{2+} ATPase	4.71	6.41*	1.52	0.06
Enkephalinase	1.12	2.81	2.77†	1.14†
Complement C3	4.56	2.70	0.77	0.03
Complement C4	6.86	2.28	1.07	0.01
Na^{+} K^{+} ATPase	2.23	5.89	0.59	0.47
Laminin beta 2	10.02	16.49**	0.77	2.80
Angioleusin-I converting enzyme	3.51	9.02*	2.98	0.15†
Ceruloplasmin	4.39	0.49	1.89	0.11†
Glycoprotein p150,95	11.87	1.90	1.24	1.92†
Breakpoint cluster gene	7.30	5.79	0.96	0.50†
Apolipoprotien B-100	7.72	11.53**	3.60	0.13
LDL receptor related protein	7.81	6.41*	0.81	0.07
Coagulation factor VIII	2.89	4.17	1.46	4.10*
Von Willebrand factor	4.28	3.18	5.58	0.63
Growth factor II receptor	5.34	11.15**	1.66	3.83
Insulin receptor precursor	9.34	3.18	3.37	0.98
Poly (ADP-ribose) polymerase	3.89	16.72**	1.30	0.75†
Fibronectin	5.25	8.02*	2.80	0.68
Embryonic myosin heavy chain	4.92	2.18	4.50	1.23
CCG-1	6.09	13.01**	5.00	0.40
Coagulation factor V	10.42	19.28**	0.76	0.89
Laminin beta 1	3.63	13.54**	0.95	4.73*
GL-1 protein	6.47	4.33	0.41	0.12

TABLE 7.3 cont/d

Leukocyte adhesion glycoprotein	2.13	14.09**	0.06	1.50
Total	167.57	232.99**	61.35	43.34
Mouse				
Dystrophin	5.76	8.05*	1.47	0.93
Complement C5	1.59	3.04	2.09	3.89*
c abl	4.66	2.45	0.59	1.18
Glucocorticoid receptor	3.85	5.62	2.91	0.02
Epidermal growth factor receptor	4.75	3.26	0.27	4.80*
Complement C3	5.34	6.43*	3.09	0.50
Complement C4	7.75	5.72	0.39	0.47
Laminin beta 2	2.62	2.17	4.50	4.46*
Insulin receptor precursor	6.38	0.42	4.73	0.61
Poly (ADP-ribose) polymerase	7.64	5.18	0.87	0.00†
Laminin beta 1	8.78	16.92**	2.66	1.07
Leukocyte adhesion glycoprotein	6.20	11.70**	4.73	0.62
Angioleusin-I converting enzyme	8.47	2.58	0.42	0.33†
Microtubule associated protein 2	11.70	4.55	2.62	0.92†
Multi-drug resistance protein	4.99	8.14*	1.68	1.15†
Total	90.48	86.23**	33.02	20.95
Rat				
Androgen receptor	7.34	5.19	0.36	1.61
α-2 macroglobulin	6.01	0.70	3.67	1.07
Ca^{2+} ATPase	9.61	6.10*	0.27†	0.31
Enkephalinase	2.50	2.34	1.86†	0.97†
Na^{+} K^{+} ATPase	7.61	10.54**	2.85	0.02†
Plasma protein inhibitor	5.64	2.45	1.48	0.80
Clathrin	10.02	13.53**	0.23	0.66
Sodium channel III	7.64	0.34	2.94	0.06
Acetyl coA carboxylase	19.66**	23.39**	3.96	0.20
Embryonic myosin heavy chain	6.87	0.03	7.15	0.00†
Fatty acid synthetase	12.04	6.77*	5.54	0.00

TABLE 7.3 cont/d

Proteoglycan core protein	13.01*	4.44	0.33	0.10
Phospholipase C-I	2.77	3.27	5.46	3.55
Guanylate cyclase 70KD subunit	6.14†	0.90	1.72	2.49
neu oncogene	4.74	3.23	0.19	0.00
Neurofilament protein	3.41	2.85	4.10†	0.29†
Total	125.01*	86.07**	42.11	12.13

Table 7.3 The χ^2 values for the four tests of codon bias.

* significant codon bias at the 5% level

** significant codon bias at the 1% level

x degrees of freedom associated with the test.

† more than 20% of expected values in chisquare test have a value of less than 5; statistic is unlikely to be χ^2 distributed.

$p<0.05$) suggesting that the two significant values for acetyl coA carboxylase and proteoglycan core protein are not the consequence of taking many samples. However there is no evidence of heterogeneity in codon bias in the human and mice C tests, or the A^T_C tests of any of the species studied.

7.3.2 Patterns of codon bias in the A^G_A and G^T_C tests

If we examine the pattern of codon bias in those genes which show significant heterogeneity in the A^G_A and G^T_C tests it becomes clear that most genes show very similar patterns of codon usage (table 7.4). In the A^G_A test Glu always shows a higher frequency of A ending codons than Gln, with Lys generally mid-way between the two; and in the G^T_C test Cys generally shows a higher frequency of T ending codons than Ser2. Not surprisingly, given the consistency of the trends in the 'significant' genes, these patterns of codon bias are found in most genes in the sample (table 7.5).

7.3.3 Bias in other reading frames and on the complementary strand

It is of great interest to see whether these patterns of bias exist in other reading frames and on the anti-sense, or complementary, strand (see discussion). In what follows the 123 frame refers to the original sequence and the 312 frame as the sequence read starting at the third base pair. The complementary sequence was formed so that the third position of a codon in the 123 frame was the third position on the complementary strand. This is equivalent to forming the complementary strand and reading it 5' to 3' from the second base pair.

Table 7.5 shows the number of genes which show a particular pattern of codon bias in the 312 reading frame and on the complementary strand. Most genes in the sample show the pattern of bias found in the 123 frame on both the complementary strand and in the 312 reading frame. However the number of genes showing the pattern of bias in the 312 frame is always less than the number of genes showing the pattern of bias in the 123 frame and on the complementary strand, often significantly so.

The number of genes showing the pattern of bias on the complementary strand is also usually less than that in the 123 frame. This difference is significant for two human relationships: for

TABLE 7.4

Part(a)

Gene	Gln	Lys	Glu
Human			
Dystrophin	0.45	0.52	0.61
Androgen receptor	0.17	0.42	0.45
Complement C5	0.42	0.70	0.71
Angioleusin-I converting enzyme	0.07	0.15	0.27
Ca^{2+} ATPase	0.47	0.67	0.71
Laminin beta 2	0.23	0.38	0.54
Apolipoprotein B-100	0.43	0.58	0.55
LDL receptor related protein	0.08	0.17	0.16
Growth factor II receptor	0.20	0.42	0.43
Poly (ADP-ribose) polymerase	0.03	0.28	0.48
Fibronectin	0.30	0.36	0.48
CCG-1	0.26	0.49	0.55
Coagulation factor V	0.33	0.49	0.67
Laminin beta 1	0.34	0.60	0.59
Leukocyte adhesion glycoprotien	0.41	0.63	0.73
Mouse			
Dystrophin	0.50	0.69	0.68
Complement C3	0.29	0.25	0.41
Laminin beta 1	0.22	0.53	0.48
Leukocyte adhesion glycoprotein	0.34	0.58	0.65
Multi-drug resistance protein	0.34	0.45	0.61
Rat			
Clathrin	0.28	0.52	0.55
Ca^{2+} ATPase	0.34	0.49	0.59
Na^{+} K^{+} ATPase	0.20	0.19	0.46
Acetyl coA carboxylase	0.22	0.40	0.59
Fatty acid synthetase	0.16	0.23	0.31

TABLE 7.4 cont/d

Part (b)

Gene	Species	Cys	Ser2
Androgen receptor	Human	0.47	0.12
Coagaulation factor VIII	Human	0.31	0.56
Laminin Beta 1	Human	0.49	0.31
Complement C5	Mouse	0.62	0.35
Epidermal Growth factor receptor	Mouse	0.49	0.27
Laminin Beta 2	Mouse	0.50	0.30

Table 7.4 The codon bias in those genes which show significant heterogeneity. Codon bias is calculated on data corrected for 4th base pair frequencies. In part (a) the frequency of A at the 3rd position of Gln, Lys and Glu is shown. In part (b) the frequency of T in the third position of Cys and Ser2 is shown.

TABLE 7.5

Relationship	Frame	Human	Rodent
	123	31**	30**
Gln<Glu	Comp	26**	29**
	312	21*	23**
	Attenuation	11.92**	6.37*†
	123	20	25**
Lys<Glu	Comp	23**	23**
	312	14	22*
	Attenuation	2.34	0.79
	123	31**	25**
Gln<Lys	Comp	22*	25**
	312	20	18
	Attenuation	13.37**	3.72
	123	20	24**
Ser2<Cys	Comp	17	21*
	312	15	16
	Attenuation	1.64	4.51*

Table 7.5 The number of genes for which the relationship given in the first column holds, in the various reading frames and orientations of the sequence. The relationships refer to the frequency of A in the 3rd position for Gln, Lys and Glu, and the frequency of T in the 3rd position of Cys and Ser2. All figures are out of a total of 31 genes. The attenuation row gives the χ^2 value calculated by comparing the number of genes showing the relationship in the 123 and 312 frames.

* Significantly different from 15.5 (i.e 31/2) at the 5% level

** Significantly different from 15.5 (i.e 31/2) at the 1% level

† 2 out of 4 cells had expected frequencies of less than 5. Statistic is not χ^2 distributed.

Gln<Glu p=0.026, and for Gln<Lys p=0.001 as measured by Fisher's exact test. This difference may be due to the fact that the third position on the complementary strand is in reality a mixture of two-fold and four-fold sites in the proper 123 frame. This will tend to affect the nucleotide frequencies observed.

The G^T_C results warrant further comment. The χ^2 tests show that several genes have significant codon bias with respect to Cys and Ser2 but it remains unclear as to whether these results are the consequence of sampling many genes. Table 7.5 shows that in rodents a significant number of genes show the same pattern of codon bias suggesting that the bias detected in the χ^2 tests is not a statistical 'artifact'. There is also no evidence that the strength of the bias differs between rats and mice since the number of genes showing the pattern of bias is very similar: 12 out of 15 mouse genes show the frequency of T ending codons to be greater in Cys than Ser2, 11 out of 16 genes in rats. However in humans the situation remains very unclear as to whether there is any overall bias between Cys and Ser2 since only 20 out of 31 genes (p>0.10) show Cys to have a higher frequency of T ending codons than Ser2.

7.3.4 Pattern of codon bias in the Rat C test

Table 7.6 shows the codon usage tables for rat acetyl coA carboxylase and proteoglycan core protein. In both cases Ser4 has an odd codon usage pattern, but it is difficult to identify any other major similarities. The vast majority of the χ^2 is generated by the A ending codons of Ser4, and the T and A ending codons of Thr in acetyl coA carboxylase. In proteoglycan it is the T ending Ser4 codons which contribute most to the χ^2 value, with some help from the A ending codon.

Because no major trend in codon bias can be identified the χ^2 tests were performed on the 312 frame and complementary (Comp) sequence. The χ^2 values for acetyl coA carboxylase are 14.73 (Comp, P<0.025) and 8.61 (312, not significant); for proteoglycan they are 21.78 (Comp, P<0.005) and 11.11 (312, not significant). The pooled χ^2 are 140.17 (Comp, df=96 p<0.005) and 100.09 (312, df=96 not significant). Thus it would seem that significant codon biases do exist on the complementary sequence, but are attenuated or non-existent in the 312 reading frame.

TABLE 7.6

	Acetyl coA carboxylase			Proteogylcan Core protein		
	T	C+G	A	T	C+G	A
Ser4	0.42	0.43	0.15	0.50	0.31	0.19
Pro	0.37	0.30	0.33	0.34	0.34	0.32
Thr	0.24	0.36	0.40	0.30	0.40	0.30
Ala	0.41	0.31	0.27	0.33	0.43	0.25

Table 7.6 Codon usage in the rat acetyl coA carboxylase, and rat proteoglycan core protein C tests. Figures show the nucleotide frequencies in the 3rd position of Ser4, Pro, Thr and Ala. Frequencies were calculated from data corrected for 4th base pair frequencies.

7.4 DISCUSSION

There are several ways in which the codon bias could have been generated: selection, neighbouring base mutational effects extending beyond the adjacent base, non-silent substitution and trends in G+C content along the gene.

7.4.1 Non-silent substitution

Non-silent substitution can generate heterogeneity in codon bias by interconverting amino acids with different codon usage patterns; patterns which could be the consequence of dinucleotide effects. However there are good examples of genes with high levels of codon bias heterogeneity but low levels of non-synonymous substitution. For instance the human Ca^{2+} ATPase gene has not gained or lost any Gln, Lys or Glu codons since it diverged from rodents, has a high rate of silent substitution and shows significant codon bias heterogeneity in the A^{G}_{A} test. Furthermore the presence of bias on the complementary

strand suggests that amino acid substitution cannot be responsible for the biases observed.

7.4.2 Trends in G+C content

Many genes in the sample showed trends in third position G+C content along their length. The basis of these trends is unknown, but coupled to a very non-uniform amino acid distribution they could generate codon bias. For instance, the human epidermal growth factor receptor gene shows a very sharp 30% increase in G+C content three quarters of the way along its length. If cysteine tended to cluster at one end of the gene and serine at the other, heterogeneity in codon usage would be produced. There are however examples of genes which show no trend in G+C content but show significant codon bias heterogeneity: e.g human androgen receptor, mouse complement C3, rat clathrin and human angiotensin-I converting enzyme.

7.4.3 Selection

It is difficult to think of any form of selection that would lead to significant bias appearing on the complementary strand and in other reading frames, but which is attenuated in the 312 reading frame. Selection during translation, to match the commonest tRNA's for instance, would only give bias in the 123 frame; selection upon motifs important for mRNA secondary structure would give bias in all reading frames but not on the complementary sequence; and selection on DNA structure would give equally strong bias in all reading frames and the complementary sequence.

7.4.4 Mutation bias

In both *E.coli* and *S.cerevisiae* it has been observed that the mutation pattern is influenced by bases several positions upstream and downstream of the site in question (Bulmer 1990). The biases detected here could be a consequence of such neighbouring base effects, and this view is supported by the attenuation of the bias in the 312 reading frame. Let us imagine that A is disfavoured by mutation when preceded by CA but not when preceded by GA. CAN codons will have a low frequency of A at the third position compared to GAN codons if there are no C$\leftrightarrow$G substitutions at the first position. However as the rate of C$\leftrightarrow$G substitutions increases so the frequency

of A in the third position of CAN and GAN codons will become more similar. In the 123 frame the first two positions in the codon are highly constrained so CAN and GAN codons differ in their synonymous codon usage. However in the 312 frame, where the first codon position is in reality a third position, and thus subject to frequent substitution, the synonymous codon usage of CAN and GAN codons will be very similar.

The reason for the attenuation of the bias in the 312 frame has important implications for the testing of the mutation hypothesis using non-coding DNA. Whenever the two 5' sites change at a comparable, or greater, rate than the site in question the pattern of bias will be attenuated. Thus we would not expect to see strong biases in non-coding DNA.

7.4.5 Summary

Mammalian genes show patterns of codon usage which are superficially consistent with the action of selection at silent sites. However these patterns also appear on the complementary strand and in other reading frames suggesting that the patterns detected are caused by bases two positions removed affecting the pattern of mutation at the site in question.

7.5 APPENDIX

The aim of the fourth base pair correction is to ensure that each amino acid in the contingency table in a test (e.g C^{T}_{C} test) has the same fourth base pair frequencies. This can be achieved by randomly sampling a subset of the codons from a gene without replacement. To do this we need to know the maximum sample size we can take of each amino acid followed by each fourth base pair. For instance, imagine that we are doing a G^{T}_{C} test and that Cys is followed by A far less often than Ser2 is. We will be able to use all the Cys codons followed by A in the contingency table but only some of the Ser2 codons which are followed by A.

Imagine we have a test involving p amino acids each with q synonymous codons. Let C_{ijk} be the number of amino acid i, with synonymous codon j and fourth base k in the sequence. The frequency

with which amino acid i is followed by fourth base k is

$$B_{ik} = \frac{\sum_{j=1}^{q} C_{ijk}}{\sum_{j=1}^{q} \sum_{k=1}^{4} C_{ijk}} \qquad (7.1)$$

and the minimum frequency with which any of the p amino acids is followed by the kth 4th base is:

$$M_k = \min_{i=1}^{p} B_{ik} \qquad (7.2)$$

So the maximum sample size we take of amino acid i followed by fourth base pair k so the fourth base pair frequencies for all amino acids are the same is

$$N_{ik} = M_k \sum_{j=1}^{q} \sum_{k=1}^{4} C_{ijk} \qquad (7.3)$$

One could use these sample sizes to actually perform a resampling event. However because such a procedure often leads to very small expected frequencies in the chi-square calulation the expected values from such a resampling event were calculated instead. So the expected number of amino acid i with synonymous codon j after such a resampling event is

$$E_{ij} = \sum_{k=1}^{4} N_{ik} \frac{C_{ijk}}{\sum_{j=1}^{q} C_{ijk}}$$

$$= \sum_{k=1}^{4} \frac{M_k \, C_{ijk}}{B_{ik}} \qquad (7.4)$$

These are the expected numbers used in the χ^2 tests.

CHAPTER 8

GENERAL DISCUSSION

8.1 Mutation biases

The central aim of this thesis has been to gain some understanding of synonymous codon use in mammals. I believe some progress has been made on this front, for although we still have no positive evidence of selection it seems that we can probably exclude mutation as the sole determinant of codon use in mammals. The evidence for this comes from two principle sources. Firstly the differences in silent site, intron and isochore G+C contents are not predicted *a priori* by mutation models although one can sometimes think up *ad hoc* explanations for them; in particular the difference between intron and isochore G+C content is very difficult to explain under replication and recombination hypotheses. And secondly mutation models appear to be unable to explain the relationships observed between the substitution rate and G+C content at silent sites (chapters 2, 4, 5 and 6). There are also other lines of evidence which suggest that particular sources of mutational bias are not responsible for the silent site G+c content variance: (1) the lack of any distinction between the G+C contents of early and late replicating genes seems to suggest that replication has little to do with the differences in codon use between genes (chapter 3); and (2) the parameter space over which DNA repair can generate a large enough range of G+C contents to explain the data is very small (chapter 4).

Despite this evidence it is still not possible to totally exclude mutation as either a minor or major determinant of synonymous codon use in mammals since there may be other pathways by which mutations are formed which we have not yet considered, and it seems difficult to exclude combinations of any of the hypotheses so far examined. Furthermore great weight has been attached to the substitution rate/G+C content relationship at silent sites, although it is evident that there is a G+C content bias associated with the substitution rate measurement. However this latter problem can be tackled since we know the distributions of both p, the proportion of

sites which differ between two sequences, and b, the maximum level of (uncorrected) divergence (both binomial)(see equations 5.1 and 5.2).

Of course the argument against mutation would be that much stronger with further independent evidence against it. Unfortunately there appear to be few other ways to approach the problem. Two possible strategies are (1) to examine the substitution patterns of introns and intergenic DNA and (2) to investigate the frequency of dinucleotide mutations. The first strategy would allow an assessment of whether one can attribute the differences between silent site, intron and isochore DNA to the fact that the introns and intergenic DNA should be able to fix insertions, deletions and dinucleotide mutations which would generally be deleterious at silent sites due to their effect on the surrounding non-silent sites.

The second investigation would reveal whether dinucleotide mutation is a reasonable explanation of the association between the silent and non-silent substitution (Fitch 1980, Lipman and Wilbur 1985); the alternative explanation being selection. At face value the simultaneous mutation of two nucleotides seems an unlikely explanation of the association between the substitution rates since Lipman and Wilbur (1985) found that the silent substitution rate was 1.6 times higher in codons which had undergone non-silent substitution than in other codons, which suggests that the rate of dinucleotide mutation is of the same order of magnitude as the rate of mononucleotide mutation; i.e if we assume that non-silent substitutions do not cluster in a sequence then this increase under the dinucleotide mutation hypothesis is entirely attributable to substitution at the second codon position. So assuming, conservatively, that the rate of substitution is the same in the first two codon positions the rate of silent substitution in those codons which have undergone substitution at the second position is 2.2 times that in other codons.

Although this is a rough estimate it does suggest that the rates of dinucleotide and mononucleotide mutation must be comparable. Not only does this seem surprising but it does not fit with the observation that silent site substitution rates are very similar to those in pseudogenes (Miyata and Hayashida 1981, Wolfe *et al.* 1989, Li and Graur 1991 chapter 4) since we would expect the silent substitution rate to be partially constrained by the selection on

non-silent sites. However the subject does need more investigation; Lipman and Wilbur (1985) probably underestimated the relative rate of non-silent substitution and the pseudogene evidence is hardly strong when the possibility of intra-genomic mutation rate variance exists. A study of pseudogene substitution should allow one to estimate the relative rates of mono and dinucleotide mutation.

8.2 Selection

Although it would be useful to test the mutation hypothesis further, the most profitable line may be to start testing selection hypotheses. To do this it seems likely that we will have to abandon the substitution rate/G+C content relationship since selection has many complications; for instance models will behave differently if selection is stabilising or directional, or if there is negative or positive epistasis. Instead it seems that we will have to test for particular features of selection as we did in chapter 7.

Selection on mRNA secondary strucure would seem to be capable of explaining the differences in silent site G+C content between mammalian genes: the high (or very low) G+C content genes are those which require a stable secondary structure, while those of intermediate G+C content are those for which selection is not as strong, or for which a reduction in secondary structure is important. A general strategy for testing for selection upon secondary structure is suggested by the work of Huynen *et al.* (1992). They compared the minimum free energy of several histone mRNA's to that in sequences with the third codon positions randomly scrambled (i.e the third position composition was the same but the order of nucleotides was different). This procedure is not entirely statistically sound, for testing selection at silent sites at least, because such reshuffling of the third position changes the amino acid sequence; e.g glutamic acid (GAA, GAG) may become aspartic acid (GAT, GAC). This problem is very easily overcome by resampling the codon usage table of the gene without replacement such that the amino acid sequence of the gene is preserved but the order of the synonymous codons is altered. By forming many such sequences and measuring whatever statistic one is interested in, say minimum free energy, one can build up the expected distribution of the statistic under the hypothesis that selection is not acting upon secondary structure at silent sites. The trick is of

course to find the right statistic since the character upon which selection might act is none too obvious. For instance it might be acting to optimise translation by eliminating secondary structure, and yet again it might be acting to decrease the elongation rate to regulate gene expression or help protein folding (e.g see Zama 1989).

If selection is acting to increase secondary structure as one might suppose in genes of very high G+C content then the prediction of Huynen *et al.* (1992), that the G/(G+C) ratio should be about 50%, should also hold. It would therefore be interesting to see if the the G/(G+C) ratio is correlated to the silent site G+C content.

8.3 Isochores

Besides the variation in G+C content between the silent sites of different mammalian genes there is also the variation in G+C between introns and isochores to be explained. Furthermore the correlation between all three is yet to be understood.

The mechanism by which isochores of different G+C content are maintained is the subject of keen debate. Bernardi and colleagues (Bernardi and Bernardi 1986, Bernardi *et al.* 1988) have emphasised the role of selection, whereas Filipski (1987) and a group at Trinity College Dublin (Wolfe *et al.* 1989, Wolfe 1991) have attributed the differences to mutation biases. The two views are not mutually exclusive since it is possible for selection to act upon the factors which influence the mutation pattern, like free nucleotide concentrations and chromatin structure. However some doubt must be expressed as to whether selection would ever be strong enough to modify the mutation pattern locally, selection being as it is only as strong as the mutation rate.

The mutation hypotheses so far proposed to explain the differences in G+C content between isochores have been replication (Wolfe *et al.* 1989) and repair (Filipski 1987), although there seems no reason to exclude CpG effects. The lack of any great distinction between the G+C contents of early and late replicating genes, and therefore isochores, suggests that DNA replication is not responsible for the variation in isochore G+C content; although one might be able to attribute a certain fraction of the variation to replication since there is a change in the G+C content of newly replicated DNA through S phase (Tobia *et al.* 1970, Flamm *et al.* 1971, Comings 1971,

Hutchison and Gartler 1973, Holmquist *et al.* 1982). The results of chapter 3 also suggest that repair is not responsible for the variation in isochore G+C content since housekeeping and tissue-specific genes do not differ in silent site G+C content. This argument is not strong since it has never been demonstrated that housekeeping and tissue-specific genes differ in the efficiency with which they are repaired, one might simply suppose that this is the case since housekeeping genes are expressed in the germ line. Furthermore there is no compelling evidence to suggest that the relationship between silent site and isochore G+C contents is the same for housekeeping and tissue-specific genes. Thus variation in the efficiency of repair would still seem to be a viable explanation of the differences in G+C content between isochores.

However variation in the frequency of recombination seems to explain more than other hypotheses. For a start it is compatible with the results of chapter 3; one would not expect early, late, housekeeping or tissue-specific genes to be subject to different levels of gene conversion unless germ line selection is strong. If germ line selection is powerful compared to selection at the individual level then recombination might be favoured around genes expressed in the germ line (i.e housekeeping genes). Secondly the recombination hypothesis can explain the differences in gene density between G+C rich and G+C poor isochores. Thirdly chiasmata density differences between chromosomes appear to be of the right order of magnitude to explain the differences in G+C content between isochores. In figure 8.1 the equilibrium G+C content for a two allele system is plotted against the strength of gene conversion bias ($N\omega$) under various mutation patterns using equation 6.5. This should give a rough idea of the difference in bias required to generate the variation in G+C content between isochores, although a full analysis of a model including both transitions and transversions would be preferable (equation 6.9). It can be seen from figure 8.1 that as the mutation pattern becomes less biased ($U_{TC} \rightarrow 0.5$) so the variation in gene conversion required to produce a certain variation in G+C content increases. For instance when $U_{TC}=0.3$ a two fold variation in gene conversion frequency is required to generate a range of G+C contents from 42 to 48%. If $U_{CT}=0.4$ a five fold range is required.

In table 8.1 the meiotic chiasmata density and the mean silent

TABLE 8.1

Chromosome number	Chiasmata density	G+C Content	
		Silent site	Isochore
19	2.03	75.0	48.5
17	2.07	74.6	48.4
6	1.29	65.8	46.4
11	1.54	65.8	46.4
7	1.48	60.0	45.0
12	1.53	58.4	44.7
5	1.35	58.0	44.6
1	1.00	56.2	44.2
4	1.14	49.0	42.5

Table 8.1 The meiotic chiasmata density, average silent site G+C content and predicted isochore G+C content of 9 human chromosomes (those with greater than 30 sequenced genes). Meiotic chismata density information was calculated from data in Fang and Jagiello (1988)(chiasmata frequencies) and Kuhn (1976)(chromosome lengths). The average silent site G+C content values come from Ikemura and Wada (1991). The predicted isochore G+C contents were calculated using equation 3.1.

site G+C content are shown for nine chromosomes, along with the average isochore G+C content calculated using the regression equation 3.1 from the data of Aissani *et al.* (1991). Average isochore G+C contents range from 42.52% to 48.48% with chiasmata densities from 1.00 to 2.07. These results agree incredibly well with the theory: i.e the variation in chiasmata density is less than five fold, but not so small that the mutation pattern has to be very biased.

The final evidence which might be taken as supporting the case of recombination is the estimate of heteroduplex length per recombination event that one can make. From figure 8.1 it is possible to get an estimate of the average length of heteroduplex per recombination event which will allow gene conversion to affect the

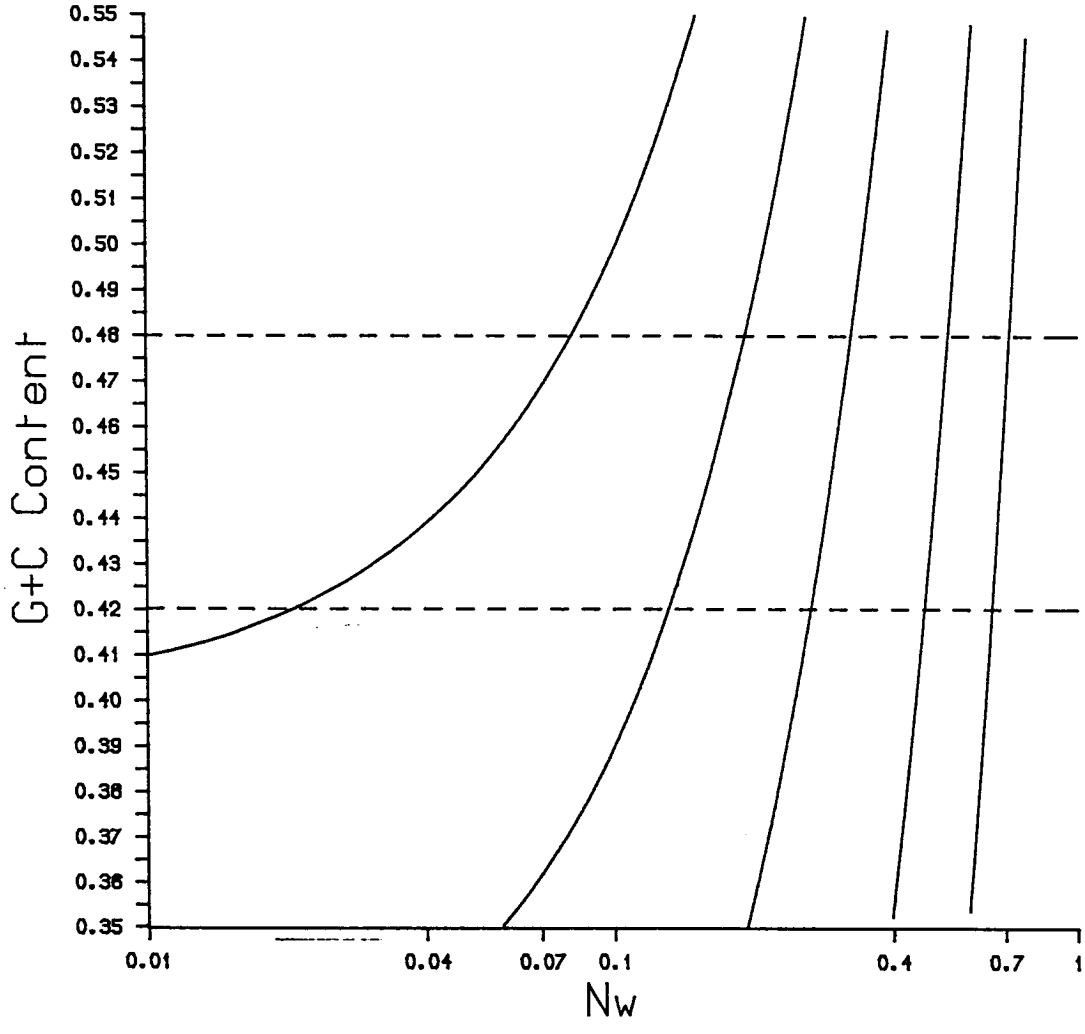


Figure 8.1 Sequence G+C content plotted against the level of gene conversion (/population size) for different mutation biases. The two dotted lines refer to the estimated average G+C content on chromosomes 19 (upper) and 4 (lower). The curves from left to right are for U_{TC}=0.40, 0.30, 0.20, 0.10 and 0.05 corresponding to sequences subject to no conversion of 40%, 30%, 20%, 10% and 5%.

G+C content of the mammalian genome. Let the average length of heteroduplex be L. Then the average amount of DNA involved in heteroduplex per meiosis is 50L since there are on average ~50 chiasmata per human diplotene spermatocyte (Fang and Jagiello 1988). The human genome is about 3.4×10^9 base pairs long so the average probability of a site being incorporated in heteroduplex (τ) is 1.47×10^{-8} L and we can estimate the value of bias in repair (ψ) as ~0.40 from Brown and Jiricny (1988) giving an overall gene conversion bias of $\omega = 5.88 \times 10^{-9} L$. The average G+C content of the human genome is ~43% which would require $N\omega$ to be about 0.2 if the observed chiasmata density and average G+C content data are to be consistent with the recombination hypothesis (see above). Thus if $N\omega = 0.2$ $L = 3.40 \times 10^7 / N$. One can only guess at the long term effective population size of the typical mammal: 10^4 or 10^5? These estimates would give the length of heteroduplex per recombination event as between 3400 and 340. This sounds reasonable and agrees well with the observation that co-conversion events can occur at least 350bp apart during mitotic recombination in mammalian cell lines (Liskay and Stachelek 1986).

8.4 Silent sites and isochores

Selection on mRNA secondary structure, and isochore maintainence via gene conversion are two interesting and hopeful working hypotheses for the future. However there remains a puzzle: if one regresses silent site G+C content upon isochore G+C content using the data of Aissani *et al.* (1991) one finds that the slope of the linear relationship is significantly greater than one (slope=2.77±0.483 with 19 degrees of freedom, $p<0.01$). This implies that a simple additive model relating isochore and silent site G+C content is insufficient; i.e we cannot split the variance at silent sites into two independent pieces, the two sources of variance must be related. For example we might think in terms of isochores being maintained by mutational biases, with the rest of the silent site G+C content variance being due to selection. The slope of least squares line suggests that the strength of selection must be related to the mutational bias (unless the two systems interact in a highly multiplicative manner). At present there is no obvious reason why such a relationship should exist.

8.4 Implications

In their 1987 paper on the molecular clock Li *et al.* noted that rodents had accumulated about 2.6 times as many silent substitutions as humans, whereas they had only accumulated 1.9 times as many non-silent substitutions. This is unexpected under a 'strict' neutral model since one would expect the ratios to be equal. (A model of 'strict' neutrality is one in which selection is either absent, or strong and purifying). If one assumes that only silent sites are strictly neutral one is led to infer that there is weak selection acting upon some of the alleles segregating at non-silent sites; alleles which in a group like primates with small long term population sizes are effectively neutral and therefore fixable. (Of course there are other explanations; for instance evolution at non-silent sites may not be mutation limited). The silent sites play a vital role in this argument since they allow one to control for differences in the mutation rate between the two groups being considered, assuming they are strictly neutral. One can also test neutrality in a very similar fashion by comparing silent and non-silent substitution rates between autosomes and sex chromosomes, since the X and Y chromosomes have different dynamics to the autosomes because of their hemizygosity in the heterogametic sex (Avery 1984, Charlesworth *et al.* 1987). Once again one has to use the silent sites to control for differences in mutation rate that might exist between different loci, and between different types of chromosome.

There are two other tests of neutrality which use comparisons between silent and non-silent sites. Probably the most popular test of neutrality, and by far the most successful, is the 'K_s/K_a' test in which an excess of non-silent substitution over silent substitution is taken to imply the action of positive natural selection, either driving alleles through the population (e.g Hill and Hastie 1987, Yokoyama and Yokoyama 1990), or maintaining genetic variation (e.g Hughes and Nei 1988, 1989, Tanaka and Nei 1989, Hughes 1991). The other, potentially more powerful test was suggested very recently by MacDonald and Kreitman (1991); under strict neutrality the ratio of K_s, the silent substitution rate, to K_a, the non-silent substitution rate, should be the same for both intra-specific and inter-specific comparisons. If there is an excess of non-silent substitution

intra-specifically then one might infer that either there are several advantageous alleles being maintained, or that slightly deleterious alleles segregate. If there is an inter-specific excess of non-silent substitution then the inference would be that advantageous alleles are being fixed at non-silent sites. Both tests appear to implicitly assume that there is no selection acting upon silent sites; certainly they are much easier to justify by making such an assumption.

So what happens to these tests if we relax the strict neutrality assumption at silent sites? In particular can one still infer departures from strict neutrality if the tests show significant results? To answer this we need only consider the situation in which strong purifying selection acts upon silent sites; any failures when weak or strong positive selection are acting will be correctly attributed to departures from strict neutrality. The crucial point to appreciate in assessing the impact of selection upon silent sites is the fact that whatever selection acts at silent sites is likely to act at non-silent sites as well (Dr M.Kreitman, University of Chicago, pers comm); for instance if selection is acting upon mRNA secondary structure at silent sites it will also affect the fate of alleles segregating at non-silent sites which do not change the protein function. Hence strong purifying selection at silent sites will reduce the number of potential neutral alleles at silent and non-silent sites by the same proportion. Such selection will therefore have no effect on the four tests described above. This is good news - the tests will tend to detect departures from strict neutrality whether or not selection acts at silent sites.

However the possibility of weak or strong positive selection acting at silent sites will complicate the interpretation of the tests. In particular if there is both weak selection upon the character which silent sites affect (e.g mRNA secondary structure) and weak selection upon protein sequence at certain sites, then the distribution of allelic effects will be different at silent and non-silent sites. This will mean that the results of all four tests are difficult to interpret. More work needs to be done on this problem.

APPENDIX

The maintainence of sexual reproduction is the subject of considerable controversy since there are many advantages and disadvantages to sexual reproduction (see Maynard Smith 1978, Michod and Levin 1988). The disadvantage usually associated with sexual reproduction is the two fold reproductive advantage that all-female populations have over their (anisogamous) sexual counterparts. Furthermore the processes of meiosis and syngamy (union of gametes) can be costly (Crow 1988), especially if one includes various forms of sexual selection in these calculations, and sex can break up coadaptation between alleles, either within a locus (overdominance) or between loci (epistasis).

There are also advantages to sex: it reduces interference between advantageous alleles (Fisher 1930, Muller 1932), generates genetic variation (Muller 1932), reduces the mutation load if there is synergistic epistasis (Kondrashov 1982, 1988, Charlesworth 1990) and stops the accumulation of genetic damage (Muller 1964). All these advantages are associated with recombination rather than segregation. There is some advantage in segregation alone in reducing the genetic load but this is small and likely to be insufficient in itself to overcome the two-fold advantage of asexuality (Charlesworth 1990).

However Kirkpatrick and Jenkins (1989) noted that in a diploid pathenogenic population an advantageous allele could only ever go to fixation if the mutation occured twice, once on each chromosome in a pair. Furthermore the second mutation had to occur in an individual carrying a chromosome with the first mutation. These are not problems for a population undergoing sexual reproduction since individuals homozygous for the mutant allele can easily be formed due to segregation. The delay between the first and second mutations exerts a load upon the asexual population since it cannot adapt as rapidly to the environment as a sexual population. Sexual populations or lineages should therefore outcompete their asexual 'relatives'.

Kirkpatrick and Jenkins (1989) quantified this load as follows. Imagine there are L loci in both sexual and asexual populations at which advantageous alleles can occur at a frequency U per locus per

generation, and that each allele has an advantage s with dominance index h (such that the heterozygote has fitness 1+hs). The rate at which loci become fixed in the heterozygous state in the asexual population is $2NULP_0$, and the rate at which such heterozygously fixed loci become homozygous is $nNUP_1$, where P_x is the appropiate fixation probability and n the number heterozygously fixed loci. If we assume that $Ns>>1$, $h\neq 0$ or 1, and $s<<1$ then $P_0\approx 2hs$ and $P_1\approx 2s(1-h)/(1+hs)$ (Fisher 1958). It is then simple to show that the equilibrium number of heterozygously fixed loci in the asexual population is

$$\bar{n} \approx \frac{2Lh(1+hs)}{1-h} \qquad (A.1)$$

So the advantage of the sexual population relative to that of the asexual population is

$$W_s = \left[\frac{1+s}{1+hs} \right]^{\bar{n}} \qquad (A.2)$$

Kirkpatrick and Jenkins (1989) evaluated equations A.1 and A.2. However it is possible to combine and simplify the equations further by realising that s and hs are both $<<1$ so we can use the approximation $1+a\approx e^a$. Thus

$$W_s \approx \exp(2Lhs(1+hs)) \qquad (A.3)$$

which can be rearranged to give an expression for L, the number of loci at which advantageous mutations can occur:

$$L \approx \frac{\text{Ln } W_s}{2Lhs(1+hs)} \qquad (A.4)$$

This expression is evaluated in table A.1. As Kirkpatrick and Jenkins (1989) showed the sexual population can obtain a substantial

TABLE A.1

hs	W_s 2	5	10	50
0.0001	$3x10^3$	$8x10^3$	$1x10^4$	$2x10^4$
0.001	$3x10^2$	$8x10^2$	$1x10^3$	$2x10^3$
0.01	30	80	$1x10^2$	$2x10^2$
0.1	3	7	10	18

Table A.1 The number of potential advantageous alleles required to give a sexual population a fitness advantage over an asexual population. Values are calculated using equation A.4.

advantage over the asexual population with 'reasonable' parameter values. For instance if there are just seven loci at which dominant alleles with a 10% advantage can segregate then the sexual population will have a five fold advantage over the asexual population. Although little is known about the number of potential advantageous alleles and the selection which acts upon them, it seems unlikely that L will exceed the total number of genes in an organism: about 10^4 in *Drosophila* and 10^5 in mammals (Kondrashov 1988). Thus selection has to be quite strong for the sexual population to gain an advantage.

However there may be a practical problem with this hypothesis: recombination. Parthenogenesis is generally achieved by the suppression of meiosis or pre-meiotic doubling of chromosome number with sister chromosome pairing (Maynard Smith 1978). (Note that parthenogenesis via post-meiotic fusion of pro-nuclei, or pre-meiotic doubling with non-sister chrmomosome doubling leads rapidly, if not instantaneously, to homozygosity at all loci and is therefore very rare (Maynard Smith 1978)). Although in both cases every daughter generally recieves an exact copy of the mother's genome, recombination and gene conversion can lead to a heterozygous mother giving homozygous offspring. Let the frequency at which heterozygous mothers give homozygous offspring be 2α, so that assuming there is no

bias in this process the rate at which heterozygously fixed loci become homozygous in the asexual population is $n(U+\alpha)NP_1$. One can then show that the equilibrium number of heterozygously fixed loci in the asexual population is:

$$\bar{n}^* = \bar{n} \times \frac{U}{U + \alpha} \qquad (A.5)$$

and the number of potential advantageous alleles is

$$L^* \approx L \frac{(U+\alpha)}{U} \qquad (A.6)$$

Thus one can see that if the rate at which heterozygous mothers give homozygous offspring is much greater than the rate of mutation, the number of potential advantageous alleles required to give the sexual population an advantage also increases. In fact it increases in a roughly linear fashion since $(U+\alpha)/U \approx \alpha/U$ when $\alpha >> U$. For instance if the rate at which heterozygous mothers give homozygous offspring is an order of magnitude greater than the rate of mutation, then the number of potential advantageous alleles must be about ten times greater to allow a sexual population to enjoy the same advantage as it would if there was no mitotic recombination or gene conversion. Since L is unlikely to be greater than about 10^4 or 10^5 at the very most, depending on thr group concerned, one can see that the parameter range over which the sexual population can have an advantage over the asexual population is very small; essentially there has to be very strong selection at a large number of loci.

Heterozygous parthenogenic individuals can produce homozygous offspring by either gene conversion or reciprocal recombination at the four strand stage of mitosis. Estimates of mitotic recombination vary from 10^{-3} events per gamete in Diptera (Gethmann 1988) to 10^{-8} events per cell generation in yeast (Lichten and Haber 1989). In mammalian cell lines the frequency is estimated to be about 10^{-6} to 10^{-8} per cell generation (Subramani and Seaton 1988, Bollag *et al.* 1989). Studies of mosaism in mice also suggest that mitotic

recombination is not uncommon (Gruneberg 1966). In both yeast and mammalian cell lines 80% of the recombination events are gene conversions, whereas studies of somatic mosaicism in Diptera suggest that mitotic recombination occurs at the four strand stage of mitosis (Rubini *et al.* 1980). Bearing in mind that per cell generation estimates should be multiplied by the number of cell divisions that occur per generation in the germ line one can estimate α to be greater than 10^{-6} for most multicellular organisms, probably considerably so.

Obtaining accurate estimates of the mutation rate is notoriously difficult. The rate of silent substitution would of course provide an excellent measure if we could guarantee that selection was not acting. Estimates from pseudogenes suggest mammalian mutation rates are about 10^{-9} events per generation (Li and Graur 1991), and estimates from silent substitution rates from a wide range of genera are always less than 10^{-7} (Sharp 1989). Thus it would seem likely that the rate of mitotic recombination is greater, possibly much greater, than the rate of mutation, and as such sexual populations may never gain sufficient advantage from segregation to offset the two fold advantage of an all female population.

REFERENCES

Aissani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C,Bernardi G (1991) The compositional properties of human genes. J Mol Evol 32:493-503

Aota S-I, Ikemura T (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acid Res 14:6345-6355

Avery P.J (1984) The population genetics of haplo-diploids and X-linked genes. Genet Res 44:321-341

Benetzen J.L, Hall B.D (1982) Codon selection in yeast. J Biol Chem 257:3026-3031

Bernardi F, Ninio J (1978) The accuracy of DNA replication. Biochimie 60:1083-1095

Bernardi G (1989) The isochore organization of the human genome. Ann Rev Genet 23:637-661

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J,Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm blooded vertebrates. Science 228:953-958

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1-11

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7-18

Bird A (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209-213

Blake R.D, Hess S.T, Nicholson-Tuell J (1992) The influence of nearest neighbours on the rate and pattern of spontaneous mutation. J Mol Evol 34:189-200

Bohr V.A, Okumoto D.S, Ho L, Hanawalt P.C (1986) Characterisation of a DNA repair domain containing the dihydrofolate reductase gene in CHO cells. J Biol Chem 261:16666-16672

Bohr V.A, Philips D.H, Hanawalt P.C (1987) Heterogeneous DNA damage and repair in the mammalian genome. Cancer Research 47:6426-6436

Bollag R.J, Waldman A.S, Liskay R.M (1989) Homologous recombination in mammalian cells. Ann Rev Genet 23:199-225

Brown T.C, Jiricny J (1988) Different base/base mispairsare corrected with different efficiencies and specificities in monkey kidney cells. Cell 54:705-711

Brown T.C, Jiricny J (1989) Repair of base-base mismatches in simian and human cells. Genome 31:578-583

Brown T.C, Brown-Leudi M.L (1989) G/U lesions are efficiently corrected to G/C in SV40 DNA. Mut Res 227:233-236

Bulmer M (1986) Neighbouring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322-329

Bulmer M (1987) A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol Biol Evol 4:395-405

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by a mutation-selection balance? J Evol Biol 1:15-26

Bulmer M (1989) Cocodn usage and secondary structure of MS2 phage RNA. Nucleic Acid Res 17:1839-1843

Bulmer M (1990) The effects of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acid Res 18:2869-2873

Bulmer M (1991a) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-907

Bulmer M (1991b) Strand symmetry of mutation ratesin the beta-globin region. J Mol Evol 33:305-310

Bulmer M (1991c) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol Biol Evol 8:868-883

Bulmer M, Wolfe K.H, Sharp P.M (1991) Synonymous nucleotide substitution rates in mammalian genes: implicationsfor the molecular clock and the relationships of the mammalian orders. Proc Natl Acad Sci USA 88:5974-5978

Calza R.E, Eckhardt L.A, DelGiudice T, Schildkraut C.L(1984) Changes in gene position are accompanied by a change in time of replication. Cell 36:689-696

Charlesworth B (1990) Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet Res 55:199-221

Charlesworth B, Coyne J.A, Barton N.H (1987) The relative rates of evolution of sex chromosomes and autosomes. Am Nat 130:113-146

Clarke B.C (1970) Darwinian evolution of proteins. Science 168:1009-11

Comings D.E (1971) The replicative heterogeneity of mammalian DNA. Exp Cell Res 71:106-112

Comings D.E (1978) Mechanisms of chromosome banding and implications for chromosome structure. Ann Rev Genet 12:25-46

Coulondre C, Miller J.H, FaraboughP.J, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775-780

Cox E.C (1972) On the organizations of higher chromosomes. Nature New Biol 239:133-134

Crow J.F (1988) The importance of recombination: in *The evolution of sex*. eds Michod R.E, Levin B.R. Sinauer Associates, Massachusetts.

Davisson M.T, Lalley P.A, Doolittle D.P, Hillyard A.L,Searle A.G. Report of the comparative subcommittee for human and mouse homologies. Cytogenet Cell Genet 55:434-456

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acid Res 12:387-395

D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage and amino acid composition. J Mol Evol 32:504-510

Eyre-Walker A (1991) An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33:442-449

Eyre-Walker A (1992a) Evidence that both G+C rich and G+C poor isochores are replicated early and late during the cell cycle. Nucleic Acid Res 20:1497-1501

Eyre-Walker A (1992b) The effect of constraint on the rate evolution in neutral models with biased mutation. Genetics 131:233-234

Eyre-Walker A (1992c) The role of DNA replication and isochores in generating mutation and sustitution rate variance in mammals. Genet Res 60:(in press).

Fang J-S, Jagiello G.M (1988) An analysis of the chromomere map and chiasmata characteristics of human diplotene spermatocytes. Cytogenet Cell Genet 47:52-57

Fersht A.R (1979) Fidelity of replication of phage ΦX174 DNA by DNA polymerase III holoenzyme: spontaneous mutation by misincorporation. Proc Natl Acad Sci USA 76:4946-4950

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. Febs Lets 217:184-186

Filipski J (1988) Why the rate of silent substitution is variable within the a vertebrate's genome. J.Theor Biol 134:159-164

Fisher R.A (1930) *The Genetical Theory of Natural Selection*. Oxford University Press.

Fisher R.A (1958) *The Genetical Theory of Natural Selection*. 2nd edition. Dover,New York.

Fitch W (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin mRNAs. J Mol Evol 16:153-209

Flamm W.G, Bernheim N.J, Brubaker P.E (1971) Density gradient analysis of newly replicated DNA from synchronised mouse lymphoma cells. Exp Cell Res 64:97-104

Friedberg E.C (1985) *DNA repair*. Freeman, New York.

Garel J.P (1974) Functional adaptation of tRNA population. J Theor Biol 43:211-225

Gethmann R.C (1988) Crossing over in males of higher diptera (*Brachycera*). J Heredity 79:344-350

Gojobori T, Li W-H, Graur D (1982) Patterns of mutation in pseudogenes and functional genes. J Mol Evol 18:360-369

Goldman M.A (1988) The chromatin domain as a unit of gene regulation. Bioessays 9:50-55

Goldman M.A, Holmquist G.P, Gray M.C, Caston L.A, Nag A (1984) Replication timing of genes and middle reptitive sequences. Science 224: 686-692

Goodman M.F (1988) DNA replication fidelity: kinetics and thermodynamics. Mutat Res 200:11-20

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

Gouy M, Gautier C, Attimmonelli M, Lanave C, di Paola G (1985) ACNUC-a portable retrievel system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1:167-172

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

Gruneberg H (1966) The case for somatic crossing over in the mouse. Genet Res 7:58-75

Hanai R, Wada A (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J Mol Evol 27:321-325

Hasegawa M, Yasunaga T, Miyata T (1979) Secondary strucure of MS2 phage RNA and bias in code word usage. Nucleic Acid Res 7:2073-2079

Hastings I.M (1989) Potential germline competition in animals and its evolutionary implications. Genetics 123:191-197

Hatton K.S, Dhar V, Brown E.H, Mager D, Schildkraut C.L (1988) The replication program of active and inactive multigene families in

mammalian cells. Mol Cell Biol 8:2149-2158

Hill R.E, Hastie N.D (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326:96-99

Hillyard A.L, Doolittle D.P, Davisson M.T, Roderick T.H (1991) Locus map of the mouse. Mouse Genome 89:16-30

Holmquist G.P (1987) Role of replication time in the control of tissue specific gene expression. Amer J Hum Genet 40:151-173

Holmquist G.P (1989) Evolution of chromosome bands: molecular ecology of non-coding DNA. J Mol Evol 28:469-486

Holmquist G.P, Caston L.A (1986) Replication time of interspersed repetitive sequences. Biochim Biophys Acta 868:164-177

Holmquist G.P, Gray M, Porter T, Jordan J (1982) Characterisation of Giemsa dark- and light-band DNA. Cell 31:121-129

Hughes A.L (1991) Circumsporozoite protein genes of Malaria parasites (*Plasmodium* spp): evidence for positive selection on immunogenic regions. Genetics 127:345-353

Hughes A.L, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Hughes A.L, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA 86:958-962

Hutchison H.T, Gartler S.M (1973) Buoyant densities of early and late replicating DNA in mammalian diploid and heteroploid cell cultures. Tex Rep Biol and Med 31:321-329

Huynen M.A, Konings D.A.M, Hogeweg P (1992) Equal G and C contents in histone genes indicate selection pressures on mRNA secondary strucure. J Mol Evol 34:280-291

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurence of the respective codons in its protein genes. J Mol Biol 146:1-21

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurence of the respective codons in protein genes. Differences in synonymous codon choice paterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573-598

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

Ikemura T, Aota S (1988) Global variation in G+C content along vertebrate genome DNA: possible correlation with chromosome band structures. J Mol Biol 203:1-13

Ikemura T, Wada A (1991) Evident diversity of coson usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acid Res 19:4333-4339

Iqbal M.A, Plumb M, Stein G, Schildkraut C.L (1984) Coordinate replication of members of the multigene family of core and H1 human histone genes. Proc Natl Acad Sci USA 81:7723-7727

Iqbal M.A, Chinsky J, Didamo V, Schildkraut C.L (1987) Replication of proto-oncogenes early during S phase in mammalian cell lines. Nucleic Acid Res 15:87-103

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115:605-610

Kimura M (1957) Some problems of stochastic processes in genetics. Ann Math Stat 28:882-901

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press.

Kimura M, Ohta T (1974) On some principles governing molecular evolution. Proc Natl Acad Sci USA 71:2848-2852

Kirkpatrick M, Jenkins C.D (1989) Genetic segregation and the maintainence of sexual reproduction. Nature 339:300-301

Kohalmi S.E, Glattke M, McIntosh E.M, Kunz B.A (1991) Mutational specificity of DNA precursor pool imbalances in yeast arising from deoxycytidylate deaminase deficiency or treatment with thymidylate. J Mol Biol 220:993-946

Kondrashov A.S (1982) Selection against harmful mutations in large sexual and asexual populations. Genet Res 40:325-332

Kondrashov A.S (1988) Deleterious mutations and the evolution of sexual reproduction. Nature 336:435-440

Kuhn E.M (1976) Localization by Q-banding of mitotic chiasmata in cases of Bloom's syndrome. Chromosoma 57:1-11

Kunkel T.A, Sabatino R.D, Bambara R.A (1987) Exonucleolytic proofreading by calf thymus DNA polymerase delta. Proc Natl Acad Sci USA 84:4865-4869

Leeds J.M, Slabaugh M.B, Mathews C.K (1985) DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and the cytoplasm of mammalian cells. Mol Cell Biol 5:3443-3450

Lewis J, Bird A.P (1991) DNA methylation and chromatin structure.

Febs Lett 205:155-159

Li W-H (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymouscodons. J Mol Evol 24:337-345

Li W-H, Wu C-I, Luo C-C (1984) Non-randomness of point mutation as reflected in nucleotide substitutions inpseudogenes and its evolutionary implications.J Mol Evol 21:58-71

Li W-H, Tanimura M, Sharp P.M (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330-342

Li W-H, Graur D (1991) *Fundamentals of molecular evolution*. Sinauer Associates, Massachusetts.

Lichten M, Haber J.E (1989) Position effects in ectopic and allelic mitotic recombination in *Saccharomyces cerevisiae*. Genetics 123:261-268

Lipman D.J, Wilbur W.J (1985) Interaction of silent and replacement changes in eukaryotic coding sequences. J Mol Evol 21:161-167

Liskay R.M, Stachelek J.L (1986) Information transfer between duplicated chromosomal sequences in mammalian cells involves contiguous regions of DNA. Proc Natl Acad Sci USA 83:1802-1806

Matthews C.K, Slabaugh M.B (1986) Eukaryotic DNA metabolism: are deoxyribonucleotides channeled to replication sites? Exp Cell Res 162:285-295

Maynard Smith J (1978) *The evolution of sex*. Cambridge University Press.

McAlpine P.J, Stranc C.C, Boucheix C, Shows T.B (1990) The 1990 catalog of mapped genes and report of the nomenclature committee. Cytogenet Cell Genet 55:5-76

McCormick P.J, Danhauser L.L, Rustim Y.M, Bertram J.S (1983) Changes in ribo- and deoxyribonucleoside triphosphate pools within the cell cycle of a synchronised mouse fibroblast cell line. Biochim Biophys Acta 755:36-40

McDonald J.H, Kreitman M (1991) Adaptive evolution at the ADH locus in *Drosophila*. Nature 351:652-654

Mellon I.M, Bohr V.A, Smith C.A, Hanawalt P.C (1986) Preferential DNA repair of an active gene in human cells. Proc Natl Acad Sci USA 83:8878-8882

Meuth M (1989) The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. Exper Cell Res 181:305-316

Michod R.E, Levin B.R (1988) eds: *The evolution of sex*. Sinauer Associates, Massachusetts.

Mita M, Sachiko I, Misuo Z, Tharappel C.J, (1988) Specific codon usage pattern and its implications on the secondary structure of silk fibroin mRNA. J Mol Biol 203:917-925

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739-5743

Miyata T, Kuma K, Iwabe N, Hayashida H, Yasunaga T (1990) Different rates of autosome-,X chromosome- and Y chromosome-linked genes: hypothesis of male driven molecular evolution: in *Population Biology Of Genes And Molecules*. eds Takahata N, Crow J.F. Baifukan.

Mouchiroud D, Fichant G, Bernardi G (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. J Mol Evol 26:198-204

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol 27:311-320

Muller H.J (1932) Some genetic aspects of sex. Am Nat 66:118-138

Muller H.J (1964) The relation of recombination to mutational advance. Mutat Res 1:2-9

Nagylaki T (1983) Evolution of a finite population under gene conversion. Proc Natl Acad Sci USA 80:6278-6281

Nishimura S (1978) Modified nucleosides and isoaccepting tRNA: in *Transfer RNA*. ed Altman S. MIT Press, Massachusetts.

Newgard C.B, Nakano K, Hwang P.K, Fletterick R.J (1986) Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue specific codon usage. Proc Natl Acad Sci USA 83:8132-8136

Okumoto D.S, Bohr V.A (1987) DNA repair in the metallothionein gene increases with transcriptional activity. Nucleic Acid Res 15:10021-10030

Phear G, Meuth M (1989a) A novel pathway for transversion mutation induced by dCTP misincorporation in a mutator strain of CHO cells. Mol Cell Biol 9:1810-1812

Phear G, Meuth M (1989b) The genetic consequences of DNA precursor pool imbalance: sequence analysis of mutations induced by excess thymidine at the hamster *aprt* locus. Mutat Res 214:201-206

Radman M, Wagner R (1986) Mismatch repair in *Escherichia coli*. Ann Rev Genet 20:523-538

Rubini P.G, Vecchi M, Franco M.G (1980) Mitotic recombination in *Musca domestica* L. and its influence on mosaicism, gynandromorphism

and recombination in males. Genet Res 35:121-130

Sharp P.M (1989) Evolution at 'Silent' sites in DNA: *Evolution and Animal Breeding: reviews on molecular and quantitative approaches in honour of Alan Robertson*. eds Hill W.G, Mckay T. Wallingford CAB International.

Sharp P.M, Li W-H (1986a) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28-38

Sharp P.M, Li W-H (1986b) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare'codons. Nucleic Acid Res 14:7737-7749

Sharp P.M, Li W-H (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon bias. Mol Biol Evol 4:222-230

Sharp P.M, Shields D.C, Wolfe K.H, Li W-H (1989) Chromosomal location and the evolutionary rate variation in enterobacterial genes. Science 246:808-810

Shields D.C, Sharp P.M (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acid Res 15:8023-8040

Shields D.C, Sharp P.M, Higgins D.G, Wright F (1988) Silent sites in *Drosophila* are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716

Subramani S, Seaton B.L (1988) Homologous recombination in mitotically dividing mammalian cells: in *Genetic Recombination* eds Kucherlapati R, Smith G.R. American Society For Microbiology, Washington D.C.

Suoeka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582-592

Suoeka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653-2657

Suoeka N (1992) Directional mutation pressure, selective constraints and genetic equilibria. J Mol Evol 34:95-114

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci USA 87:4692-4696

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269-285

Tanaka T, Nei M (1989) Positive Darwinian selection abserved at the variable-region genes of immunoglobulins. Mol Biol Evol 6:447-459

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J Mol

Evol 28:286-298

Tobia A.M, Schildkraut C.L, Maio J.J (1970) Deoxyribonucleic acid replication in synchronised cultured mammalian cell lines I. Time of synthesis of molecules of different average guanine+cytosine content. J Mol Biol 54:499-515

Topal M.D, Fresco J.R (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285-289

Travers AA, Klug A (1987) The bending of DNA in nucleosomes and its wider implications. Phil Trans R Soc Lond B 317:537-561

Von Heijne G, Blomberg C, Lijenstrom H (1987) Theoretical modelling of protein synthesis. J Theor Biol 125:1-14

Wada A, Suyama A (1985) Third letters in codons counterbalance the (G+C) content of their first and second letters. Febs Lett 188:291-294

Weissenbach J, Dirheimer G (1978) Pairing properties of the methylester of 5-carboxymethyl uridine in the wobble position of yeast tRNA. Biochim Biophys Acta 518:530-534

Wells D, Bains W, Kedes L (1986) Codon usage in histone gene families reflects functional rather phylogenetic relationships. J Mol Evol 23:224-241

Wilkins R.J, Hart R.W (1974) Preferential repair in human cells. Nature 247:35-36

Wright S (1969) *Evolution and the genetics of populations. Vol II. The theory of gene frequencies.* University of Chicago Press.

Wolfe K (1991) Mammalian DNA replication: mutation biases and the mutation rate. J Theor Biol 149:441-451

Wolfe K, Sharp P.M, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283-285

Yokoyama S, Yokoyama R (1990) Molecular evolution of visual pigment genes and other G-Protein-Coupled receptor genes: in *Population Biology Of Genes And Molecules*.eds Takahata N, Crow J.F. Baifukan.

Zama M (1990) Codon usage patterns in a2(I) chain domain of chicken type I collagen and its implications for the secondary structure of the mRNA and the synthesis pauses of the collagen. Biochem Biophys Res Commun 167:772-776

PUBLISHED PAPERS

Eyre-Walker A (1990) A recombination. Nature 345:673

Eyre-Walker A (1991) An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33:442-449

Eyre-Walker A (1992) Evidence that both G+C rich and G+C poor isochores are replicated early and late during the cell cycle. Nucleic Acid Res 20:1497-1501

Eyre-Walker A (1992) The effect of constraint on the rate evolution in neutral models with biased mutation. Genetics 131:233-234

Eyre-Walker A (1992) The role of DNA replication and isochores in generating mutation and sustitution rate variance in mammals. Genet Res 60:(in press).

Written permission has been obtained from the publishers for photocopying the relevant papers.

J Mol Evol (1991) 33:442–449

Journal of
Molecular Evolution

An Analysis of Codon Usage in Mammals: Selection or Mutation Bias?

Adam C. Eyre-Walker

Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

Summary. A new statistical test has been developed to detect selection on silent sites. This test compares the codon usage within a gene and thus does not require knowledge of which genes are under the greatest selection, that there exist common trends in codon usage across genes, or that genes have the same mutation pattern. It also controls for mutational biases that might be introduced by the adjacent bases. The test was applied to 62 mammalian sequences, and significant codon usage biases were detected in all three species examined (humans, rats, and mice). However, these biases appear not to be the consequence of selection, but of the first base pair in the codon influencing the mutation pattern at the third position.

Key Words: Humans — Mouse — Rat — Codon usage — Mutation bias — Selection

Introduction

The term "silent sites" reflects the early belief that the genetic code was truly degenerate—that certain codons were informationally equivalent. However, this has been demonstrated to be untrue in a number of species over the last 10 years. Selection has been shown to act on silent sites in bacteria (Ikemura 1981; Sharp and Li 1986; Shields and Sharp 1987; Sharp 1990), yeast (Ikemura 1982), and *Drosophila* (Shields et al. 1988) (for review see Sharp 1989). The basis of this selection appears to be the optimization of translation, which is largely achieved by the use of codons with common tRNAs, or codons with certain codon/anticodon binding characteristics (Gouy and Gautier 1982; Ikemura 1985; Bulmer 1988). However, the potential exists for selection to act on silent sites via DNA folding (see Travers and Klug 1987), mRNA secondary structure (Clarke 1970), and sequences controlling gene expression.

The evidence for selection on silent sites in mammals is less convincing. Miyata and Hayashida (1981) suggested that silent sites evolve more slowly than pseudogenes and thus are subject to selection. However, a subsequent analysis by Wolfe et al. (1989) failed to support these findings. Fitch (1980) and Lipman and Wilbur (1985) have both shown that codons that have undergone amino acid substitution have higher rates of silent substitution, a finding that the latter authors interpreted as being due to selection. However, this could be explained by the simultaneous mutation of two adjacent nucleotides, a phenomenon for which there is some evidence (P.M. Sharp and M. Averoff, personal communication).

There are major problems in detecting selection at silent sites in mammals using the techniques developed in the study of other taxonomic groups. The basic approach has been to identify trends in codon bias across genes with different levels of gene expression, ultimately relating this, where possible, to some selective cause. However, in mammals there are three problems. First, it is very difficult to identify genes that are likely to be under the greatest selection at silent sites. Second, the tissue-specific nature of some gene expression, including tRNA genes, reduces the likelihood of there being common trends across genes for what in other species is a major selective factor, tRNA concentration. And third, even if common trends exist, they will be confounded by variation in G+C content among different chromosomal regions (i.e., isochores) (Bernardi et al. 1988). Doubts have also been voiced as

Table 1. The number of times the codons of cysteine and serine-2 are used by human dystrophin and human androgen receptor genes

	Human dystrophin third position		Human androgen receptor third position	
	T	C	T	C
Ser2	55	40	4	30
Cys	15	20	13	14

Cys and Ser2 show different codon usage patterns in androgen ($\chi^2 = 9.91$, $P < 0.005$) but not in dystrophin ($\chi^2 = 2.33$, $P > 0.10$)

to whether mammals have large enough population sizes to make selection at silent sites effective (Sharp 1989).

In this paper a statistical method is presented that looks at the codon usage within a gene, comparing codon usage in one group of synonymous codons with that in others. By restricting the analysis to a single gene at a time one can largely control for the isochore structure of the genome and allow selection pressures particular to a gene to be revealed. Significant codon biases are detected. An investigation is made into whether these biases exist in other reading frames and on the complementary strand to determine whether the biases are caused by selection or mutation. In the Discussion various alternative explanations for the biases observed are considered.

Methods

Codon Bias Tests. If there is no selection acting at silent sites the third position of a codon will simply reflect mutational processes. Codons of similar degeneracy should therefore show similar third position nucleotide frequencies. However, Bulmer (1986) has shown that the rate of base mutation is dependent upon the adjacent nucleotides, so the third position nucleotide frequencies are expected to be influenced by both the second base in the codon and the first base of the distal codon. Correction for this source of error at the second position can be achieved by restricting comparisons to those amino acids that have the same nucleotide at the second position. It was thus possible to compare the third position nucleotide frequencies of the following amino acids using a chi-square independence test.

1) The C test: Alanine, threonine, proline, and the fourfold degenerate codons of serine (Ser4) were compared. These amino acids all have C at the second position and any base at the third. Note that the fourfold degenerate codons of serine are not connected to the twofold serine codons by a single base substitution, so substitution between them is very rare. Because these amino acids all have C at the second position and therefore very small G frequencies at the third it was necessary to combine data. The frequencies of G and C were combined as this will not remove heterogeneity due to selection for intermediate codon/anticodon binding strength (Gouy and Gautier 1982), which appears to act in yeast (Bulmer 1988).

2) The A^G_A test: Glutamic acid, lysine, and glutamine were compared; these are twofold degenerate codons with A at the second position and purines (A and G) at the third.

Table 2. An example of correcting for fourth base pair frequencies

		Uncorrected fourth base pair		Corrected fourth base pair	
		A	G	A	G
Cys	TGT	4	4	3.5	4
	TGC	7	1	6.1	1
	Total	11	5	9.6	5
	Frequency	0.69	0.31	0.66	0.34
Ser2	AGT	11	21	11	16.1
	AGC	20	0	20	0
	Total	31	21	31	16.1
	Frequency	0.60	0.40	0.66	0.34

First of all the frequency with which each amino acid is followed by each base is calculated; e.g., Ser2 is followed by A 60% of the time. The minimum frequency with which each fourth base follows each amino acid is found. For A this is 0.60; for G it is 0.31. By multiplying the total number of each amino acid by these minimum frequencies one obtains the maximum sample size one is allowed to draw of that amino acid followed by that fourth base. So we can use $0.60 \times 61 = 9.6$ Cys codons followed by A, and $0.31 \times 16 = 5.0$ Cys codons followed by G. On average, when 9.6 Ser2 codons followed by G are drawn from the sequence there will be 3.5 TGT codons and 6.1 TGC codons. By summing these corrected totals over all fourth bases one gets the expected number of each codon after resampling. In the case of TGT this would be $3.5 + 4 = 7.5$. It is these numbers that are used in the χ^2 tests. Note that once the correction has been performed each amino acid has the same fourth base pair frequencies

3) The A^T_C test: Tyrosine, histidine, aspartic acid, and asparagine were compared. These are twofold degenerate codons with A at the second position and pyrimidines (T and C) at the third.

4) The G^T_C test: Cysteine and the twofold degenerate codons of serine (Ser2) were compared. These are twofold degenerate codons with G at the second position and pyrimidines at the third.

As an example consider the codon usage tables of Cys and Ser2 (G^T_C test) for two genes (Table 1). If a standard 2×2 chi-square independence test is calculated, the χ^2 value equals 2.33 for human dystrophin and 9.91 for human androgen. This would suggest that Ser2 and Cys are subject to the same processes (mutation, selection) in dystrophin but not in androgen.

Distal Base Effects. The potential bias introduced by the first base of the distal codon (henceforth known as the fourth base) can be eliminated by randomly drawing codons from the sequence without replacement in such a way that each amino acid involved in a comparison, say an A^G_A test, has the same fourth base frequencies. After resampling, if selection is not acting and mutation biases do not extend beyond the adjacent base, each amino acid in a test should have the same third position nucleotide frequencies. Actually resampling the data theoretically leaves the statistic calculated on the resultant contingency table χ^2 distributed. However, the reduction in sample size leads to many cells having small expected frequencies, which leads to departures from a χ^2 distribution. So as an approximation the expected frequencies from a resampling event were calculated (see Table 2 for a worked example, and Appendix 1). Simulations with large samples show that the approximation is unbiased and accurate to within 10% for large χ^2 values, but becomes an underestimate for small χ^2 values: i.e., the test is slightly conservative.

Table 3. The χ^2 values for the four tests of codon bias

Gene	C (df = 6)	A^G_A (df = 2)	A^T_C (df = 3)	G^T_C (df = 1)
Human				
Dystrophin	3.41	16.11[b]	2.22	1.46
Androgen receptor	5.43	8.02[a]	2.72[c]	8.82[b]
Complement C5	7.96	10.86[b]	2.88	1.80
c abl	4.05	3.45	1.68	0.44
α-2 macroglobulin	3.85	5.82	0.41	0.62
Epidermal growth factor receptor	3.41	0.08	2.86	1.83
Glucocorticoid receptor	2.71	4.08	1.76[c]	1.14
Ca^{2+} ATPase	4.71	6.41[a]	1.52	0.06
Enkephalinase	1.12	2.81	2.77[c]	1.14[c]
Complement C3	4.56	2.70	0.77	0.03
Complement C4	6.86	2.28	1.07	0.01
Na^+-K^+ ATPase	2.23	5.89	0.59	0.47
Laminin beta 2	10.02	16.49[b]	0.77	2.80
Angioleusin-I converting enzyme	3.51	9.02[a]	2.98	0.15[c]
Ceruloplasmin	4.39	0.49	1.89	0.11[c]
Glycoprotein p150, 95	11.87	1.90	1.24	1.92[c]
Breakpoint cluster gene	7.30	5.79	0.96	0.50[c]
Apolipoprotein B-100	7.72	11.53[b]	3.60	0.13
LDL receptor-related protein	7.81	6.41[a]	0.81	0.07
Coagulation factor VIII	2.89	4.17	1.46	4.10[a]
Von Willebrand factor	4.28	3.18	5.58	0.63
Growth factor II receptor	5.34	11.15[b]	1.66	3.83
Insulin receptor precursor	9.34	3.18	3.37	0.98
Poly (ADP-ribose) polymerase	3.89	16.72[b]	1.30	0.75[c]
Fibronectin	5.25	8.02[a]	2.80	0.68
Embryonic myosin heavy chain	4.92	2.18	4.50	1.23
CCG-1	6.09	13.01[b]	5.00	0.40
Coagulation factor V	10.42	19.28[b]	0.76	0.89
Laminin beta 1	3.63	13.54[b]	0.95	4.73[a]
GL-1 protein	6.47	4.33	0.41	0.12
Leukocyte adhesion glycoprotein	2.13	14.09[b]	0.06	1.50
Total	167.57	232.99[b]	61.35	43.34
Mouse				
Dystrophin	5.76	8.05[a]	1.47	0.93
Complement C5	1.59	3.04	2.09	3.89[a]
c abl	4.66	2.45	0.59	1.18
Glucocorticoid receptor	3.85	5.62	2.91	0.02
Epidermal growth factor receptor	4.75	3.26	0.27	4.80[a]
Complement C3	5.34	6.43[a]	3.09	0.50
Complement C4	7.75	5.72	0.39	0.47
Laminin beta 2	2.62	2.17	4.50	4.46[a]
Insulin receptor precursor	6.38	0.42	4.73	0.61
Poly (ADP-ribose) polymerase	7.64	5.18	0.87	0.00[c]
Laminin beta 1	8.78	16.92[b]	2.66	1.07
Leukocyte adhesion glycoprotein	6.20	11.70[b]	4.73	0.62
Angioleusin-I converting enzyme	8.47	2.58	0.42	0.33[c]
Microtubule associated protein 2	11.70	4.55	2.62	0.92[c]
Multidrug resistance protein	4.99	8.14[a]	1.68	1.15[c]
Total	90.48	86.23[b]	33.02	20.95
Rat				
Androgen receptor	7.34	5.19	0.36	1.61
α-2 macroglobulin	6.01	0.70	3.67	1.07
Ca^{2+} ATPase	9.61	6.10[a]	0.27[c]	0.31
Enkephalinae	2.50	2.34	1.86[c]	0.97[c]
Na^+-K^- ATPase	7.61	10.54[b]	2.85	0.02[c]
Plasma protein inhibitor	5.64	2.45	1.48	0.80

Table 3. Continued

Gene	C (df = 6)	A^G_A (df = 2)	A^T_C (df = 3)	G^T_C (df = 1)
Clathrin	10.02	13.53[b]	0.23	0.66
Sodium channel III	7.64	0.34	2.94	0.06
Acetyl coA carboxylase	19.66[b]	23.39[b]	3.96	0.20
Embryonic myosin heavy chain	6.87	0.03	7.15	0.00[c]
Fatty acid synthetase	12.04	6.77[a]	5.54	0.00
Proteoglycan core protein	13.01[a]	4.44	0.33	0.10
Phospholipase C-I	2.77	3.27	5.46	3.55
Guanylate cyclase 70 kd subunit	6.14[c]	0.90	1.72	2.49
neu oncogene	4.74	3.23	0.19	0.00
Neurofilament protein	3.41	2.85	4.10[c]	0.29[c]
Total	125.01[a]	86.07[b]	42.11	12.13

[a] Significant codon bias at the 5% level
[b] Significant codon bias at the 1% level
[c] More than 20% of expected values in chi-square test have a value of less than 5; statistic is unlikely to be χ^2 distributed

The Data Set. A set of 62 large genes from humans, rats, and mice was taken from the GenBank and Embl databases using the GCG (Devereux et al. 1984) and ACNUC (Gouy et al. 1985) packages. Human and rodent genes were used as they are the best represented mammalian groups in the sequence databases. A full set of accession numbers is available on request from the author. Only genes with more than 600 codons were included in the analysis to avoid the expectations in the χ^2 tests becoming too small through insufficient sample size.

Results

Testing Codon Bias

Table 3 shows the results of the four-codon bias tests performed on the 62 mammalian sequences. In the analysis that follows, human and rodent sequences will be analyzed separately so that different trends in the two groups may be revealed and the analysis is not complicated by covariances introduced by genes represented in both groups. These covariances arise because genes represented in both samples are not independent as they share a relatively recent common ancestor.

It is clear that there are significant amounts of heterogeneity in the A^G_A comparison. Not only are there several significant values but the overall level of heterogeneity is inconsistent with similar codon usage in Gln, Lys, and Glu: the pooled χ^2 value for humans is 232.99 ($P < 0.005$) and for rodents it is 172.30 ($P < 0.005$).

There are also several significant values in the G^T_C test. However, it is not immediately clear that this is not simply due to taking 62 samples. Because 1.55 tests are expected to fail at the 5% level in a sample of 31, the probability of getting two or more failures in humans is ~0.46 (using a Poisson approximation to the binomial distribution). Similarly, the probability of getting one or more values significant at 0.5% (human androgen) is ~0.15. Hence, the overall probability or two values significant at 5% and one value significant at 0.5% is ~0.07. For rodents the chance of getting three or more values significant at 5% is ~0.19. However, all the rodent significant values come in the mouse data set, and the probability of this is 0.04. It thus appears that there is evidence of significant heterogeneity in codon bias in mice, a suggestion of heterogeneity in humans, but nothing in rats.

The overall pooled χ^2's are not significant in any species. However, it should be noted that many of the G^T_C tests have very small expected frequencies and are thus not likely to be χ^2 distributed, and that the fourth base pair correction leads to an underestimate of the true χ^2 value (see Methods). As a result the pattern of codon bias in the G^T_C test was investigated further (see below).

The overall pooled χ^2 for the rat C tests is significant ($\chi^2 = 128.86$; $P < 0.05$), suggesting that the two significant values for acetyl coA carboxylase and proteoglycan core protein are not the consequence of taking many samples. However, there is no evidence of codon bias in the human and mice C tests, or the A^T_C tests of any species studied.

Patterns of Codon Bias in the A^G_A and G^T_C Tests

If the pattern of codon bias is examined in those genes that show significant heterogeneity in the A^G_A and G^T_C tests it becomes clear that most genes show very similar patterns of codon usage (Table 4). In the A^G_A test Glu always shows a higher frequency of A-ending codons than Gln, with Lys generally midway between the two; and in the G^T_C test Cys generally shows a higher frequency of T-ending codons than Ser2. Not surprisingly, given the consistency of the trends in the significant genes, these

Table 4. The codon usage in those genes that show significant heterogeneity

Gene	Gln	Lys	Glu	Cys	Ser2
Human					
Dystrophin	0.45	0.52	0.61		
Androgen receptor	0.17	0.42	0.45	0.47	0.12
Complement C5	0.42	0.70	0.71		
Angioleusin	0.07	0.15	0.27		
Ca^{2+} ATPase	0.47	0.67	0.71		
Laminin beta 2	0.23	0.38	0.54		
Apolipoprotein	0.43	0.58	0.55		
LDL receptor	0.08	0.17	0.16		
Growth factor II	0.20	0.42	0.43		
Poly (ADP) polymerase	0.03	0.28	0.48		
Fibronectin	0.30	0.36	0.48		
CCG-1	0.26	0.49	0.55		
Factor V	0.33	0.49	0.67		
Laminin beta 1	0.34	0.60	0.59	0.49	0.31
Leukocyte glycoprotein	0.41	0.63	0.73		
Factor VIII				0.31	0.56
Mouse					
Dystrophin	0.50	0.69	0.68		
Complement C3	0.29	0.25	0.41		
Laminin beta 1	0.22	0.53	0.48		
Leukocyte glycoprotein	0.34	0.58	0.65		
Multidrug resistance	0.34	0.45	0.61		
Complement C5				0.62	0.35
Epidermal growth				0.49	0.27
Laminin beta 2				0.50	0.30
Rat					
Clathrin	0.28	0.52	0.55		
Ca^{2+} ATPase	0.34	0.49	0.59		
Na^{+}-K^{+} ATPase	0.20	0.19	0.46		
Acetyl coA carboxylase	0.22	0.40	0.59		
Fatty acid synthetase	0.16	0.23	0.31		

Codon bias is calculated on data corrected for fourth base pair frequencies. The frequency of A at the third position of Gln, Lys, and Glu is shown, and the frequency of T in the third position of Cys and Ser2 is shown

Table 5. The number of genes for which the relationship given in the first column holds, in the various reading frames and orientations of the sequence

Relationship	Frame	Human	Rodent
Gln < Glu	123	31[a]	30[a]
	Comp	26[a]	29[a]
	312	21[b]	23[a]
	Attenuation	11.92[a]	6.37[a,c]
Lys < Glu	123	20	25[a]
	Comp	23[a]	23[a]
	312	14	22[a]
	Attenuation	2.34	0.79
Gln < Lys	123	31[a]	25[a]
	Comp	22[b]	25[a]
	312	20	18
	Attenuation	13.37[a]	3.72
Ser2 < Cys	123	20	24[a]
	Comp	17	21[b]
	312	15	16
	Attenuation	1.64	4.51[b]

The relationships refer to the frequency of A in the third position for Gln, Lys, and Glu, and the frequency of T in the third position of Cys and Ser2. All figures are out of a total of 31 genes. The attenuation row gives the χ^2 value calculated by comparing the number of genes showing the relationship in the 123 and 312 frames. Comp, complementary strand

[a] Significant at the 1% level

[b] Significant at the 5% level

[c] Two out of four cells had expected frequencies of less than 5. The statistic is not χ^2 distributed

patterns of codon bias are found in most genes in the sample (Table 5).

Bias in Other Reading Frames and on the Complementary Strand

It is of great interest to see whether these patterns of bias exist in other reading frames and on the antisense, or complementary, strand (see Discussion). In what follows the 123 frame refers to the original sequence and the 312 frame as the sequence read starting at the third base pair. The complementary sequence was formed so that the third position of a codon in the 123 frame was the third position on the complementary strand. This is equivalent to forming the complementary strand and reading it 5′ to 3′ from the second base pair.

Table 5 shows the number of genes that show a particular pattern of codon bias in the 312 reading frame and on the complementary strand. Most genes in the sample show the pattern of bias found in the 123 frame on both the complementary strand and in the 312 reading frame. However, the number of genes showing the pattern of bias in the 312 frame is always less than the number of genes showing the pattern of bias in the 123 frame and on the complementary strand, often significantly so. The number of genes showing the pattern of bias on the complementary strand is usually less than that in the 123 frame. This difference is significant for two human relationships: for Gln < Glu $P = 0.026$, and for Gln < Lys $P = 0.001$, as measured by Fisher's exact test.

The $G^{T}{}_{C}$ results warrant further comment. The χ^2 tests showed that several genes had significant codon bias with respect to Cys and Ser2 but it remained unclear as to whether these results were the consequence of sampling many genes. Table 5 shows that in rodents a significant number of genes show the same pattern of codon bias, suggesting that the bias detected in the χ^2 tests is not a statistical artifact. There is also no evidence that the strength of the bias differs between rats and mice as the number of genes showing the pattern of bias is very similar: 12 out of 15 mouse genes show the frequency of

T-ending codons to be greater in Cys than Ser2, and 11 out of 16 genes in rats.

The situation remains very unclear in humans as to whether there is any overall bias between Cys and Ser2 as only 20 out of 31 genes show Cys to have a higher frequency of T-ending codons than Ser2.

Pattern of Codon Bias in the Rat C Test

Table 6 shows the codon usage tables for rat acetyl coA carboxylase and proteoglycan core protein. In both cases Ser4 has an odd codon usage pattern, but it is difficult to identify any other major similarities. The vast majority of the χ^2 is generated by the A-ending codons of Ser4, and the T- and A-ending codons of Thr in acetyl coA carboxylase. In proteoglycan it is the T-ending Ser4 codons that contribute most to the χ^2 value, with some help from the A-ending codon.

Because no major trend in codon bias could be identified the χ^2 tests were performed on the 312 frame and complementary sequence. The χ^2 values for acetyl coA carboxylase were 14.73 [Comp (complementary strand sequence), $P < 0.025$] and 8.61 (312, not significant); for proteoglycan they were 21.78 (Comp, $P < 0.005$) and 11.11 (312, not significant). The pooled χ^2 were 140.17 (Comp, df = 96, $P < 0.005$) and 100.09 (312, df = 96, not significant). Thus, it would seem that significant codon biases do exist on the complementary sequence, but are attenuated or nonexistent in the 312 reading frame.

Table 6. Codon usage in the rat acetyl coA carboxylase and rat proteoglycan core protein C tests

	Acetyl coA carboxylase			Proteoglycan core protein		
	T	C + G	A	T	C + G	A
Ser4	0.42	0.43	0.15	0.50	0.31	0.19
Pro	0.37	0.30	0.33	0.34	0.34	0.32
Thr	0.24	0.36	0.40	0.30	0.40	0.30
Ala	0.41	0.31	0.27	0.33	0.43	0.25

Figures show the nucleotide frequencies in the third position of Ser4, Pro, Thr, and Ala. Frequencies were calculated from data corrected for fourth base pair frequencies

Discussion

The Causes of Bias

There are several ways in which the codon bias could have been generated: selection, neighboring base mutational effects extending beyond the adjacent base, nonsilent substitution, and trends in G+C content along the gene.

Nonsilent Substitution

Nonsilent substitution can generate heterogeneity in codon bias by interconverting amino acids with different codon usage patterns; patterns that could be the consequence of dinucleotide effects. However, there are good examples of genes with high levels of codon bias heterogeneity but low levels of nonsynonymous substitution. The human Ca^{2+} ATPase gene has not gained or lost any Gln, Lys, or Glu codons since it diverged from rodents, has a high rate or silent substitution, and shows significant codon bias heterogeneity in the $A^G{}_A$ test.

Trends in G+C Content

Many genes in the sample showed trends in third position G+C content along their length. The basis for these trends is unknown, but coupled to a very nonuniform amino acid distribution these trends could generate codon bias. For instance the human epidermal growth factor receptor gene shows a very sharp increase in G+C content of about 30% three-quarters of the way along its length. If cysteine tended to cluster at one end of the gene and serine at the other, heterogeneity in codon usage would be produced. There are however examples of genes that show no trend in G+C content but show significant codon bias heterogeneity, e.g., human androgen receptor, mouse complement C3, rat clathrin, and human angiotensin-I converting enzyme.

Selection

It is difficult to think of any form of selection that would lead to significant bias appearing on the complementary strand and in other reading frames, but which is attenuated in the 312 reading frame. Selection during translation, to match the commonest tRNAs for instance, would only give bias in the 123 frame; selection on RNA secondary structure would give bias in all reading frames but not on the complementary sequence; and selection on DNA structure would give equally strong bias in all reading frames and the complementary sequence.

Mutation Bias

In both *Escherichia coli* and yeast it has been observed that the mutation pattern is influenced by bases several positions upstream and downstream of the site in question (Bulmer 1990). The biases detected here could be a consequence of such neighboring base effects, and this view is supported by the attenuation of the bias in the 312 reading frame. Let us imagine that A is disfavored by mutation when preceded by CA but not when preceded by

GA. CAN codons will have a low frequency of A at the third position compared to GAN codons if there are no C ↔ G substitutions at the first position. However, as the rate of C ↔ G substitutions increases so the frequency of A in the third position of CAN and GAN codons becomes similar. In the 123 frame the first two positions in the codon are highly constrained so CAN and GAN codons differ in their codon usage. However, in the 312 frame, where the first codon position is in reality a third position, and thus subject to frequent substitution, the codon usage of CAN and GAN codons is homogeneous.

The reason behind the attenuation of the bias in the 312 frame has important implications for the testing of the mutation hypothesis using noncoding DNA. Whenever the two 5′ sites change at a comparable, or greater, rate than the site in question, the pattern of bias will be attenuated. Thus we would not expect to see strong biases in noncoding DNA.

Other Forms of Selection

Although there is no evidence of selection in these data there remains the possibility that other selective forces act at silent sites in mammals. Studies in bacteria suggest that about 47% of the variation in silent substitution rate can be explained by codon bias and 14% by distance from *oriC* (Sharp et al. 1989). The 39% unexplained variation could be due to other selection pressures acting, say, on mRNA secondary structure, or it could be due to the complex and nonlinear relationship one expects between codon bias and silent substitution rate. The development of new techniques is required to resolve this question further.

Acknowledgments. I am very grateful to Paul Sharp for his patience, encouragement, discussion, and criticism; to Peter Keightley, Andy Leigh Brown, Bill Hill, Frank Wright, Nick Barton, and two anonymous referees for their comments on this manuscript; and to Michael Bulmer for helpful discussion. I also thank SERC for the use of the SEQNET computing facility and for supporting me financially. Part of this work was carried out using the facilities of the Irish National Centre for Bioinformatics.

Appendix 1

Imagine we have a test involving p amino acids each with q synonymous codons. Let the number of codon j of amino acid i with fourth base k in our sequence be C_{ijk}. Then the frequency with which amino acid i is followed by fourth base k is

$$B_{ik} = \frac{\sum_{j=1}^{q} C_{ijk}}{\sum_{j=1}^{q}\sum_{k=1}^{4} C_{ijk}}$$

We now need to find the minimum frequency with which any of the p amino acids is followed by the kth fourth base:

$$M_k = \min_{i=1}^{p} B_{ik}$$

The expected number of codon i, j on resampling will then be

$$E_{ik} = \sum_{k=1}^{4} \frac{M_k \times C_{ijk}}{B_{ik}}$$

These expected numbers are used in the χ^2 tests.

References

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bulmer M (1986) Neighbouring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by a mutation–selection balance? J Evol Biol 1:15–26

Bulmer M (1990) The effects of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res 18:2869–2873

Clarke BC (1970) Darwinian evolution of proteins. Science 168: 1009–1011

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acid Res 12:387–395

Fitch W (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin mRNAs. J Mol Evol 16:153–209

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

Gouy M, Gautier C, Attimmonelli M, Lanave C, di Paola G (1985) ACNUC—a portable retrievel system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1:167–172

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–598

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Lipman DJ, Wilbur WJ (1985) Interaction of silent and replacement changes in eukaryotic coding sequences. J Mol Evol 21:161–167

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743

Sharp PM (1989) Evolution at 'silent' sites in DNA. In: Hill WG, Mackay TFC (eds) Evolution and animal breeding: reviews on molecular and quantitative approaches in honour of Alan Robertson. Wallingford CAB International, pp 23–31

Sharp PM (1990) Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol Microbiol 4:119–122

Sharp PM, Li W-H (1986) An evolutionary perspective on

synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Shields DC, Wolfe KH, Li W-H (1989) Chromosomal location and the evolutionary rate variation in enterobacterial genes. Science 246:808–810

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 15:8023–8040

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

Travers AA, Klug A (1987) The bending of DNA in nucleosomes and its wider implications. Phil Trans R Soc Lond B 317:537–561

Wolfe K, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Received January 30, 1991/Revised May 13, 1991

Nucleic Acids Research, Vol. 20, No. 7 1497–1501

Evidence that both G+C rich and G+C poor isochores are replicated early and late in the cell cycle

Adam Eyre-Walker
Institute of Cell Animal and Population Biology, University of Edinburgh, EH9 1JT, UK

Received January 31, 1992; Revised and Accepted March 12, 1992

ABSTRACT

Since the G+C content of a gene is correlated to that of the isochore in which it resides, and early replicating isochores are thought to be relatively G+C rich, early replicating genes should also be rich in G+C. This hypothesis is tested on a sample of 44 mammalian genes for which replication time data and sequence information are available. Early replicating genes do not appear to be more G+C rich than late replicating genes, instead there is considerable variation in the G+C content of genes replicated during both halves of S phase. These results show that both G+C rich and poor fractions of the genome are replicated early and late in the cell cycle, and suggest that isochores are not maintained by the replication of DNA sequences in compositionally biased free nucleotide pools.

INTRODUCTION

A paradox seems to have gone unnoticed. It is believed that G+C rich isochores and housekeeping genes replicate early in the cell cycle, with G+C poor isochores and some tissue specific genes replicating late (1–3). Since the G+C content of a gene is correlated to the isochore in which it lies (4,5) housekeeping genes should be G+C rich compared to tissue specific sequences. However there appears to be no difference in the G+C contents of housekeeping and tissue specific genes (6).

The link in this paradox I wish to focus on is the early replication of G+C rich isochores, since the evidence for it can be interpreted in two ways. Evidence for the early replication of G+C rich DNA comes from the 3–5% difference in G+C content that has been measured between the early and late replicating fractions of the genome (7–11) and the coincidence of chromosome bands produced by G+C content sensitive methods, such as quinacrine staining, and replication time bands (1,12). The simplest and most popular interpretation of these observations is that most, if not all of the isochores replicated early in the cell cycle have a higher G+C content than those replicated late in S phase (3,12–14). However the observations are equally consistent with the replication of all fractions of the genome both early and late in the cell cycle, with the early replicating DNA only being on average slightly more G+C rich. The difference is very important since it has implications for our understanding of chromosome structure and evolution; in particular how isochores are maintained. It also seems inappropiate to talk about early replicating DNA being more G+C rich if in fact most of the variation in G+C content is within the early and late replicating fractions, not between them.

In order to distinguish between these alternatives a set of genes for which there are replication time data and sequence information was compiled. If we assume that gene and isochore G+C contents are correlated the G+C contents of the genes will give an insight into the range of isochore G+C contents replicated during the two halves of S phase. The data will also allow us to eliminate the possibilities that the paradox has arisen, (i) because many tissue specific genes replicate early in the cell cycle, and (ii) because the relationship between isochore and gene G+C contents is different for housekeeping and tissue specific sequences.

MATERIALS AND METHODS

Replication Time Data

Data on the replication time of specific genes was taken from Holmquist (3) with minor modifications (see below). His list is a compilation of data from references 15–20. These studies suggest that all genes expressed in a tissue are replicated early, but that unexpressed genes may replicate at any time during S phase. If a gene does replicate late in one cell type it does so in most other cell lines in which it is not expressed. Therefore genes which were found to replicate late in most cell types, in which they were not expressed, were classified as late replicating. All other genes were classed as early replicating. The classification was the same as that given by (3) except for albumin, which appears to have been misclassified, and complement C4 which was not included in the compilation. It is worth pointing out that the replication time of most genes was coincident over several cell types from several species. In particular there were no genes with different replication times between the two species groups (rodents and primates).

Sequence Information

Sequence information was taken from the Genbank (Release 68) and Embl (Release 27) databases using the GCG sequence analysis package (20). Accession numbers are available on request. Human and mouse sequences were extracted for all genes for which replication time data were available. If a mouse gene was not available the rat sequence was used instead. The G+C contents of mouse and rat genes are very similar (21,22) so mixing rat and mouse genes should not lead to substantial bias.

If human sequences were not available other primate sequences were used. The classification as to housekeeping or tissue specific expression was taken from Holmquist (3) who gave no details as to how the classification was performed.

Our primary interest in the G+C contents of early and late replicating genes is not the compositions of the sequences themselves, but what they tell us about the isochore in which they reside; i.e we are interested in whether G+C rich and G+C poor isochores replicate both early and late. It is therefore important to ensure that a particular isochore is only represented once in the data set, by including only one gene from a set of linked, or recently diverged, genes. Since isochores are thought to be at least 300kb in length (12) genes within this distance of one another were regarded as representing the same isochore. The average G+C content over a set of linked genes was not used because it is possible for a linkage group to traverse two or more isochores of different compositions. Instead the longest sequence was used. Small scale physical distance information (<300kb) was taken from references 17 and 19. Large scale linkage was also checked using HGM10.5 (23,24) and a mouse map (25). No genes were excluded on the basis of this information because of the scale and the lack of accuracy involved. However suffice it to say that only six genes were found to be 'linked' (within a centimorgan or in the same chromosome band). Beta-globin and c-Ha-ras map to the same human chromosome band, 11p15.5, but are some 18 centimorgans apart in mice; immunoglobulin kappa constant and variable map to the same chromosome band, 2p12, in humans and the same centimorgan in mice, 6.32; and arginine succinate synthetase and c-abl are the same distance along chromosome 2 in mice.

If several sequences from a dispersed multigene family were available with replication time information (e.g beta and gamma actin), only one sequence was used since any recently diverged members will tend to correlate with the G+C content of the 'parental' isochore, not that of their present location. Such sequences will therefore tend to contribute information about the same isochore.

One further problem with multigene families is identifying which member the replication time is actually known for, especially if some members of the family are quite disimilar to each other. For instance the rodent and primate placental lactogen genes have very different G+C contents which suggests that they are paralogous, and such paralogous genes could have different replication times. However this source of error is only relevent if the paralogous genes have different G+C contents, of which there is little evidence in the data set (table 1) and most of the sequences used are probably single copy genes. Therefore any errors should be small.

Testing The Data

Differences in the distribution of G+C contents, say of early and late replicating genes, were tested with a Mann-Whitney test. This tests whether two sets of data could have come from the same distribution, and fails if the medians are different, or if the medians are the same but the shapes of the distributions are asymmetrical and different.

Such tests ask whether two sets of data could have come from the same distribution, whereas we want to ultimately ask a slightly more subtle question: are the gene G+C contents we observe consistent with all the early replicating isochores being more G+C rich than the late replicating isochores? In order to do this we need to take into account the less than perfect correlation between gene and isochore G+C contents; i.e it is possible for all early replicating isochores to be more G+C rich than late replicating isochores and yet for there still to be some overlap in the G+C contents of early and late replicating genes. The approach taken was as follows: isochore G+C contents were randomly generated from gene G+C contents in a way consistent

Table 1. The expression status, replication time and G+C content of a set of primate and rodent genes.

Gene	Rep Time[a]	Exp[b]	Time known[c]	Primate 3[d]	Time known[c]	Rodent 3[d]
HPRT	E	H	√	39.6	√	41.5
APRT	E	H	√	81.6	√	74.3
CAD	E	H	√	71.5	√	NA
DHFR	E	H	√	42.5	√	47.8
Argininesuccinate synthetase	E	H	√	74.7	√	67.9
Glucose-6-phosphate dehydrogenase	E	H	√	85.1		62.5
β-tubulins	E	H	√	81.8	√	71.8
Phosphogylcerate kinase 1	E	H	√	55.8		54.3
Tyrosine aminotransferase	E	H	√	NA	√	62.7
β-actin	E	H	√	84.5		73.0
Metallothionein I	E	H	√	80.0	√	88.3
c-myc	E	H	√	76.7	√	75.8
c-Ha-ras	E	H	√	81.4	√	NA
c-ki-ras	E	H	√	32.6	√	43.2
c-fos	E	H		71.5	√	68.1
c-raf	E	H	√	37.3	√	58.0
Histone H2A.1	E	H	√	67.4	√	94.6
α-globin	E	T	√	88.8	√	67.9
c-sis	E	T	√	77.9		NA
c-myb	E	T		45.5	√	55.6
c-fes/fps	E	T		80.3	√	NA
c-rel	E	T		26.7	√	NA
c-mos	E	T	√	74.2	√	67.3
c-fms	E	T		74.8	√	67.5
Apolipoprotein AI	E	T	√	85.0		71.3
Thy-1	E	T		79.4	√	76.4
Placental lactogen	E	T	√	74.5		43.9
Complement C4	E	T		70.8	√	65.1
Immunoglobulin Kappa constant	E	T		NA	√	59.4
Albumin	E	T		39.0	√	57.0
N-ras	E	?	√	45.7		55.7
c-abl	E	?		71.5	√	68.4
Skeletal muscle actin	L	T	√	89.1		77.1
β-globin	L	T	√	66.4	√	66.9
α1-antitrypsin	L	T	√	68.8	√	64.2
β-casein	L	T		53.0	√	49.6
Phenylalanine hydroxylase	L	T	√	52.5		56.0
Factor IX	L	T	√	35.4		31.8
Fibronectin	L	T	√	50.2		53.4
Myosin heavy α-cardiac	L	T		85.8	√	80.7
N-myc	L	T	√	79.8	√	75.3
α-amylase 1	L	T		34.1	√	40.0
Major urinary proteins	L	T		NA	√	41.3
Immunoglobulin kappa variable	L	T		56.6	√	45.2

[a]The replication time of the gene: E-early and L-late.
[b]The expression status of the gene: H-housekeeping, T-tissue specific and ?-unknown.
[c]Replication time known in primate/rodent cell line.
[d]Third postion G+C content.

with the available data for each gene. The isochore G+C contents so produced were then compared to see if any overlap existed in the range of early and late replicating fractions.

Of all the relationships between gene and isochore G+C content that have been published the best, in terms of sample size and correlation coefficient, is that given by Aissani et al (5) for human third position G+C contents. Aissani et al chose to leave out two genes from their regression analysis because one of the genes had very biased amino acid composition, and the other had a very low G+C content. Since the second of these reasons appears to be arbitary and the first is not relevent to the third position G+C content both genes were included in this study. The relationship between isochore and gene G+C content obtained by least squares linear regression is:

(1) Isochore G+C = 31.3 + 0.229×Third Position G+C

Since the error terms (residuals) appear to be normally distributed, and unrelated in magnitude or sign to the third position G+C contents, the predicted isochore G+C content for a gene of G+C content X_0 is t-distributed with N−2 degrees of freedom, a mean of Y_0, the isochore G+C content given by the regression line (1), and a standard deviation of

$$(2) \qquad S\left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}\right]^{\frac{1}{2}}$$

where S is the standard deviation of the residuals and N the sample size of the data used in the regression. Thus by sampling at random from a t-distribution with the appropiate parameters it is possible to convert gene G+C contents into isochore G+C contents in a way consistent with the data of Aissani et al. The isochore G+C contents so produced can then be examined to see if early and late isochores overlap in G+C content. By repeating this procedure many times it is possible to assess how much overlap there must be between the G+C contents of early and late replicating isochores. For instance if we found that only 0.5% of a very large number of randomly produced isochore sets showed no overlap between early and late replicating fractions, then we would be able to reject the null hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at the 0.5% level. This test was only applied to human genes because there are far fewer rodent genes for which isochore location is known. The test would therefore be much less powerful.

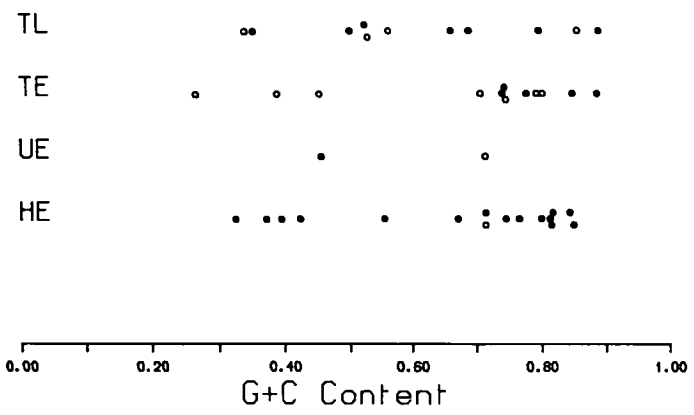


Figure 1. The third postion G+C contents of human housekeeping, tissue-specific, early and late replicating genes. **HE**−early replicating housekeeping genes. **UE**−early replicating genes of unknown expression. **TE**−Early replicating tissue specific genes. **TL**−Late replicating tissue specific genes. Filled circles are those genes whose replication time is known in a human cell line.

RESULTS

The G+C content, replication time and expression status of the 44 genes in the data set are given in table 1, and represented graphically in figures 1 and 2. Only third position G+C contents are given since the correlation between third position and isochore G+C contents is much better than for other positions (4,5). Since there is no evidence that replication times differ between rodents and primates (table 1), and no evidence of differences in the G+C contents of genes whose replication time is known and those whose replication time is inferred from another group (table 2a), it was assumed in all subsequent analyses that replication times were identical in primates and rodents. The results are not qualitatively affected by this assumption.

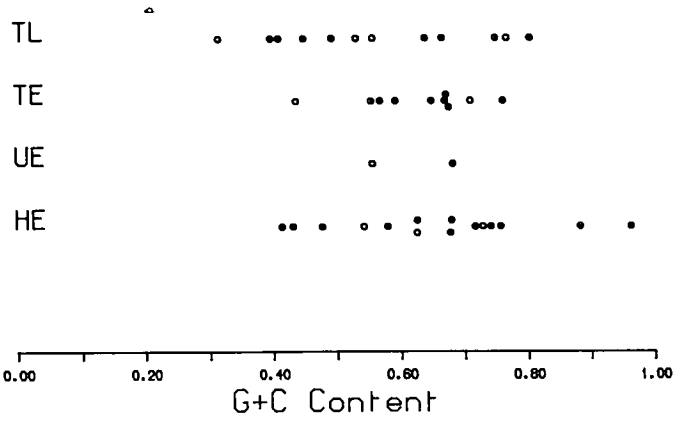


Figure 2. The third postion G+C contents of mouse housekeeping, tissue-specific, early and late replicating genes. Symbols as in figure 1. Filled circles are those genes whose replication time is known in a rodent cell line.

Table 2. Testing for differences in G+C content.

Data set	Primates	Rodents
(a)		
H	–	0.72
TE	0.14	0.90
TL	0.78	0.93
E	0.21	0.37
T	0.25	0.53
(b)		
H v TE	0.91	0.62
TE v TL	0.41	0.28
H v T	0.72	0.28
E v L	0.41	0.17

Figures in the body of the table show the probability of the two data sets being more dissimilar than they are by chance alone in a Mann-Whitney test. In part (a) the G+C contents of genes whose replication time is known in a group (e.g primates) are compared against those whose replication is inferred from another group (e.g rodents). In part (b) genes with different characteristics are compared. H−housekeeping, T−tissue specific, E−early and L−late replicating genes. The test cannot be performed for primate housekeeping genes due to insufficient sample size.

Confirming the result of Mouchiroud et al (6) figures 1 and 2 show that there is no difference in the distribution of G+C contents of housekeeping and tissue-specific genes. Mann–Whitney tests confirm this (table 2b). More importantly there is also little difference in the distributions of early and late replicating genes. The early replicating genes appear to be slightly more G+C rich than the late replicating genes but this difference is not significant (table 2b).

To illustrate how inconsistent these results are with the replication of only G+C rich isochores early, and G+C poor isochores late in S phase, isochore G+C content is plotted against third position G+C content for a set of 21 human genes (5) in figure 3. There is no horizontal line which would split the data so they look like the patterns in figures 1 and 2. For instance let us imagine that all isochores above 43% replicate early in S phase with the rest replicating late: there is almost no overlap between the G+C contents of early and late replicating genes.

It is possible to make this argument more quantitative by converting gene G+C contents to isochore G+C contents in a way consistent with the data of Aissani et al (5,figure 3), as detailed in the materials and methods. In 10000 simulated sets of isochores generated from the human early and late sets of genes, there was not a single case when the most G+C rich late replicating isochore was less G+C rich than the least G+C rich early replicating isochore; i.e there was always some overlap between the early and late replicating isochores. We can therefore reject the hypothesis that all early replicating isochores are more G+C rich than late replicating isochores at 0.05% significance or lower. Furthermore in 9999 cases the upper quartile (the value above which 25% of the observations lie) of the late replicating genes was greater than the lower quartile (the value below which 25% of the observations lie) of the early replicating genes. In other words the overlap was always substantial. When the test was repeated on just early and late replicating tissue specific genes there was always an overlap of G+C contents, and in all but 22 cases the early replicating lower quartile was less than the late replicating upper quartile.

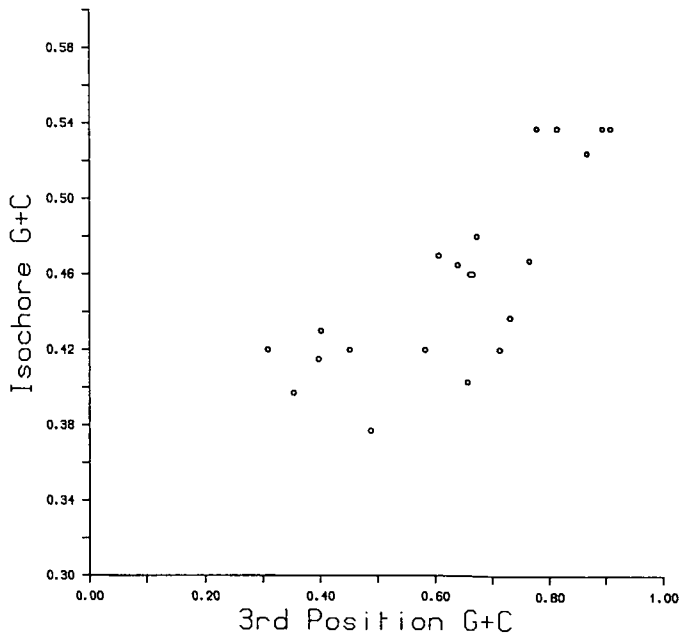


Figure 3. Isochore G+C content against third position G+C content for 21 human genes. Data from Aissani et al (5).

DISCUSSION

These results demonstrate that there is considerable heterogeneity in the G+C content of isochores replicated early and late during S phase. It is unclear however whether the replication of G+C rich and poor DNA is temporally separated, but over a much shorter time scale than the length of S phase, or whether sequences of different G+C content are simultaneously replicated. Several groups have looked at the G+C content of DNA being replicated at hourly intervals during S phase. Comings (9) found in a hamster cell line that the average G+C content of replicating DNA changed continuously from relatively A+T rich, to G+C rich before decreasing again to G+C poor. Thus there was an overlap in the G+C content of sequences replicated early and late in S phase. In contrast Tobia et al (7) and Flamm et al (8) found that in a mouse cell line the G+C content of replicating DNA decreased monotonically during S phase. However in all analyses the range of average G+C contents replicated at different times (~5%) was insufficient to cover the range of isochore G+C contents (~9–15%). It therefore seems likely that sequences of very different G+C content are replicated simultaneously during S phase.

Isochore Replication Times

Further evidence that G+C rich and poor isochores replicate in both halves of S phase is provided by a few genes for which isochore location and replication time are known (table 3). In mice there appears to be no relationship between replication time and isochore class, and in humans early replicating genes are located in all isochore classes except the very G+C poorest.

The Maintainence Of Isochores

The fact that sequences of very different G+C contents may be replicated simultaneously has some implications for the

Table 3. Gene replication time and isochore location

Gene	Isochore G+C	Replication time
Human		
Factor IX	39.7	L
β-globin	40.3	L
HPRT	41.5	E
c-mos	43.7	E
c-myc	46.7	E
Glucose-6-dehydrogenase	52.4	E
c-Ha-ras	53.7	E
α-globin	53.7	E
c-sis	53.7	E
Mouse		
IgK Variable	40.5	L
IgK Constant	42.0	E
β-globin	42.0	L
α-globin	49.1	E
Skeletal actin	49.1	L
c-abl	49.1	E

Replication time data comes from references cited in the methods section. Isochore location data comes from (5) for humans, and from (4) for mice.

maintainence of isochores. The mechanism by which isochores are maintained is the subject of considerable debate (13,22,26). The simplest and most cogent hypothesis has been put forward by Wolfe and colleagues (13,14). They proposed that different replicons are replicated in free nucleotide pools of different compositions which biases the pattern of mutation, thus producing replicons/isochores of different G+C contents. This very neatly explains the relationship between replication time and G+C content that was originally thought to exist, since it had been shown that the free nucleotide pool composition changed through the cell cycle (27,28). The fact that isochores of different G+C contents appear to replicate simultaneously poses something of a problem for this hypothesis, unless the free nucleotide pools are spatially heterogeneous. Paradoxically one is loath to drop the Wolfe/Li/Sharp hypothesis because it provides a very elegant explanation of the correlation between gene and isochore G+C contents, one of the observations which led to the original paradox. The correlation arises under this hypothesis, because although selection and DNA repair may vary across a replicon, all sequences in a replicon have the same pattern of misincorporation which is different to other replicons replicated under different conditions. Therefore sequences within a replicon are expected to have correlated compositions.

It should be appreciated that the conclusions reached via table 1 are only strictly applicable to the cell lines in which the gene replication times were studied. The conclusions do not neccessarily extend to the germ-line, which is the relevent tissue when discussing the origins and maintainence of isochores. It is possible that the pattern of replication is quite different in germ and somatic cell lines. However it is clear from the present work that in certain cell lines both G+C rich and G+C poor isochores replicate early and late in the cell cycle.

ACKNOWLEDGEMENTS

I would like to thank Paul Sharp, Peter Keightley, Bill Hill and two anonymous referees for their encouragement and comments on this manuscript. I would also like to thank the SERC for their financial assistance.

REFERENCES

1. Comings,D.E. (1978) *Ann Rev Genet.*, **12**, 25–46
2. Goldman,M.A. (1988) *Bioessays* **9**, 50–55
3. Holmquist,G.P. (1989) *J Mol Evol* **28**, 469–486
4. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science* **228**, 953–958
5. Aissani,B., D'Onofrio,G., Mouchiroud,D., Gardiner,K., Gautier,C. and Bernardi,G. (1991) *J Mol Evol* **32**, 493–503
6. Mouchiroud,D., Fichant,G. and Bernardi,G. (1987) *J Mol Evol* **26**, 198–204
7. Tobia,A.M., Schildkraut,C.L. and Maio,J.J. (1970) *J Mol Biol* **54**, 499–515
8. Flamm,W.G., Bernheim,N.J. and Brubaker,P.E. (1971) *Exp Cell Res* **64**, 97–104
9. Comings D.E (1971) *Exp Cell Res* **71**, 106–112
10. Hutchison,H.T. and Gartler,S.M (1973) *Tex Rep Biol and Med* **31**, 321–329
11. Holmquist,G.P., Gray,M., Porter,T. and Jordan,J. (1982) *Cell* **31**, 121–129
12. Bernardi,G. (1989) *Ann Rev Genet* **23**, 637–661
13. Wolfe,K.H., Li,W-H. and Sharp,P.M. (1989) *Nature* **337**, 283–285
14. Wolfe,K.H. (1991) *J Theor Biol* **149**, 441–451
15. Goldman,M.A., Holmquist,G.P., Gray,M.C., Caston,L.A. and Nag,A. (1984) *Science* **224**, 686–692
16. Calza,R.E., Eckhardt,L.A., DelGiudice,T. and Schildkraut,C.L. (1984) *Cell* **36**, 689–696
17. Iqbal,M.A., Plumb,M., Stein,G. and Schildkraut,C.L. (1984) *Proc Natl Acad Sci USA* **81**, 7723–7727
18. Iqbal,M.A., Chinsky,J., Didamo,V. and Schildkraut,C.L. (1987) *Nucleic Acid Res* **15**, 87–103
19. Hatton,K.S., Dhar,V., Brown,E.H., Mager,D. and Schildkraut,C.L. (1988) *Mol Cell Biol* **8**, 2149–2158
20. Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acid Res* **12**, 387–395
21. Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) *J Mol Evol* **27**, 311–320
22. Bernardi,G., Mouchiroud,D., Gautier,C. and Bernardi,G. (1988) *J Mol Evol* **28**, 7–18
23. McAlpine,P.J., Stranc,C.C., Boucheix,C. and Shows,T.B. (1990) *Cytogenet Cell Genet* **55**, 5–76
24. Davisson,M.T., Lalley,P.A., Doolittle,D.P. and Hillyard,A.L. (1990) *Cytogenet Cell Genet* **55**, 434–456
25. Hillyard,A.L., Doolittle,D.P., Davisson,M.T. and Roderick,T.H. (1991) *Mouse Genome* **89**, 16–30
26. Filipski,J., Salinas,J. and Rodier,F. (1987) *DNA* **6**, 109–118
27. McCormick,P.J., Danhauser,L.L., Rustim,Y.M. and Bertram,J.S. (1983) *Biochim Biophys Acta* **755**, 36–40
28. Leeds,J.M., Slabaugh,M.B. and Matthews,C.K. (1985) *Mol Cell Biol* **5**, 3443–3450

Letter to the Editor

The Effect of Constraint on the Rate of Evolution in Neutral Models With Biased Mutation

One of the central axioms of the neutral theory of molecular evolution is that the rate of evolution of a sequence is inversely related to its importance (KIMURA and OHTA 1974; KIMURA 1983). The reason is simply that important sequences have few potential neutral alleles. This prediction was made assuming that the pattern of mutation is uniform. However, as I will show below there are biased mutation patterns under which a reduction in the number of potential neutral alleles can lead to an increase in the rate of evolution. The approach taken is to consider the rate of evolution at single sites, comparing the rates at sites which have two, three or four neutral alleles (with all other alleles being very deleterious). Note that a site with one neutral allele does not undergo substitution.

Following WRIGHT (1969, chapter 3) let us consider a single site in a DNA molecule at which n alleles can segregate. Normally n will have a maximum value of four corresponding to the four bases which can occupy a site. Let the frequency of the ith allele be f_i, the probability of a mutation from allele i to j be U_{ij} and the probability of fixing a newly arising j mutant in a population of i be P_{ij}. Let $K_{ij} = 2NU_{ij}P_{ij}$ where N is the population size of a diploid organism. If we assume $NU_{ij} \ll 1$, then $K_{ij} = U_{ij}$ (KIMURA 1968). Furthermore the population will generally be monomorphic for one of the alleles and the expected frequency of allele i ($\bar{f}_i$) (*i.e.*, the relative amount of time for which the population is fixed for i) can be obtained from a consideration of the flux between the n alleles: *i.e.*, by solving $n - 1$ simultaneous equations of the form:

$$\Delta f_i = -f_i \sum_{j \neq i} K_{ij} + \sum_{j \neq i} f_j K_{ji} = 0. \quad (1)$$

The average rate of substitution at the site is then

$$R_n = \sum_i \bar{f}_i \sum_{j \neq i} K_{ij}. \quad (2)$$

The equilibrium frequencies for two, three and four allele systems are given by WRIGHT (1969, chapter 3). As one might expect the expressions for R_3 and R_4 are complex and do not yield readily to further analysis. However, there is an informative simplification that can be made. If we consider the mutation pattern symmetrical about allele 4 such that $U_{12} = U_{21} = U_{13} = U_{31} = U_{23} = U_{32} = U$, $U_{14} = U_{24} = U_{34} = U_{\cdot 4}$ and $U_{41} = U_{42} = U_{43} = U_{4\cdot}$ the rates of evolution in the following systems of two, three and four neutral alleles become:

$$R_4 = \frac{6U_{4\cdot}(U + U_{\cdot 4})}{3U_{4\cdot} + U_{\cdot 4}} \quad (3a)$$

$$R_3(*,*,4) = \frac{2U_{4\cdot}(U + 2U_{\cdot 4})}{2U_{4\cdot} + U_{\cdot 4}} \quad (3b)$$

$$R_3(*,*,*) = 2U \quad (3c)$$

$$R_2(*,*) = U \quad (3d)$$

where $*$ refers to one of alleles 1, 2 or 3. Using the expressions in (3) it is simple to show that

$$R_3(*,*,*) > R_4 \quad \text{when} \quad U > 3U_{4\cdot} \quad (4a)$$

$$R_2(*,*) > R_4 \quad \text{when} \quad U > \frac{6U_{4\cdot}U_{\cdot 4}}{U_{\cdot 4} - 3U_{4\cdot}} \quad (4b)$$

$$\text{and} \quad U_{\cdot 4} - 3U_{4\cdot} > 0$$

$$R_2(*,*) > R_3(*,*,4) \quad \text{when} \quad U > 4U_{4\cdot}. \quad (4c)$$

In other words, there are mutation patterns under which a reduction in the number of potential neutral alleles at a site leads to an increase in the average rate of evolution. Note, however, how biased the mutation patterns must be for this to occur. Essentially the elimination of allele 4 in each case leads to an increase in the time for which the site is occupied by alleles 1, 2 and 3. So if the mutation rate between alleles 1, 2 and 3 is high, elimination of allele 4 increases the overall flux. As an extreme example, consider a mutation pattern which is very biased so the site is almost permanently fixed for allele 4. Clearly there can be little substitution. Removal of allele 4 from the system allows the other alleles to occupy the site alternately thereby increasing the rate of substitution at the site. Of course eliminating an allele from a system also removes several mutation pathways (*e.g.*, those between alleles 1 and 4). This explains why the conditions under which three-allele systems evolve faster than four-allele systems are rather less stringent than those for other comparisons.

It turns out that some of the relationships can be applied to mutation patterns in general, not just those in which $U_{12} = U_{21} \ldots$, etc., by using average mutation rates. Thus if we define

$$U = \frac{(U_{12} + U_{21} + U_{23} + U_{32} + U_{13} + U_{31})}{6} \quad (5a)$$

$$U_{\cdot 4} = \frac{U_{14} + U_{24} + U_{34}}{3}, \quad (5b)$$

$$U_{4\cdot} = \frac{U_{41} + U_{42} + U_{43}}{3}, \quad (5c)$$

the relationships 4a and 4c almost always hold. This was demonstrated by sampling mutation rates at random from a uniform distribution. In less than 2% of the 50,000 mutation patterns sampled was either $R_3(*,*,*) > R_4$ and $U < 3U_{4\cdot}$, or $R_2(*,*) > R_3(*,*,4)$ and $U < 4U_{4\cdot}$. However, in 20–30% of the cases where the mutation pattern inequality held (*i.e.*, $U > 3U_{4\cdot}$ or $U > 4U_{4\cdot}$) the rate inequality did not hold.

A general expression for the case when a two-allele system evolves faster than a four-allele system alluded this investigator. Suffice it to say that the conditions under which a two-allele system will evolve faster than a four-allele system are likely to be rather more stringent than the conditions under which a three-allele system evolves faster than a four-allele system. This is because in reducing a four-allele site to a two-allele site one loses five of the six possible mutation pathways, compared to the three that are lost in the reduction to a three-allele site.

So far the analysis has dealt with only single sites. Although it is evident that a sequence with fewer potential alleles can evolve faster than one with more potential neutral alleles, the conditions under which this occurs are more restricted for a number of reasons. First, a reduction in the number of potential alleles may be brought about by an increase in the number of one allele sites, not a reduction, say of four-allele sites to three-allele sites. Second, the inequalities 4a and 4c are partially exclusive. For a three-allele site to evolve faster than a four-allele site requires that allele 4 is removed to form the three-allele site. However, a three-allele site must contain allele 4 if a two-allele site is to evolve faster than it. Finally different alleles may be removed from different sites: *e.g.*, at one three-allele site allele 4 may be removed, whereas at another, allele 3 may have been eliminated. However, it is worth appreciating that a sequence of say three-allele sites, with alleles removed at random, can evolve faster than a sequence of four-allele sites. To see this, consider the rate at four- and three-allele sites under the symmetrical mutation pattern used above, when $U_{4\cdot} = 0$. The rate of evolution at three- and four-allele sites containing allele 4 is zero (3a, 3b), whereas it is nonzero at the three-allele site missing allele 4 (3c). Therefore, the average rate of evolution over a set of three-allele sites is greater than that at a four-allele site. Of course the mutation patterns must be very biased for this to occur; although if such patterns did exist they would overcome the partial exclusivity of inequalities 4a and 4c.

Hence in both sequences and sites, a reduction in the number of potential neutral alleles can increase the rate of evolution. However, this does depend on there being extreme bias in the mutation pattern. Although it is unclear what mutation patterns one might expect to find in the natural world, the very limited data available suggest that the patterns are not very biased (GOJOBORI, LI and GRAUR 1982; LI, WU and LUO 1984).

The present findings suggest some caution should be exercised in deducing the action of positive natural selection when the rate of substitution at nonsilent sites exceeds that at silent sites, especially when few sites are involved or there is extreme codon bias; codon bias being an indicator of biased mutation.

I am very grateful to JOHN BROOKFIELD, EDDY HOLMES, ANDY LEIGH BROWN, IAN HASTINGS, BILL HILL and W.-H. LI for helpful discussions and criticism. I also thank the Science and Engineering Research Council for their financial support.

ADAM EYRE-WALKER
Institute of Cell, Animal and
Population Biology
University of Edinburgh
Edinburgh, EH9 3JT, Great Britain

LITERATURE CITED

GOJOBORI, T., W.-H. LI and D. GRAUR, 1982 Patterns of mutation in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press.

KIMURA, M., and T. OHTA, 1974 On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA **71:** 2848–2852.

LI, W.-H., C.-I. WU and C.-C. LUO, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21:** 58–71.

WRIGHT, S., 1969 *Evolution and the Genetics of Populations. Vol. II. The Theory of Gene Frequencies.* University of Chicago Press.

Genet. Res., Camb (1992), **00**, *pp.* 000–000 *With 5 text-figures Printed in Great Britain*

The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals

ADAM EYRE-WALKER
Institute of Cell Animal and Population Biology, University of Edinburgh, Edinburgh, EH9 3JT, Great Britain

(Received 23 October 1991 and in revised form 26 February 1992)

Summary

It has been suggested that isochores are maintained by mutation biases, and that this leads to variation in the rate of mutation across the genome. A model of DNA replication is presented in which the probabilities of misincorporation and proofreading are affected by the composition and concentration of the free nucleotide pools. The relationship between sequence $G+C$ content and the mutation rate is investigated. It is found that there is very little variation in the mutation rate between sequences of different $G+C$ contents if the total concentration of the free nucleotides remains constant. However, variation in the mutation rate can be arbitrarily large if some mismatches are proofread and the total concentration of free nucleotides varies. Hence the model suggests that the maintenance of isochores by the replication of DNA in free nucleotide pools of biased composition does not lead *per se* to mutation rate variance. However, it is possible that changes in composition could be accompanied by changes in concentration, thus generating mutation rate variance. Furthermore, there is the possibility that germ-line selection could lead to alterations in the overall free nucleotide concentration through the cell cycle. These findings are discussed with reference to the variance in mammalian silent substitution rates.

1. Introduction

There is considerable variation in the rate of silent substitution within all mammalian species so far studied (Li *et al.* 1987, Wolfe *et al.* 1989; Bulmer *et al.* 1991). Although selection is known to act at silent sites in a number of organisms (review by Sharp, 1989) there is little evidence of it doing so in mammals (Eyre-Walker, 1991). This suggests that the variation in silent substitution rate is caused by differences in the rate of mutation between genes.

Recently two mechanisms by which the mutation rate might come to vary across the genome have been proposed (Filipski 1988; Wolfe *et al.* 1989; Wolfe, 1991). Filipski (1988) suggested that the silent substitution rate variance was caused by differences in the level of repair between genes. Such variation in repair across the genome has been observed for various types of DNA damage, but not for base mismatches (Bohr *et al.* 1987).

The idea of Wolfe *et al.* (1989) came out of attempts to explain the isochore structure of the vertebrate genome (review by Bernardi 1989). They noted three observations: that the rate and pattern of mutation during DNA replication was dependent upon the free nucleotide concentrations (Phear and Meuth, 1989*a*, *b*; Meuth, 1989; Kohalmi *et al.* 1991), that the relative concentrations of the free nucleotides varied through the cell cycle (McCormick *et al.* 1983; Leeds *et al.* 1985) and that different DNA sequences were replicated at different times. As a consequence early and late replicating DNA should come to differ in their $G+C$ contents, and isochores (blocks of DNA with homogeneous $G+C$ content) should form. Wolfe *et al.* (1989) further suggested that isochores differ in their mutation rate as a result of this process.

The possible involvement of isochores and their maintenance in the generation of mutation rate differences is suggested by the fact that much of the variation in silent substitution rate can be explained by $G+C$ content (Filipski, 1988; Wolfe *et al.* 1989, Ticher & Graur, 1989; Bulmer *et al.* 1991), and that the $G+C$ contents of genes are correlated to that of the isochore in which they reside (Bernardi *et al.* 1985; Ikemura & Aota, 1987; Aissani *et al.* 1991). In quantitative terms there is approximately a twofold variation in silent substitution rate explained by $G+C$ content, with the overall variation being somewhat larger.

This paper considers whether the replication of DNA in free nucleotide pools of different composition and concentration produces variation in the rate of

mutation. A recent theoretical analysis of this problem by Wolfe (1991) suggested that this was indeed generally the case. In this paper his model is extended to more fully quantify the variation in mutation rate produced. The effects of changing the overall concentrations of free nucleotides are also investigated.

2. The Model

Let us consider a very general model of DNA replication, in which a free nucleotide collides with the DNA polymerase and is then either incorporated into the growing DNA sequence or rejected. The probability that the base, z (where z can be A, T, C or G), collides with the polymerase is P_z, the relative concentration of the free tri-phosphate nucleotide dZTP. Given that a collision has occurred let the probability that the nucleotide is subsequently incorporated be j if the nucleotide is correct and k_a if it is incorrect (where a is i for a type I mismatch and ii for a type II mismatch). Type I mismatches are those which if left unproofread and unrepaired give mutations altering G–C content: i.e. $C \leftrightarrow T$ and $A \leftrightarrow G$ transitions, and $C \leftrightarrow A$ and $G \leftrightarrow T$ transversions. If type II mismatches are left unaltered they give $C \leftrightarrow G$ and $A \leftrightarrow T$ mutations. The frequency with which y is misincorporated instead of z is then:

$$T_{zy} = k_i P_y / (jP_z + k_i P_m + k_i P_y + k_{ii} P_n)$$

Type I mutation, (1*a*)

$$T_{zy} = k_{ii} P_y / (jP_z + k_i P_m + k_i P_n + k_{ii} P_y)$$

Type II mutation, (1*b*)

where bases m and n also form mismatches. Under biologically realistic conditions where the probability of incorporating a mismatch is very small (i.e. $k_a \ll j$) equations 1*a* and 1*b* reduce to the form used by Wolfe (1991): $T_{zy} = \alpha P_y / P_z$ where α is a constant. It would of course be possible to further divide type I mutations into transitions and transversions, and to treat them separately. However, transitions and type I transversions turn out to have the same dynamics in this model, and were thus treated together.

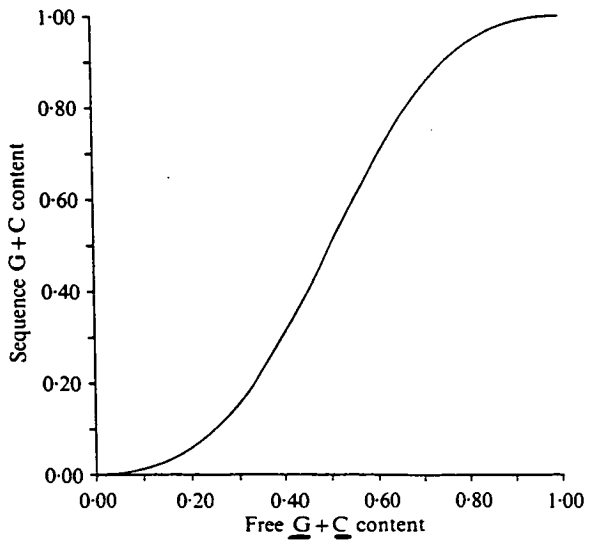


Fig. 1. The equilibrium sequence $G+C$ content.

Once a mismatch is incorporated it may be removed by proofreading or mismatch repair. We will only consider elimination by proofreading in this model since the interaction between repair and replication is likely to be non-trivial. Let the probability of not proofreading a mismatch be N_a. Then the probability with which z will mutate to y per cycle of replication will be:

$$M_{zy} = T_{zy} N_a \quad \text{where } a \text{ is } i \text{ or } ii \text{ as appropriate.} \quad (2)$$

(i) *Equilibrium G+C content*

Now consider the change in the frequency of a nucleotide in a sequence each time the sequence is replicated. For instance the change in the frequency of G:

$$\Delta f_G = -f_G(M_{GC} + M_{GA} + M_{GT}) + f_C M_{CG} + f_A M_{AG} + f_T M_{TG}, \quad (3)$$

where f_z is the frequency of nucleotide z in the sequence being replicated.

Let us make the simplifying assumption that the concentrations of dCTP and dGTP are the same (i.e. $P_C = P_G$) and the $P_T = P_A$. Quite clearly $f_C = f_G$ and $f_T = f_A$, so $P_A = (1-2P_G)/2$ and $f_A = (1-2f_G)/2$. If we let $p = P_G$ and $f = f_G$ then (3) simplifies to

$$\Delta f = 2k_i N_i \left[\frac{f(2p-1)}{2pB + 2k_i} + \frac{p(2f-1)}{2pB - j - k_{ii}} \right], \quad (4)$$

where $B = j - 2k_i + k_{ii}$. Solving $\Delta f = 0$ to get the equilibrium frequency of G (or C) in a sequence, gives under biologically realistic conditions (i.e. $k_a \ll j$),

$$f = 2p^2/(8p^2 - 4p + 1). \quad (5)$$

So the equilibrium frequency of G (or C) in the sequence is independent of proofreading and the probabilities of incorporation; essentially because of the symmetry in the model. As expected, when the free nucleotide pool is either all $A+T$, all $G+C$ or half and half ($p = 0, \frac{1}{2}$ and $\frac{1}{4}$) the equilibrium sequence $G+C$ content is equal to the pool $G+C$ content. Figure 1 shows the equilibrium frequency of $G+C$ plotted against the free nucleotide concentration of $G+C$. The relationship between the two variables is sigmoidal, so that at intermediate $G+C$ contents small changes in the free nucleotide concentrations have large effects on the equilibrium $G+C$ content of the sequence (e.g. a sequence of 80% $G+C$ is replicated in a pool of only 70% $G+C$). This non-

linearity arises because the probability of misincorporating a nucleotide is dependent on the probability of a nucleotide being incorrect per collision, and the number of collisions that occur, both of which are dependent upon the pool composition.

(ii) *Proofreading*

The probability that a mismatch will be proofread depends on how long it takes to replicate the next position in the sequence. Once replication of the distal base has occurred the mismatch cannot be proofread, and must instead be corrected by other mechanisms which we shall not consider here. Let the average probability of proofreading a mismatch between collisions by V_a (where a is i or ii for proofreading type I and type II mutations respectively); and let us imagine that the polymerase is waiting to replicate nucleotide z distal to a mismatch. Assuming that proofreading cannot occur until the polymerase is ready to incorporate the next nucleotide, the probability that neither replication not proofreading has occurred after t collisions is $(1-V_a)^{t+1}(1-jP_z)^t$. So the probability that proofreading never occurs when a mismatch is followed by z is

$$jP_z(1-V_a)\sum_{t=0}^{\infty}((1-V_a)(1-jP_z))^t = \frac{P_z}{E_a+P_z}, \quad (6)$$

where $E_a = V_a/(j(1-V_a))$. Therefore the average probability of not repairing a mismatch is

$$N_a = \sum \frac{f_z P_z}{E_a+P_z} \quad (7)$$

as given by Wolfe (1991). In a sequence at equilibrium this becomes

$$N_a = \frac{p(2p-1)(8p^2-4p+1)+E_a(12p^2+6p-1)}{(8p^2-4p+1)(p(2p-1)+E_a(2E_a-1))} \quad (8)$$

E_a is a measure of the proofreading stringency. When there is no proofreading $E_a = 0$, and when proofreading is stringent $E_a \to \infty$. It is worth noting here that as proofreading becomes very stringent (i.e. $E_a \to \infty$)

$$(1+4E_a)N_a \to (24p^2-12p+2)/(8p^2-4p+1). \quad (9)$$

(iii) *Mutation rate*

The average mutation rate per nucleotide per replication of a sequence is

$$\mu = f_G \sum_{z \neq G} M_{Gz} + f_C \sum_{z \neq C} M_{Cz} + f_A \sum_{z \neq A} M_{Az} + f_T \sum_{z \neq T} M_{Tz}, \quad (10)$$

which for a sequence at equilibrium simplifies under biologically realistic conditions (i.e. $k_a \ll j$), to

$$\mu = \underbrace{\frac{8p(1-2p)(k_i N_i)}{j(8p^2-4p+1)}}_{\text{Type I mutations}} + \underbrace{\frac{k_{ii}N_{ii}}{j}}_{\text{Type II mutations}}. \quad (11)$$

Since we are interested in the relative, rather than absolute, mutation rate, let us divide μ by the rate of mutation in an equilibrium sequence of 50% $G+C$ content: i.e. $2k_i/(1+4E_i)+k_{ii}/(1+4E_{ii})$. The relative rate of mutation is

$$R = \frac{w_i(1+4E_i)4p(1-2p)N_i}{8p^2-4p+1} + w_{ii}(1+4E_{ii})N_{ii}, \quad (12)$$

where

$$w_i = \frac{2k_i/(1+4E_i)}{2k_i/(1+4E_i)+k_{ii}/(1+4E_{ii})},$$

and

$$w_{ii} = \frac{k_{ii}/(1+4E_i)}{2k_i/(1+4E_i)+k_{ii}/(1+4E_{ii})},$$

w_i and w_{ii} are the proportion of mutations which are of type I and type II in a sequence of 50% $G+C$ content.

3. Results

(i) *Type II mutations*

Let us consider the frequency of type II mutations alone. By setting $k_i = 0$ in equation 11 we obtain

$$R_{ii} = N_{ii}(1+4E_{ii}) \quad (13)$$

an expression which is solely dependent upon p, the concentration of dGTP or dCTP, and E_{ii} a measure of the proofreading stringency. R_{ii} is plotted against the equilibrium $G+C$ content of a sequence in figure 2 (dashed line) for various levels of proofreading. Remember that as proofreading becomes stringent ($E_a \to \infty$) the expression $(1+4E_a)N_a$ becomes independent of E_a (see equation 9).

When there is no proofreading sequences of all $G+C$ contents have the same rate of type II transversion mutations. However as the stringency of proofreading increases so sequences of extreme $G+C$ content have higher mutation rates than sequences of intermediate $G+C$ content. The reason: at extreme $G+C$ contents the polymerase only has to try on average two nucleotides before the correct one is found, compared to the four that must be tested at intermediate $G+C$ contents. Therefore, the probability of not proofreading (when proofreading is stringent) at extreme $G+C$ contents is twice that at intermediate $G+C$ contents. This means the mutation rate is twice as great at extreme $G+C$ contents.

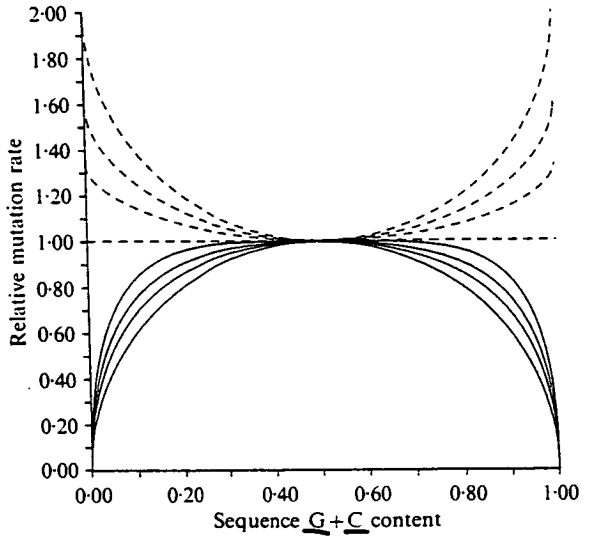


Fig. 2. The effect of proofreading on the relative rates of type I and type II mutations. Figure shows the relative mutation rates for type I (solid lines) and type II (dashed lines) mutations for various levels of proofreading. In each case from bottom to top E_a, the strength of proofreading, is 0, 0·25, 0·75 and ∞. Under these values of E_a the probability of proofreading a mismatch in a sequence of 50% $G+C$ content is 0, 0·5, 0·75 and $\rightarrow 1$ respectively.

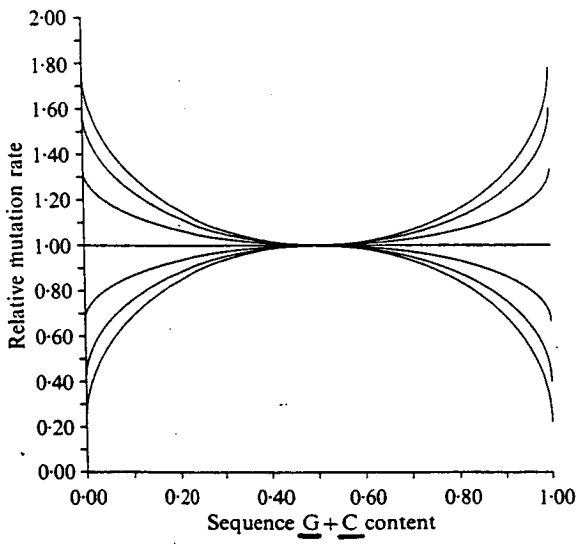


Fig. 3. The relative mutation rate for a sequence undergoing unproofread type I mutation and stringently proofread type II mutation. The curves represent different ratios of type I and type II mutations. The ratio of unproofread type I mutations to proofread type II mutations in a sequence of 50% $G+C$ content is, from top to bottom, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4 and 1:8.

(ii) *Type I mutations*

Now consider type I mutations alone. Setting $k_{tt} = 0$ in equation 12 we obtain

$$R_t = \frac{N_t(1+4E_t)4p(1-2p)}{(8p^2-4p+1)} \qquad (14)$$

an expression dependent upon p and E_t. Figure 2 (solid lines) shows R_t plotted against the equilibrium sequence $G+C$ content. For all levels of proofreading sequences of intermediate $G+C$ content always have higher rates of transition (and type I transversion) mutations than sequences of extreme $G+C$ content. The reason is that sequences of extreme $G+C$ content are replicated in nucleotide pools which are deficient in the free nucleotides required to make transition mismatches. As the stringency of proofreading increases the curves become much flatter so that eventually the differences in mutation rates of sequences at 50% and 70% (or 30%) $G+C$ content are negligible. Flattening occurs because proofreading elevates the mutation rate of sequences at extreme $G+C$ content compared to sequences at intermediate $G+C$ content (see figure 2 dashed lines).

(iii) *The overall mutation rate*

In graphical terms the overall mutation rate relative to that in a sequence of 50% $G+C$ content (equation 12), is simply the average of the curves shown in figure 2. For instance if at 50% $G+C$ content there are two unproofread type I mutations to every proofread type II mutation, then two of the bottom curves in figure 2 should be added to the top curve and the result divided by three. Thus the maximum variation in the mutation rate is achieved when all mutations arise via unproofread type I mismatches or stringently proofread type II mismatches.

Let us consider sequences in which both type I and II mutations can occur starting with the cases when type II mutations are not proofread. Since the rate of unproofread type II mutations is independent of the sequence $G+C$ content sequences of intermediate $G+C$ content will always have higher mutation rates than sequences of extreme $G+C$ content whether or not the type I mismatches are proofread. Qualitatively the curves will be similar to those given in figure 2 (solid lines) only flatter. Quantitatively the gradient at each point will be λ times the original gradient and the curve will bisect the abscissa at λ, where λ is the fraction of mutations which are type-I mutations in an equilibrium sequence of 50% $G+C$ content.

Consider now the case when type II mutations are stringently proofread and type I mutations are not proofread at all. Under these conditions equation 12 reduces to

$$R = \frac{8w_t p(1-2p)+2w_{tt}(12p^2-6p+1)}{8p^2-4p+1}, \qquad (14)$$

which can be shown to have a local maximum when more than half ($w_t > \frac{1}{2}$), and a local minimum when less than half ($w_{tt} < \frac{1}{2}$), of the mutations in a sequence of 50% $G+C$ content are type I mutations. Equation 14 is plotted against the equilibrium sequence $G+C$ content in figure 3. The rate of mutation is only weakly dependent upon sequence $G+C$ content unless type I mutations are much more common than type II

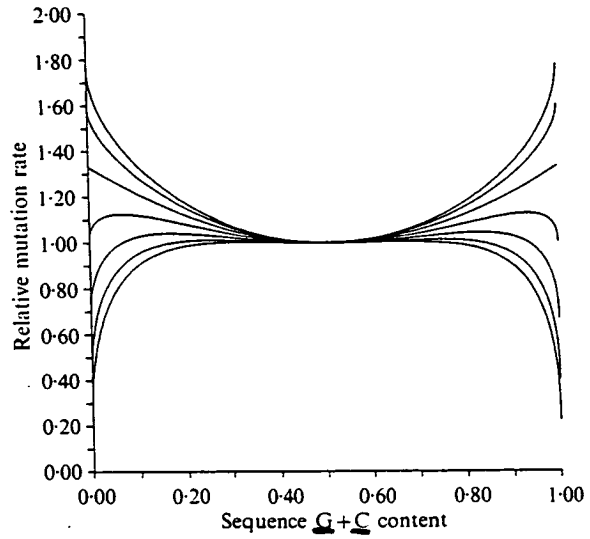


Fig. 4. The relative mutation rate for a sequence undergoing stringently proofread type I and type II mutation. Each curve represents a different ratio of type I and type II mutations. The ratio of proofread type I mutations to proofread type II mutations in a sequence of 50% $G+C$ content is, from top to bottom, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4 and 1:8.

mutations, or vice versa. This is even more the case when all mismatches are stringently proofread (figure 4).

(iv) *Changing the overall concentration of free nucleotides*

Up till now we have only considered the effect of varying the relative concentrations of the free nucleotides. However variations in the overall concentration of free nucleotides will affect the probability of proofreading, under this model, if the polymerase is not saturated by the free nucleotides.

Let the rate at which correct nucleotides collide with the polymerase be α and the rate at which proofreading occurs be β. The time t between collisions is distributed exponentially and the probability of proofreading during a particular time interval is a Poisson process with mean βt. Therefore the probability of not proofreading a mismatch between collisions is

$$1-V=\int_0^\infty e^{-\beta t}\alpha e^{-\alpha t}\,dt=\frac{\alpha}{\alpha+\beta} \qquad (18)$$

and $E=\beta/(j\alpha)$. α is proportional to the overall concentration of the nucleotides, so doubling the concentration leads to a doubling of α. When proofreading is slow (β and E small) changes in the overall concentration of free nucleotides have little effect on the probability of proofreading (see equation 7). However, when proofreading is stringent (β and E large) the probability of proofreading becomes a linear function of the free nucleotide concentration: i.e. $N\approx\alpha j(12c^2-6c+1)/(\beta(16c^2-8c+2))$. Hence it is possible for fluctuations in the overall concentration of nucleotides to cause large variations in the probability of proofreading, and the rate of mutation.

4. Discussion

It is important to appreciate that the $G+C$ values of the model do not necessarily correspond to the $G+C$ contents of the real world in either value or scale. This is because the mutation pattern was highly simplified and several assumptions were made about the free nucleotide pools. However, the general lack of mutation rate variance that DNA replication produces across sequences of different $G+C$ contents in the model is expected to be a robust result. Note in particular how the rates of misincorporation and proofreading act against each other for type I mutations. If there is an increase in the number of free nucleotides the polymerase has to try before finding the correct one, then the rate of misincorporation increases. However, the rate of DNA replication slows and the probability of proofreading rises. Also note how the relationship between $G+C$ content and the frequency of type II mutations opposes or dampens the relationship for type I mutations.

(i) *Isochore $G+C$ content*

Isochore $G+C$ contents vary from ~ 38% to ~ 55% in humans, with the range being somewhat narrower in mice and rats (Bernardi *et al.* 1985). Over this sort of range, assuming that the $G+C$ content scale of the model corresponds roughly to that in the real world, there is very little variation in the rate of mutation unless the total concentration of free nucleotides changes. Note in particular that if the frequency of type I and type II misincorporations are within an order of magnitude of each other the variation in the mutation rate across sequences of all $G+C$ contents is extremely limited. The variation appears to be too limited to explain either the $G+C$ content related, or total, silent substitution variance.

(ii) *The mutation pattern*

There is very little data on the pattern of misincorporation in mammals. Some data are available from the analysis of substitution patterns in pseudogenes (Gojobori *et al.* 1982; Li *et al.* 1984). However, the probability of repair appears to differ amongst mismatches (Brown and Jiricny, 1988, 1989) and across the genome (Bohr *et al.* 1987; Filipski, 1988) so substitution patterns may not correspond to the patterns of misincorporation.

(iii) *Overall changes in the free nucleotide concentrations*

The analysis above suggests that the maintenance of isochores by replication in biased free nucleotide

pools does not *per se* generate mutation rate variance. However, changes in the overall concentration of free nucleotides can affect the mutation rate (under the present model) if at least one type of mismatch is proofread and the polymerase is unsaturated by free nucleotides. These conditions appear to be met in the real world, although proofreading may be restricted to certain mismatches and only one polymerase (Matthews & Slabaugh 1986; Kunkel *et al.* 1987; Meuth, 1989).

Changes in the overall concentration of free nucleotides might come about as a consequence of variations in the composition maintaining isochores, or via germ-line selection (Hastings 1989). Germ-line selection might lead to changes in the concentration because genes expressed in a tissue appear to be replicated early in that tissue (Holmquist, 1987, 1989; Goldman, 1988). So a decrease in the overall concentration might be favoured by selection to reduce the mutation rate in those genes being expressed in the germ-line; or alternatively an increase in the concentration might be selected for to increase the dosage of the early replicating genes by speeding up replication.

(iv) *Replication and repair*

The range of gene $G+C$ contents, especially at the 3rd position thereof ($\sim 35\%$ to $\sim 95\%$ in humans), is considerably greater than that across isochores. It seems likely that the mechanism responsible for generating this vast range of gene $G+C$ contents is also in part responsible for the variation in silent substitution rate. In particular the case has been recently made that biased DNA repair is the responsible party (Filipski, 1988). Although the interaction between repair and replication is non-trivial it seems unlikely that the maintenance of isochores by DNA replication would be a major contributor to mutation rate variance unless there are large changes in the total free nucleotide pool concentrations. Such overall changes in concentration might accompany changes in free nucleotide pool composition, or be produced by germ-line selection.

I am very grateful to Ken Wolfe for giving me access to his work prior to publication, for helpful discussion and encouragement. I thank Bill Hill, Peter Keightley and Prof A. J. Jeffreys for comments on this manuscript and to SERC for their financial assistance.

References

Aissani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K. & Bernardi, G. (1991). The compositional properties of human genes. *Journal of Molecular Evolution* **32**, 493–503.

Bernardi, G. (1989). The isochore organization of the human genome. *Annual Review of Genetics* **23**, 637–661.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. (1985). The mosaic genome of warm blooded vertebrates. *Science* **228**, 953–958.

Bohr, V. A., Phillips, D. H. & Hanawalt, P. C. (1987). Heterogeneous DNA damage and repair in the mammalian genome. *Cancer Research* **47**, 6426–6436.

Brown, T. C. & Jiricny, J. (1988). Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**, 705–711.

Brown, T. C. & Jiricny, J. (1989). Repair of base–base mismatches in simian and human cells. *Genome* **31**, 578–583.

Bulmer, M., Wolfe, K. H. & Sharp, P. M. (1991). Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationships of the mammalian orders. *Proceedings of the National Academy of Science USA* **88**, 5974–5978.

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

Filipski, J. (1988). Why the rate of silent substitution is variable within the *a* vertebrate's genome. *Journal of Theoretical Biology* **134**, 159–164.

Gojobori, T., Li, W.-H. & Graur, D. (1982). Patterns of mutation in pseudogenes and functional genes. *Journal of Molecular Biology* **18**, 360–369.

Goldman, M. A. (1988). The chromatin domain as a unit of gene regulation. *Bioessays* **9**, 50–55.

Hastings, I. M. (1989). Potential germline competition in animals and its evolutionary implications. *Genetics* **123**, 191–197.

Holmquist, G. P. (1987). Role of replication time in the control of tissue specific gene expression. *American Journal of Human Genetics* **40**, 151–173.

Holmquist, G. P. (1989). Evolution of chromosome bands: molecular ecology of non-coding DNA. *Journal of Molecular Evolution* **28**, 469–486.

Ikemura, T. & Aota, S. (1988). Global variation in $G+C$ content along vertebrate genome DNA: possible correlation with chromosome band structures. *Journal of Molecular Biology* **203**, 1–13.

Kohalmi, S. E., Glattke, M., McIntosh, E. M. & Kunz, B. A. (1991). Mutational specificity of DNA precursor pool imbalances in yeast arising from deoxycytidylate deaminase deficiency or treatment with thymidylate. *Journal of Molecular Biology* **220**, 933–946.

Kunkel, T. A., Sabatino, R. D. & Bambara, R. A. (1987). Exonucleolytic proofreading by calf thymus DNA polymerase delta. *Proceedings of National Academy Science USA* **84**, 4865–4869.

Leeds, J. M., Slabaugh, M. B., Mathews, C. K. (1985). DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and the cytoplasm of mammalian cells. *Molecular and Cellular Biology* **5**, 3443–3450.

Li, W.-H., Wu, C.-I. & Luo, C.-C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution* **21**, 58–71.

Li, W.-H., Tanimura, M. & Sharp, P. M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *Journal of Molecular Evolution* **25**, 330–342.

Matthews, C. K. & Slabaugh, M. B. (1986). Eukaryotic DNA metabolism: are deoxyribonucleotides channelled to replication sites? *Experimental Cell Research* **162**, 285–295.

McCormick, P. J., Danhauser, L. L., Rustim, Y. M. & Bertram, J. S. (1983). Changes in ribo- and deoxyribonucleoside triphosphate pools within the cell cycle of a synchronised mouse fibroblast cell line. *Biochima Biophysica Acta* **755**, 36–40.

Meuth, M. (1989). The molecular basis of mutations induced

by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. *Experimental Cell Research* **181**, 305–316.

Phear, G. & Meuth, M. (1989*a*). A novel pathway for transversion mutation induced by dCTP misincorporation in a mutator strain of CHO cells. *Molecular and Cellular Biology* **9**, 1810–1812.

Phear, G. & Meuth, M. (1989*b*). The genetic consequences of DNA precursor pool imbalance: sequence analysis of mutations induced by excess thymidine at the hamster *aprt* locus. *Mutation Research* **214**, 201–206.

Sharp, P. M. (1989). Evolution at 'silent' sites in DNA. In *Evolution and Animal Breeding: Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson.* (ed. W. G. Hill, T. F. C. Mackay). Wallingford CAB International 1989, pp. 23–31.

Ticher, A. & Grauer, D. (1989). Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein coding genes. *Journal of Molecular Evolution* **28**, 286–298.

Wolfe, K. (1991). Mammalian DNA replication: mutation biases and the mutation rate. *Journal of Theoretical Biology* **149**, 441–451.

Wolfe, K., Sharp, P. M. & Li, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285.