Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 3560–3564, June 1982
Genetics

Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites

(population genetics/evolution/restriction enzymes)

JULIAN ADAMS* AND E. D. ROTHMAN†

*Division of Biological Sciences and Department of Human Genetics, and †Department of Statistics and Human Genetics, University of Michigan, Ann Arbor, Michigan 48109

Communicated by James V. Neel, January 4, 1982

ABSTRACT **The distributions of cleavage sites and their related sequences have been analyzed for 54 restriction endonucleases in the genome of human mitochondrial DNA; in three papova viruses, BK, simian virus 40, and polyoma; and in three bacteriophages, ϕX174, fd, and G4. The results show that the cleavage sites and related sequences for most of the restriction enzymes tested are distributed nonrandomly. These results (*i*) constitute *prima facie* evidence for the action of natural selection, either direct or indirect on the restriction sites, and (*ii*) suggest that estimates of phylogenetic relationship, based on a phenetic approach using restriction enzyme data, will be biased.**

Reports have appeared that use data on the restriction patterns of DNA, usually from mitochondria, to infer evolutionary relationships (1–4). Such analyses have intrinsic interest that is not possessed by similar studies using protein sequences (e.g., ref. 5) or allele frequencies (e.g., ref. 6). The use of restriction endonucleases allows a direct assay of the DNA without the confounding factors of mutation to synonymous codons, mutations involving no charge change in the protein, and so on. Indeed the claim has been made that evolutionary relationships inferred from restriction patterns are superior to those inferred from other data (1).

Two different approaches may be used for the reconstruction of phylogenetic relationships using restriction patterns. The first, which we term the cladistic approach, reconstructs phylogenetic relationships much in the same way that inversions were used by Dobzhansky and his school (refs. 4 and 7; A. R. Templeton, personal communication). The second approach, which we may term the phenetic approach, involves the determination of the proportion of shared restriction sites for any two species or taxonomic units and the consequent generation of a distance metric (8–11). In this paper we limit ourselves to a consideration of the phenetic approach.

To estimate phylogenetic relationships by a phenetic approach, the same basic assumption of randomness must be made for evolutionary analyses of DNA restriction patterns as is required for evolutionary analyses with other types of data (12, 13). For data involving DNA restriction patterns, this assumption requires that the distribution of bases and of base changes within the DNA is completely random.

Given the importance of DNA restriction patterns to evolutionary studies, it is important to examine in some detail the veracity of the assumptions required for these data. Conversely, results showing that the distributions of restriction sites are nonrandom also would be important as this would constitute *prima facie* evidence for the action of natural selection, either on the restriction sites themselves or on related sequences. Previous workers have concluded from the migration patterns of DNA fragments in electrophoresis that restriction endonuclease sites are distributed randomly throughout the genome, at least for a limited sample of restriction endonucleases and DNA sequences (3, 14, 15). However, such data have a number of shortcomings for testing the assumption of randomness. For example, the method does not distinguish between different fragments of the same size that comigrate; moreover, very small fragments are usually ignored (2) because of experimental difficulties (cf. ref. 4).

The recent advent of rapid techniques for DNA sequence determinations has made it possible to sequence the entire genomes of a number of different viruses and of human mtDNA. In this paper we use these data (*i*) to test critically the assumption of randomness required to construct phylogenetic trees from restriction data with a phenetic approach and (*ii*) to examine the restriction sites for evidence of natural selection. The logical sequence to use for this study is that of the human mitochondrion (16), as this genome has been most extensively used for the evaluation of phylogenetic relationships (1–4). To examine the sequences of a group of evolutionarily related species, we have analyzed in addition the sequences of the genomes of three related papova viruses, simian virus 40 (17, 18), BK (19), and polyoma (20), and the sequences of three related single stranded bacteriophages, ϕX174 (21), fd (22), and G4 (23). We have analyzed these sequences for 54 restriction enzymes documented by Roberts (24).

EVOLUTIONARY CHANGE OF RESTRICTION SITES AND TESTS OF RANDOMNESS

The estimation of phylogenetic relationships from restriction data by a phenetic approach involves the explicit or implicit assumption that an equilibrium exists where the nucleotides are completely randomly distributed throughout the genome (8–11). In this section we examine the implications of this assumption by considering the relationship between the dynamics of the mutational changes in the nucleotides and the change in restriction sites.

For a given restriction enzyme, the expected number of restriction sites and their allied sequences in a genome will change over time in accordance with the following series of differential equations.

$$\frac{de_{00}}{dt} = \lambda\left[\frac{q}{2}e_{1,0} + pe_{1,1} - re_{00}\right], \qquad [1]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

$$\frac{de_{ij}}{dt} = \lambda \left\{ p(r - i + 1)e_{i-1,j-1} + q(r - i + 1)e_{i-1,j} + \frac{q}{2}(i - j + 1)e_{i,j-1} + q(j + 1)e_{i,j+1} + \frac{q}{2}(i + 1 - j)e_{i+1,j} + p(j + 1)e_{i+1,j+1} - [(r - i + j) + q(i - j)]e_{i,j} \right\}, \quad [2]$$

$$i = 1, 2, \ldots, r, \; 0 \le j \le i,$$

in which e_{00} is the expected number of recognition sites of length r in the genome; e_{ij} is the expected number of sequences that are i base substitutions different from the restriction site, where j of those differences are due to transition mutations, the other $i - j$ being due to transversions; λ is the mutation rate; and p is the probability that the mutation is a transition, with $q = 1 - p$ being the probability that the mutation is a transversion. These equations assume that (*i*) the probability of mutation per nucleotide base per unit time, λ, is constant for all bases within the sequence, (*ii*) reciprocal mutational events are equally likely (e.g., the probability of mutation of T → A is equal to the probability of mutation of A → T), (*iii*) no selection is operating on the mutational changes occurring, and (*iv*) addition and deletion mutations do not contribute to evolutionary change. Although these equations are only appropriate for enzymes that recognize unique sequences, it is easy to derive equivalent equations for those enzymes with multiple recognition sequences. The implications of these equations, which we discuss below, hold for both categories of enzymes.

The e_{ij} converge to an equilibrium given by

$$e_{ij} = N \frac{r!}{(r - i)!\, j!\, (i - j)!} \left(\tfrac{1}{4}\right)^{r-i+j} \left(\tfrac{1}{2}\right)^{i-j}, \; 0 \le j \le i \le r, \quad [3]$$

where N is the length of the genome in nucleotides. By summing over the index j, at equilibrium the expected number of sequences that are i base substitutions different from the restriction site e_{i+} is given by

$$e_{i+} = N \binom{r}{i} \left(\tfrac{1}{4}\right)^{r-i} \left(\tfrac{3}{4}\right)^{i}. \quad [4]$$

This equilibrium solution of the model has three main implications for the unbiased estimation of phylogenetic relationships. (*i*) At equilibrium the model predicts a completely random distribution of nucleotides throughout the genome so that each of the sequences of length r is equally likely. Thus, for sequences of length $r = 1$ (namely, the individual nucleotides), these should occur with frequency 0.25. For human mtDNA and for many other species, the frequencies of these nucleotides show an apparently close agreement with this expected equilibrium frequency. However, for sequences of length $r = 2$, it is well known (e.g., refs. 17, 25–29) that the frequencies of certain dinucleotides, in particular C-G in eukaryotes, differ markedly from the expected equilibrium frequency of 0.0625, suggesting that the equilibrium solution of the model is not a good approximation of reality. (*ii*) Although the rate of approach to equilibrium will depend on the relative frequency of transitions and transversions, the equilibrium solution is not a function of these frequencies. Therefore, it is not necessary to consider the relative frequency of transversions and transitions in any test of the equilibrium solution of the model. (*iii*) The equilibrium specifies constraints on the expected number of related sequences e_{i+}, $i \ge 0$, not just on the number of restriction sites themselves (cf. ref. 10). A random distribution of the observed number of sequences e_{i+}, $i \le 0 \le r$, would imply that these constraints have been satisfied.

Distribution of Numbers of Fragments. As N is typically large and the probabilities used to compute e_{00} and e_{i+} are small, it is reasonable to assume that the observed numbers of sites and related sequences are Poisson-distributed variables. Using this fact, one can compare the observed results with the expectations to test the reasonableness of the model. Sequences two or more base substitutions removed from the restriction site become increasingly unrelated to the restriction sites themselves. Therefore, we have chosen to limit our attention to the distribution of the restriction sites and to those sequences one base substitution removed from the restriction site (hereafter, potential sites). It should be noted that the observed numbers of sites and related sequences are only approximately Poisson distributed, as their distributions in the genomes are "r-dependent." That is, the probability of occurrence of a site or related sequence is dependent on the occurrence of the same sequence occurring at neighboring positions, as the sequences will overlap. For the majority of the restriction endonucleases, this r-dependence will reduce the variation below that predicted by the Poisson distribution. Therefore, for the most part, the tests are somewhat conservative.

The expected number of restriction sites and potential sites, if we assume a complete random distribution of the nucleotides, may be simply calculated from the product of the appropriate nucleotide frequencies. However, because there is evidence that nucleotides may not be distributed randomly within base pairs (25–29) or triplets (28, 30), we also consider two specific alternatives to the hypothesis of mutual independence—namely, that the DNA sequence is described by a one-step or two-step Markov chain. Calculation of the expected number of sites under these hypotheses is straightforward.

Distribution of Fragment Sizes. A second important component of randomness is the position of those sites within the genome. A nonrandom distribution of fragment sizes may result in an overabundance of exceptionally large or small (or both) fragments, both of which may be undetected in some electrophoretic procedures and may cause underestimates of the amount of differences between the organisms or taxonomic units being compared. There is a certain amount of *a priori* evidence that would suggest such a nonrandom distribution of cleavage sites. For example, certain untranslated regions of the papova viruses are A+T-rich (17), and this would reduce the occurrence of certain restriction sites (e.g., *Hha* I) and increase the probability of others (e.g., *Hin*dIII) within them. Randomness of the position of restriction sites or potential sites may be tested with the U_N^2 statistic of Watson (31).

If the positions of both the sites and the potential sites are nonrandom, the bias introduced by this nonrandomness may be minimized if the distributions of the two categories of sites can be shown to be drawn from the same underlying distribution. This hypothesis can be tested using a two-sample equivalent of the U_N^2 test (32). A summary of the characteristics of all the tests is given in Table 1.

RESULTS AND DISCUSSION

Tables 2 and 3 summarize the tests for randomness of the distribution of the restriction sites for 54 enzymes in human mtDNA (Table 2) and the papova viruses (Table 3). For space considerations, the results for the bacteriophages are not shown. The most striking aspect of the results is that the great majority of the 54 restriction endonucleases show an extremely nonrandom pattern of cleavage for all three groups of organisms.

Table 1. Summary of tests for randomness

Test	Statistic	Model	Test statistic
0A	Observed number of restriction sites	Mutual independence	Based on cumulative Poisson
0B		One-step Markov chain	
0C		Two-step Markov chain	
1A	Number of potential restriction sites	Mutual I independence	
1B		One-step Markov chain	
1C		Two-step Markov chain	
2	Position of restriction sites	Uniform distribution	U^2_N (see ref. 33)
3	Position of potential restriction sites		
4	Position of restriction sites and potential sites	Identical distributions	$U^2_{N_1,N_2}$ (see ref. 34)

By considering only one species for mtDNA, the sites for only 7 out of 54 enzymes fulfill all the criteria for randomness. For the three papova viruses, only 9 of 54 meet these criteria, whereas for the three bacteriophages, only 14 out of 54 enzymes (results not shown) meet all criteria for randomness.

Closer inspection of the results reveals that the number of restriction enzymes that are nonrandom in their cleavage pattern is much higher when the number and distribution of potential sites are considered. For example, for the human mitochondrial genome, 17 of 54 enzymes have a significantly different number of cleavage sites than expected under the assumption of random distribution of individual nucleotides. Yet when potential sites are considered under the same hypothesis, an additional 20 enzymes show significant nonrandomness. The same trend is seen when the position of sites is considered. For example, in human mtDNA only three enzymes show a significantly nonrandom location of sites, whereas, when the locations of potential sites are considered, an additional nine enzymes are significantly nonrandom. The patterns of the results are similar for both the bacteriophages and the papova viruses. Thus, a consideration of only the extant cut sites can lead to a major underestimate of the degree of nonrandomness of the restriction enzymes.

It is clear that the numbers of restriction sites in the genomes deviate significantly from those expected under the most restrictive assumption—mutual independence of the individual nucleotides. Most of these significant differences disappear under the less stringent assumptions—namely, consideration of the genome as a one-step or two-step Markov chain. However, many enzymes that have a significantly different number of cleavage sites from those expected when the genome is considered to be a one-step Markov chain show the same significant difference under the two-step Markov chain consideration, indicating that sequences longer than three nucleotides are distributed nonrandomly.

Certain enzymes show an extremely nonrandom distribution of cleavage sites for all three groups of organisms. For example, the observed number of cleavage sites for the enzyme *Mbo* I is significantly less ($P < 0.001$) than the expected number for all seven genomes tested. This result strongly suggests that the recognition sequence of *Mbo* I is under strong selection, either direct or indirect. The existence of this deficiency in genomes as diverse as single-stranded coliphages and the human mitochondrion implicates this sequence in a basic role in cell function. Other restriction enzymes show a highly nonrandom cleavage pattern for one or two groups of organisms, but not all. Thus, the enzyme *Bbv* I shows a highly nonrandom pattern of cuts for both the papova viruses and the bacteriophages but not in human mtDNA. Similarly *Xho* II is highly nonrandom for mtDNA only and *Fnu*DIII is very nonrandom for the papova viruses only. These results also imply selection, either direct or indirect, on the recognition sequences but suggest that their

Table 2. Distribution of restriction sites in human mtDNA

Restriction enzyme	Sites, no.	e_{00}†	Tests for randomness*								
			0A	0B	0C	1A	1B	1C	2	3	4
Acc I	8	16	1			2					
Alu I	63	63						1	1		
Asu I	35	40				3	3	3			
Asu II	7	5				2					
Ava I	3	10	1	1	2	1	3	3			1
Ava II	8	22	3	3	3	3	3	3	1	3	
Ava III	3	5				1					
Avr II	12	3	3			3					
Bal I	6	3					2	1			
*Bam*HI	1	3				3	3	3			
Bgl II	0	5	1	1		1	1				
*Bst*EII	2	3				1	3	3			
Cla I	1	5				3					
Dde I	71	63						2			
*Eco*RI	3	5				2	1				
*Eco*RII	16	22		3	3		3	3	1		1
*Fnu*DII	6	40	3			3					
*Fnu4*HI	29	40				3					
Hae I	25	12	2			2					
Hae III	52	40					2				
Hga I	11	22	1			3					
*Hgi*AI	8	12				3	3	3			
Hha I	17	40	3			3		1			
*Hind*II	12	16				2				1	
*Hin*fI	36	63	3	1	1	3	1	1		1	
Hpa I	3	5								2	
Hpa II	23	40	2	2		3	3	1		2	
Hph I	55	28	3	3		3	3				
Kpn I	3	3				3	3	3			
Mbo I	23	63	3	3	3	2					
Mbo II	41	35			3		2	3			
Mnl I	202	101	3	3		3	3	3		1	
Mst I	0	3				3					
Pst I	2	3					1				
Pvu I	0	3				3					
Pvu II	1	3								1	
Rsa I	34	63	3	3	1	2	2				
Sac I	2	3						2			
Sac II	2	2				2				1	
Sal I	0	3				3					
*Sfa*NI	23	28				2					
Sma I	0	2				2	3	3			
Taq I	29	63	3			3		1			
Xba I	5	5				3	3	3			
Xho I	1	3				3	1	3			
Xho II	1	15	3	3	3	3	3	3			
Xma III	1	2				3				2	

Enzymes *Acy* I, *Bbv* I, *Bcl* I, *Bgl* I, *Hae* II, *Hind*III, and *Sph* I showed a random distribution of cleavage sites according to all tests. Ratings: 3, significantly nonrandom at $\alpha < 0.001$; 2, significant at $\alpha < 0.01$; and 1, significant at $\alpha < 0.05$.

* See Table 1.

† Assumes mutual independence, model A.

Table 3. Distribution of restriction sites in the papova viruses

Restriction enzyme	Tests for randomness*											
	BK				SV40				Polyoma			
	0A	1A	2	3	0A	1A	2	3	0A	1A	2	3
Acc I		2				3						
Acy I		1		1		1		3		2		
Alu I	1				2	1			1			
Asu I		2		2				3	1			
Asu II		3				2						
Ava II				1				1				
Ava III								2				
Bal I						1						1
Bbv I	3	2	3	1	3	3	2	1	2	2		2
Bcl I		1										
*Bst*EII										1		
Cla I		2				3				3		1
*Eco*RI								2				
*Eco*RII	3	3		2	3	2			1	3		
*Fnu*DII	3	3		3	3	3		3	3	3		3
*Fnu*4HI	3		3	1	3	1	2	3				3
Hae I	3				3	1	2			2		
Hae II								1				
Hae III	3			2	2				1			
Hga I	1	3		2	2	3		3		3		1
*Hgi*AI								1				
Hha I	2	3		1	1	3		2	3	3		
*Hind*II	1	3				1				2		
*Hind*III					1	1						
*Hinf*I	3	3			1	3						
Hpa I		1								1		
Hpa II				2	2			2	1			
Hph I												1
Kpn I	2											
Mbo I	3	3			2	3			2			
Mbo II			1							2		
Mnl I		2	2		3	2			3	1		
Mst I										1		
Pst I		3				3			1	2		
Pvu I		3				3				3		
Pvu II		3				3		3		2		
Rsa I		1				3				2		
Sac I			1									
Sac II		2				1						
Sal I		3				3				1		
*Sfa*NI						2		1				2
Taq I	3	3			3	3			3	3		
Xho I		2										
Xho II		1										
Xma III		1						2		2		1

Enzymes *Ava* I, *Avr* II, *Bam*HI, *Bgl* I, *Bgl* II, *Dde* I, *Sma* I, *Sph* I, and *Xba* I showed a random distribution of cleavage sites for the papova viruses according to all tests. Ratings: 3, significantly nonrandom at $\alpha < 0.001$; 2, significant at $\alpha < 0.01$; and 1, significant at $\alpha < 0.05$. SV40, simian virus 40.
*See Table 1.

involvement in the replication and processing of the DNA is more specific than that for *Mbo* I.

Role of Selection in Directly Influencing the Frequency and Distribution of Target Sites. The first and most obvious explanation for these results is that the restriction sites are themselves under some form of selection. This may be a reasonable explanation for those enzymes that still show a significantly nonrandom number of cuts, even when the distribution of two-base doublets and three-base triplets is taken into consideration. Three possible functions have been ascribed to restriction enzymes *in vivo*: (*i*) breakdown of foreign DNA from potential parasites or pathogens, (*ii*) the promotion of cross-specific gene flow through heterologous recombination, and (*iii*) site-specific recombination such as that seen in host-specificity determination in the phages P1 and Mu and in phase variation in *Salmonella* (J. A. Shapiro, personal communication). The host-controlled restriction and subsequent modification of the DNA of many phages by transmethylation and glucosylation (33) and the existence of antirestriction proteins (34) are evidence for the direct role of restriction enzymes in the breakdown of foreign DNA. Protection of the phage DNA from restriction may be achieved by DNA modification but also by selection for a reduced number of restriction sites. We may expect to find evidence for such selection in the host–parasite system defined by the coliphages ϕX174, G4, and fd and their potential hosts. Although the host range of these phages is fairly narrow, for the purposes of this argument, it is reasonable to consider all members of the Enterobacteriaceae as potential hosts. The hypothesis would predict that the numbers of restriction sites in the coliphage chromosome would be significantly *lower* than those expected from a random distribution of nucleotides. At the time of writing, 10 restriction enzymes have been isolated from members of the enterobacteriaceae; *Eca* I (*Bst*EII), *Eco*RI, *Eco*RII, *Pal* I (*Hae* III), *Kpn* I, *Pst* I, *Pvu* I, *Ecc* I (*Sac* II), and *Sma* I (ref. 24; the prototype isoschizomers are in parentheses). However, neither the number of sites nor the potential number of sites for these enzymes are significantly reduced below random expectation for the three coliphages. Thus, there is no evidence that reduction in the number of restriction sites has been a significant adaptive strategy in response to host-controlled restriction systems.

Although the data do not allow us to examine the significance of the role of restriction enzymes in heterologous or site-specific recombination, possible other selective factors may be inferred from trends in the data. For example, the restriction enzymes with four-base recognition sites tend to exhibit a much higher degree of nonrandomness than those enzymes that recognize longer sequences. Although this pattern could represent a particular property of enzymes that recognize sequences of four-base length, it is more likely that the increased nonrandomness reflects the higher number of expected cleavage sites, thereby increasing the power of the statistical tests for these enzymes. Support for this conclusion can be obtained when the results of the tests for the randomness of the number of restriction sites are compared with the results of the tests of the randomness of the potential sites. For example, in the papova virus BK, two-thirds of the four-base-recognizing restriction enzymes (8 of 12) show a significantly different number of restriction sites from that expected under test 0A (see Table 2), compared with the figure of <1/10th (3 of 34) for the same test with six-base-recognizing restriction enzymes. When potential sites are considered, where the number of sites is much larger, the difference between the four-base-recognizing enzymes (8 of 12; 67%) and the six-base-recognizing enzymes (12 of 34; 35%) becomes much less. Clearly the *in vivo* functions of the enzymes may be different for each enzyme, and a variety of selective forces may affect the overall distribution of sites, making the detection of individual effects refractory.

Role of Selection in Indirectly Influencing the Frequency and Distribution of Target Sites. The extensive nonrandomness of the results also may be explained by the action of selection—not on the restriction sites themselves but on other sequences that may include or may be a subset of the recognition sequences of the restriction endonucleases. For example, when the frequencies of dinucleotides are taken into account by considering the genome to be described by a single-step Markov chain, many of the recognition sites appear to be ran-

domly distributed. The restriction endonucleases that exhibit the most nonrandomness tend to be those that recognize sites with a high frequency of the nucleotides cytosine and guanine—for example, *Eco*RII, *Fnu*DII, *Fnu*4HI, and *Hae* I. The observed numbers of restriction sites for these enzymes are lower in general than the expected numbers under the hypotheses of complete randomness. Thus, the nonrandom distribution of restriction sites may reflect the nonrandom distribution of the G and C nucleotides. In simian virus 40, the C-G dinucleotide is much more common in the control regions than in the translated regions of the genome, suggesting that it may serve some regulatory function (17). Therefore, the nonrandom distribution of the restriction endonucleases that recognize sites incorporating the C-G dinucleotide may reflect selection for regulatory sequences rather than selection for the enzyme sites themselves. However, none of the 54 restriction enzymes (24) recognize either of two known regulatory sequences: the "Pribnow box" (35) or the canonical "Hogness box" or any of its close relatives found in polyoma (36).

A further explanation for the nonrandomness of the restriction sites is that it may simply reflect the nonrandom codon usage that has been reported for human mtDNA (10), the papova viruses (17, 37), and the single-stranded coliphages (23) or even a nonrandom use of amino acids (37).

CONCLUSIONS

Our results show that the distributions of restriction sites and potential sites are highly nonrandom for the majority of enzymes in all genomes examined. Thus, the equilibrium solution of the model is not a good approximation of reality, and some of the assumptions of the model are violated for these data. The most important assumption violated is that of selective neutrality of the mutational changes, though other assumptions of the model may also be violated. For example, the probability of mutation may not be constant for all nucleotides, as it is well known that local regions of high mutational activity exist within some DNA sequences (38). In addition, it is clear that addition and deletion mutations must contribute to the evolutionary pool.

A consequence of these results is that estimates of phylogenetic relationships using a phenetic approach with restriction data will, in general, be biased. Moreover, the direction of the bias is difficult to predict, being dependent on (*i*) the genomes under study, (*ii*) the distributions of the restriction sites, (*iii*) the distributions of the potential sites, and (*iv*) the enzymes chosen. Such bias could be minimized or even eliminated by using only those data from enzymes that have a random distribution of cleavage sites and by avoiding those that show extreme nonrandomness. However, the choice of suitable enzymes becomes severely limited, particularly when we consider human mtDNA in which only seven enzymes (*Acy* I, *Bbv* I, *Bcl* I, *Bgl* I, *Hae* II, *Hin*dIII, and *Sph* I) cleave randomly in all respects. Furthermore, none of these enzymes has a four-base recognition sequence and only one has a five-base recognition sequence—the two categories of enzymes that provide the possibility of producing reasonable amounts of data. Therefore, the use of restriction enzymes to estimate phylogenetic relationships will generally dictate that enzymes with nonrandom numbers and distributions of sites be used. However, it must be remembered that the assumption of randomness is also commonly violated when protein sequence or similar data are used to generate evolutionary trees. Therefore, we conclude that restriction enzyme data are subject to the same limitations and deficiencies as protein data, and, given the limited number of sequences that may be assayed, may at present be less suitable than protein data for inferring evolutionary relationships.

We thank W. M. Brown, R. B. Helling, J. Hixson, P. Smouse, and E. White for helpful comments on the manuscript. This work was supported in part by Department of Energy Contract ACO-2-76-EVO2828.

1. Avise, J. C., Lansman, R. A. & Shade, R. O. (1979) *Genetics* **92**, 279–295.
2. Brown, W. M., George, M., Jr., & Wilson, A. C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1967–1971.
3. Upholt, W. B. & Dawid, I. B. (1977) *Cell* **11**, 571–583.
4. Ferris, S. D., Wilson, A. C. & Brown, W. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2432–2436.
5. Goodman, M. (1976) in *Molecular Evolution*, ed. Ayala, F. J. (Sinauer, Sunderland, MA), pp. 123–140.
6. Ward, R. H. (1972) *Ann. Hum. Genet.* **36**, 21–43.
7. Dobzhansky, T. (1970) *Genetics of the Evolutionary Process* (Columbia Univ. Press, New York).
8. Upholt, W. B. (1977) *Nucleic Acids Res.* **4**, 1257–1267.
9. Gotoh, O., Hayashi, J.-I., Yonekawa, H. & Tagashira, Y. (1979) *J. Mol. Evol.* **14**, 301–310.
10. Kaplan, N. & Langley, C. H. (1979) *J. Mol. Evol.* **13**, 295–304.
11. Nei, M. & Li, W.-H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
12. King, J. L. & Jukes, T. H. (1969) *Science* **164**, 788–798.
13. Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967) *Am. J. Hum. Genet.* **19**, 233–257.
14. von Hippel, P. H. (1979) in *Biological Regulation and Development*, ed. Goldberger, R. F. (Plenum, New York), Vol. 1, pp. 279–347.
15. Nei, M. & Tajima, F. (1981) *Genetics* **97**, 145–163.
16. Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. C. (1981) *Nature (London)* **290**, 457–465.
17. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volkaert, G. & Ysbaert, M. (1978) *Nature (London)* **273**, 113–120.
18. Van Heuverswyn, H. & Fiers, W. (1979) *Eur. J. Biochem.* **100**, 51–60.
19. Yang, R. C. A. & Wu, R. (1979) *Science* **206**, 456–462.
20. Deininger, P. L., Esty, A., LaPorte, P., Hsu, H. & Friedman, T. (1980) *Nucleic Acids Res.* **8**, 855–860.
21. Sanger, F., Coulson, A. R., Friedmann, T., Air, G. N., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., III, Slocombe, P. M. & Smith, M. J. (1978) *J. Mol. Biol.* **125**, 225–246.
22. Beck, E., Sommer, R., Anerswald, E. A., Kurz, C., Zink, B., Osterburg, G. & Schaller, H. (1978) *Nucleic Acids. Res.* **5**, 4475–4503.
23. Godson, G. N., Barrell, B. G., Staden, R. & Fiddes, J. C. (1978) *Nature (London)* **276**, 236–247.
24. Roberts, R. J. (1980) *Nucleic Acids Res.* **8**, r63–r80.
25. Subak-Sharpe, H., Bürk, R. R., Crawford, L. V., Morrison, J. M., Hay, J. & Keir, H. M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31**, 737–748.
26. Morrison, J. M., Keir, H. M., Subak-Sharpe, H. & Crawford, L. V. (1967) *J. Gen. Virol.* **1**, 101–108.
27. McGeoch, D. J., Crawford, L. V. & Follett, E. A. C. (1970) *J. Gen. Virol.* **6**, 33–40.
28. Barrell, B. G., Anderson, S., Bankier, A. T., deBruijn, M. H. L., Chen, E., Coulson, A. K., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3164–3166.
29. Nussinov, R. (1981) *J. Mol. Biol.* **149**, 125–131.
30. Bonitz, S. G., Berlani, R., Coruzzi, G., Li, M., Macino, G., Nobrega, F. G., Nobrega, M. P., Thalenfeld, B. E. & Tzagoloff, A. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3167–3170.
31. Watson, G. S. (1961) *Biometrika* **48**, 109–114.
32. Watson, G. S. (1962) *Biometrika* **49**, 57–63.
33. Stent, G. S. & Calendar, J. R. (1978) *Molecular Genetics, An Introductory Narrative* (Freeman, San Francisco).
34. Spoerel, N., Herrlich, P. & Bickle, T. A. (1979) *Nature (London)* **278**, 30–34.
35. Proudfoot, N. J. (1979) *Nature (London)* **279**, 376.
36. Soeda, E., Arrand, J. R., Smolar, N., Walsh, J. E. & Griffin, B. E. (1980) *Nature (London)* **283**, 445–453.
37. Clarke, B. (1970) *Nature (London)* **228**, 159–160.
38. Benzer, S. (1961) *Proc. Natl. Acad. Sci. USA* **47**, 403–415.