11 June 1971, Volume 172, Number 3988 SCIENCE

Fitting Discrete Probability Distributions to Evolutionary Events

The probability of fixing nucleotide substitutions varies over codons of a cistron.

Thomas Uzzell and Kendall W. Corbin

Determination of amino acid sequences of proteins has led several authors to reexamine the role of random events in evolution (*1*, *2*), and data have been summarized to suggest that many changes in DNA and in the primary structure of proteins are neutral with respect to natural selection (*2*). Such possibly neutral changes include nucleotide base-pair substitutions that result in a change in the amino acid encoded (provided that the new amino acid is functionally equivalent to the former amino acid at that position) as well as nucleotide base-pair substitutions that result in synonymous amino acid code words and therefore do not change the primary structure of the polypeptide encoded by the mutant cistron. The fact that the frequency distribution of phenetically inferred fixations of amino acid substitutions at certain amino acid positions sampled from a series of homologous polypeptides can be described by means of a Poisson distribution was used to support the assumption that changes in protein structure are selectively neutral (*2*). We (*3*) and others (*4*) have, however, presented evidence that the majority of nucleotide base-pair substitutions, and thus amino acid substitutions, that are not eliminated by chance are eliminated by natural selection. Indeed, we believe that the majority of the amino acid substitutions fixed are selectively advantageous relative to most if not all other amino acid kinds at their respective positions.

In this article we consider (i) plausible discrete probability distributions of evolutionary events (nucleotide base-pair substitutions, the fixation of such substitutions in populations, and the changes in amino acid sequence that result from such substitutions); (ii) the improbability that the frequency distributions of fixations of substitutions at amino acid positions are the same for all polypeptides and (iii) the appropriateness and meaning of describing the number of fixations of amino acid substitutions estimated for different amino acid positions by means of the Poisson distribution. We describe the frequency of inferred fixations of nucleotide base-pair substitutions in cytochromes c by means of the negative binomial distribution, and show that the goodness of fit is better than that of the Poisson distribution.

We also examine the ways in which the assumptions underlying the use of the Poisson distribution either are or are not met when applied to molecular evolutionary data. The assumptions of the Poisson distribution are as follows.

1) The probability of an event is small but nearly identical over all occurrences. For mutations, the probability of a nucleotide base-pair substitution is equal over the genome. For fixation of synonymous nucleotide base-pair substitutions and of substitutions that result in amino acid substitutions, the probability of fixing any substitution is equal over the genome, regardless of the probability of a mutation.

2) There is no form of contagion. The occurrence of a mutation does not change the probability of an additional mutation occurring at that position. The fixation of a nucleotide base-pair substitution does not alter the probability of another substitution being fixed at that position, regardless of the probability of a mutation at that codon.

3) Events observed are independent of one another. The occurrence of a nucleotide base-pair substitution does not alter the probability of another substitution occurring elsewhere in either the cistron or the genome. The fixation of a substitution does not alter the probability of an additional fixation elsewhere in the cistron or the genome, regardless of varying probabilities of mutations at other codons.

As an alternative to the Poisson distribution, the negative binomial distribution may be used to describe molecular evolutionary events. This distribution is compatible with varying probabilities of mutations or of fixations, both between codons of a cistron and at individual codons over time.

Molecular Evolutionary Events

The unitary event in most protein evolution is the substitution of one nucleotide base pair for another (*5*). Such substitutions may result from mispairing of nucleotide bases during DNA replication and normally require two DNA replications for incorporation into a genome. Substitutions may be either transitions or transversions and may start from a guanine·cytosine (G · C) or an adenine · thymine (A · T) pair. The probability of a mutation occurring at a particular locus depends on several factors. The following may be important: the nature of the original

Dr. Uzzell is assistant professor of biology and assistant curator of the herpetological collections of the Peabody Museum of Natural History of Yale University, New Haven, Connecticut; Dr. Corbin was research associate in biology at Yale University and is now curator of systematics in the James Ford Bell Museum of Natural History at the University of Minnesota, Minneapolis.

Table 1. Distribution of numbers of inferred nucleotide base-pair substitutions required to account for observed changes at amino acid positions of several polypeptides.

Number of changes	Hemoglobin α (10, 141)*	Hemoglobin β (11, 146)	Cytochromes c (11, 104)	Insulins A and B (11, 51)	Fibrinopeptides A and B selected taxa† (11, 35)
0	93 (1 × 9)‡	88 (1 × 12)	87	33	6 (2)
1	14 (1)	18 (1, 1)	11	11	4 (2, 2)
2	17 (1, 1)§	13 (1 × 5)	4	3	7 (3,2,2)
3	9	14 (1, 1, 1, 3)		2	5 (5,3,2)
4	5 (1)	6	1	1	2 (1)
5	1	2 (1)	1	1	2 (1, 3)
6	2	1			4
7		3			2
8					2 (3)
9					
10		1			
11					1
	Cumulative evolutionary time (years)				
	532×10^6	589×10^6	738×10^6	660×10^6	641×10^6

* Number of taxa, number of positions. † Selected to match as closely as possible the taxa used for hemoglobins, cytochromes c, and insulins; including *Bos, Ovis, Lama, Sus, Equus caballus, E. asinus, Oryctolagus, Rattus, Macaca,* and *Homo.* ‡ One taxon missing at nine positions. § One taxon missing at one position, another taxon missing at a second position.

nucleotide base pair (A · T or G · C), the type of substitution (transition or transversion), the nature of adjacent nucleotide base pairs (because of the triple hydrogen bond uniting G · C base pairs, long adjacent stretches of such pairs may stabilize DNA), and the efficacy of DNA polymerases. Since most mutations are deleterious (*3*), it seems likely that natural selection will adapt the DNA polymerase of each species to its ratio of A · T to G · C so that species having a high G · C content will have DNA polymerases that replicate G · C bonds more faithfully than A · T bonds and vice versa. Although there are some data bearing on the extent to which these various factors influence mutation rates, one could conclude, as a first approximation, that the mutation rate in terms of nucleotide base-pair substitutions is constant throughout the genome except for individuals with poorly adapted DNA polymerases. If this were true, mutation would be the most likely of the three classes of molecular evolutionary events (mutation, fixation of all mutations, and fixation of mutations that result in amino acid substitutions) to form a Poisson distribution; the probability of a mutation would be small but essentially uniform over all nucleotide base pairs, the occurrence of a nucleotide base-pair substitution would not alter the probability of another substitution at either the same or another nucleotide, and the number of substitutions expected in a given-sized sample of nucleotide base pairs would be a small but finite number.

Some loci, however, may be more mutable than others (*6, 7*). Hot spots may, in certain cases, represent short sequences of nucleotide base pairs being scored as though only a single nucleotide base pair were involved, but there is evidence that the phenomenon as originally conceived by Benzer (*6*) does involve only single nucleotide base pairs that are more mutable than others (*7, 8*). Mutator genes have been reported (*9*) and do not necessarily affect loci equally. Such mutator genes may be mutant DNA polymerase cistrons which specify DNA polymerases that are not optimally functional for the ratio of A · T to G · C of the species (*10*); their differential effect on loci may be a reflection of varying ratios of A · T to G · C among cistrons. Provided that the reconstructed phylogenies of cytochromes c and fibrinopeptides A and B are correct, there is a significant excess of guanine to adenine transitions without a corresponding increase in the complementary replacements of cytosines by uracils (*11*). If all or a majority of these amino acid substitutions were fixed as though they were selectively neutral, then this is additional evidence suggesting that mutations are not all equally likely. If some nucleotide base-pair substitutions are indeed more likely than others, then at least one of the assumptions underlying the use of the Poisson distribution to describe the frequency distribution of mutations is not met.

A discrete probability distribution might be used to describe the fixation of nucleotide base-pair substitutions. Kimura (*1*) showed that the probability of fixing selectively neutral ($N_e s \ll 1$) substitutions in populations is a function of the mutation rate. If all nucleotide base-pair substitutions are equally likely, if all nucleotide base-pair substitutions are selectively neutral, and if there are a fixed number of cell divisions per year along the germ lines of the species sampled, then the assumptions underlying the Poisson distribution are met and the distribution of fixations of such substitutions may be appropriately described by this function.

Evidence cited above suggests that nucleotide base-pair substitutions are not equally likely. Similarly, nucleotide base-pair substitutions apparently are not equally likely to be fixed. The different efficacy (in vitro) of codons at incorporating amino acids into polypeptides of various organisms (*12*) may be due to an adjustment by natural selection of the frequency with which codons are used and with which different transfer RNA's are produced. Such differentials might, for instance, help regulate protein synthesis in different cells (*13*), but certainly must be adjusted closely to allow protein synthesis to proceed at all. Thus even synonymous codons may not be equivalent in terms of selection. The majority of mutations are deleterious; of those that reach high enough frequencies in populations for selection to affect them, about 79 percent are eliminated by natural selection (*3*). The probability of fixing nucleotide base-pair substitutions, even those resulting in synonymous codons, apparently is not constant either throughout a single genome or among the genomes of different taxa.

The probability of substituting any amino acid for any other cannot be uniform over all amino acid positions. Of the 190 possible amino acid pairs, only 75 are such that a codon of one member of a pair differs by a single nucleotide base-pair substitution from a codon of the other, whereas 101 require at least two base-pair substitutions, and 14 require three. Alternatively, there are four to seven (mean, 5.8) other amino acids that can replace a given ancestral amino acid and its codon due to the fixation of single nucleotide substitutions (*14*). Also, about 24 percent of all mutations are synonymous (*15*) since the number of codons per amino acid varies from one to six. Clearly, the probability of certain amino acids replacing others is not a constant.

There is evidence of contagion of amino acid substitutions fixed during evolution. Within the ribonucleases of pig and cow, several substitutions form

Table 2. Comparison of inferred number of nucleotide base-pair substitutions fixed with minimum replacement distances (MRD, range and mean) for variable positions of several polypeptides of mammalian species. Inferred fixations based on phylogeny of species compared. To correct for undetected synonymous fixations, *both* numbers of fixations and minimum replacement distances should be multiplied by 1.3 (*15*).

Number of fixations	Insulins A and B		Cytochromes c		Hemoglobins				Fibrinopeptides A and B*		Totals	
					α-chains		β-chains					
	Codons (No.)	MRD	Codons (No.)	MRD	Codons (No.)	MRD	Codons (No.)	MRD	Codons (No.)	MRD	Codons (No.)	MRD
1	11	1 (1.0)	11	1 (1.0)	15	1 (1.0)	18	1 (1.0)	3	1 (1.0)	58	1 (1.00)
2	3	2 (2.0)	4	1–2 (1.5)	16	1–2 (1.6)	12	1–2 (1.7)	7	1–2 (1.7)	42	1–2 (1.66)
3	2	1–2 (1.5)	0		9	1–3 (2.1)	15	1–3 (1.9)	6	1–3 (2.3)	32	1–3 (2.03)
4	1	1 (1.0)	1	2 (2.0)	5	2–3 (2.4)	6	1–4 (2.5)	2	3 (3.0)	15	1–4 (2.40)
5	1	3 (3.0)	1	4 (4.0)	1	3 (3.0)	2	3–5 (4.0)	2	3–4 (3.5)	7	3–5 (3.57)
6					2	3 (3.0)	1	5 (5.0)	4	3–6 (4.2)	7	3–6 (4.00)
7							3	3–5 (4.0)	3	4–7 (5.3)	6	3–7 (4.66)
8							0		1	4 (4.0)	1	4 (4.00)
9							0		0		0	
10							1	6 (6.0)	0		1	6 (6.00)
11									1	7 (7.0)	1	7 (7.00)

* Selected taxa, see footnote †, Table 1.

pairs, with one substitution within each pair probably being a compensatory change for the other (*16*). Fitch and Markowitz provided theoretical support for the contagious distribution of amino acid substitutions in their discussion of concomitantly variable amino acid positions (*17*). Yanofsky, Ito, and Horn provided evidence of contagion related to the function of mutant forms of the A protein of tryptophan synthetase (*18*). A loss of activity due to a specific substitution in tryptic peptide 3 can be regained by specifically altering the amino acid sequence in tryptic peptide 8.

Finally, the number of cell divisions per year along the germ line varies significantly among taxa (*19*). For taxa that differ in this respect, the probability of a substitution being fixed within a certain period of time within one population will not be the same as that within another population.

For each of the above reasons, the Poisson distribution is inappropriate for describing nucleotide base-pair fixation data, whether or not the substitutions are synonymous.

Number of Fixations Inferred

Recently we constructed phylogenies for several mammalian proteins (*3*). The taxa included are those for which complete amino acid sequences were available for hemoglobin α- and β-chains, cytochromes c, insulins A and B, and fibrinopeptides A and B (*20*). The phylogeny is based on the classification of the mammals and on the paleontological record (*21*). Times were obtained from the Geological Society Phanerozoic time-scale (*22*).

On the basis of this phylogeny, it is possible to determine the minimal number of fixations required at each amino acid position to account for the distribution of amino acids observed. A revised frequency distribution of inferred fixations per position is tabulated for the hemoglobins, cytochromes c, insulins A and B, and a selected sample of fibrinopeptides (Table 1). The fibrinopeptide sequences chosen include taxa that match as nearly as possible the taxa represented by the other polypeptide chains. The number of fixations shown in Table 1 includes synonymous mutations (required six times), parallel fixations of identical or synonymous codons (required 90 times), fixations of reverse mutations (required 11 times), and alternative routes (*23*) for fixing amino acids in different parts of the phylogeny. The number of fixations that we infer is conservative, since, on the average, at least 1.3 nucleotide base-pair substitutions are fixed per fixation of an amino acid substitution (*15*).

The number of mutations fixed during the evolution of a series of polypeptides was estimated by others (*2, 24*) as the minimum number of nucleotide base-pair substitutions required at each amino acid position to account for the different amino acids observed. This approach is phenetic, rather than phyletic, and fails to detect numerous parallel fixations of amino acid substitutions. Such phenetic data can be described adequately by a single Poisson model with an optimum number of invariable sites (*2*). In contrast, our phyletic data cannot be described adequately by such a model.

In Table 2 we compare the range and mean of the minimum replacement distance with the number of fixations inferred by us. These data show that if we use the minimum replacement distance we consistently underestimate the total number of fixations inferred except when only one fixation is inferred; the underestimate increases slowly as the number of fixations inferred increases. This striking discrepancy is primarily due to parallel fixations of identical substitutions and, to a much lesser extent, to fixations of reverse substitutions, to fixations of synonymous substitutions, and to alternative fixation of identical substitutions. No theoretical model exists by which the number of parallel fixations at an amino acid position can be determined in the absence of a phylogeny although the data in Table 2 provide an empirical conversion. For these reasons the minimum replacement distance cannot accurately reflect the evolutionary history at amino acid positions.

Frequency Distributions of Fixations in Polypeptides

Frequency distributions of phenetically inferred fixations for several polypeptides have been constructed (*2, 24*); the differences between these have in some cases (*2*) been discounted as the effect of chance. Nevertheless, such phenetically based frequency distributions are irrelevant to the evolutionary history of these polypeptides in that they do not provide an adequate estimate of the number of fixations that must have occurred in the phylogeny of the taxa considered. The distributions of the number of fixations phyletically inferred per amino acid position are similar for equivalent samples of homologous polypeptides, with the exception of the fibrinopeptides (Table 3). To demonstrate the improbability that

the fibrinopeptide data were drawn from the same population of amino acid positions from which the other sets of data were drawn, we computed χ^2 values for each of the four pairs of comparisons involving the fibrinopeptide data (Table 3). To avoid small expected values, we used three categories for the number of fixations: no fixation, one or two fixations, and three or more fixations. For each polypeptide, a certain number of residues were considered to be invariable (positions that cannot vary); these are omitted from the comparisons, and include six residues for the globins (*2*), 32 for the cytochromes c (*17*), five for insulins A and B (*25*), and two for fibrinopeptides A and B (*26*). With two degrees of freedom, it is highly improbable that the frequency distribution of the number of fixations inferred per amino acid position among the fibrinopeptides is the same as the distributions for the other four polypeptides.

The differences between these various frequency distributions may be explained by one or more of the following considerations. (i) The phylogeny used in the determination of the minimum number of fixations is incorrect and produces erroneous estimates of the number of fixations at each amino acid position. (ii) Differences in the cumulative evolutionary time, the number of taxa sampled, or the number of unknown amino acids bias the results. (iii) The mutation rates vary significantly between the genes controlling the synthesis of the fibrinopeptides as compared to those encoding the other peptides. (iv) The fibrinopeptides are the most highly adaptive molecules represented in the comparisons, and the other polypeptides are evolving by random fixation of neutral amino acid substitutions. (v) Selection limits the amino acid types tolerated at most variable amino acid positions.

It is difficult to demonstrate which of the above explanations is correct. We believe, however, that selection strongly limits the variety of amino acids at most, if not all, amino acid positions.

Differences Due to Sampling Error

The first explanation is unlikely for several reasons. The same phylogeny (*3*) is used for each polypeptide compared. Errors introduced for one polypeptide should be balanced by errors introduced for each other polypeptide. Such errors, however, cannot contribute significantly to the χ^2 values obtained in comparing the fibrinopeptide data with the frequency distribution of the other polypeptides because most amino acid positions sampled are unvaried, except within insulins A and B and the fibrinopeptides. The major contribution to χ^2 comes from this unvaried category, and no change in the phylogeny used would alter this category.

Table 3. Comparison of number of inferred fixations at "variable" amino acid positions for several polypeptides*. Values of χ^2 (d.f. = 2) for the comparison of each polypeptide with fibrinopeptides are given; all have $P < .001$.

	Number of fixations inferred			
	0†	1 or 2	3 or more	χ^2
Fibrinopeptides A and B‡	4	10	19	
Hemoglobin				
α-chains	87	31	17	39.0
β-chains	82	31	27	27.1
Cytochromes c	55	15	2	51.5
Insulins A and B	12	14	4	14.3

* Only data pertaining to mammalian species used. † Excluding "invariable" sites; two for fibrinopeptides, six each for α- and β-chains of hemoglobin, 32 for cytochromes c, and 25 for insulins. ‡ Selected taxa, positions A-1, A-2, A-3, A-8, B-1, B-2, B-8, are omitted.

Although differences in the cumulative evolutionary time, the number of taxa sampled, or the number of unknown amino acids could bias some of the results in Table 3, we do not believe this to be the case. In some instances (cytochromes c and insulins A and B) the cumulative evolutionary time is longer than it is for the selected sample of fibrinopeptide sequences (Table 1). The number of missing amino acids in the cytochromes c and in insulins A and B is so small that this absence cannot account for the discrepancy observed. We submit that all of the polypeptides listed in Table 3 have been adequately sampled with respect to the selected fibrinopeptides. Since more amino acids are missing in the fibrinopeptides (often a result of deletion, Table 1), the comparison is conservative in this respect. For these reasons we reject the second explanation.

As we mentioned above, it is conceivable that there are differences in the mutability of the two kinds of nucleotide base pairs, but regardless of the mechanism that results in this differential it is unlikely that nucleotide base pairs would be grouped in such a way that all or most codons of one cistron are made more mutable than those of another. (For models of protein evolution requiring neutral mutations, nucleotide pairs of different mutability would be randomly distributed over cistrons.) The third explanation is rejected for this reason.

Selection versus Chance

During evolution, more amino acid substitutions have been fixed in fibrinopeptides A and B (*27*) than in the hemoglobins, cytochromes c, or insulins A and B. At the 13 fibrinopeptide positions that have changed most often, there is a tendency for the amino acid substitutions to be nonconservative (*28, 29*). We argued (*3*) that this indicates these positions to be selectively neutral (those positions at which all 20 protein amino acids function equally well), although some might argue that this is a reflection of the highly adaptive nature of these fibrinopeptides. We believe, however, that stringent selection associated with adaptiveness limits the kinds of amino acids observed at most positions, even of fibrinopeptides A and B. The generally acidic nature of these small peptides probably functions to prevent polymerization by keeping fibrinogen molecules separate (*30*). Furthermore, the sequence of fibrinopeptide A appears to be related to species-specific enzymes that cleave the peptides from fibrinogen (*30*). Fibrinopeptide B has a marked effect on the contractility of smooth and cardiac muscle (*31*), a phenotypic expression that should be affected by natural selection, even though the tyrosyl-aspartyl dipeptide that produces this effect appears not to occupy an identical position in all of the fibrinopeptides B that have it (*20*).

In addition, the organization of the genetic code is such that most mutations result in either a synonymous code word or in a code word for a chemically similar amino acid (*32*). Of the changes that have occurred in polypeptides, most are more conservative with respect to the variety of chemical structures observed than expected on the basis of chance alone (*29*). Epstein showed that his indices of difference for amino acid substitutions changed less than expected by chance for a variety of proteins and that the index of difference increased significantly among amino acid types when they are external on globins (*29*). Substitutions involving internally placed

residues tend to be more conservative. This agrees with the observation that most internal globin residues are nonpolar (*33*). Sneath (*34*) examined a series of analogs of oligopeptidic hormones and demonstrated that the effectiveness of the analogs as hormones fell as more groups were changed, or with greater change at any position. These considerations suggest that few, if any, variable amino acid positions are neutral in the sense that all 20 of the protein amino acids function equally well at them. It is much more likely that, for most amino acid positions, only one or a small number of amino acid types are functional.

In summary, various lines of evidence indicate (i) that because the number of amino acid types tolerated by natural selection appears to vary from one position to another, the probability of fixing nucleotide base-pair substitutions is not equivalent over all codons of a cistron; (ii) that the distribution of fixations of nucleotide base-pair substitutions inferred per amino acid position is significantly different for the various kinds of polypeptides sampled; and (iii) that there is contagion at least at the amino acid level in the form of compensatory substitutions. Clearly, therefore, the Poisson is an inappropriate distribution for describing the frequency distribution of fixations of nucleotide base-pair substitutions. As has been indicated in the discussion, data that fit a single Poisson distribution, even with an optimum number of invariable codons, are misleading.

Negative Binomial Distribution

We have no a priori conviction as to the actual distribution of the fixations of nucleotide base-pair substitutions at codon positions. However, each of the arguments given above for rejecting the use of the Poisson distribution suggests that the distribution of fixations of nucleotide base-pair substitutions may be described by the negative binomial distribution. This distribution is based on the assumption that the probability of fixing an additional nucleotide base-pair substitution at any given codon of a cistron is not equivalent for all codons, but that it is drawn at random from a gamma-distributed universe of probabilities with a mean probability equal to the constant, fixed probability of the Poisson distribution that best describes the data. Thus, the negative binomial distribution assumes that the probability of fixing nucleotide base-pair substitutions varies over codons of a cistron and at individual codons; contagion and interaction could produce such effects. As a result, the majority of codons sustain few fixations while the tail of the distribution is long with a few codons sustaining many fixations.

The negative binomial distribution can be specified in terms of the parameters $\bar{x}$ and k (*35*). As k increases from one to infinity, the negative binomial distribution converges to a Poisson distribution; small values of k necessarily indicate that a distribution follows the negative binomial rather than the Poisson (*36*).

To estimate the number of fixations of nucleotide base-pair substitutions that have occurred during the evolution of 33 cytochromes c from 32 taxa, we have constructed a phylogeny of these taxa (Fig. 1). This phylogeny is based on the paleontological record (*21*), on morphological studies of extant species, and on an educated guess. For the purposes of this discussion the times of branching of the different phyletic lines are irrelevant, whereas the sequence of the branch points is important. This phylogeny differs in several respects from the phenogram used either by Fitch and Margoliash (*37*), Fitch and Markowitz (*17*), or Dayhoff (*14*). On the basis of this phylogeny, we inferred fixations for 104 positions of these polypeptides (those that are present in the amniotes sampled). We thus have omitted up to eight amino acid positions at the amino-terminal end of the cytochromes c of nine species. These terminal sequences cannot be compared in a meaningful way since positions homologous to them are not present in the cytochromes c of the other species examined (*38*). Of the 104 positions that are homologous, 70 vary. The distribution of fixations inferred is given in Table 4, along with three negative binomial distributions fitted to these data (*35*).

Figure 2 is based on the distribution of fixations inferred. The number inferred per position varies from 0 to 17, giving 18 categories (Table 4). We partitioned the positions not observed to vary into those that belong

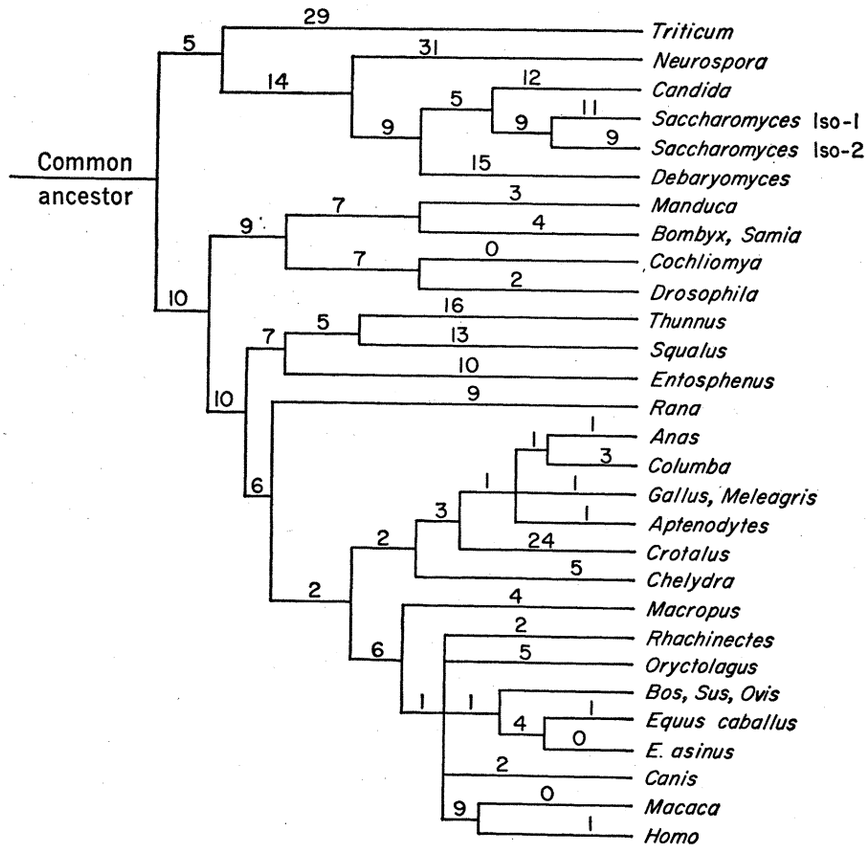

Fig. 1. Phylogeny for 32 taxa for which cytochrome c sequences are considered. The branching sequence, based on a variety of evidence, is plausible, but the lengths of the branches are arbitrary. The number of fixations inferred for 104 amino acid positions is indicated above each branch.

to the negative binomial distribution and those that do not by finding the minimum value of χ^2. The minimum χ^2 value is obtained by assigning all possible numbers (0 to 34) to the zero-fixation category (Fig. 2). For the ungrouped data a minimum χ^2 value (21.06) occurs when 11 of the 34 positions not observed to vary in this sample of data are assigned to the negative binomial distribution ($k = 1.65$), and the remaining 23 are excluded from it (upper curve, open squares). For the middle curve (open circles) categories 9 to 11 (8 to 10 fixations per codon) and 12 to 18 (11 to 17 fixations per codon) were grouped, giving a minimum χ^2 of 9.74. For the lower curve (closed circles) categories 9 and 10 (8 or 9 fixations per codon) and 11 to 18 (10 to 17 fixations per codon) were grouped giving a minimum χ^2 of 6.73. These groupings were made to avoid expected values of less than 5. The contribution of each fixation category to the overall χ^2 value at the inflection point of each curve is indicated in Table 4.

The best-fit negative binomial distribution ($k = 2.05$) is associated with a χ^2 of either 6.62 or 9.75, depending on the grouping used. With six degrees of freedom (ten comparisons, three degrees of freedom lost for fitting $\bar{x}$, k, and the zero category, one lost for χ^2) these χ^2 values are associated, respectively, with probabilities between .5 and .3 or between .2 and .1. These probabilities compare favorably with a probability of less than .2 for Fitch and Markowitz' (*17*) optimum two-Poisson fit.

Discussion

The fit of a Poisson distribution to phenetically inferred fixations has been used as an argument that many acid substitutions fixed are selectively equivalent to the amino acids they replace (*2*). Other authors (*39*) have argued, as we do in this paper and elsewhere (*3*), that most evolutionary events are controlled by natural selection, and that the fit of fixation data to a Poisson distribution does not necessarily mean that the substitutions are a result of neutral mutations. In fact, however, the phyletically inferred fixation data for cytochromes c (*40*) show that substitutions have not occurred uniformly over the variable codons of the cytochrome c cistron. The likelihood of fixing such substitutions cannot be uniform unless the probability of mutations varies essentially according to a gamma distribution (*41*).

Nevertheless, Poisson distributions have been used several times to describe the frequency of inferred fixations of nucleotide base-pair substitutions that occurred during the evolution of various polypeptides. Fitch and Margoliash (*42*) used this technique to estimate the number of invariable codons in cytochromes c. They reasoned that amino acid positions of cytochrome c fall into either an invariable group or a variable group with nearly equal probability of varying. The positions that were not observed to vary were partitioned into these groups by allowing the zero-change category to vary from 0 to 35, calculating the best-fit Poisson distribution for all 36 combinations, and determining which zero-change value was associated with a minimum χ^2 value. They estimated that 27 to 29 of the amino acid positions that had not been observed to vary were invariable because of functional requirements of the molecule, whereas the other six to eight were variable, but, because of sampling error, had not yet been observed to vary. Fitch and Margoliash found an appreciable number of hypervariable positions, those that had sustained more changes than expected by chance, given the best-fit Poisson distribution. They concluded that codons were of three sorts with respect to the probability of fixing an amino acid substitution: those with zero probability, the majority with nearly equal probability, and a small minority with high probability. It is these last that make a single Poisson distribution with an optimum number of invariable codons a poor model for describing the phyletically inferred fixations in cytochromes c.

Later, Jukes and co-workers (*2, 24*) also described the estimated number of changes by means of a Poisson distribution. They adjusted the zero category to give a minimum χ^2 value, and extended the technique to other polypeptides (globins and immunoglobulins G). King and Jukes (*2*) found no hypervariable category; they used the absence of a hypervariable category of codons and the excellent fit of their calculated Poisson distributions to the number of fixations inferred to argue (i) that the likelihood of fixing an amino acid substitution is essentially uniform over all variable amino acid positions, and (ii) that this is most consistent with the proposition that protein evolution is largely a result of fixing selectively neutral mutations. However, the method that Jukes and colleagues used to estimate the number of fixations is phenetic rather than phyletic. Fitch and Markowitz (*17*) pointed out that this procedure does not detect parallelisms [or, for that matter, back mutations or alternative routes (*23*) for fixing synonymous base-pair substitutions in different parts of a phylogeny]. Parallelisms as well as back mutations and alternative fixations of synonymous substitutions account for the hypervariable positions; as was noted above, we find that parallelisms are numerous in our phylogenies for several mammalian polypeptides (Tables 1 and 4). For the cytochromes c data upon which this article is based, parallelisms amount to 25 percent of all phyletically inferred fixations.

Table 4. Best-fit negative binomial distributions of inferred fixation data for cytochromes c.

Fixations per codon (No.)	Codons observed (o) (No.)	Codons expected (e) (No.)	$\frac{(o-e)^2}{e}$	$\frac{(o-e)^2}{e}$
		Ungrouped data		
0	10	9.21	0.07	
1	8	11.29	0.96	
2	14	10.97	0.84	
3	7	9.74	0.77	
4	11	8.23	0.93	
5	10	6.75	1.56	
6	4	5.42	0.37	
7	5	4.28	0.12	
8	1	3.35	1.65	
9	2	2.59	0.13	
10	0	1.99	1.99	
11	0	1.52	1.52	
12	2	1.16	0.61	
13	1	0.88	0.02	
14	2	0.66	2.72	
15	2	0.50	4.50	
16	0	0.37	0.37	
17	1	0.28	1.85	
				$\chi^2 = 21.06$
				$.2 > P > .1$
				14 d.f.
		Grouped data		
0	7	7.22	0.01	0.01
1	8	10.04	0.41	0.41
2	14	10.45	1.21	1.21
3	7	9.66	0.74	0.74
4	11	8.38	0.82	0.82
5	10	6.97	1.32	1.32
6	4	5.64	0.48	0.48
7	5	4.46	0.07	0.07
8	1	3.48		1.62
9	2	2.68	3.30	
10	0	2.04		
11	0	1.54		
12	2	1.16		
13	1	0.86		0.06
14	2	0.64	1.40	
15	2	0.47		
16	0	0.35		
17	1	0.26		
		$\chi^2 = 9.74$		$\chi^2 = 6.73$
		$.2 > P > .1$		$.5 > P > .3$
		6 d.f.		6 d.f.

Fitch and Markowitz (*17*) further explored the possibility of describing codon variability by means of the Poisson distribution. They estimated that the best-fit single Poisson distribution is associated with a probability of $< 10^{-26}$, and that a single Poisson with an optimum number of invariable positions is associated with a probability of $< 10^{-15}$. They therefore used two Poisson distributions as well as an optimum number of invariable positions. One Poisson distribution describes the majority of variable amino acid positions, which have a relatively low probability of sustaining fixations, and the other describes the hypervariable positions. This model describes the distribution of inferred fixations rather well ($P < .2$).

There is little mathematical basis for choosing between the two-Poisson plus invariable amino acid positions model used by Fitch and Markowitz and the negative binomial plus invariable amino acid positions model that we use. In each case, three parameters are fitted: the number of invariable positions and two means for the Fitch and Markowitz two-Poisson model (*16*); the number of invariable positions, $\bar{x}$, and k for our model. The probability associated with the negative binomial model ($.5 > P > .3$) is greater as discussed above. We argue throughout our article that it is unlikely that molecular evolutionary events have uniform probabilities over either all variable codons of cistrons or all variable amino acid positions of polypeptides. Fitch and Markowitz (*17*) come to the same conclusion although they prefer the two-Poisson model. We believe, however, that the arguments we and others have given against uniform probabilities of fixing amino acid substitutions provide a very credible biological basis for utilizing the negative binomial distribution to describe evolutionary events.

Why some unvaried amino acid positions should be excluded from either the two-Poisson model or the negative binomial model is not clear. The invariable category is subject to sampling effects since the percentage of the gene that appears invariable increases markedly as the average minimum replacement distance between sequences decreases [figure 2 in (*17*)]. This may either be due to changing numbers of taxa considered or to restrictions imposed by the genetic affinities of the taxa examined.

The most plausible interpretation for the exclusion of certain unvaried amino acid positions from either the two-Poisson model or the negative binomial model is that, although the number of sequences of cytochromes c presently known may be sufficient, the genetic diversity of the species represented is inadequate to detect variation at all of the variable sites. For example, the 29 cytochrome c sequences used by Fitch and Markowitz represent only two animal phyla; 19 sequences are from vertebrates. On the basis of comparisons of procaryotic and eucaryotic cytochromes c, Fitch and Markowitz (*17*) suggested that the number of invariable residues may be as few as seven out of the 75 amino acid positions available for comparison between pro- and eucaryotes. Thus, roughly ten of the eucaryote amino acid positions may be invariable, rather than the 32 (Poisson) or 27 (negative binomial) that are excluded on the basis of probability distributions. Whether examination of amino acid sequences from a greater diversity of species of eucaryotes will be sufficient to reveal variation at these excluded but presumably variable positions remains to be discovered. Variation at some of these positions among eucaryotes may be highly improbable if numerous amino acid substitutions in other parts of the molecule are required to convert these excluded positions to positions at which alternative amino acid types are selectively acceptable [concomitantly variable positions (*17*)].

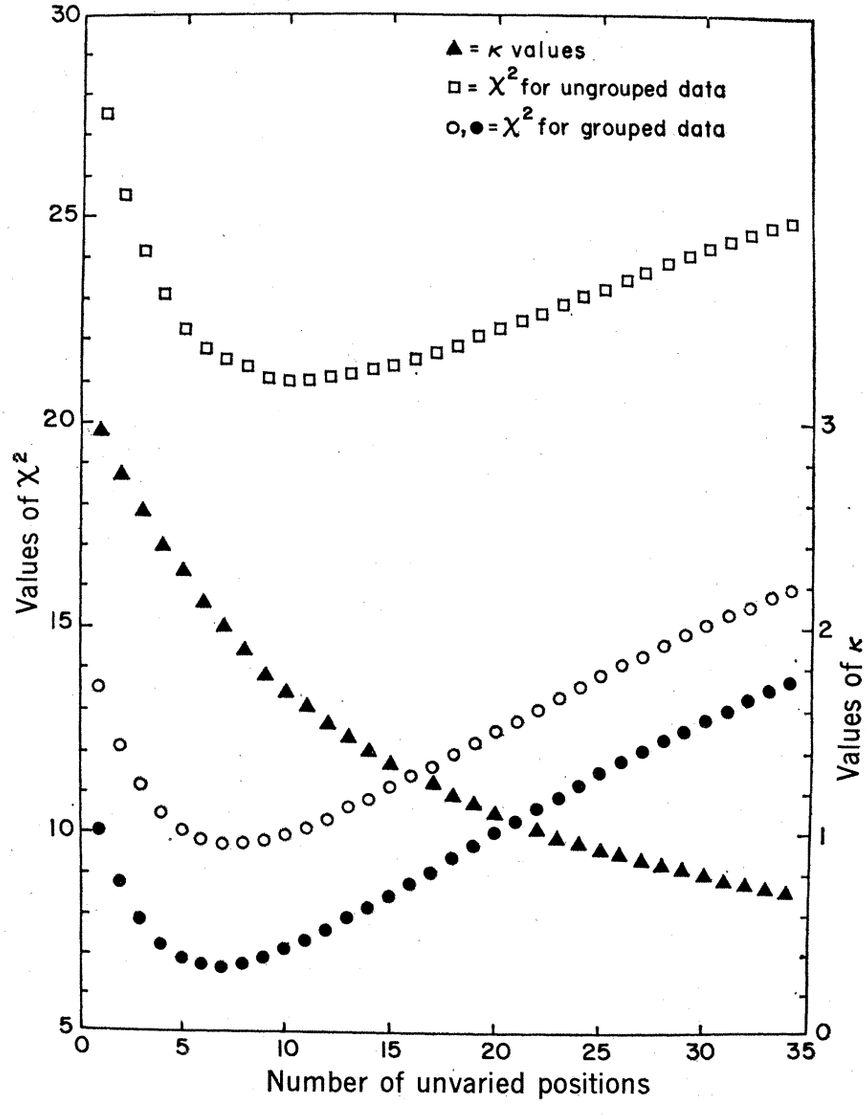

Fig. 2. Relation of χ^2 values of k values to the number of unvaried positions included in the zero-fixation category of 34 negative binomial distributions. Chi-squared values measure how well the calculated negative binomial distributions fit the frequency distribution of fixations inferred to account for the amino acid substitutions in cytochromes c.

Summary and Conclusions

The assumptions underlying the use of the Poisson distribution are essentially that the probability of an event is small but nearly identical for all occurrences and that the occurrence of an event does not alter the probability of recurrence of such events. These assumptions do not seem to be met for evolutionary events since (i) the probability of fixing nucleotide codon substitutions is not equal for all substitutions at a codon, and probably varies for the same substitution in different lineages; (ii) the probability of fixing codon substitutions varies among positions of a cistron; and (iii) the fixation of a nucleotide codon substitution at one position in a cistron modifies, and may even promote, the fixation of a codon substitution elsewhere along the cistron. Natural selection presumably is the causative factor that acts to modify the probability of a nucleotide codon substitution's being fixed in a population.

The use of the negative binomial distribution is consistent with the evidence that selective pressure on amino acid or nucleotide codon positions varies both among codon positions of a cistron and at a particular position during evolutionary time.

If the number of fixations of nucleotide codon substitutions per position of cistrons encoding cytochromes c are phyletically inferred (phylogeny based on a paleontological record) rather than phenetically inferred (based on paired comparisons of extant species' differences in the absence of a phylogeny) the distribution of these fixation data cannot be described adequately by a single Poisson distribution. The fit of these same data to a negative binomial distribution is very satisfactory.

It has been argued that the fit of phenetically inferred fixation data, which do not take account of parallel or reverse fixations, to the Poisson distribution was supportive evidence for the hypothesis that protein evolution results from the fixation of selectively neutral codon substitutions. This argument now appears to be undercut by the evidence that data on nucleotide codon fixation are more probably distributed according to the negative binomial distribution.

The fact that fixation data can be described by a particular discrete probability distribution does not of itself provide insight into the mechanisms of the evolutionary process. However, the facts—(i) that the assumptions underlying the use of the negative binomial distribution adequately deal with the varying probability of fixing amino acid or nucleotide codon substitutions at and among the positions of a cistron and (ii) that the negative binomial distribution provides an excellent fit for the phyletically inferred fixation data—suggest that the negative binomial is a very appropriate discrete probability distribution for describing evolutionary events.

Amino acids or their nucleotide codon *substitutions* may be fixed at a position of a cistron as though selectively neutral relative to the codon being replaced, even though the codon *position* will not be selectively neutral, since many amino acids cannot function there. The negative binomial distribution treats this situation well whereas a single Poisson distribution could only be satisfactory if all codon *positions* that could vary were selectively neutral.

References and Notes

1. M. Kimura, *Nature* **217**, 624 (1968).
2. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969).
3. K. W. Corbin and T. Uzzell, *Amer. Natur.* **104**, 37 (1970).
4. King and Jukes argued in favor of selective neutrality on the basis of the distribution of phenetically estimated fixations of amino acid substitutions in immunoglobulins G, cytochromes c, and globins. Nevertheless, they note that, relative to fibrinopeptides, at least 90 percent of the substitutions in cytochromes c have been eliminated by natural selection during evolution (*2*, p. 792).
5. A second large class of mutations results from frameshifts. We purposely ignore this complication since most mutations observed to be fixed in higher organisms during evolution are a result of base-pairing errors and not deletions or additions of codons. A third class of mutations results from genic and chromosomal duplication.
6. S. Benzer, *Proc. Nat. Acad. Sci. U.S.* **47**, 403 (1961).
7. J. W. Drake, *The Molecular Basis of Mutation* (Holden-Day, San Francisco, 1970).
8. S. P. Champe and S. Benzer, *Proc. Nat. Acad. Sci. U.S.* **48**, 532 (1962); R. E. Koch and J. W. Drake, unpublished results.
9. P. T. Ives, *Evolution* **4**, 236 (1950); B. McClintock, *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13 (1951); H. P. Treffers, V. Spinelli, N. O. Blesser, *Proc. Nat. Acad. Sci. U.S.* **40**, 1064 (1954).
10. J. E. Speyer, *Biochem. Biophys. Res. Commun.* **21**, 6 (1965); J. F. Speyer, J. D. Karam, A. B. Lenny, *Cold Spring Harbor Symp. Quant. Biol.* **31**, 693 (1966); E. C. Cox and C. Yanofsky, *Proc. Nat. Acad. Sci. U.S.* **58**, 1895 (1967); B. Commoner, *Nature* **220**, 334 (1968).
11. W. M. Fitch, *J. Mol. Biol.* **26**, 499 (1967); E. Margoliash and W. M. Fitch, *Ann. N.Y. Acad. Sci.* **151**, 359 (1968); W. M. Fitch, personal communication cited in C. Nolan and E. Margoliash, *Annu. Rev. Biochem.* **37**, 727 (1968); W. M. Fitch and E. Margoliash, *Brookhaven Symp. Biol.* **21**, 217 (1968).
12. R. E. Marshall, C. T. Caskey, M. Nirenberg, *Science* **155**, 820 (1967).
13. B. L. Strehler, D. D. Hendley, G. P. Hirsch, *Proc. Nat. Acad. Sci. U.S.* **57**, 175 (1967).
14. Calculations based on nucleotide codon assignments given in M. O. Dayhoff, *Atlas of Protein Sequence and Structure 1969*, (National Biomedical Research Foundation, Silver Spring, Md., 1969), p. 90.
15. M. Kimura, *Genet. Res.* **11**, 247 (1968).
16. H. W. Wyckoff, *Brookhaven Symp. Biol.* **21**, 252 (1968).
17. W. M. Fitch and E. Markowitz, *Biochem. Genet.* **4**, 579 (1970).
18. C. Yanofsky, J. Ito, V. Horn, *Cold Spring Harbor Symp. Quant. Biol.* **31**, 151 (1967).
19. Taking 19 as the number of cell divisions along the germ line of the human female (*3*) and a generation time in man of 14 years [J. Buettner-Janusch, *Origins of Man* (Wiley, New York, (1966)] gives 1.3 cell divisions per year along the germ line of the human female. A comparable estimate for the female rat would be 30 cell divisions per year along the germ line based on a total of 15 cell divisions along the germ line and a generation time of about 0.5 years for wild animals —2^{15} =32,768, the number of primary oocytes in the rat is about 40,000 [D. L. Ingram, in *The Ovary*, S. Zuckerman, A. M. Mandl, P. Eckstein, Eds. (Academic Press, New York, 1962), vol. 1, p. 255].
20. M. O. Dayhoff and R. V. Eck, *Atlas of Protein Sequence and Structure*, 1967–1968 (National Biomedical Research Foundation, Silver Spring, Md., 1968).
21. G. G. Simpson, *Bull. Amer. Mus. Natur. Hist.* **85**, 1 (1945); J. A. Lillegraven, *Paleontol. Contrib. Univ. Kansas* **50**, 1 (1969); L. Van Valen, *Evolution* **23**, 96 and 118 (1969).
22. W. B. Garland, A. G. Smith, B. Wilcock, Eds., *Q. J. Geol. Soc. London* **120**, suppl., 260 (1964).
23. Identical amino acids may be fixed in different parts of a phylogeny as a result of unique series of fixations. These we term alternative routes.
24. T. H. Jukes, *Biochem. Genet.* **3**, 109 (1969); ——— and C. R. Cantor, in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Press, New York, 1969), vol. 3, p. 21.
25. Calculated from inferred number of fixations based on data in (*20*).
26. The carboxy terminal arginine at which each peptide is cleaved from fibrinogen is considered invariable.
27. Fibrinopeptides A and B are cleaved from the amino terminal ends of fibrinogen α- and β-chains, respectively. Fibrinogen also contains a third kind of polypeptide, γ-chains. The seven amino terminal residues for γ-chains are known for *Homo, Bos*, and *Sus*. The only difference among these three is that *Sus* has glutamic rather than aspartic acid at position 6. These data suggest that the amino terminal ends of the α-, β-, and γ-chains are evolving at different rates. H. Pirkle, A. Henschen, A. Potapous, *Nature* **223**, 400 (1969).
28. C. J. Epstein, *Nature* **215**, 355 (1967).
29. E. L. Smith, *Harvey Lect. Ser.* **62**, 231 (1968).
30. N. Chandrasekhar and K. Laki, in *Fibrinogen*, K. Laki, Ed. (Dekker, New York, 1968), p. 117.
31. K. Laki, *ibid.*, p. 1; J. A. Gladner, *ibid.*, p. 87.
32. A. L. Goldberg and R. E. Wittes, *Science* **153**, 420 (1966).
33. M. F. Perutz, J. C. Kendrew, H. C. Watson, *J. Mol. Biol.* **13**, 669 (1965).
34. P. H. A. Sneath, *J. Theoret. Biol.* **12**, 157 (1966).
35. C. I. Bliss and R. A. Fisher, *Biometrics* **53**, 176 (1953). We use the method given in this reference to find the best-fit, negative binomial distribution for our phyletically inferred fixation data.
36. H. Seal, personal communication.
37. W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).
38. The amino terminal sequences of *Drosophila melanogaster, Cochlyomya hominivorax, Samia cynthia, Bombyx mori, Manduca sexta, Triticum* sp., *Neurospora crassa, Saccharomyces oviformis, S. cerevisiae, Candida krusei*, and *Debaryomyces* sp. were omitted from the comparison. The cytochrome c amino acid sequences of *Debaryomyces* (K. Titani) and *Saccharomyces* Iso-2 (E. Margoliash) are unpublished.
39. R. Richmond, *Nature* **225**, 1025 (1970); B. Clarke, *Science* **168**, 1009 (1970).
40. The number of inferred fixations along lineages is similar for our data and those of Fitch and Markowitz (*17*) but both differ considerably from Dayhoff's estimates (*20*).
41. We thank Masatoshi Nei for bringing this fact to our attention.
42. W. M. Fitch and E. Margoliash, *Biochem. Genet.* **1**, 65 (1967).
43. Computer time was paid for out of funds available from NSF grant GB 8769 to T.U.