J Mol Evol (1994) 39:123–128

Journal of Molecular Evolution

Estimating Selection by Comparing Synonymous and Substitutional Changes

J. Maynard Smith

School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom

Received: 14 January 1994 / Revised and accepted: 9 March 1994

Abstract. A higher ratio of substitutional to synonymous changes in between-species than in within-species comparisons has been taken as evidence for positive selection changing amino acids. A model is presented in which a difference of this kind arises as a result of purely neutral mutations, provided that the "species" compared are sufficiently different to approach a steady state between forward and backward mutation. In *Neissseria,* substitutions are twice as frequent, relative to synonymous changes, in between-species comparisons: it is shown that the data are consistent with the neutral model. The argument does not invalidate evidence for positive selection, for example in *Drosophila,* when the species compared are fairly similar.

Key words: Substitutions — Synonymous mutations — Estimating selection

Introduction

Following Kreitman (1983), it has become popular to argue that, if the ratio of substitutional to synonymous changes is greater in between-species than in within-species comparisons, then there has been selection favoring amino acid changes. The purpose of this paper is to describe a context in which this argument does not hold. A simple neutral model is presented which predicts that, if the difference between species is large, the between-species ratio should be approximately twice the within-species ratio, and data from the bacterial genus, *Neisseria,* are shown to be consistent with the model. The argument does not invalidate Kreitman's evidence for positive selection at the ADH locus of *Drosophila,* but the potential dangers of using his method uncritically should be recognised.

Let R_W be the ratio of substitutions to synonymous changes within a species and R_b the same ratio when comparing different species. Section 2 describes a neutral model which predicts $R_b \simeq 2R_w$. Essentially, the reason for this is as follows. R_w depends on the relative rates of substitutional and synonymous mutations. Since the latter are usually transitions, and the former often transversions, this causes a relative increase in the frequency of synonymous changes. The model assumes that, in the between-species comparisons, the two populations have reached a steady state between forward and backward mutation: they have reached saturation. Mutation rates, therefore, are no longer relevant, and R_b depends only on the number of possible neutral alternatives of the two kinds. The argument does not invalidate Kreitman's evidence for selection, because, in *Drosophila,* species are very far from saturation.

The rest of this paper spells out this argument and illustrates it with data from *Neisseria.* The model is presented in section 2. Section 3 briefly describes the source of the *Neisseria* data. Section 4 analyzes synonymous differences within species. The data are used to estimate the relative rates of transitions and transversions, and it is shown that the data can be fully accounted for, assuming this relative rate, and the observed amino acid frequencies and codon biases.

Section 5 analyzes synonymous differences between species. The differences approach, but do not quite reach, those predicted by the steady-state model.

Section 6 considers nonsynonymous changes. The ratio of amino acid to synonymous changes is twice as great in the between-species as in the within-species

comparisons. This is the result predicted by the model: it is shown that the data are consistent with the assumptions of the model.

2. The Model

Consider the following simple model. A gene consists of N codons, of which a fraction p code for two selectively equivalent amino acids. The average mutation rate of one amino acid to the another is m_a. A fraction S of the $3N$ nucleotides can exist as either of two synonymous bases, between which the average mutation rate is m_s. R_w is the expected ratio of amino acid to synonymous changes within species. In the short term, this will depend on the mutation rates. That is:

$$R_w = pm_a/3Sm_s$$

In the long term, when the steady state has been reached, approximately $Np/2$ of amino acid sites will differ between strains, and one-half of the potentially polymorphic synonymous sites will differ. Hence the ratio of substitutions to synonymous changes is

$$R_b = p/3s(1 - p/2)$$

Hence, since p is small, $R_b/R_w \simeq m_s/m_a$.

It is shown in section 6 that, for *Neisseria*, $m_s \simeq 2m_a$. Hence the model predicts that $R_b \simeq 2R_w$.

If the above model is to be a reasonable approximation, two things must be true:

1. At most loci at which an amino acid change has occurred, only one neutral alternative is possible.
2. Most amino acid changes were brought about by a single base substitution.

It is shown in section 6 that the data are consistent with these two assumptions.

3. The PBP2 Gene in *Neisseria*

The PBP2 gene in *Neisseria* codes for an enzyme concerned with cell-wall synthesis. In susceptible strains, penicillin binds to this enzyme and inactivates it. The coding region of the gene, specifying 581 amino acids, has been sequenced for 24 strains from seven "species" (Spratt et al. 1992). Some of these strains (e.g., susceptible strains of the two pathogenic species, *N. meningitidis* and *N. gonorrhoeae*) are very similar to one another, differing at <2% of nucleotides. Other strains (e.g., *N. meningitidis* and the penicillin-resistant commensal *N. flavescens*) differ by approximately 25% of nucleotides. In still other strains (in particular, resistant strains of the two pathogens), the gene consists of a mosaic of regions similar to the sensitive species, and regions similar to one of the naturally resistant commensals: these are thought to have arisen by horizontal transfer of short regions of DNA by transformation.

The complete data set affords an opportunity to compare the types of difference that exist between rather similar, and hence recently diverged, DNA regions with those between very different DNA regions that presumably diverged much earlier. These will be referred to as "within-species" and "between-species" comparisons. These phrases do *not* refer to named species, but to the extent of DNA divergence. "Within species" refers to genes or regions differing by <5% of nucleotides (usually <2%) and "between species" refers to genes or regions differing by >20% of nucleotides.

4. Synonymous Differences Within Species

Regions of DNA can be classified into four classes:

1. "*Meningitidis*-like": penicillin-susceptible strains of *N. meningitidis, N. gonorrhoeae,* and *N. lactamica* (except for a central region that differs from the other two species by 15% of nucleotides), and varying regions of the gene in resistant strains of these species that have not been replaced by blocks of DNA from other species.
2. "*Flavescens*-like": *N. flavescens,* and blocks of similar DNA in resistant strains of other species.
3. "*Cinerea*-like": *N. cinerea,* and blocks of similar DNA in resistant strains of other species.
4. "*Lactamica*-like": blocks of DNA resembling the central region of *N. lactamica.*

Within each class, nucleotide differences are always <5%, and usually <2%.

Table 1 shows the numbers of different types of difference between strains, within classes. Synonymous differences are those between codons specifying the same amino acid. Since up to 12 strains were compared in a single comparison, it was important not to count the same substitution twice. If a site was occupied by one of only two nucleotides, no difficulty arose: a single difference was counted. In a few cases, three nucleotides were present. This probably represents two changes, but it is not known which two. Therefore, the type strain of each species was taken as a "master," and the two differences from that master were counted: this may not be correct in individual cases, but it should not introduce a bias into the estimates.

Transitions are more common than transversions, presumably reflecting the higher mutation rate of the former. The actual numbers depend not only on relative rates of mutation, but also on the numbers of different amino acids in the protein, and on any codon biases. For example, a protein consisting of only Phe would permit only T-C synonymous changes: one consisting entirely of Val would permit twice as many transversions as transitions if the four codons were equally used, but would permit only A-G changes if only GTA and GTG were used. Table 2 gives the numbers of different amino

Table 1. Within-species differences[a]

	AG	CT	AC	GT	AT	CG	Transitions	Transversions	Total
All differences	54	96	13	10	6	16	150	45	195
Synonymous	46	82	11	4	5	12	128	32	160
Expected	59.8	68.2	9.3	7.3	8.1	7.3	128	32	160

[a] The expected numbers (see text) allow for amino acid frequencies, but not for codon bias, and for a ratio of transition to transversion rates of 4.46. Chi^2 (4 df) = 10.8, $0.05 > P > 0.02$

Table 2. Numbers of amino acids in the PBP2 gene of *N. Gonorrhoea*

Met 15	Gln 21	Cys 1	Ala55
Trp 3	Asn 23	Ile 27	Gly 50
Phe 21	Lys 38	Val 45	Leu 56
Tyr 17	Asp 29	Pro 34	Ser 31
His 9	Glu 36	Thr 33	Arg 37

Table 3. Codon bias[a]

Lys	AAA,43	AAG,16		
Glu	GAA,34	GAG,19		
Ile	ATT,23	ATC,20	ATA,4	
Gly	GGT,41	GGC,37	GGA,10	GGG,6
Pro	CCT,14	CCC,18	CCA,4	CCG,24
Leu	CTT,17	CTC,5	CTA,0	CTG,24
			TTA,7	TTG,37
Arg	CGT,16	CGC,23	CGA,1	CGG,15
			AGA,4	AGG,8

[a] In counting codon numbers, at each site each codon that occurs at least once in some strain was counted as a single case. There was no statistically significant bias for the remaining amino acids

acids in the PBP2 gene of *N. gonorrhoeae,* and Table 3 lists those amino acids for which there is significant ($P < 0.01$) evidence of codon bias.

Ignoring codon bias for the moment, the expected relative frequencies of the six types of difference can be calculated as follows. Imagine a long protein with the amino acids in the observed frequencies. For each amino acid, choose a codon at random, and count the number of single synonymous changes that are possible. For example, Phe is coded for by TTT or TTC. Thus with $P = 0.5$, it is coded for by TTT, and if so only one change, T − C, is possible: with $P = 0.5$, it is coded for by TTC, and only a C − T change is possible. Since direction of change is unknown, we conclude that, for each Phe, 0.5 + 0.5 = 1.0 T–C difference is possible. For Ile (ATT, ATC or ATA), there is a one-third chance that the codon will be ATT, and if so two changes, T − C and T − A, are possible: hence the expected number of possible changes, for each Ile residue, is (T − C and T − A)/3 + (C − T) and C − A)/3 + (A − T and A − C)/3, or 2/3 TC + 2/3 AT + 2/3 AC. These expected numbers, per residue, are then weighted by the observed frequencies of the amino acids to give the expected relative numbers of the six kinds of difference. Let S_i and S_v be the probabilities that a possible synonymous transition or transversion, respectively, will in fact be established as a difference between two strains. In the short term, double changes can be ignored, so S_i and S_v will be proportional to the relative mutation rates of specific transitions and transversions: they will also depend on the time since divergence, and perhaps on population size, but since these will be the same for all kinds of change they will not affect the relative numbers.

Note that a change from T → C in the coding strand could arise from a T → C mutation in that strand, or from an A → G mutation in the complementary strand. We therefore cannot use the data to distinguish between the rates of these two kinds of transition. Further, we cannot distinguish between the rates of T → C and C → T. Thus there are only four rates that could in principle be distinguished using data of this kind: (T ↔ C + A ↔ G), (A ↔ C + G ↔ T), A ↔ T, and G ↔ C. These four rates are bracketed in Tables 1 and 5.

The expected proportions of transitions and transversions, calculated in this way, allowing for the observed amino acid frequencies but not for codon bias, are $589S_i$:$657S_v$. Since the observed numbers were 128:32, the best estimate of S_i/S_v is $4 \times 657/589 = 4.46$. This is an estimate of the relative mutation rates for specific transitions and transversions: e.g. A ↔ G cf. A ↔ C.

Given this estimate, we can compare the observed numbers of synonymous changes with those expected (Table 1), allowing for the observed amino acid frequencies, and a transition/transversion rate of 4.46, but not for codon bias. The discrepancy is just significant ($P = 0.02$), and arises primarily from an excess of T − C and a deficiency of A − G. For the reasons given above, this could not be caused by a difference in mutation rates. However, Table 3 shows that these are precisely the discrepancies to be expected, given the observed codon bias. All the observed biases, except for Ile, would lead to fewer A − G differences, and the biases in Ile, Gly, and Arg to an excess of T − C differences. It is not easy to estimate how large this effect would be, but it should be sufficient to explain the observed discrepancy.

Hence the observed within-species synonymous changes can be explained if we allow for observed amino acid frequencies and codon biases, and for a mu-

Table 4. Synonymous differences between species (%)[a]

	AG	CT	AC	GT	AT	CG	Total	% transitions	% nucleotide divergence
Hypotheses									
A	17.1	27.8	15.8	9.1	15.8	14.4	100	44.9	27.1
B	22.2	27.5	14.5	9.4	13.9	12.5	100	49.7	25.0
C	17.6	34.8	9.9	12.3	11.1	14.3	100	52.4	23.1
D	19.3	37.7	10.5	13.0	8.0	11.5	100	57.0	21.4
Observed	18.3	42.5	10.1	10.1	4.9	14.1	100	60.8	15.7

[a] The four hypotheses give the expected proportions of different types of change, at a steady state between forward and backward mutation, according to assumptions explained in the text

tation rate of specific transitions 4.46 times that of specific transversions. There is no evidence of any other factor influencing the data.

5. Synonymous Differences Between Species

Between-species differences have been calculated by summing the differences between five pairs of strains, as follows:

1. *N. meningitidis* vs *N. flavescens:* whole coding region.
2. *N. cinerea* vs . *flavescens:* whole coding region.
3. *N. mucosa* vs *N. cinerea:* upstream region.
4. *N. mucosa* vs *N. meningitidis:* upstream region.
5. *N. lactamica* vs *N. flavescens:* central region.

These comparisons were chosen because, in all cases, the percent nucleotide difference was greater than 20% (average, 23.2%): counting synonymous changes only, the observed difference was always greater than 14%, with a mean of 15.7%. A single strain was taken as representative of each species. In summing all differences, there is a likelihood that the same evolutionary events have been counted twice. However, it will be shown that these strains are not very far from the steady-state divergence expected when forward and backward mutations are in equilibrium: since this is the hypothesis being tested, it is appropriate to treat differences between different pairs as independent events. The results are given in Table 4.

The steady-state hypothesis is equivalent to assuming that, for each amino acid for each strain, a codon has been chosen at random from the appropriate set. Expectations have been calculated according to four hypothese, in increasing order of plausibility:

A. No codon bias. Amino acids present in proportion to the number of possible codons: for example, Leu and Arg are six times as common as Met or Trp, and so on.
B. No codon bias. Amino acids present in the observed frequencies (Table 2).
C. Codon bias as observed (Table 3). Amino acids in observed frequencies.
D. As C, but only single nucleotide changes allowed in serine codons. Thus each serine is coded for either by a TCN codon or by AGT or AGC: as discussed later, this is in fact the case.

The expectations from these four hypotheses are compared in Table 4. The expected numbers from hypothesis D, assuming a total of 574 differences, are compared to the observed in Table 5. The observed nucleotide divergence at synonymous sites is lower than the expected—15.7% as compared to 21.4%—indicating that even the most divergent strains have not reached saturation. The relative numbers of the different kinds of change also differ ($P < 0.01$). The discrepancy arises mainly from the deficiency of A − T changes. The reasons for this is unclear. The striking feature, however, is how close the observations come to the expected values. In particular, the greater numbers of expected A − G and T − C changes does *not* depend on any assumption concerning the relative rates of transition and transversion mutations but only on one concerning the relative number of transitions that are synonymous: in other words, the high frequency of these changes depends not on relative mutation rates but on the nature of the code.

6. Nonsynonymous Changes

Table 6 compares the numbers of synonymous and nonsynonymous changes for within- and between-species comparisons. The between-species data are the summed numbers for the five comparisons listed in the last section. The within-species data are based on four pairwise comparisons, as follows:

1. *N. meningitidis* vs *N. gonorrhoeae:* whole coding region.
2. *N. cinerea* vs *N. meningitidis*-resistant strain NmR6: residues 199–584.
3. *N. flavenscens* vs *N. meningitidis*-resistant strain NmR4: residues 182–519.

Table 5. Synonymous differences between species (%)[a]

	AG	CT	AC	GT	AT	CG	Total	% nucleotide difference
Observed	105	244	58	58	28	81	574	15.7
Expected (D)	110.6	216.3	60.1	75.0	46.2	65.8	574	21.4

[a] Chi^2 (5 df) = 18.43, $P < 0.01$

Table 6. A comparison of synonymous and nonsynonymous differences within and between species

	Nonsynonymous				Synonymous		
	Transitions	Transversions	Total	Amino acid changes	Transitions	Transversions	Total
Within species	12	9	21	20	72	19	91
Between species	215	242	457	263	349	225	574

4. *N. lactamica* vs *N. polysaccharea:* residues 282–400.

For each group of strains, the longest region available for comparison has been taken.

There is a lower ratio of amino acid to synonymous changes in the within-species comparisons (1:4.5) than in the between-species comparisons (1:2.2). The difference is significant ($P < 0.001$). Thus the data fit the model prediction that $R_b \simeq 2R_w$.

A second feature of the data that accords with the idea that the different species are approaching saturation is as follows. The proportion of amino acid differences caused by a single nucleotide substitution is 19/20 in the within-species comparisons and substantially less than one-half in the between-species comparisons. This is expected, since in the latter case there will have been time for several substitutions to occur in a single codon.

The model assumed that

1. At most loci at which an amino acid change has occurred, only one neutral alternative is possible.
2. Most amino acid changes were brought about by a single base substitution.

The evidence favoring these assumptions is now reviewed: it is strong for the second, but only moderate for the first.

Of the 190 pairs of amino acids, only 76 are "connected," in the sense that it is possible to convert a codon of one into a codon of the other by a single base substitution: for example, Val and Ala (GTN and GCN) are connected, but Val and Thr (GTN and ACN) are not. Of the 76 connected pairs, only 26 are connected by a transition.

In the full data set, 97/581 of the amino acid sites are polymorphic. At 69 of these, only two amino acids are present, of which 63 pairs are connected, and six are not. At 25 sites, three amino acids are present: in 10 cases, all three possible pairs are connected, and in 13 further cases, two of the three pairs are connected. In only two cases is one amino acid not connected to the other two. Finally, at three sites there are more than three amino acids: in none of these is there an amino acid unconnected to the others.

Considering the first assumption above, in approximately two-thirds of all polymorphic sites only two amino acids are actually found. Of course, this does not prove that there is no other equivalently neutral residue that could exist at the site. However, for the 28/97 loci at which more that two amino acids were found, it does not follow that, at any one time, there was more than one possible neutral substitution; it could be that A → B was a neutral substitution, and that later, after substitutions had occurred elsewhere, B → C was neutral. The assumption of a single neutral alternative at any one time is probably not far from the truth.

Turning to the second assumption, the predominance of single nucleotide changes is confirmed by the fact that in only eight cases is there an amino acid not connected to others at the site. This implies that, typically, evolution has involved only a single nucleotide change at a time (there have also been a small number of insertions or deletions of whole codons, not discussed in this paper). Almost certainly, the eight exceptions are only apparent. For example, three of the eight exceptions involve a Val-Thr polymorphism (GTN-ACN). However, Val-Ala and Ala-Thr changes (both connected) are common at other sites. Hence the change between Val and Thr probably occurred through an Ala intermediate that is absent from the data set. The other five exceptions may be explicable in a similar way.

That evolution has involved only single base substitutions is confirmed by the usage of Ser codons. There are 19 sites at which Ser is coded for by more than one codon. At 11 of these the codons are TCN, and at eight they are AGC and AGT: in no case do TCN and AGN codons occur at the same site.

Turning to the relative rates of synonymous and non-

synonymous mutation, m_a and m_s, this can be estimated as follows. All codons capable of synonymous change (with the exception of the ATA codon of Ile, which is rarely used) can do so by a transition: in about half the cases, the codon can also change by two alternative transversions. Hence $m_s \simeq m_i + m_v$, where m_i and m_v are the rates of mutation of specific transitions and transversions, respectively.

Things are more complex for amino acid changes. In the present data set, the two most common situations are:

1. A single transition can cause the observed change: e.g., Val-Ala, GTN-GCN.
2. A single transversion can cause the observed change: e.g., Gln-Lys, CA(A or G)-AA(A or G).

There are, in fact, 89 cases in which we can be reasonably confident about the nature of the amino acid change that occurred (63 cases in which there are two connected amino acids only at a site, plus 13 cases in which there are three amino acids, but only two connected pairs). These 89 cases involve 35 pairs of amino acids, of which 18 are connected only by a transversion, and 17 by a transition. Hence $m_a \simeq (m_i + m_v)/2$. It follows that, $m_s \simeq 2m_a$.

Hence if most changes, both synonymous and nonsynonymous, were selectively neutral, we expect that the ratio R of amino acid to synonymous changes will be approximately twice as great in the between-species comparisons, as was in fact observed.

Conclusions

In the case of *Neisseria,* the ratio of substitutions to synonymous changes is twice as great in between-species as in within-species comparisons. But this is consistent with the assumption that all changes are selectively neutral. The discrepancy arises because the "species" of *Neisseria* diverged so long ago that they approach saturation for forward and backward mutation. This possibility should be kept in mind when using comparisons of this kind to demonstrate positive selection.

References

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature 304:412–417

Spratt BG, Bowler LD, Zhang Q-Y, Zhou J, Maynard Smith J (1992) Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J Mol Evol 34:115–125