J Mol Evol (1985) 21:182–191

Journal of Molecular Evolution

Codon Equilibrium II: Its Use in Estimating Silent-Substitution Rates

W. John Wilbur

Mathematical Research Branch, National Institute of Arthritis, Diabetes and Digestive and Kidney Disease, National Institutes of Health, Building 31, Room 4B-54, 9000 Rockville Pike, Bethesda, Maryland 20205, USA

Summary. We study the equilibrium in the use of synonymous codons by eukaryotic organisms and find five equations involving substitution rates that we believe embody the important implications of equilibrium for the process of silent substitution. We then combine these five equations with additional criteria to determine sets of substitution rates applicable to eukaryotic organisms. One method employs the equilibrium equations and a principle of maximum entropy to find the most uniform set of rates consistent with equilibrium. In a second method we combine the equilibrium equations with data on the man-mouse divergence to determine that set of rates that is most neutral yet consistent with both types of data (i.e., equilibrium and divergence data). Simulations show this second method to be quite reliable in spite of significant saturation in the substitution process. We find that when divergence data are included in the calculation of rates, even though these rates are chosen to be as neutral as possible, the strength of selection inferred from the nonuniformity of the rates is approximately doubled. Both sets of rates are applied to estimate the human-mouse divergence time based on several independent subsets of the divergence data consisting of the quartet, C- or T-ending duet, and A- or G-ending duet codon sets. Both rate sets produce patterns of divergence times that are shortest for the quartet data, intermediate for the CT-ending duets, and longest for the AG-ending duets. This indicates that rates of transitions in the duet-codon sets are significantly higher than those in the quartet-codon sets; this effect is especially marked for A ↔ G, the rate of which in duets must be about double that in quartets.

Key words: Synonymous codon — Codon equilibrium — Silent substitution — Fixation rates — Human-mouse divergence time — Eukaryotes — Simulation — Maximum entropy — Minimum selectivity

I. Introduction

The proposal that silent substitutions might provide the basis for a useful evolutionary clock seems to have first been put forward by Jukes (1980) and Miyata et al. (1980). Our purpose in this article is to investigate the relation of this proposal to detailed substitution rates that could form the basis for such a clock. We have argued in Part I of this paper that the equilibrium properties of eukaryotic coding sequences make them uniquely suited to serve as molecular clocks. We will express these equilibrium properties as equations derived from the codon-frequency distribution for 40 eukaryotic sequences as reported by Grantham et al. (1981) (see also Table 4 of Part I). In this form they provide the foundation on which the detailed rates may be constructed.

Our analysis consists of three steps. In Section II we investigate the information about rates that may be obtained from equilibrium considerations. In the ideal situation a complete determination should be possible. Imperfections in the eukaryotic equilibrium, however, limit the useful constraints on rates—in this case to five equations. In Section III these five equations are combined with additional criteria to determine complete sets of silent-substitution rates. In one determination we make use of maximum entropy inference as pioneered by Jaynes (1957a, b) and applied in a setting similar to ours by Holmquist and Cimino (1980). This determines the most uniform set of rates consistent with the equilibrium data. In addition to epistemological

reasons for this approach (see Jaynes 1978), this quality of uniformity seems desirable because many studies have been done in the past assuming that all rates are equal [see Holmquist and Cimino (1980) for discussion]. In a second determination of rates we combine the equilibrium constraints with data based on five homologous sequence pairs from human and mouse (or rat) to determine rates of special significance for the human-mouse divergence. The consistency of the rate sets found is examined in Section IV. Both rate sets are applied to estimate the human-mouse divergence time from the complete data and from three independent subsets of the data. Significant discrepancies are found in the estimated times. Such discrepancies, along with the nonuniformity of the rates determined, lead to the conclusion that strong selective constraints act against silent changes even over short divergence times. Implications for the silent clock are discussed.

II. Equilibrium Constraints

In Part I of this paper we applied a constrained method of least squares to find best-fitting nonnegative solutions to the Eqs. (3) of Part I. Although this method is useful and efficient for the study of equilibria, it is not useful for finding an overall best set of rates. The 12 equations of form (3) of Part I generally yield best least-squares rate solutions, the corresponding rates of which are not proportional to each other unless the system is in exact equilibrium. Frequently in each of the 12 solutions a number of rates are set to zero. The reason for this is discussed in the Results section of Part I and an example is given in Table 8 there. The example shows that the distortion produced by the least-squares process does not produce anything we can recognize as a best overall set of rates for the eukaryotes. We must therefore take a different approach.

Our approach is first to define a fitting function F(R), where R is any nonnegative rate vector with 12 components. Given such a rate set R, we determine equilibrium distributions for the synonymous-codon sets of each amino acid with a degenerate code. For each such amino acid the total number of occurrences in the data is distributed in the equilibrium distribution and for each codon the difference between the actual and the equilibrium counts is computed. The sum of the absolute values of each of these differences then serves as the measure of fit for that amino acid. The sum of measures of fit for all 18 amino acids with degenerate codes is the value F(R). Thus F(R) is a measure of how well R agrees with the codon counts in the experimental data. We use the absolute value rather than the square of differences so that estimates of R will be more robust (see Efron 1979).

Now the constraint the data places on R is that F(R) should be small, and our problem seems reduced to a simple minimization of F(R). Both advantages and disadvantages, however, attend the use of the function F. A clear advantage is that minimization of F(R) no longer entails a tendency to minimize the components of R, because F takes the same value on any two vectors R that are proportional to each other. Disadvantages of the use of F(R) are that F is a complicated nonlinear function of R, the minimization of which is a lengthy calculation, and that the value of F(R) when near its minimum tends to be somewhat insensitive to changes in R. This later difficulty is related to the lack of exact equilibrium in the data and suggests limitations on our ability to infer the complete R from the data. The function F(R) nevertheless provides a useful measure of nearness to equilibrium in the data with respect to a given R. To illustrate this we have carried out minimizations of F for several of the data sets analyzed in Part I, with the following results:

min F(R) = 206 for the 16 weakly expressed bacterial sequences;
= 213 for the 40 eukaryotic sequences;
= 438 for the 13 highly expressed bacterial sequences;
= 381, 498, 422, 467, 470 for the data for the 16 weakly expressed bacterial sequences randomized uniformly among synonymous codons for each amino acid as in SM1 (see Methods in Part I) and repeated five times.

Minima were found by the conjugate gradient method of Powell (1977) and all calculations (including those elsewhere in this paper) were made on the DEC KL-10 computing facility of the National Institutes of Health. Unfortunately, such calculations are lengthy and not suitable for large simulations such as were discussed in Part I of this paper.

We point out that the maximum value of F possible for these data sets is approximately 2000. This is true because the number of codons in each data set is close to 1000 and the worst R could do would be to dictate placement of all codons where none were previously; the sum of absolute values of error would then count each codon twice, yielding an F(R) close to 2000. In practice this theoretical maximum cannot be obtained for the data, because each codon is represented more than zero times. Such a value is possible, but very unlikely, for the random trials. The calculated minimum values shown are consistent with the balances found in Part I for the 16

Table 1. Total codon counts from 40 eukaryotic sequences as reported by Grantham et al. (1981)[a]

Amino acid	Third base			
	A	C	G	T
Quartets				
Ser	9	18	2	16
Thr	11	28	6	15
Pro	10	17	5	14
Ala	14	38	6	28
Gly	16	32	11	22
Val	5	21	33	9
Total counts	65	154	63	104
CT-ending duets				
Ser		21		12
Asn		28		8
His		21		10
Asp		24		16
Tyr		23		10
Cys		13		10
Phe		28		13
Total counts		158		79
AG-ending duets				
Lys	19		49	
Gln	10		28	
Glu	21		34	
Total counts	50		111	

[a] See also Table 4 of Part I. Codon counts are added to produce an average quartet and average duets. Data used by permission of authors and publisher

Table 2. Rates as relative substitution probabilities[a]

Rate	R_I	R_{II}	R_{III}	Simulation mean (SD)
r_{AC}	0.099	0.322	0.311	0.327 (0.027)
r_{AG}	0.107	0.058	0.058	0.033 (0.010)
r_{AT}	0.107	0.076	0.086	0.099 (0.015)
r_{CA}	0.068	0.171	0.176	0.180 (0.011)
r_{CG}	0.043	0.023	0.017	0.023 (0.007)
r_{CT}	0.054	0.035	0.035	0.031 (0.005)
r_{GA}	0.048	0.027	0.027	0.015 (0.005)
r_{GC}	0.120	0.093	0.101	0.112 (0.016)
r_{GT}	0.128	0.060	0.052	0.044 (0.015)
r_{TA}	0.066	0.024	0.017	0.021 (0.008)
r_{TC}	0.111	0.068	0.068	0.060 (0.011)
r_{TG}	0.048	0.044	0.053	0.054 (0.008)

[a] Rates were determined under the following constraints: R_I, equilibrium data and maximum entropy; R_{II}, equilibrium and human–mouse data and maximum entropy; R_{III}, equilibrium and human–mouse data and minimum selectivity. See text for discussion of the different rate sets and the simulation

weakly expressed bacterial sequences and the 40 eukaryotic sequences and with the nonequilibrium of the 13 highly expressed bacterial sequences, which appear at the same level as do random distributions.

Though minimum F(R) produced in this way are of interest because of their relevance to the question of equilibrium, the R corresponding to a minimum can only under ideal circumstances be considered the most appropriate estimate for the substitution rates for the given class of sequences. Neither the equilibrium nor the data are perfect, and there exist many R for a given data set that produce F(R) values close to the minimum. For this reason we shall impose only the strongest requirements of the equilibrium data on R and complete the determination of R by some independent criterion. We believe this is the only realistic way of choosing among those possible R that come close to minimizing F(R).

We are specifically concerned with the eukaryotes, though the method is applicable to other cases. We wish to consolidate the data and impose their strongest requirements on R. For this purpose we note that the six codons for serine may be considered to consist of a quartet (four each beginning with UC) and a duet (two each beginning with AG). Then Ser, Thr, Pro, Ala, Gly, and Val yield six quartets that, if equilibrium were perfect, would each impose the same conditions (equations) on the rate set R. We add the counts from all six quartets to produce an average quartet. Likewise, from the seven C- or T-ending duets we produce an average CT-ending duet, and from the three A- or G-ending duets we produce an average AG-ending duet. Table 1 shows how the addition of the codon counts within each group is performed. It is clear that the average for each type of codon within a group is a weighted average weighted by the frequencies of occurrence of the different codons. There is no need to divide the sums by a total, because only relative size is important. From the average quartet and the two average duets we derive five independent equations that R is required to satisfy exactly. After some manipulation, these five equations take the forms

$$r_{GA} = 0.45r_{AG}, \quad (1)$$

$$r_{TC} = 2.00r_{CT}, \quad (2)$$

$$r_{GC} = 2.44(r_{CA} + r_{CG}) - 1.03r_{AC} - 0.86r_{CT}, \quad (3)$$

$$r_{TA} = 0.63(r_{AC} + r_{AT}) + 0.35r_{AG} - 1.48r_{CA}, \quad (4)$$

$$r_{TG} = 1.48r_{CA} + 0.61r_{GT} - 0.63r_{AC} - 0.35r_{AG} - 0.52r_{CT}. \quad (5)$$

Details of their derivation are contained in the Appendix. In the next section we combine these conditions with additional criteria to determine R, and use F(R) as a measure of the fit of such R to the equilibrium data.

III. Substitution Rates

We shall use two different criteria to determine a set of rates R compatible with Eqs. (1)–(5). The first approach is to apply the principle of maximum entropy of Jaynes (1957a, b) in a manner similar to

its use by Holmquist and Cimino (1980). In this method we combine with Eqs. (1)–(5) the requirements that the rates be normalized to add up to 1 and that under these six constraints an R be determined that maximizes the entropy expression

$$E = \sum r_{XY} \log r_{XY}, \quad (6)$$

where X and Y are the bases of the substitutions $X \rightarrow Y$. The calculation is accomplished by using the six constraining equations on R to write the 12 elements of R in terms of six independent variables and then maximizing Eq. (6) in terms of these independent variables by the conjugate gradient method of Powell (1977). The resultant rate set (R_I) for the eukaryotes is listed in Table 2. For this R, F(R) is 230, which is close to the absolute minimum of 213 mentioned in the previous section and consistent with the extraction of maximum information about R from the eukaryotic equilibrium data. As discussed in the Introduction, maximizing Eq. (6) corresponds to finding the most uniform set of rates consistent with the data. [For a further discussion of the general approach of maximum entropy inference, see Jaynes (1978).]

The second approach to determination of R also employs Eqs. (1)–(5), but is designed to fit the silent changes perfectly in five human–mouse sequence pairs. For this purpose we considered eight homologous gene pairs of human and mouse (or rat) and chose the five that were most nearly in equilibrium for analysis. These five gene pairs are coding regions for the Ig gamma chain C region, insulin, growth hormone, prolactin, and Ig kappa chain C region. The actual sequences were taken from the Genbank data bank or the National Biomedical Research Foundation data bank.

The five human-mouse sequence pairs were processed as follows: For each pair of homologous sequences an alignment was made at the protein level, and if an amino acid replacement had taken place then the codons at that position were removed from their respective DNA sequences. Likewise, nucleotides appearing in a gap were eliminated. When this process of elimination was completed each pair of sequences differed only by silent mutations. All the sequences of each organism were then strung together in the same order end-to-end to make two long sequences. These two resultant sequences contain all the data we shall consider and they differ at corresponding positions only by silent changes. We refer to the two sequences composed in this way as *summary* sequences for the human-mouse comparison.

The summary sequences may be used to write additional equations for R. Let $N_{X(Y)}$ denote the time-averaged count of base X in the two summary sequences since their divergence, where X is counted only in the third codon positions and then only if substitution $X \rightarrow Y$ would be silent. Because the sequences are near equilibrium throughout this time period, $N_{X(Y)}$ is well defined for each pair of bases XY. To determine $N_{X(Y)}$ we used the synonymous-codon distributions from the 40 eukaryotic sequences as reported by Grantham et al. (1981), but corrected for the actual amino acid proportions in the summary sequences; this proved, however, to be a very small correction. Let N_{XY} denote the number of pairs of X and Y in either order that occur when paired (homologous) bases in the third codon position are examined throughout the aligned summary sequences. For short [less than 100 million years (Myr)] periods since divergence, saturation of silent changes is not as great as for longer periods, and we shall assume that changes develop linearly with time. Thus we may write

$$N_{X(Y)}r_{XY} + N_{Y(X)}r_{YX} = N_{XY}. \quad (7)$$

We shall subsequently examine the error resulting from making this linearity assumption for the case we are investigating.

To help clarify the terms in Eq. (7) it is useful to consider the concrete example where X = A and Y = G. Then $N_{A(G)}$ is the average number of occurrences of A in both summary sequences where A is a third base of a codon and $A \rightarrow G$ would be silent. Assuming the summary sequences are at equilibrium, $N_{A(G)}$ should not change with time and is not affected by the differences that develop between the two summary sequences over time. The number $N_{G(A)}$ has a similar meaning, but need not equal $N_{A(G)}$. On the other hand, N_{AG} ($=N_{GA}$) is the number of third-codon-position differences of the form A vs G or G vs A between the two present-day summary sequences. Over short time periods N_{AG} should increase linearly with time beginning at time zero. The increase is produced by r_{AG} acting on the $N_{A(G)}$ sites and r_{GA} acting on the $N_{G(A)}$ sites in both summary sequences. This gives rise to Eq. (7). Time is not included as a factor on the left side of Eq. (7) because for simplicity we have assumed one unit of time to equal the time since divergence of the two sequences. This allows the correct determination of relative rates to be made. Determination of absolute rates would require setting the time scale by methods that are not our concern here (cf. Wilson et al. 1977).

There are six equations of the form (7), which together with Eqs. (1)–(5) leave only one degree of freedom for R. The form of Eqs. (7) makes it generally impossible to add a separate condition $\sum r_{XY} = 1$, because this will prove impossible to satisfy. Thus, R is almost completely determined. We have used the maximum entropy principle to complete the determination of R; this is reported in normalized

form as rate set R_{II} in Table 2. However, for our special purposes we have chosen to study a different method of completing the determination of R. If silent changes were completely neutral, we would have r_{AC} and r_{TG} equal, because r_{AC} on one DNA strand would represent the sum of A → C changes on that strand and T → G changes on the opposite strand, and r_{TG} would be the same sum with the strands reversed. The larger of the ratios r_{AC}/r_{TG} and r_{TG}/r_{AC} thus becomes a measure of the selective force acting. Five other reciprocal pairs of rate ratios are related in the same way. For any given R, all of whose components are positive, let I(R) denote the largest of the twelve ratios of form r_{AC}/r_{TG}, r_{TG}/r_{AC}, r_{CA}/r_{GT}, r_{GT}/r_{CA}, etc. Then I(R) is a function of R and is always greater than or equal to 1 because the reciprocal of each ratio considered is also a ratio considered. It is convenient to refer to I(R) as the *index* of selectivity of the rate set R. Clearly, by minimizing I(R) we make all ratios as close to 1 as possible. By standard linear programming techniques we have found that R consistent with Eqs. (1)–(5) and (7) for which I(R) is a minimum. This R is then the most selectively neutral rate set consistent with Eqs. (1)–(5) and (7). In its standard normalized form (in which the sum of rates is 1 but the time scale is adjusted to allow this) it is listed in Table 2 as rate set R_{III}. For this set of rates, I(R) is the ratio r_{GC}/r_{CG}, which is 5.94, the smallest possible value consistent with all of the data. For this set of rates the fitting function F(R) takes the value 238, which is again a reasonably good fit to the data of Grantham et al. (1981) on codon usage for eukaryotes.

It is convenient to term the method of determining R by minimizing I(R) the method of *minimum selectivity,* as opposed to the method of maximum entropy. By way of comparison, $I(R_{II}) = 7.32$. Thus R_{III} is considerably more neutral than R_{II}, but the two entropies are $E(R_{II}) = 3.0675$ and $E(R_{III}) = 3.0611$, a difference of only 0.0064 bits and not significant by most standards (note that this calculation depends on using more decimal places than are listed in Table 2). For this reason and because we wish to ascertain the lowest level of selection required by the data, we focus our attention on the minimum selectivity method of determining R.

As a test of the method of minimum selectivity described here, we used the set of rates R_{III} based on Eqs. (7) and the human-mouse summary sequence data to simulate a sequence divergence, and in turn applied the same method to the simulated data in an attempt to reproduce the original rates. We first mutated one of the 589 codon summary sequences 10,000 times to bring it to equilibrium. We then took two identical copies and mutated them according to the human-mouse rate set R_{III} of Table 2 until 218 differences were produced (the summary sequences differ in 218 positions). The numbers N_{XY} were then determined and used in Eqs. (7), and that set of rates was determined that satisfied Eqs. (1)–(5) and (7) and minimized the index of selectivity. This was repeated 100 times. The average and standard deviation of the 100 resultant rate sets are shown in Table 2. In the trials we found that on the average 457 mutations were required to produce just 218 differences between the two sequences. The ratio of total to observed substitutions predicted is 2.10 (=457/218). This may be compared with a ratio of 1.65 for the human-mouse beta hemoglobin gene pair, which may be calculated using nonrandom REH theory by the methods of Holmquist et al. (1982, p. 304). The lower figure may be explained partly by the fact that nonrandom REH theory employs a single substitution intensity for all three codon positions, which may underestimate somewhat the substitution intensity at the third position, and partly by the fact that the hemoglobin genes have proven to be not as close to equilibrium as the molecules we used to compose the summary sequences of our study.

It is evident that the mutational process must be viewed as highly saturated. In spite of this level of saturation, the recovery of the original rates is remarkably good. The reason this is possible with the linearity assumption of Eqs. (7) is that all of the numbers N_{XY} tend to be affected to the same extent by the saturation. To the extent that the effect is uniform it does not produce an error in the relative rates. For example, examination of Eqs. (1)–(5) and either the maximum entropy or minimum selectivity method shows that if all numbers N_{XY} in Eqs. (7) are reduced by 10%, all components of a solution R will be reduced by 10%. Such a reduction would have no effect on the ratios of the different rates and would be important only in determining the absolute time scale of the rates, which is not of interest here. Only when saturation reduces different N_{XY} values by different fractions is error introduced into the relative rates. The simulation shows that this problem is not too severe in our case. Only the rates A → G and G → A are significantly in error, both being relatively underestimated. As we shall see subsequently, this is a fault not of the method, but of attempting to represent the differences between the summary sequences as produced by a single set of rates. This difficulty has little if any effect on the index of selectivity, which averaged 6.1 (1.2) for the simulation, as compared with 5.9 for the human-mouse rate set R_{III}. It is useful to point out that for divergences over longer periods and where saturation is a critical problem, Eq. (7) may still be applied if one of the available methods is used to correct the N_{XY} for the effects of saturation.

IV. Consistency

We have determined several sets of substitution rates that may be applied to the human-mouse divergence. We now wish to compare the predictions of rate sets R_I and R_{III} for the human-mouse divergence time and to study their internal consistency.

Let the summary sequences for human and mouse be denoted by S1 and S2, respectively. We wish to estimate a parameter proportional to the time of divergence. We can expect only proportionality and not equality because the numbers in Table 2 are relative, not absolute, rates. We assume that the summary sequences are in equilibrium under the rate sets R of Table 2. Then from a given R and the amino acid sequence that underlies both S1 and S2 we may determine a relative mutation rate M for S1 or S2. M may be approximated from S1 by adding each element of R that may be applied to a given codon and summing over all codons in S1. Nearly the same M would be obtained in this way from S2. We determine M from the exact equilibrium codon distributions that would ideally hold for both S1 and S2. Then the mutation process must go on long enough to produce the number D of differences seen. To find the number of mutations required to produce D we start with two identical copies of a sequence obtained from S1 or S2 that is in equilibrium. We then mutate the copies alternately according to the rates R until the number of differences D has been produced. This process is repeated 100 times and the mean number of mutations required to produce D is found and divided by the mutation rate M. The result is an estimate of the relative time since divergence. To test the internal consistency of the method, we perform this analysis not only for the two summary sequences but also for the subsequences derived from these consisting of just the quartets, the CT-ending duets, or the AG-ending duets. The relative times derived from these three subsequences are calculated from completely independent data, but should agree if all our assumptions are correct.

Results of relative time calculations for the human-mouse divergence time performed using both rate set R_I and rate set R_{III} of Table 2 are given in Table 3. The independent relative times calculated from quartets, CT-ending duets, and AG-ending duets, also listed in Table 3, reveal considerable internal inconsistency. It is important to know whether this is due to error in the calculations or to a fault in the underlying assumptions. There are several possible sources of error in the calculations. One source of error could be small sample size for the summary sequences used and another could be saturation of a sequence by mutations during the simulation process. Both of these sources of error

Table 3. Relative human–mouse (rat) divergence times[a]

Rate set used	Relative divergence times for			
	Summary sequences	Quartets	CT-ending duets	AG-ending duets
R_I	4.10 (0.30)	3.67 (0.37)	4.55 (0.64)	8.71 (1.69)
R_{III}	5.41 (0.47)	4.34 (0.66)	7.23 (1.15)	15.56 (3.38)

[a] Numbers are means of divergence times based on 100 simulations of the relative time, with standard deviations given in parentheses. R_I and R_{III} are rate sets from Table 2

Table 4. P values for the actual agreement of relative times given the simulation results of Table 3

	Quartets vs CT-ending duets	CT-ending duets vs AG-ending duets
R_I	0.13	0
R_{III}	0.005	0.002

are necessarily reflected in the variation in times obtained for the individual trials in a simulation. The times obtained in a single simulation are not distributed normally or in any simple distribution as far as we can determine. For short times, at which one is far from saturation of mutations, the distribution of times is close to normal, but it develops a tail to the right similar to that of a gamma distribution as one moves to larger times and consequently closer to saturation. A sharp cutoff of the distribution always exists on the left side because there is an absolute minimum number of mutations needed to produce the required number of differences in the two sequences. Because of the uncertainty in the nature of the distributions, the standard deviations in Table 3 cannot be relied on to support the judgment that the average times are inconsistent, though they may be a useful indication of this. To assess the consistency of the times reported in Table 3, we assumed the distribution of 100 randomly generated times to be representative of the actual distribution of such times for each case. Using these distributions, we then calculated the probability that when $t_A > t_B$ (average times), the change represented by t_A did not take longer to occur than the change represented by t_B. We compared quartet and CT-ending duet times and also CT-ending duet and AG-ending duet times in this way; the results are given in Table 4. These results are based on simulations of 100 trials each, but little change could be expected if one used simulations with larger numbers of trials. Table 4 confirms the apparent inconsistency in Ta-

ble 3 and that this inconsistency cannot be due to sample size or the effect of saturation.

Another source of possible error has to do with the derivation of the summary sequences. If in the original sequences from which a summary sequence is derived there are multiple replacement fixations at a particular codon site that finally result in the same amino acids occurring in both sequences at that site, then this site will be included in the summary sequences constructed and may distort the data. How often might this occur? In the most extreme case only 40% of amino acid positions are replaced between the sequences we considered. It seems conservative to expect not more than half of these to have been doubly replaced. If replacement by any amino acid were equally likely, only about 1 in 100 sites could be expected to revert back to the same amino acid and appear as a simple silent-mutation site. Even if this result is multiplied by 2 or 3 to allow for the unevenness in replacement probabilities for shorter times in the Dayhoff et al. (1978) PAM matrix, the effect is insignificant.

Finally, it could be that the sequences chosen are not in good equilibrium. This would reflect the existence of selective forces acting in a nonuniform way and could make time calculations unreliable. The unexplained preponderance of CUG coding for Leu in eukaryotic organisms (see Grantham et al. 1981) is the most unbalanced aspect of eukaryotic codon usage, but Leu is not included in the quartets or duets we used as a test for internal consistency. In selecting the sequences for the human vs mouse comparison, we selected from eight pairs the five most balanced pairs for construction of the summary sequences. We have compared the balances of the quartet and CT- and AG-ending duet subsequences coming from these summary sequences with the expected levels of balance for sequences of these lengths. In each case the original sequence was mutated 1000 times to bring it into equilibrium and then mutated another 1000 times while its distance from equilibrium was recorded at each step. All four duet subsequences were found to be in better than average equilibrium for the mutation rates of Table 2. [Note that both R_I and R_{III} satisfy Eqs. (1)–(5), and so yield the same equilibrium for the quartets and duets.] Thus, in these cases any deviation from equilibrium could be more than accounted for by the lengths of the sequences. The quartet subsequences were found to require about a 7% change to be brought within one standard deviation of the exact average distance from equilibrium for sequences of their length. Thus, it is clear that lack of equilibrium in the summary sequences cannot account for the inconsistency of the CT- and AG-ending duet times in Table 2, and it appears unlikely to have been a major factor in the quartet times.

V. Discussion

Few attempts to determine detailed substitution rates appear to have been made in the past. Kimura (1981a) and Takahata and Kimura (1981) have studied models in which several rates are required to be equal and used such models to estimate the silent-mutational distance between homologous DNA sequences, but have not reported detailed rates involved in such calculations. Holmquist and Cimino (1980) employed three equations defining equilibrium among base frequencies and the maximum entropy principle of Jaynes (1957a, b) to obtain the 12 rates of fixation for base mutations. They made rate determinations for genes from several different protein families in each codon position; the results were different for the different types. For the third codon position such variability may be credited at least partially to their treating silent and replacement changes on the same basis. Not only do replacement changes vary between sequence classes, but they generally differ in rate from the silent changes. Such differences raise the question of whether base frequencies in the third codon position are in equilibrium for any possible set of rates. There is no evident way to test for such equilibrium. Holmquist (1983) has extended the method to incorporate transition/transversion ratios at various divergence times and has considered other possible approaches to calculating substitution rates, but has reported no rates. In contrast to the studies mentioned, we have taken a more restricted setting and attempted a more detailed analysis for this setting.

We have calculated by three different methods sets of fixation rates for eukaryotic nucleic acid sequences. The results are listed in Table 2. One method employs the equilibrium constraints (1)–(5) and the principle of maximum entropy to determine a set of fixation rates, R_I, completely independent of any homologous sequence comparison. The other methods also employ the Eqs. (1)–(5) but were designed to fit perfectly the silent changes in five human-mouse (rat) homologous-sequence pairs in the form of Eqs. (7). The rate set R_{II} was included only for the purpose of comparison with R_{III}. These two rate sets are very close, as might be expected from the fact that each is determined to satisfy a uniformity condition (though the conditions are different). The rate set R_{III} has the special quality of being the most neutral rate set consistent with the data in the form of Eqs. (1)–(5) and (7). Because Eqs. (7) assume a linear accumulation of substitutions with time and this is not strictly true, we performed a simulation based on R_{III} that showed that if one employs the components of R_{III} as substitution rates in the usual sense then the method for determining R_{III} will reproduce R_{III} if applied to the simulated

divergence (see Table 2 and text). This provides a degree of validation for Eqs. (7) and the method of determining R_{III}. We employed rate sets R_I and R_{III} to estimate relative divergence times for human and mouse. The results show considerable discrepancies, as reported in Tables 3 and 4.

What conclusions can fairly be drawn from these results? The construction of rates based on the human-mouse divergence and the estimates of the divergence time for human and mouse are both based on the standard model of evolutionary divergence of the human and mouse sequences from common ancestral sequences by the accumulation of point mutations. If this model is accepted, then several conclusions seem appropriate. First, the change C → G is strongly selected against. This is evident already in the rate set R_I determined by maximum entropy, for which $r_{GC}/r_{CG} = 2.79$ (and no model of divergence is used). The selection appears much stronger in the rate set R_{III} based on divergence, for which $r_{GC}/r_{CG} = 5.9$. For this latter set of rates we also have $r_{AC}/r_{TG} = 5.9$, showing that in the human-mouse data T → G is selected against approximately as strongly as is C → G. If we used the rate set R_{II} the selection would appear even stronger. The discrepancies between the divergence times as estimated by the rate sets R_I and R_{III} are perhaps not of great importance, but illustrate that the proper choice of rates can have a very noticeable effect on the calculation. What is important is the internal inconsistency among divergence times, which is of roughly the same pattern for either of the two rate sets. This indicates a significantly higher rate of transitions in the duet-codon sets than in the quartet-codon sets. The same rates cannot then be applicable to the two codon sets; this is especially marked for the A ↔ G transitions, for which the discrepancy is more than two-fold. This accounts for the systematic error in recovering the A ↔ G rates from simulations based on the human-mouse rate set R_{III} (see Table 2 and accompanying remarks). Transitions must be relatively selected against in quartets.

How generally applicable are these conclusions? We have made a similar comparison of human vs cow, with essentially the same results. Further, D. Lipman (unpublished data) has surveyed a larger number of sequence pairs and concluded that in general in a nonreplacement site of homologous sequences the probability of a silent change in quartets is 0.378, whereas in duets it is 0.229. This quartet/duet change ratio of 1.65 differs from the ratio of 3 assumed in the silent-site corrections of Perler et al. (1980) and Miyata and Yasunaga (1980) and is consistent with slower rates of mutation in quartets. Hastings and Emerson (1983) have examined codon usage in muscle and liver and concluded that quartet-codon fixation is under significant selective pressure, whereas duet-codon selection appears much less constrained. This corroborates the differential rates we have proposed.

Kimura (1977, 1981a, b) has argued, based on the preponderance of synonymous changes among all changes and on the limiting of all substitution rates approximately by the silent-substituion rate, that silent changes must, at least over short times, be very close to neutral. At the same time Efstratiadis et al. (1980), Perler et al. (1980), Kimura (1981a), and Miyata et al. (1982) have all questioned the validity of the silent clock over long times, and Perler et al. have suggested that the difficulty is caused by saturation of truly silent changes after a short period of perhaps 85–100 Myr, followed by selected changes occurring at about one-seventh the initial rate. If one holds to the point-mutational model of evolution, our results suggest some modifications of or additions to this picture:

1. Strength of selection: When divergence data are included in the rate calculations, the index of selectivity of the resultant rate set is over twice as large as that obtained for the most uniform (maximum entropy) rate set that is consistent with the nonuniform codon distributions found in sequence data. Thus selective pressure appears to be at a higher level than would be predicted on the basis of codon frequencies alone (see Kimura 1981b). Evidence for constraints acting on divergence at silent sites over short time periods has also been presented by Miyata and Hayoshida (1981) and Sheppard and Gutman (1981).

2. Degree of saturation: Our simulations on the 589 codon summary sequences required, on the average, 457 mutations to produce 218 changes. This suggests that saturation becomes a factor well before 85 Myr elapses. Support for this point is also available from calculations made by Holmquist et al. (1982) comparing human, mouse, and rabbit hemoglobin genes by nonrandom REH theory.

3. Inadequacy of a single set of rates: A single set of rates is inadequate to deal consistently with silent change in the codon sets of different amino acids. This claim also finds support in Part I of this article, where those eukaryotic rates involving the base G are shown not to be homogeneous but to vary among the different synonymous-codon sets.

It is unclear whether rates based on equilibrium [Eqs. (1)–(5)] and divergence data [Eq. (7)] and restricted to a particular class of sequences such as α-globins, β-globins, or insulins would show more internal consistency than the rate sets we have calculated and tested using the human-mouse data. It seems likely, however, that such rates would shed light on the nature of the constraints specific to particular families of sequences.

Acknowledgments. The author would like to thank Ray Mejia and David Lipman for helpful suggestions and the latter for making available unpublished results.

References

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. Vol 5, Suppl 3, Silver Spring, Maryland, pp 345–352

Efron B (1979) Computers and the theory of statistics: thinking the unthinkable. SIAM Rev 21:460–480

Efstratiadis A, Posakony J, Maniatis T, Lawn R, O'Connell C, Spritz R, DeRiel J, Forget B, Weissman S, Slightom J, Blechl A, Smithies O, Baralle F, Shoulders C, Proudfoot N (1980) The structure and evolution of the human β-globin gene family. Cell 21:653–668

Grantham C, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl Acids Res 9:143–174

Hastings KEM, Emerson CP (1983) Codon usage in muscle and liver genes. J Mol Evol 19:214–218

Holmquist R (1983) Transitions and transversions in evolutionary descent: an approach to understanding. J Mol Evol 19:134–144

Holmquist R, Pearl D, Jukes T (1982) In: Goodman M (ed) Macromolecular sequences in systematic and evolutionary biology. Plenum, New York, pp 281–315

Holmquist R, Cimino JB (1980) A general method for biological inference: illustrated by the estimation of gene nucleotide transition probabilities. Biosystems 12:1–22

Jaynes E (1978) Where do we stand on maximum entropy? In: Levin D and Tribus M (eds) The Maximum entropy formalism. MIT Press, Cambridge, Massachusetts, pp 15–118

Jaynes E (1957a) Information theory and statistical mechanics. Phys Rev 106:620–630

Jaynes E (1957b) Information theory and statistical mechanics II. Phys Rev 108:171–190

Jukes TH (1980) Silent nucleotide substitutions and the molecular evolutionary clock. Science 210:973–978

Kimura M (1981a) Estimation of evolutionary distance between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kimura M (1981b) Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci USA 78:5773–5777

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

Miyata T, Hayoshida H, Kikuno R, Hasogawa M, Kobayashi M, Koike K (1982) Molecular clock of silent substitution: at least six-fold preponderance of silent changes in mitochondrial genes over those in nuclear genes. J Mol Evol 19: 28–35

Miyata T, Hayoshida 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, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23–36

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Powell MJD (1977) Restart procedures for the conjugate gradient method. Math Prog 12:241–254

Sheppard HW, Gutman GA (1981) Allelic forms of rat κ chain genes: evidence for strong selection at the level of nucleotide sequence. Proc Natl Acad Sci USA 78:7064–7068

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641–657

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann Rev Biochem 46:573–639

Received October 3, 1983/Revised August 25, 1984

Appendix

We give here the derivations of Eqs. (1)–(5). We denote the codon counts for the average quartet (see Table 1) by QA, QC, QG, and QT, where QA is the total number of A-ending codons in quartets, DC and DT are totals for CT-ending duets, and DA and DG are totals for AG-ending duets. Balance equations that would apply at equilibrium to individual quartet or duet synonymous-codon sets must also apply to the average codon sets of the same type. This allows us, to write

$$DA \cdot r_{AG} = DG \cdot r_{GA}, \tag{1A}$$

$$DC \cdot r_{CT} = DT \cdot r_{TC}, \tag{2A}$$

$$QA \cdot (r_{AC} + r_{AG} + r_{AT}) = QC \cdot r_{CA} + QG \cdot r_{GA} + QT \cdot r_{TA}, \tag{3A}$$

$$QC \cdot (r_{CA} + r_{CG} + r_{CT}) = QA \cdot r_{AC} + QG \cdot r_{GC} + QT \cdot r_{TC}, \tag{4A}$$

$$QG \cdot (r_{GA} + r_{GC} + r_{GT}) = QA \cdot r_{AG} + QC \cdot r_{CG} + QT \cdot r_{TG}. \tag{5A}$$

Equation (1A) expresses the fact that at equilibrium the flow from all A-ending AG duet codons to all G-ending AG duet codons must equal the flow in the reverse direction. Equation (2A) expresses the same balance for CT-ending duet codons. Equation (3A) states that among all quartet codons the flow away from A-ending codons must equal the flow to A-ending codons at equilibrium. Equations (4A) and (5A) express the same condition for C- and G-ending codons; a like equation for T-ending codons is not included because it is not independent but can be derived from the equations for A-, C-, and G-ending codons.

To obtain Eqs. (1)–(5) some manipulation of Eqs. (1A)–(5A) is required. First Eqs. (1A)–(2A) can be rewritten as

$$r_{GA} = (DA/DG) \cdot r_{AG}, \tag{6A}$$

$$r_{TC} = (DC/DT) \cdot r_{CT}. \tag{7A}$$

When the correct values for DA, DC, DG, and DT are obtained from Table 1, Eqs. (6A) and (7A) become Eqs. (1) and (2), respectively. A simple rearrangement of Eq. (4A) yields

$$r_{GC} = (QC/QG) \cdot (r_{CA} + r_{CG} + r_{CT}) - (QA/QG) \cdot r_{AC} - (QT/QG) \cdot r_{TC}. \tag{8A}$$

By substituting in Eq. (8A) for r_{TC} using Eq. (7A) we eliminate r_{TC} from Eq. (8A). When terms containing r_{CT} are collected, we obtain

$$r_{GC} = (QC/QG) \cdot (r_{CA} + r_{CG}) - (QA/QG) \cdot r_{AC} + [QC/QG - (QT \cdot DC)/(QG \cdot DT)] \cdot r_{CT}. \tag{9A}$$

Substitution from Table 1 then produces Eq. (3). A virtually identical rearrangement of Eq. (3A) followed by a substitution for r_{GA} using Eq. (6A) produces

$$r_{TA} = (QA/QT)\cdot(r_{AC} + r_{AT}) - (QC/QT)\cdot r_{CA} + [QA/QT - (QG\cdot DA)/(QT\cdot DG)]\cdot r_{AG}. \quad (10A)$$

Substitution from Table 1 produces Eq. (4). The final equation is obtained by substracting the left side of Eq. (4A) from the right side of Eq. (5A) and the right side of Eq. (4A) from the left side of Eq. (5A). The resulting equation is solved for r_{TG}, and Eqs. (6A) and (7A) are used to eliminate r_{GA} and r_{TC} from the result to obtain

$$r_{TG} = (QC/QT)\cdot r_{CA} + (QG/QT)\cdot r_{GT} - (QA/QT)\cdot r_{AC} + [(QG\cdot DA)/(QT\cdot DG) - QA/QT]\cdot r_{AG} + [QC/QT - DC/DT]\cdot r_{CT}. \quad (11A)$$

Finally, substitution from Table 1 produces Eq. (5).