Statistical Methods of DNA Sequence Analysis: Detection of Intragenic Recombination or Gene Conversion[1]

J. Claiborne Stephens
Center for Demographic and Population Genetics, University of Texas at Houston

Simple but exact statistical tests for detecting a cluster of associated nucleotide changes in DNA are presented. The tests are based on the linear distribution of a set of s sites among a total of n sites, where the s sites may be the variable sites, sites of insertion/deletion, or categorized in some other way. These tests are especially useful for detecting gene conversion and intragenic recombination in a sample of DNA sequences. In this case, the sites of interest are those that correspond to particular ways of splitting the sequences into two groups (e.g., sequences A and D vs. sequences B, C, and E–J). Each such split is termed a *phylogenetic partition.* Application of these methods to a well-documented case of gene conversion in human γ-globin genes shows that sites corresponding to two of the three observed partitions are significantly clustered, whereas application to hominoid mitochondrial DNA sequences—among which no recombination is expected to occur—shows no evidence of such clustering. This indicates that clustering of partition-specific sites is largely due to intragenic recombination or gene conversion. Alternative hypotheses explaining the observed clustering of sites, such as biased selection or mutation, are discussed.

Introduction

Gene conversion has recently become an important consideration in the evolutionary analysis of multigene families (Arnheim 1983). The globin gene family in primates (Lauer et al. 1980; Slightom et al. 1980; Zimmer et al. 1980; Shen et al. 1981; Scott et al. 1984) and heat-shock loci in *Drosophila melanogaster* (Leigh Brown and Ish-Horowicz 1981) are well-studied examples of small multigene families in which interlocus gene conversion seems to have occurred, while ribosomal genes in hominoids (Arnheim et al. 1980) and loci of the mammalian histocompatibility complex (Hood et al. 1975; Mellor et al. 1983) provide evidence of gene conversion in larger multigene families. The phenomena of interallelic gene conversion and intragenic recombination have also been recognized and documented (Chovnick et al. 1971; Kreitman 1983), although appropriate data were hard to obtain until DNA sequencing and restriction-site mapping became available.

However, current methods of detecting recombination or gene conversion among DNA sequences are largely intuitive, and it is not always clear whether unusual similarity between two DNA sequences is due to chance, gene conversion, or functional constraints. In view of this situation, I have developed rigorous statistical methods by which one can identify and characterize gene conversion and other recombinational events. The mathematical theory presented below is useful for analyzing the linear

1. Key words: gene conversion, recombination, DNA sequence analysis, evolution.

Address for correspondence and reprints: Dr. J. Claiborne Stephens, Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225.

Mol. Biol. Evol. 2(6):539–556. 1985.

0737-4038/85/0206-0366$02.00

distribution of any group of s sites in a sequence of n sites. For instance, the distribution of variable sites, sites of insertion/deletion, restriction-enzyme recognition sites, or simply particular bases, may be tested for clustering. Of special importance for detection of intragenic recombination and gene conversion are those variable sites that yield identical phylogenetic information.

Theory

In principle, any group of s sites in a sequence of total length n sites may be distinguished by some criterion, and then the linear distribution of these s sites among the n sites may be tested for departures from a random distribution. I will use several tests in two different contexts. First, I will determine whether or not the variable (polymorphic) sites have a nonrandom distribution. Second, clusters of variable sites associated with particular groupings or partitionings of the DNA sequences will be identified. The latter context is important in inferring the events of gene conversion or recombination in phylogenetic analysis.

Concept of a Phylogenetic Partition

Variable nucleotide sites create phylogenetic partitions (henceforth *partitions*) in the sense that all sequences that share one variant at a particular site are distinguished from those sequences that share an alternative nucleotide (fig. 1). Any particular site can partition a sample of sequences into a maximum of four groups, corresponding to the four possible nucleotides. In the analysis of polymorphic sequences within species, however, it is usually sufficient to consider partitions leading to only two groups of sequences. For m different sequences, there are $2^{m-1} - 1$ different ways of dividing the sequences into two groups.

I will treat gaps arising from insertion/deletion as segregating sites; hence, insertion/deletion events also create partitions. Restriction-site polymorphisms also generate partitions among DNA sequences, but in this case some modifications of the theory are required. Therefore, I shall consider this problem in a subsequent paper.

Consider mutations occurring at two sites, as shown in figure 2. Two unique mutations occurring in the same branch (fig. 2*A*.) will be found in all descendants from that point onward and will not be found elsewhere. Thus sites i and j create the same phylogenetic partition ABCDE/FGH. All sites creating the same partition will be called *congruent* sites, and I will refer to them as being congruent to the partition as well as to each other.

Some mutations might not be unique and occur several times in parallel (e.g., mutations at site i in fig. 2*B*.). In this case unrelated sequences will appear to be related on the basis of this site. Similarly, a mutation may revert to a previously existing

Sequence										
A	A	G	C	T	G	A	C	G	T	G
B	A	A	C	T	A	G	C	A	T	A
C	A	G	C	C	G	A	T	G	T	A
D	A	G	C	C	G	A	T	A	T	A
E	A	G	C	C	G	A	T	A	T	A
Partition:		2		a	2	2	a	b		1

FIG. 1.—Examples of some partitions created by the comparison of five nucleotide sequences. Partition 1, A/BCDE; 2, B/ACDE; a, AB/CDE; b, AC/BDE.

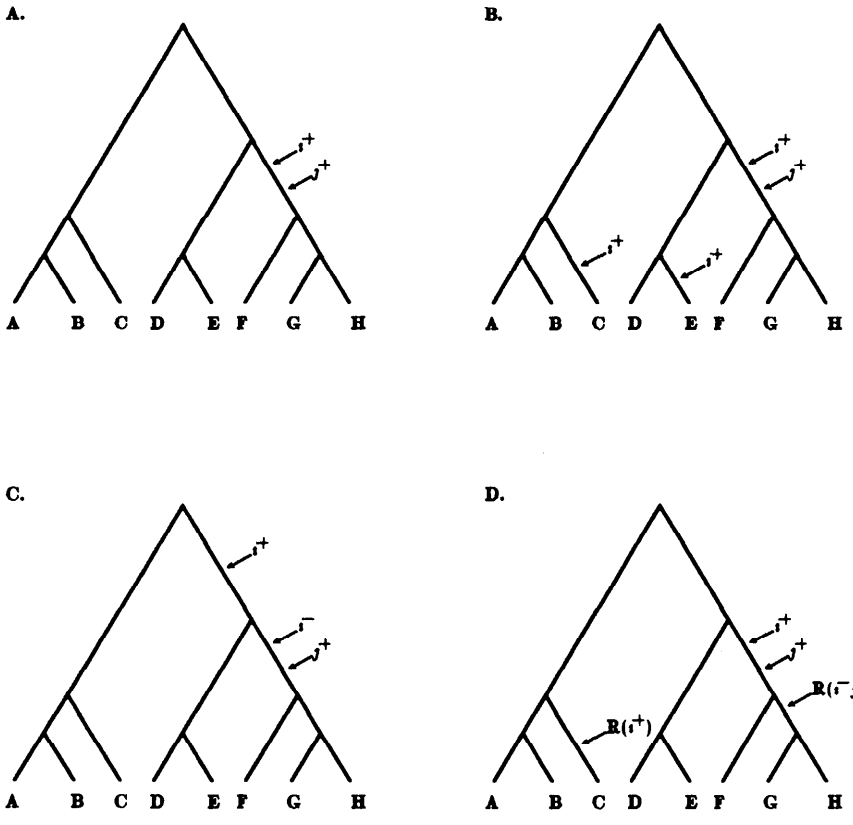

FIG. 2.—Effect of different mutational events on labeling of partitions. A–H are sequences; i and j are nucleotide sites. *A.*, Unique mutations occur at sites i and j. Partitions: i = ABCDE/FGH; j = ABCDE/FGH. *B.*, Parallel mutations occur at site i; unique mutation occurs at site j. Partitions: i = ABD/CEFGH; j = ABCDE/FGH. *C.*, Reverse or backward mutation occurs at site i; unique mutation occurs at site j. Partitions: i = ABCFGH/DE; j = ABCDE/FGH. *D.*, Recombination may affect sites i and j differently. In one case, $R(i^-j^-)$, the lineage leading to sequences G and H loses both mutations i^+ and j^+. In another case, $R(i^+)$, the lineage leading to sequence C gains the i^+ mutation. Partitions: i = ABDEGH/CF; j = ABCDEGH/F.

nucleotide (e.g., mutation at site i in fig. 2*C*.). In such cases, mutations at sites i and j create different partitions.

In the case of intraspecific nucleotide sequence comparisons, recombination or gene conversion subsequent to the mutational events in figure 2*A*. could transfer one or both of these mutations to different lineages (fig. 2*D*.). Hence, if recombinational events occur with an appreciable frequency, two mutational events that occur in the same lineage in the same time interval, such as those in figure 2*A*., might not yield the same phylogenetic partition. In this case sites i and j would only be congruent if there were no recombination between them.

Given these observations, it will be useful to call a partition *primary* if it corresponds to unique, undisturbed mutations, as in figure 2*A*., whereas a *secondary* partition would be one created by multiple events occurring at a given site (e.g., site i in figs. 2*B*. and 2*C*., sites i and j in fig. 2*D*.). While it is usually difficult to know which partitions are primary and which are secondary, it is possible to infer which are primary under certain conditions, as will be discussed below.

In the simple situation of no recombination and no parallel or backward mutations, two mutations such as those in figure 2*A*. must create the same partition, but these mutations need not occur physically near each other within the linear sequence of the gene. I will show below that clustering of sites that are congruent to secondary partitions indicates that such partitions were created by recombinational events.

Statistical Tests of Clustering

Spatial clustering of any set of s sites may be tested under the assumption that such sites are randomly distributed along the DNA sequence. Let us start with a pair of sites i and j ($j > i$) in a sequence of n bp, i.e., $s = 2$, and let $d = j - i$ be the distance between the pair of sites. Obviously, there is only one possible way in which $d = n - 1$, two ways in which $d = n - 2$, and so on down to $n - 1$ ways in which $d = 1$, i.e., the sites are adjacent. Therefore, for $s = 2$ the distribution of d is triangular and given by $f_d = n - d$, $1 \leqq d \leqq n - 1$. The absolute frequency f_d is correct if each of the $\binom{n}{2}$ possible combinations of pairs of sites is equally likely. This is not necessarily true for a pair of variable sites because different regions of the DNA sequence are often under different functional constraints. Fortunately, "hot" and "cold" spots of variability can be detected with some of the tests described later. I will show later that the average distance between two sites is $(n + 1)/3$.

Figure 1 shows two sites corresponding to partition a (AB/CDE). A test is needed to determine whether or not these sites are significantly clustered under the assumption that they could have occurred anywhere in the sequence with equal probability. For this purpose, I use the probability that the distance d between a random pair of sites is less than or equal to the observed distance (d_o) between the two s-sites,

$$P(d \leqq d_o) = \sum_{d=1}^{d_o} (n - d) \Big/ \binom{n}{2} = [nd_o - d_o(d_o + 1)/2] \Big/ \binom{n}{2}$$

$$= 2\left(\frac{d_o}{n-1}\right) - \left(\frac{d_o}{n-1}\right)\left(\frac{d_o+1}{n}\right) \cong 2x - x^2 \text{ for } n \text{ large}, \tag{1}$$

where $x = d_o/(n - 1)$. This indicates that two sites ($s = 2$) separated by a distance $d_o \leqq 0.025(n - 1)$ are significantly clustered at the 5% level. On the other hand, $P(d > d_o) = 1 - P(d \leqq d_o) \cong (1 - x)^2$, so that $(1 - x)^2 = 0.05$ for $P(d > d_o) = 0.05$. Therefore, two sites separated by a distance $d_o \geqq 0.776(n - 1)$ are significantly far apart at the 5% level. For example, in a sequence of 2,000 bases, two congruent sites are significantly close together at the 5% level if they are within 50 bp of each other, whereas they are significantly far apart if they are separated by more than 1,552 bp.

For $s = 3$ (e.g., the three sites congruent to partition 2 in fig. 1), consider the three sites as the ordered triplet i, j, k, where $1 \leqq i < j < k \leqq n$. There are several possible ways to proceed, but note that for three points on a straight line, the average distance, the sum of pairwise distances, and the largest pairwise distance all have a distribution of the same form. This is seen by noting that $(k - i) + (k - j) + (j - i) = 2(k - i)$, which is just twice the largest distance. Therefore, a test for clustering of the three sites can be based on the length of the interval spanned by the three sites, that is, on the distance $d = k - i$. There are $n - d$ intervals of length d, with the intermediate site j occupying any of the $d - 1$ possible positions between i and k. For

$s > 3$ sites, there are $\binom{d-1}{s-2}$ ways of choosing the $s - 2$ intermediate sites, where d is again the length of the interval spanned by all s sites. Hence, in general,

$$f_d = (n - d)\binom{d-1}{s-2} \qquad s - 1 \leqq d \leqq n - 1; \tag{2}$$

$f_d = 0$ otherwise. It can be shown that

$$\sum_{d=1}^{n-1} f_d = \binom{n}{s}, \tag{3}$$

i.e., the number of ways of choosing s of n possible nucleotide sites. The method of proof of equation (3) is followed in the determination of $P(d \leqq d_o)$. Letting $l = s - 2$, note that

$$\begin{aligned}\sum_{d=1}^{d_o} f_d &= \sum_{d=1}^{d_o} (n - d)\binom{d-1}{l} = n \sum_{d=1}^{d_o} \binom{d-1}{l} - (l + 1) \sum_{d=1}^{d_o} \binom{d}{l+1} \\ &= n \sum_{j=0}^{d_o-1} \binom{j}{l} + (l + 1)\binom{0}{l+1} - (l + 1) \sum_{d=0}^{d_o} \binom{d}{l+1}.\end{aligned} \tag{4}$$

The second term above is 0 by the convention that $\binom{j}{l} = 0$ for $j < l$, while the summations in the first and third terms are $\binom{d_o}{l+1}$ and $\binom{d_o+1}{l+2}$, respectively (Feller 1968, p. 64). From equations (2), (3), and (4),

$$\begin{aligned}P(d \leqq d_o) &= \sum_{d=1}^{d_o} f_d \Big/ \binom{n}{s} = \left\{n\binom{d_o}{l+1} - (l + 1)\binom{d_o + 1}{l + 2}\right\} \Big/ \binom{n}{l+2} \\ &= [s - (s - 1)(d_o + 1)/n] \prod_{j=0}^{s-2} \frac{d_o - j}{n - j - 1}.\end{aligned} \tag{5}$$

Equation (5) is useful for testing whether or not s sites are significantly clustered relative to the regions flanking them. Note that equation (5) yields equation (1) as a special case for $s = 2$. If $s \ll d_o$, equation (5) allows the useful approximation $P(d \leqq d_o) \cong sx^{s-1} - (s - 1)x^s$, where $x = d_o/(n - 1)$. In the Appendix, the mean of the distribution of d is shown to be

$$E(d) = \sum_{d=1}^{n-1} d f_d \Big/ \binom{n}{s} = (s - 1)(n + 1)/(s + 1). \tag{6}$$

Note that d is expected to be nearly the length of the sequence when s is large. When $s = 2$, this becomes $(n + 1)/3$, as noted earlier.

The probability presented in equation (5) is useful for testing whether or not a set of s specified sites is clustered relative to the regions flanking them. In many applications, it is important to determine whether subsets of the sites are clustered. For $s = 3$, the test is simple. In this case, I assume that the intermediate site may occupy any of the $d_o - 1$ positions between the other two sites with equal probability. Therefore,

the probability that the intermediate site is as close as or closer than the observed distance (d_E) to either of the other two sites is

$$P(d \leqq d_E) = 2d_E/(d_o - 1), \qquad 1 \leqq d_E \leqq (d_o - 1)/2. \tag{7}$$

The situation becomes more complicated for $s > 3$ sites. Figure 3 shows two examples in which the test in equation (5) fails to show clustering yet clustering of some of the sites is apparent. The feature underlying the apparent clustering in these and other distributions is that a group of s-sites is bounded on either or both sides by many consecutive non–s sites. There is a simple test to see whether the segments of consecutive non–s sites are unusually long.

The length $d_o + 1$ consists of two terminal s-sites, with $s - 2$ of the s sites dispersed among $r \equiv d_o - (s - 1)$ non–s sites (fig. 3*A*.). The s sites of interest and these other r sites correspond to a simple occupancy problem in probability theory (Feller 1968). There are $s - 1$ "spaces" between the s sites, into which the other r sites can be "placed." The probability that there will be exactly k of the r-sites between a randomly chosen pair of consecutive s sites is (Feller 1968, p. 61)

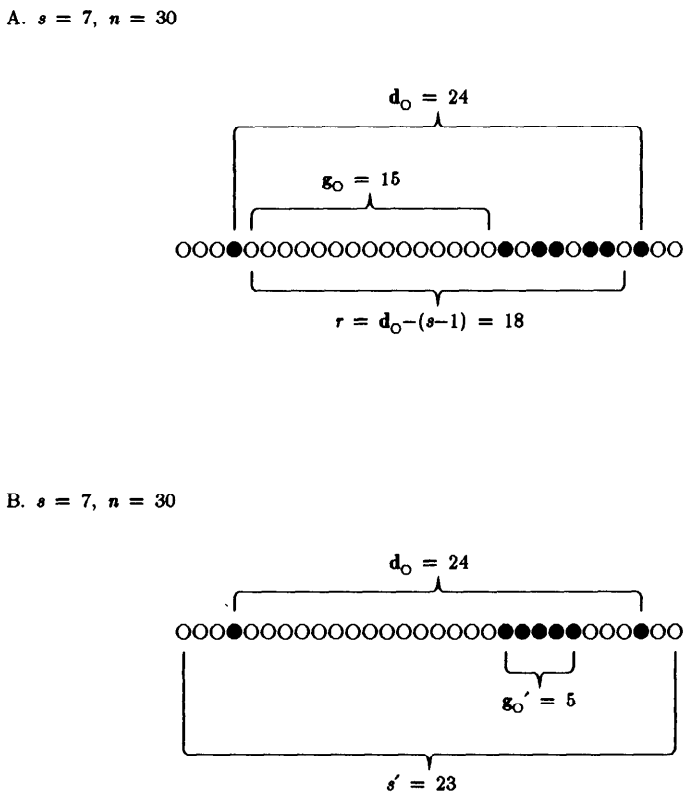

FIG. 3.—Two examples of clustering that would be undetectable by eq. (5), since $P(d \leqq d_o) = 0.5667$ in each case. A., A segment of r-sites (open circles, $g_o = 15$) is too long, since $P = 0.0099$ by eq. (10). This is also an indication that the six s-sites (black dots) on the right are clustered. B., Slight modification of case A. in which an apparent cluster of s-sites can be tested directly. The probability of having five or more adjacent s-sites by chance is $P' = 0.0047$, which indicates that these sites are significantly clustered.

$$g_k \equiv P(g = k) = \binom{s + r - k - 3}{r - k} \Big/ \binom{s + r - 2}{r}, \tag{8}$$

where $0 \leqq k \leqq r$. Note that for $s > 3$, $g_0 > g_1 > \cdots > g_r$, whereas for $s = 3$, $g_0 = g_1 = \cdots = g_r = 1/(r + 1)$, as was assumed in the derivation of equation (7). Equation (8) may be used to evaluate the probability that a random segment of consecutive r-sites is as long as or longer than the longest observed segment (length g_o),

$$P_m \equiv P(g \geqq g_o) = \sum_{k=g_o}^{r} g_k. \tag{9}$$

This probability is then used to evaluate the probability that at least one of $s - 1$ random, independently observed segments would be as long as or longer than g_o:

$$P \equiv 1 - (1 - P_m)^{s-1} \cong (s - 1)P_m. \tag{10}$$

Equation (10) is primarily useful when P_m is small and it can be assumed that the $s - 1$ segments actually observed are roughly independent. For example, in figure 3*A*, $s = 7$, $r = 18$, and $g_o = 15$. By equations (9) and (10), $P_m = 0.0017$ and $P = 0.0099$, which indicates that the six s-sites on the right are clustered, since one of six segments of consecutive r-sites is unusually long. Note that P_m corresponds to a randomly sampled segment and that P corresponds to a random sample of $s - 1$ segments. In general, P will be more useful than P_m, since I target the longest observed segment as the one to test. If s and r are relatively large, the entire distribution of r-site segments may be tested against the probabilities in equation (8) by using a standard goodness-of-fit test. Brown and Clegg (1982) tested the clustering of variable nucleotide sites in *Zea mays* knob heterochromatin sequences, basing their test on an underlying geometric distribution for all unvaried segment lengths. The geometric probabilities are, in fact, the limit of those in equation (8) for large r and s (e.g., Feller 1968).

Finally, for extreme cases of clustering (several adjacent s sites), it is useful to reverse the correspondence of r-sites and s-sites in equation (8), (9), and (10). For example, in figure 3*B*, five s-sites are adjacent, and it is desirable to directly evaluate the probability of a run of s-sites this long. To avoid confusion, I will add primes to s, r, g, g_o, P_m, and P when testing segments of consecutive s-sites. Thus $s' = 23$, $r' = 7$, $g'_o = 5$, $P'_m = 2.14 \times 10^{-4}$, and $P' = 0.0047$, which indicates that the segment of consecutive s-sites is significantly long. I now turn to an illustration of these different tests and a discussion of the biological meaning of the detected nonrandomness.

Applications

Two sets of published data were chosen to illustrate the statistical methods. The first set, i.e., the nucleotide sequences from three human γ-globin genes (Slightom et al. 1980), appears to be one of the best-documented cases of gene conversion. My second application is to the homologous mitochondrial DNA segments obtained from five hominoid species (Brown et al. 1982). The latter sequences are my "control," in that no recombination would be expected among mitochondrial DNA sequences, especially those taken from different species. It is my intention to show that it is the clustering of partition-specific sites that gives an indication of recombination or gene

conversion. It is, however, necessary to apply the tests first to the entire set of variable sites to detect "hot" and "cold" spots of variation.

Human γ-Globins

Three human γ-globin genes from the same individual ($^{G}\gamma$-I, from chromosome I; $^{A}\gamma$-I, the duplicate gene from the same chromosome; and $^{A}\gamma$-II, an allelic variant of $^{A}\gamma$-I from chromosome II) have been sequenced (Slightom et al. 1980). Nucleotide sequence comparisons among these and two homologous gorilla sequences (Scott et al. 1984) have identified regions in which sequence identity is considerably greater than would be expected under the assumption that the gene duplication is at least 35 Myr old. Several gene conversions during hominoid evolution have been invoked as an explanation of this observation. Furthermore, a region of simple DNA sequence called the "hot spot" was apparently involved in these conversions and in multiple substitutions and length variations. Of particular interest for my purpose is the most recent of these conversions, i.e., that of $^{A}\gamma$-I by $^{G}\gamma$-I, which was identified by Slightom et al. (1980).

Nucleotide sequence variations were categorized according to the phylogenetic partition each produced (fig. 4 and table 1). Note that in this case there are only three possible partitions, which are labeled 1, 2, and 3. At no site outside the putative hot spot (sites 1,080–1,099) were all three sequences different from each other. My treatment of the hot spot, however, recognizes the $^{A}\gamma$-I hot spot sequence as ancestral, with $^{G}\gamma$-I and $^{A}\gamma$-II each having a unique insertion in this region (fig. 5). This treatment seems reasonable, since the gorilla $^{A}\gamma$ is identical to $^{A}\gamma$-I in this region (Scott et al. 1984). The precise location of these insertions is ambiguous, because the sequences of the insertions and the surrounding regions are all very similar (fig. 5). Fortunately, the statistical tests used here are not sensitive to slight ambiguities in the precise location of variations. For my analysis, I have treated the two hot spot insertions as corresponding to both partitions 1 and 3 at site 1,080 (fig. 4).

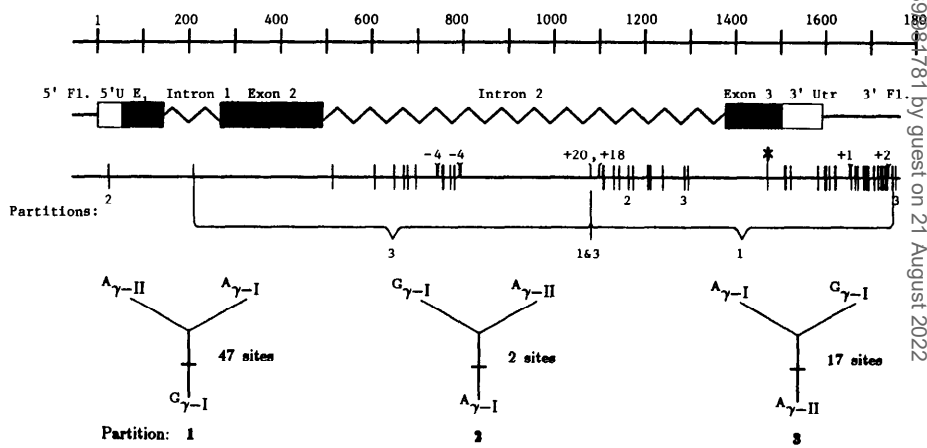

FIG. 4.—Locations of variable sites (vertical lines) and the corresponding phylogenetic partitions in the human γ-globin nucleotide sequences. *Top,* Scale is that used by Slightom et al. (1980), extended to include additional sequence as published by Scott et al. (1984). Five sites of length variation are labeled with the length corresponding to the partition. The glycine → alanine difference between $^{G}\gamma$ and $^{A}\gamma$ is labeled with an asterisk. The "hot spot" sequences, sites 1,080–1,099, are regarded here as two unique insertions relative to $^{A}\gamma$-I. Site 1,080 is treated as congruent to both partitions 1 and 3 to reflect this consideration. *Bottom,* Phylogenetic interpretation and number of sites corresponding to each of the three partitions.

Table 1
Clustering of Variable Sites in Human γ-Globin Nucleotide Sequences

Partition[a]	s	Range[b]	d_o[c]	$P(d \leq d_o)$ by eq. (5)
1	47	1,080–1,748	648	5.0×10^{-20}
2	2	25–1,163	1,113	0.86
3	17	210–1,754	1,518	0.24
Total	65[d]	25–1,754	1,703	0.15

NOTE.—n = 1,793.
[a] See fig. 4.
[b] Position of 5′-most and 3′-most sites.
[c] Calculated as the distance covered by the range, less any excess insertion lengths covered by range.
[d] Site 1,080 corresponds to both partitions 1 and 3.

I have retained the position numbering used by Slightom et al. (1980) (−56 to −1, 1 to 1,592), extended through position 1,763 by including the additional 3′ sequence of these genes as published by Scott et al. (1984). All lengths (e.g., n, d_o, g_o) including insertions should be shortened, however, because of the sequence gaps caused by insertion/deletion. This was done by regarding each gap as being one nucleotide long. For instance, for tests using equation (5), as in table 1, $n = 56 + 1{,}763 - (3 + 3 + 19 + 0 + 1) = 1{,}793$, where the five numbers in parentheses are each one less than the number of nucleotides involved in the five gaps covered by this study (see fig. 4).

Before considering the partitions individually, it is useful to first identify regions of extremely high or low variability. Application of equation (5) to the entire set of variable sites (s = 65, n = 1,793, and d_o = 1,703) did not indicate clustering (table 1). However, tests of the distribution of observed lengths of consecutive unvaried sites by equation (8) showed that regions of very high and/or very low variability were present (table 2). Extremes of variability may be tested individually by using equation (10). For example, in the top of table 3 are shown the probabilities of observing segments of consecutive unvaried sets as long as or longer than each of the six longest segments observed. The probability of observing the longest of these (g_o = 303, covering exon 2 and parts of introns 1 and 2) in a random sample of 64 segments is very low, under the assumption of random distribution of varied and unvaried sites. It is therefore justifiable to exclude this segment from further analyses and to recalculate the probability of observing a segment as long as or longer than the next largest segment observed (g_o = 285, r = 1,336, s = 64). Note that in this procedure r becomes $r - g_o$ and s becomes $s - 1$, since each unvaried segment is terminated by a variable site. The rightmost column of table 3 contains the probabilities (P) obtained by this procedure. The five longest segments show a very low probability of being observed under

	1070	1080	1090	1100	1110
$^{G}\gamma$–I	TGTGTGTGTG	TGTGCGCGCGTGTGTTTGTG		TGTGTGTGAGA	
$^{A}\gamma$–I	TGTGTGTGTG	----------		TGTGTGTGTCA	
$^{A}\gamma$–II	TGTGTGTGTG	CGCGCGCGCGTGTG--TGTG		TGTGTGTGTCA	

FIG. 5.—Sequence of the hot spot and surrounding region, taken from Slightom et al. (1980). Although homology and alignment of two "inserts" between sites 1,080 and 1,099 is reasonable, these are regarded as a 20-bp insert in $^{G}\gamma$-I and an 18-bp insert in $^{A}\gamma$-II.

Table 2
Fit of Observed Lengths of Unvaried Site Segments to Those Expected: Human γ-Globin Nucleotide Sequences

Length	Observed Frequency (O_1)	Expected Frequency[a] (E_1)	$\frac{(E_1 - O_1)^2}{E_1}$	Length	Observed Frequency (O_2)	Expected Frequency[b] (E_2)	$\frac{(E_2 - O_2)^2}{E_2}$
0–2	23	6.85	38.07	0–2	20	12.94	3.86
3–5	6	6.13	0.00	3–5	6	9.99	1.59
6–8	9	5.48	2.26	6–8	9	7.70	0.22
9–12	7	6.41	0.05	9–12	7	7.59	0.05
13–16	2	5.52	2.25	13–16	2	5.35	2.19
17–21	4	5.84	0.58	17–21	4	4.51	0.06
22–27	1	5.70	3.87	22–27	1	3.33	1.63
28–34	2	5.21	1.97	>27	7	4.59	1.25
35–44	2	5.40	2.14			$X^2 =$	10.77
45–59	2	5.07	1.86				
>59	6	6.39	0.02				
		$X^2 =$	53.08				

SOURCE.—Slightom et al. (1980) and Scott et al. (1984).

NOTE.—Expected frequencies (under eq. [8]) were pooled to exceed five in each class (in most cases).

[a] Based on all variable sites, with $s = 65$, and $r = 1639$. $X^2 = 53.08$, $P(\chi^2_{(10)} > 25.19) < 0.005$.

[b] The five largest unvaried segments and the run of four consecutive variable sites were tested and excluded as being improbably long (see table 3), using eq. (10). Expectations were recalculated, using eq. (8), with $s = 57$ and $r = 602$. $X^2 = 10.77$, $P(2.17 < \chi^2_{(7)} < 14.07) = 0.90$.

Table 3
Probabilities of Runs of Extreme Lengths of Consecutive Unvaried Sites and of Varied Sites Observed in Human γ-Globin Nucleotide Sequences

Length[a]	Location	P_m	P	P[b]
g_o:				
303	Sites 211–513; intron 1, exon 2, intron 2	3.35×10^{-6}	2.1×10^{-4}	. . .
285	Sites 795–1,079; intron 2, 5′ of hot spot	7.63×10^{-6}	4.9×10^{-4}	3.2×10^{-5}
184	Sites 26–209; 5′ untranslated, exon 1, intron 1	6.16×10^{-4}	0.039	7.1×10^{-4}
175	Sites 1,296–1,470; intron 2, exon 3	9.39×10^{-4}	0.058	1.4×10^{-4}
90	Sites 515–604; intron 2	0.0305	0.862	0.023
60	Sites 1,523–1,582; 3′ untranslated	0.0999	0.999	0.162
g'_o:				
4	Sites 1,508–1,511; 3′ untranslated	1.58×10^{-6}	0.003	. . .
3	Sites 1,706–1,708; 3′ flanking	4.57×10^{-5}	0.076	0.064

[a] For g_o, $s = 65$ and $r = 1{,}639$; for g'_o, $s' = 1{,}728$ and $r' = 65$.

[b] Recalculated after omitting all lengths from previous rows. For instance, for $g_o = 184$, P was recalculated with $r = 1{,}639 - 303 - 285 = 1{,}051$ and $s = 63$. By eq. (9), $P_m = 1.15 \times 10^{-5}$, from which P is calculated by eq. (10).

the assumption of random distribution of variable sites. These segments correspond to the highly conserved regions identified by Slightom et al. (1980) (fig. 4).

Table 2 also suggests that some of the variable sites may be clustered. Actual "hot spots" of variability are somewhat harder to identify, since the shortest possible unvaried segments ($g = 0$) are expected to be the most frequent, from equation (8). As was indicated previously, the most extreme cases can be tested by reversing the correspondence between r and s in equations (8), (9), and (10). At the bottom of table 3 are shown the probabilities of obtaining runs of variable sites, now with $r' = 65$ and $s' = 1{,}793 - 65 = 1{,}728$. There is a run of four consecutive variable sites in the 3′ untranslated region and a run of three in the 3′ flanking region. By equation (10), $P' = 0.0027$ for $g'_o = 4$, and $P' = 0.064$ for $g'_o = 3$ after deleting the run of four (table 3), which means that, in this context, only the run of four variable sites is improbably long.

The preceding analysis of extreme variabilities illustrates a dilemma: is it the runs of variable sites or the runs of unvaried sites (or both) that are too long? Note that exclusion of long segments of unvaried sites will make long runs of variable sites more likely and vice versa. Therefore, one approach to resolving this dilemma is to test runs of variable sites after excluding the most significant unvaried segments and also to test the latter after excluding the most significant runs of variable sites. For example, runs of four or more variable sites are still improbable ($P'_m = 5.9 \times 10^{-5}$ and $P' = 0.035$, with $r' = 60$ and $s' = 611$) after excluding the five longest unvaried segments. Similarly, even the fifth longest unvaried segment is still significant after excluding the run of four variable sites ($P_m = 6.4 \times 10^{-4}$ and $P = 0.035$, with $r = 691$ and $s = 57$). I therefore regard the five longest unvaried segments and the run of four variable sites as being statistically significant. Exclusion of these and application of equation (8) improved the fit to expectation considerably (table 2). It will be necessary to consider these regions when analyzing the clustering of sites congruent to each partition.

Forty-seven sites are congruent to partition 1 (table 1). These sites are all confined to a 648-bp region at the 3′ end of the sequenced region (fig. 4). The probability (from equation [5]) that these sites would be clustered so tightly by chance is 5.05×10^{-20} (table 1). This calculation highlights the considerable difference between the two regions on opposite sides of the hot spot. Of course, this is precisely what Slightom et al. (1980) inferred by comparing the sequences. One reason for this difference is the general lack of variability in the region upstream of the hot spot (fig. 4). Before invoking gene conversion, it is necessary to recalculate the probability after excluding the long unvaried segments (table 3), i.e., nucleotides at positions −56–605, 792–1,099, and 1,296–1,471. I obtain $P(d \leqq d_o) = 9.1 \times 10^{-5}$ from equation (5), based on $n = 670$, $s = 45$, and $d_o = 472$. Additionally excluding the run of four adjacent variable sites (all congruent to partition 1), the probability becomes 2.3×10^{-4}. Therefore, although there is a dramatic difference in variability on opposite sides of the hot spot, it is not the only factor affecting the distribution of sites congruent to partition 1.

Unlike partition 1, partition-3 sites ($s = 17$, $d_o = 1{,}518$) were not significantly clustered when the entire set of 17 sites was considered (table 1). However, it is obvious that in this case most of these sites are on the 5′ side of the hot spot and that the two sites on the 3′ side of the hot spot are widely separated (fig. 4). I therefore test the distance between the two 3′ sites with equation (10). It is illustrative to temporarily ignore the differences in variability noticed earlier. Hence, with $r = 1{,}502$, $s = 17$, and $g_o = 461$, the distance is not significantly long, since $P = 0.066$. This calculation did

not exclude the regions previously identified as being highly conserved, and these are mostly in the 5′ region. Again excluding nucleotides corresponding to the five longest unvaried segments, I obtain $P = 0.0133$ from equation (10), with $s = 14$, $r = 649$, and $g_o = 286$. Furthermore, the effective distance between site 1,288 and site 791 is 188 bases (recall that I have excluded insertion lengths and an unvaried segment here), for which $P = 0.0047$ after excluding $g_o = 286$. The biological conclusion drawn from this statistical observation is that sites corresponding to partition 3 are too scarce on the 3′ side of the hot spot or, conversely, somewhat enriched on the 5′ side (fig. 4). A statistical difference between the sequence on the 5′ side of the hot spot and that on the 3′ side is apparent for sites congruent to partition 1 and partition 3, and the difference is not the result of more general differences in variability on opposite sides of the hot spot.

The distribution of sites congruent to partitions 1 and 3 is complementary, as was graphically demonstrated in the study by Slightom et al. (1980). The gene conversion identified by Slightom et al. (1980), i.e., conversion of the region 5′ of the hot spot in $^{A}\gamma$-I by the corresponding region of $^{G}\gamma$-I, is certainly a reasonable explanation of this observation. Furthermore, although the two sites congruent to partition 2 are not significantly close together (table 1), the lack of partition-2 sites is consistent with a recent conversion of $^{A}\gamma$-I.

Hominoid Mitochondrial DNA (mtDNA)

Brown et al. (1982) sequenced an 896-bp region of mtDNA from five different hominoids (human, chimpanzee, gorilla, orangutan, and gibbon) and showed that 284 sites had varied. There are 15 possible partitions of five sequences, and each of these was observed several times (table 4). There were, however, 106 sites at which three different nucleotides were found and one at which all four occurred. Most of these were resolvable as congruent to two (or three) species-specific partitions, i.e., those labeled 1–5. This is the same convention that I used for site 1,080 in the γ-globin sequence data. Nine of these sites, however, are less readily resolvable. These sites are congruent to one species-specific partition and one of two other partitions. Fortunately, the precise resolution of these nine sites did not seem to matter, so I only include results for one set of resolutions (see note to table 4). Where site positions are given, these are the same as those given by Brown et al. (1982), in which the lone deletion (site 560 of orangutan) is treated as if it were a nucleotide substitution.

The region sequenced by Brown et al. (1982) largely comprised open reading frames for two different proteins, with 199 contiguous base pairs coding for three tRNAs in the middle. These authors noted heterogeneity in the level of sequence variability among these different regions, so I first applied the test in equation (5) to determine whether the variable sites were clustered. The 284 variable sites are significantly clustered ($P[d \leqq d_o] = 0.025$), indicating a lack of variation in either or both of the regions flanking the variable sites. These regions contain highly conserved recognition sequences for *Hin*dIII, which played a role in Brown et al.'s decision to sequence this fragment of mitochondrial DNA. Application of equations (8), (9), and (10) indicated further heterogeneity in the overall variability of the sequenced region (table 5). By equation (10), the probability of observing a conserved segment 34 bp long or longer is 4.1×10^{-4}, yet one is observed in the sequence coding for leucine tRNA. The next two longest unvaried segments were also in a contiguous region coding for serine and leucine tRNAs, but they were not significantly long. Exclusion

Table 4
Partitions of Hominoid Mitochondrial DNA Nucleotide Sequence Variations and Tests of Clustering

PHYLOGENETIC PARTITION[a] LABEL (Split)	s[c]	d_o	g_o	SIGNIFICANCE[b] $P(d \leq d_o)$	P_m	P
1 (A/BCDE)	21	824	107	0.486	0.069	0.76
2 (B/ACDE)	26	776	98	0.116	0.037	0.61
3 (C/ABDE)	29	854	113	0.601	0.020	0.44
4 (D/ABCE)	63	867	62	0.381	0.009	0.43
5 (E/ABCD)	82	873	95	0.361	6.1×10^{-5}	0.0049[d]
a (AB/CDE)	10	794	172	0.682	0.140	
b (AC/BDE)	9	784	271	0.687	0.050	0.34[d]
c (AD/BCE)	2	139	138	0.286	. . .	. . .
d (AE/BCD)	4	360	245	0.181	0.100	0.27
e (BC/ADE)	12	803	162	0.641	0.103	0.70
f (BD/ACE)	2	287	286	0.538	. . .	. . .
g (BE/ACD)	4	273	160	0.087	0.169	0.43
h (CD/ABE)	7	788	319	0.798	0.074	0.37[d]
i (CE/ABD)	8	742	364	0.588	0.017	0.11[d]
j (DE/ABC)	32	861	101	0.642	0.022	0.50[d]
All sites	284	881	34	0.025	1.4×10^{-6}	4.1×10^{-6}[d]

SOURCE.—Brown et al. (1982).

[a] Species labels correspondence: A, human; B, chimpanzee; C, gorilla; D, orangutan; and E, gibbon.

[b] By eqq. (5), (9), and (10), respectively.

[c] Includes resolution of ambiguous sites. Sites 332, 771, and 879 were overlaps of partition 3 and either j or a; taking account of known transition bias, j was chosen since 3 could arise from j by a transition. Site 205 was an overlap of 4 and either b or g; b was chosen since 4 could arise from b by a transition. Site 265 was an overlap of 4 and either b or g; b was chosen since it is apparently more frequent. Sites 717 and 801 were overlaps of 5 and either b or f; b was chosen since it is apparently more frequent. Site 792 was an overlap of 5 and either e or c; e was chosen since it is apparently more frequent. Site 822 was an overlap of 4 and either e or d; e was chosen since it is apparently more frequent.

[d] Includes the highly conserved region identified by eq. (10).

of these three segments does not improve the fit to expectations from equation (8), since there is still an obvious excess of pairs of adjacent conserved sites (table 5). Seventy-three of these were observed, and 63 of them were accounted for as consecutive third-base substitutions in the open reading frames. The longest run of variable sites ($g'_o = 7$ in protein 5) was not significantly long ($P' = 0.1726$).

There are several interesting aspects of these sequence comparisons that are detectable by equations (5), (8), and (10), but I chose Brown et al.'s data as an example in which gene conversion and recombination were not factors, since these sequences are presumably free of such effects. In marked contrast to the human γ-globin partitions, none of the 15 sets of sites congruent to mtDNA partitions showed any tendency toward clustering by equation (5). Table 4 also shows the probabilities associated with the longest segments of consecutive r-sites for each partition. By equation (10), only partition 5 shows an exceptionally long segment. However, this segment covers the region coding for tRNAs. Exclusion of the significantly long unvaried segment reduces both g_o and r, such that g_o is no longer significantly long ($P = 0.13$).

Table 5
Test of Fit of Observed Lengths of Unvaried Site Segments to Those Expected: Hominoid Mitochondrial DNA Nucleotide Sequences

Length	Observed Frequency (O_1)	Expected Frequency[a] (E_1)	$(E_1 - O_1)^2 / E_1$	Length	Observed Frequency (O_2)	Expected Frequency[b] (E_2)	$(E_2 - O_2)^2 / E_2$
0	96	90.69	0.31	0	96	96.68	0.00
1	49	61.70	2.61	1	49	63.38	3.26
2	73	41.95	22.98	2	73	41.52	23.87
3	18	28.51	3.87	3	18	27.18	3.10
4	9	19.36	5.55	4	9	17.78	4.34
5	15	13.15	0.26	5	15	11.63	0.98
6	4	8.92	2.71	6	4	7.60	1.70
7	4	6.05	0.69	7	4	4.96	0.19
8–9	7	6.88	0.00	8–9	7	5.35	0.51
>9	8	5.80	0.83	>9	5	3.93	0.29
		$X^2 =$	39.83			$X^2 =$	38.24

SOURCE.—Brown et al. (1982).

NOTE.—Expected frequencies (under eq. [8]) were pooled to exceed five in each class (in most cases).

[a] Based on all variable sites, with $s = 284$ and $r = 598$. $X^2 = 39.83$, $P(\chi^2_{(9)} > 23.59) = 0.005$.

[b] Recalculated after excluding the three longest unvaried segments ($s = 281$ and $r = 529$). $X^2 = 38.24$, $P(\chi^2_{(9)} > 23.59) = 0.005$.

Discussion

Recent analyses of nucleotide sequence data (e.g., Chakravarti et al. 1984) show apparent heterogeneity in recombination rate along chromosomal DNA sequences. I have developed statistical tests that aid in deciding whether or not particular sets of nucleotide sites are statistically clustered along a chromosome and use this information to make qualitative inferences about recombination events, including gene conversion. For instance, application of these tests to the human γ-globin sequences shows quite strikingly that the hot spot and the highly conserved region on the 5′ side of it separate two segments of DNA with distinct evolutionary histories.

I have also used these tests to analyze DNA sequences from the *Drosophila melanogaster Adh* locus (Kreitman 1983) and gorilla and human γ-globins (Scott et al. 1984). Both of these studies indicated that recombinational events had occurred (intragenic recombination and gene conversion, respectively), and indeed my tests show many striking associations among the sets of congruent sites (J. C. Stephens, unpublished data). Specific inferences about recombination at the *Drosophila melanogaster Adh* locus have been incorporated into a phylogenetic analysis of these sequences (Stephens and Nei 1985).

Having detected strong spatial clustering among a set of nucleotides, one must exercise some caution in drawing conclusions regarding the underlying biological mechanism. Since some of these tests are sensitive to highly conserved regions of DNA, they should first be used to identify such regions prior to separate analyses of the phylogenetic partitions. I have used this strategy in my two examples. Note that highly conserved regions can bias the recombination analysis either way, as observed for partitions 1 and 3 in the analysis of the human γ-globin sequences.

These tests may also detect strong local biases of mutation, perhaps interacting with natural selection. For example, if several mutations were caused by the same event—such as removal of a transposable element— they would probably be physically associated. An improbably long run of variable sites, all congruent to the same partition, may indicate a common origin, such as compensating insertion/deletion in one of the sequences. Transient local increases in mutation rate could also lead to clustering of partition-specific mutations. Local differences in substitution rate are certainly known, as they are for the various functional regions in the genome. Constant differences in regional mutation rates, however, would not be expected to produce physical clustering of congruent sites beyond that of variable sites in general. Selection favoring particular combinations of specific mutations might also lead to some sites sharing the same partition, although it seems doubtful that there could be many such sites involved. Neither selection nor mutation would be expected to obscure the phylogenetic relationship of DNA sequences—and therefore affect the labeling of partitions—as profoundly as would recombination.

If recombinational events are relatively rare, moderately uniform mutation rates along the chromosome should not lead to statistical association of congruent sites such as was seen in the human γ-globin data. On the other hand, when recombination and gene conversion are relatively frequent, sites that mutated in the same lineage in the same time interval (fig. 2*A*.) will only be congruent if they are physically close together; otherwise they will become congruent to different partitions via recombination.

The only partitions of immediate importance for my purposes are those generated by two or more congruent sites, i.e., $s \geqq 2$. Therefore, I anticipate that these methods will work best when there are a relatively small number of unique sequences and a relatively large number of variable sites and when differentiation is no more than that between closely related species. If the number of recombination events sampled is too large, either because of high recombination rates or large sample size, each variable site may be congruent to a unique partition. In such cases it is best to treat subsamples of the sequences.

While it is not necessary to have a phylogenetic tree in order to detect statistical associations, explanations of particular associations will usually refer to an inferred tree. For instance, a cluster of inferred "parallel" mutations is much more reasonably explained by a recombination than by independent parallel mutations. Contrary to standard maximum parsimony treatments (Fitch 1977), I have recognized that even sequence-specific variations yield phylogenetic information in the present context. The necessity of considering such sites should be obvious. I treated a sample of three human γ-globin sequences above—and in some cases even subsamples of two sequences (Stephens and Nei 1985). Such applications would be virtually useless under parsimony procedures.

My emphasis on application of the statistical tests to the task of detection of gene conversion and recombination should not hide their potential utility for detecting heterogeneity in the distribution of other sites of interest. For instance, particular types of variations—e.g., insertions/deletions, transitions, or transversions—may have nonrandom distributions. Base or restriction-site distributions would be another potential application of such tests. It seems, however, that application of the tests to detection of recombination may be very powerful, for the following reasons. Of the 2^{m-1}-1 theoretically possible partitions of m different sequences, only $2m - 3$ can be primary.

For example, if the branching order of the hominoid mtDNA sequences above is gibbon, orangutan, gorilla, and chimpanzee-human, then only partitions 1–5, a, and j are primary, and the rest are secondary. This means that all of the other eight partitions correspond to parallel or backward mutations, or recombinational events. In the mtDNA data, partition-specific sites were not clustered, which indicates that the eight secondary partitions were all created by parallel and backward mutation. It is reasonable to infer that this is the case, since Brown et al. (1982) have established a strong bias toward transitional mutations in primate mtDNA sequence evolution.

In practice, one never knows which partitions are primary and which are secondary, especially when dealing with intraspecific DNA sequences. In fact, the distinction between primary and secondary partitions was not especially useful for the γ-globin case I considered. The following example should clarify the utility of this distinction. Consider five nuclear DNA sequences with partition labels 1–5 and a–j as in table 4 and suppose that 1–5, a, and j were primary. If part of sequence A is a product of intragenic recombination or gene conversion between sequences A and E, then there are several predictable consequences (J. C. Stephens, unpublished data). For now, consider the effect on partition-5 sites and partition-1 sites. Any mutations unique to sequence E's lineage prior to conversion of sequence A will now be shared by both sequences, which creates sites congruent to partition d (AE/BCD). Likewise, sites congruent to partitions 1 or 5 will be lost in the region that was converted. Similarly, sites congruent to the primary partition j create the secondary partition e in the region that was converted. The tests presented in this paper are designed to detect the presence or absence of blocks of associated nucleotide changes, such as those congruent to each partition. In studies of multicopy or intraspecific nuclear DNA sequences, there are often many secondary partitions, and these may in fact be the product of past recombinational events.

Acknowledgments

This work benefitted greatly from stimulating discussion with my colleagues and from their critical reading of earlier drafts of the manuscript. I would especially like to thank M. Nei, P. Pamilo, D. Hudson, C.-I. Wu, N. Saitou, and P. Smouse. This work was supported by research grants GM 20293 (National Institutes of Health) and BSR 83115 (National Science Foundation) to M. Nei.

APPENDIX

Mean of *d*

The mean of the distribution of d can be calculated easily. First note that from equation (2),

$$\sum_{d=1}^{n-1} df_d = \sum_{d=1}^{n-1} d(n-d)\binom{d-1}{l} = \sum_{d=1}^{n-1} (n-d)\left(\frac{d!}{(d-l-1)!l!}\right)$$

$$= \sum_{d=1}^{n-1} (n-d)(l+1)\binom{d}{l+1} = (l+1)\left[n\sum_{d=1}^{n-1}\binom{d}{l+1} - \sum_{d=1}^{n-1} d\binom{d}{l+1}\right]$$

$$= (l+1)\left[(n+1)\sum_{d=1}^{n-1}\binom{d}{l+1} - \sum_{d=1}^{n-1}\frac{(d+1)!}{(l+1)!(d-l-1)!}\right]$$

$$= (l+1)\left[(n+1)\sum_{d=1}^{n-1}\binom{d}{l+1} - (l+2)\sum_{d=1}^{n-1}\binom{d+1}{l+2}\right]$$

$$= (l+1)\left[(n+1)\binom{n}{l+2} - (l+2)\sum_{j=2}^{n}\binom{j}{l+2}\right]$$

$$= (l+1)\left[(n+1)\binom{n}{l+2} - (l+2)\binom{n+1}{l+3}\right]$$

$$= (l+1)\binom{n+1}{l+3}[(l+3) - (l+2)] = (l+1)\binom{n+1}{l+3}.$$

Hence the mean is simply

$$E(d) = \sum_{d=1}^{n-1} df_d \Big/ \binom{n}{s} = (l+1)\binom{n+1}{l+3}\Big/\binom{n}{l+2}$$

$$= (l+1)(n+1)/(l+3) = (s-1)(n+1)/(s+1).$$

LITERATURE CITED

ARNHEIM, N. 1983. Concerted evolution of multigene families. Pp. 38–61 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

ARNHEIM, N., M. KRYSTAL, R. SCHMICKEL, G. WILSON, O. RYDER, and E. ZIMMER. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA **77**:7323–7327.

BROWN, A. H. D., and M. T. CLEGG. 1982. Analysis of variation in related DNA sequences. Pp. 107–132 *in* B. S. WEIR, ed. Statistical analysis of DNA sequence data. Marcel Dekker, New York.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

CHAKRAVARTI, A., K. H. BUETOW, S. E. ANTONARAKIS, P. G. WABER, C. D. BOEHM, and H. H. KAZAZIAN. 1984. Non-uniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36**:1239–1258.

CHOVNICK, A., G. H. BALLANTYNE, and D. G. HOLM. 1971. Studies on gene conversion and its relationship to linked exchange in *Drosophila melanogaster*. Genetics **69**:179–209.

FELLER, W. 1968. An introduction to probability theory and its applications. Vol. **1**. Wiley, New York.

FITCH, W. M. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. **111**: 223–257.

HOOD, L., J. H. CAMPBELL, and S. C. R. ELGIN. 1975. The organization, expression, and evolution of antibody genes and other multigene families. Annu. Rev. Genet. **9**:305–353.

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

LAUER, J., C.-K. J. SHEN, and T. MANIATIS. 1980. The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell **20**:119–130.

LEIGH BROWN, A. J., and D. ISH-HOROWICZ. 1981. Evolution of the 87A and 87C heat-shock loci in *Drosophila*. Nature **290**:677–682.

MELLOR, A. L., E. H. WEISS, K. RAMACHANDRAN, and R. A. FLAVELL. 1983. A potential donor gene for the *bm1* gene conversion in the C57BL mouse. Nature **306**:792–795.

SCOTT, A. F., P. HEATH, S. TRUSKO, S. H. BOYER, W. PRASS, M. GOODMAN, J. CZELUSNIAK, L.-Y. E. CHANG, and J. L. SLIGHTOM. 1984. The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. Mol. Biol. Evol. **1**:371–389.

SHEN, S.-H., J. L. SLIGHTOM, and O. SMITHIES. 1981. A history of the human fetal globin gene duplication. Cell **26**:191–203.

SLIGHTOM, J. L., A. E. BLECHL, and O. SMITHIES. 1980. Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21**:627–638.

STEPHENS, J. C., and M. NEI. 1985. Phylogenetic analysis of polymorphic DNA sequences at the *Adh* locus in *Drosophila melanogaster* and its sibling species. J. Mol. Evol. (accepted).

ZIMMER, E. A., S. L. MARTIN, S. M. BEVERLEY, Y. W. KAN, and A. C. WILSON. 1980. Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Natl. Acad. Sci. USA 77:2158–2162.

WALTER M. FITCH, reviewing editor

Received April 12, 1985; revision received June 4, 1985.