Locating Regions of Differential Variability in DNA and Protein Sequences

Hua Tang* and R. C. Lewontin†

**Department of Statistics, Stanford University, Stanford, California 94305 and †Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138*

Manuscript received April 25, 1998
Accepted for publication May 17, 1999

ABSTRACT

In the comparison of DNA and protein sequences between species or between paralogues or among individuals within a species or population, there is often some indication that different regions of the sequence are divergent or polymorphic to different degrees, indicating differential constraint or diversifying selection operating in different regions of the sequence. The problem is to test statistically whether the observed regional differences in the density of variant sites represent real differences and then to estimate as accurately as possible the location of the differential regions. A method is given for testing and locating regions of differential variation. The method consists of calculating $G(x_k) = k/n - x_k/N$, where x_k is the position of the kth variant site along the sequence, n is the total number of variant sites, and N is the total sequence length. The estimated region is the longest stretch of adjacent sequence for which $G(x_k)$ is monotonically increasing (a hot spot) or decreasing (a cold spot). Critical values of this length for tests of significance are given, a sequential method is developed for locating multiple differential regions, and the power of the method against various alternatives is explored. The method locates the endpoints of hot spots and cold spots of variation with high accuracy.

A common question that arises in the comparison of related DNA or protein sequences is whether the differences between them are concentrated in some regions of the sequence and are relatively sparse in others. This problem arises commonly in three contexts: polymorphism of a gene within a species, divergence of a gene between two species, and divergence of paralogous sequences that arose originally through duplication. In all of these cases there are some clearly definable regions, which we expect, *a priori*, to be more or less variable or divergent than others, as, for example, introns compared to exons. The differences between such *a priori* regions can be detected by standard statistical tests of heterogeneity. A much more difficult statistical problem arises, however, when there are no such clearly defined *a priori* regions, but we are looking for evidence of heterogeneity of variation within, say, an intron or an exon. A solution to the problem of detecting such regions was given in an article by Goss and Lewontin (1996), in which two fairly powerful tests for heterogeneity were developed. This earlier study, however, only provided statistical tests to detect heterogeneity, but did not offer any method for locating those parts of the sequence that differ from other parts in their variation. In this article we develop an estimation procedure for locating the position along the sequence of regions of differential probability of substitution. The method we describe, using empirical cumulative distribution function (ECDF) statistics, has the property that it also provides another test of the null hypothesis that has about the same power as the Goss and Lewontin variance and extremal run length tests. The statistical question then, is whether the lengths and positions of these runs are what we might expect if the positions along the sequence have equal probabilities of substitution.

Because the estimation method arises in the context of a test of the heterogeneity, we begin our exposition by a discussion of the test, turning later to the estimation problem. The performance of the test (power) is evaluated under several alternative hypotheses. The second part of this article describes how the same algorithm is used to estimate the regions of differential variability. We assess the estimate through several measures and discuss its potential problems. The article then concludes with two numeric examples using real molecular data.

Corresponding author: R. C. Lewontin, Museum of Comparative Zoology, Cambridge, MA 02138. E-mail: lewontin@oeb.harvard.edu

A METHOD OF HYPOTHESIS TESTING

The structure of the data is fairly simple. Two or more sequences are aligned and a new resultant sequence is produced with a 0 at each position at which all the sequences are identical, and a 1 at any position where there is at least one variant among the sequences. Where only two sequences are compared, as, for example, between two species or between two paralogous sequences, the 1's mark the sites of divergence. Where multiple sequences are compared, typically in a polymorphism study, the 1's mark sites that are polymorphic without

reference to how many distinguishably different allelic forms are seen at the site. The result is a single sequence made up of runs of 0's separated by 1's. Beginning at one end of the sequence, we can describe the data as a series of "events" marked by the 1's, separated by runs of "no event," denoted by the 0's. It is the lengths and arrangements of these runs of 0's between "events" that provide the basis for statistical tests and for locating regions of high or low probability of an event. In most applications the degree of polymorphism or sequence divergence is small as compared to the total sequence length, so there will be many more 0's than 1's, and that data will appear as runs of 0's punctuated by single 1's. But we are not restricted to such cases. Where there are many divergent sites between sequences, there will often be uninterrupted stretches of multiple 1's, but two adjacent 1's are simply counted as a run of 0's of length 0, with no loss of generality. When the proportion of divergent or polymorphic sites is actually >50% we can simply reverse the definition of events. In this article we consider cases where the proportion of events is ≤45% (see the last two columns of Table 1).

It is important to note that the procedures we derive do not assume that there are only two alternatives at a site for a multiple sequence comparison. The methods are equally valid whether a site is marked as "variable" because a single sequence differs from all the others or because every sequence differs uniquely from every other at the site. What is lost by considering only two classes, variant and invariant, is the potential information contained in the frequency distribution and enumeration of all the alternative forms. Ultimately, *all other things being equal*, this represents a loss of statistical power to detect some kinds of heterogeneity. For example, the distribution of interevent distances might conform to the null hypothesis of no clumping, but all the events that appear in one region of the sequence might be the result of only a single divergent sequence at any variant site, while in another region every sequence might differ from every other one at every variant site. But to make statistical use of such information is a great deal more complex than it may appear. First, it is not clear what the null hypothesis is. The simplest would be that for all sites there is a common number of different equiprobable states, but this number cannot even be estimated from the data because of ascertainment bias, a bias that depends on the true number of alternatives, the number of sequences in the sample, and the level of polymorphism. Other *ad hoc* null hypotheses suggested *a posteriori* from the observed patterns of variation parameters suffer from the dangers of all *a posteriori* tests, while *a priori* tests contain various numbers of undetermined parameters. Second, the increase in power obtainable from more detailed classification could only be achieved by increases in sample size, because the change from an underlying binomial hypothesis to a multinomial one reduces the number of observations falling in any distinguishable observed class, thus spreading the deviations among more degrees of freedom. Thus, the two-state classification, variant and invariant, seems the only practical basis for a general search for heterogeneity.

The detection of heterogeneity is essentially a test of goodness-of-fit to a uniform null distribution. Several nonparametric methods have been developed, which may be grouped into two broad categories. One group includes the runs tests, which are based on the interval lengths between two events. For a comprehensive review and comparison of various methods in this group, readers are referred to Goss and Lewontin (1996). The other group, which includes the method discussed in this study, uses the ECDF statistics (Stephens 1986b).

The ECDF statistics: The ECDF statistics use the difference between the observed cumulative distribution of events, the ECDF, and the theoretical cumulative density function (CDF) under some null hypothesis (Stephens 1986a). In our context the meaning of the ECDF, $F_n(x)$, is as follows. We have a total of N positions along the sequence labeled sequentially from one end by x ($1 \le x \le N$). On this sequence there are n events (marked by 1's). Beginning at one end of the sequence and progressing to the end of the sequence, we record how many of the events have occurred up to and including position x. In a sample of n events (1's), $F_n(x)$ is a stepwise function, calculated from the observations

$$F_n(x) = \frac{\#\text{ of events occurring up to position } x}{n},$$

$$1 \le x \le N.$$

The CDF function, $F(x)$, is calculated from the null hypothesis. In our case, the null distribution is uniform and

$$F(x) = x/N, \quad 1 \le x \le N.$$

The quantity $n \cdot F(x)$ is then the *number* of events we expect up to and including site x in a uniform distribution. The ECDF statistic, $F_n(x) - F(x)$, the cumulative difference between the observed and expected proportion of events up to position x_k, increases over an interval that is shorter than the expected length of spacing and decreases over an interval that is longer than the expected length. Under the null hypothesis, $F_n(x) - F(x)$ has mean 0 for all x. An extremely large departure from 0 is a ground for rejecting the null hypothesis. This is the essence of the Kolmogorov-Smirnnov (K-S) test (Sokal and Rohlf 1995):

$$D = \text{maximum} \mid F_n(x) - F(x) \mid.$$

The K-S test has many desirable properties, such as being distribution free and coordinate free, and it is consistent against all alternatives (Feltz and Goldin 1992). But the test has a relatively low power for our problem (R. C. Lewontin, unpublished results). The method presented below is similar to the K-S test.

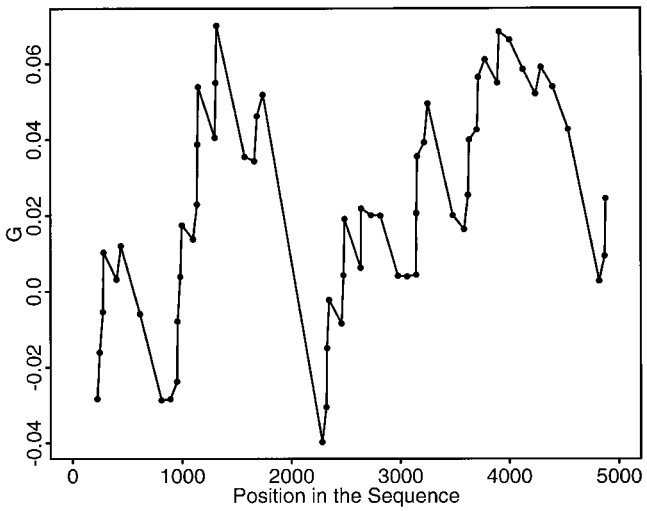

Figure 1.—An example of G under the null (uniform) distribution. $n = 60$.

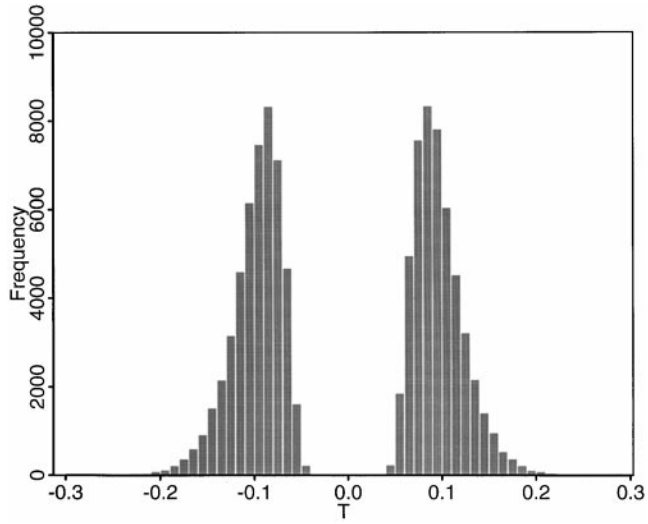

Figure 2.—Distribution of T under the null hypothesis. $n = 60$. Note the approximate symmetry with respect to 0.

The motivation for our test is as follows. If the true substitution rate at each site is the same, no region should contain unusually large or unusually small numbers of events. We expect the length of most spacings (length between two consecutive events) to be more or less close to the average, n/N, rather than being either extremely short or extremely long. Further, the longer spacings and the shorter ones should occur in a nonsystematic fashion; that is, we expect shorter spacings interspersed among longer ones, and vice versa. Hence $F_n(x) - F(x)$ moves up and down. On the other hand, if a region in the sequence has a very high substitution rate, we will observe more events occurring in that region, and thus a cluster of shorter spacings. Meanwhile, because n is fixed, there must be too few events elsewhere. When this happens ($F_n(x) - F(x)$) decreases sharply and consistently in a region of lower rate while increasing sharply in a region of higher rate.

Method: Denote the positions of the events by $X_1, X_2, \ldots X_n$, where $X_1 < X_2 < \ldots < X_n$, and calculate

$$G(x_k) = F_n(x_k) - F(x_k) \quad \text{for } k = 1, \ldots, n.$$

For example, suppose in a 5000-bp DNA sequence with 60 polymorphic sites, the 10th event falls on the 1000th site, then $X_{10} = 1000$, $F_n(x_{10}) = 10/60$, and $F(x_{10}) = 1000/5000G\ (1000) = k/n - x_{10}/N = 10/60 - 1000/5000 = -0.0333$.

In a region bounded by two events, i and j, let

$$\Delta G_{i,j} = G(x_j) - G(x_j), \quad 1 \le i \le j \le n.$$

The test statistic, T, is calculated by

$$T = \text{sign}(G) \cdot \max(|\Delta G|).$$

The maximum is taken over all intervals in which G increases or decreases monotonically. The sign of T is the same as that of ΔG. Figure 1 shows an example of the plot of G *vs.* the index of site under the null distribution.

We reject the hypothesis of uniformity if the test statistic T is greater than some critical value, T^*, a function of n, N, and α (probability of type I error). To find T^*, we did Monte Carlo simulation using programs written in C with the drand48 random number generator to produce 100,000 samples of n events by sampling sites without replacement along a sequence of 5000 sites. Figure 2 shows the results of the simulation for the null model. The distribution of T under the null hypothesis is symmetric with respect to 0. Therefore, we choose T^* to be the percentiles of $|T|$ among the 100,000 replicates, so that a two-tailed test has type I error of α. The critical values of T^* are given in Table 1 for various sequence lengths, N, numbers of events, n, and rejection levels, α.

In the above model, ΔG is calculated over intervals where G increases (or decreases) monotonically. Essentially, this method looks for the longest stretch in the sample in which every spacing is shorter (or longer, for a cold spot) than the expected length. But it is conceivable that one or more spacings may be slightly longer in a true hot spot; similarly, a spacing in a real cold spot may be slightly shorter than average. Such an atypical random spacing produces noise on the G curve (Figure 1), which we treat by smoothing. A great deal of literature on smoothing procedures is available (see Simonoff 1996). In this study, for the ease of computation, we adopt a very simple smoothing scheme. We say that G is almost monotonically increasing (or monotonically decreasing) in any interval in which any opposite change is <0.005. The relaxation of the definition of monotonicity amounts to a slight smoothing of the G curve. The value of 0.005 is chosen quite arbitrarily, and

TABLE 1

Critical values, T^*, for a rejection α of 5% and 1%, for different total sequence lengths, N, and for different numbers of events, n

		T^*									
	N:	10,000		5,000		1,000		500		200	
n	α:	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
10		0.401	0.480	0.399	0.476	0.398	0.475	0.396	0.472	0.390	0.465
20		0.289	0.349	0.289	0.349	0.286	0.346	0.282	0.340	0.275	0.330
30		0.230	0.279	0.229	0.279	0.226	0.276	0.223	0.273	0.212	0.257
40		0.191	0.234	0.193	0.234	0.189	0.230	0.185	0.226	0.170	0.210
50		0.168	0.205	0.167	0.205	0.164	0.201	0.160	0.194	0.145	0.175
60		0.149	0.182	0.149	0.182	0.145	0.178	0.141	0.171	0.128	0.157
70		0.136	0.166	0.135	0.164	0.132	0.160	0.127	0.155	0.112	0.136
80		0.124	0.151	0.124	0.150	0.121	0.147	0.116	0.141	0.100	0.123
90		0.117	0.142	0.116	0.141	0.113	0.137	0.109	0.132	0.093	0.113
100		0.109	0.133	0.108	0.132	0.104	0.126	0.100	0.122	–	–
125		0.096	0.117	0.096	0.116	0.090	0.110	0.086	0.106	–	–
150		0.088	0.107	0.088	0.107	0.085	0.103	0.081	0.099	–	–
175		0.085	0.103	0.085	0.103	0.082	0.100	0.074	0.091	–	–
200		0.083	0.102	0.082	0.101	0.076	0.093	0.066	0.081	–	–

Each value is from 100,000 replicates.

it is discussed in more detail in a later section. But as long as we use a consistent method when simulating under the null distribution, the probability of type I error will not be increased.

Alternative hypotheses: Under the null hypothesis, the probability of substitution at each site is the same. The alternative distribution can be very complex, as its parameter space is multidimensional (Goss and Lewontin 1996). Because the primary goal is to detect regions of heterogeneous substitution rate, we only consider alternative hypotheses in which each true distribution is composed of a few regions. We begin with two simple alternative distributions in which the true probability density function partitions the entire sequence into three regions. The central region is considered as the differential region. Three parameters that completely characterize an alternative hypothesis are the width (measured as a fraction of the length of the entire sequence), depth, and location of this differential region.

The differential region in alternative hypothesis A is a hot spot, and that in alternative hypothesis B is a cold spot. It should be emphasized, however, that these terms are only relative, not absolute. A cold spot only has lower probability of substitution relative to regions flanking it. Thus, distribution B can be equivalently considered as composed of two hot spots.

We first examine the performance of the test when the width, depth, and the position of the differential region vary. We then briefly discuss the change in the power when the alternative distribution contains more regions. To evaluate the power of the test we simulate samples with n events under different alternative distributions, recording the proportion of replicates in which the test statistic T, simulated under this alternative hypothesis, is more extreme than the critical value T^*. In most cases, we perform a two-tailed test and reject the null hypothesis when $|T| > T^*$. Of course, if we have prior knowledge of the direction of the deviation, the test can be one-tailed, with the critical values derived in a similar fashion.

Power of the test: Figure 3, a and b, shows the power of the test when the underlying distribution contains one hot spot or one cold spot of varying width. The ratio, r, of probability of a substitution between the differential and the constant regions is 5:1 in the case of a hot spot and 1:5 for a cold spot. The x-axis represents the width, w, of the differential region, and the y-axis represents the power (among 10,000 trials) when the rejection level $\alpha = 0.05$. The critical values of T^* are given in Table 1 for various numbers of events, n. For a hot spot and with $n = 30$, the test has reasonably good power (>0.5) when the differential region spans between 10–50% of the entire sequence. In the case of a cold spot, the differential region needs to span 30–80% of the entire sequence to have good power.

The power will, in general, be a function of the number of distinct sites of variation (n), the ratio of substitution rate between the differential and constant region (r), the width of the differential region (w), and the position of the differential region.

Figure 3, a and b, shows that for a given alternative hypothesis the power of the test increases as n, the number of distinct sites of variation, increases. This is expected because an increased sample size provides ad-

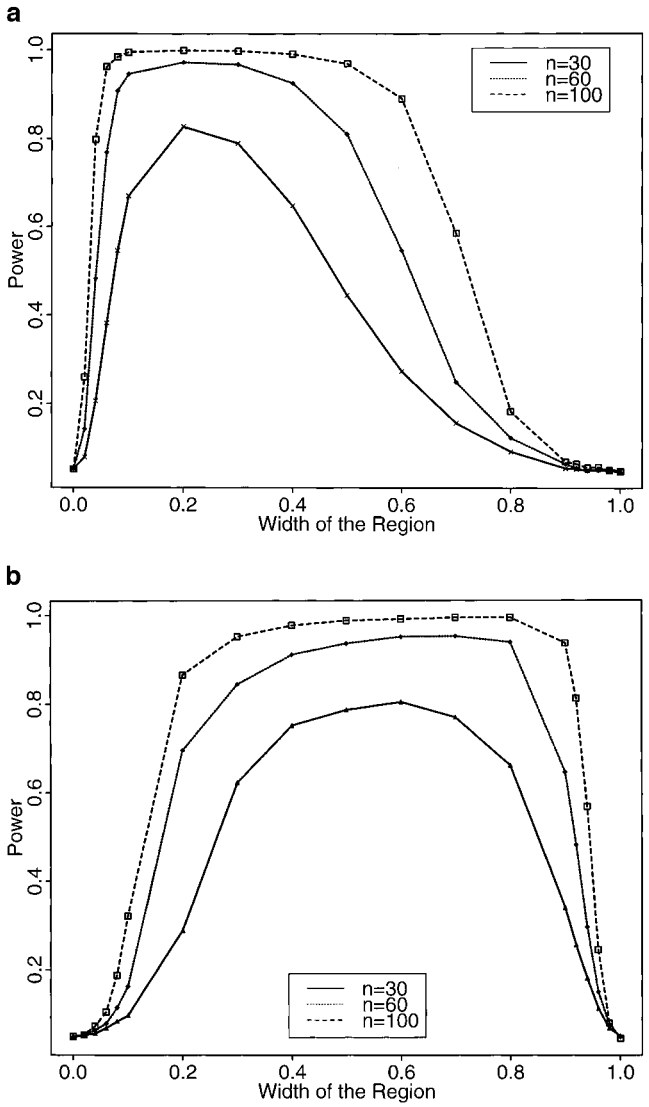

Figure 3.—In a and b, the *x*-axis shows the width, *w*, of a spot (hot or cold) located in the center of the sequence. (a) Power to detect a hot spot, $r = 5{:}1$; (b) power to detect a cold spot, $r = 1{:}5$.

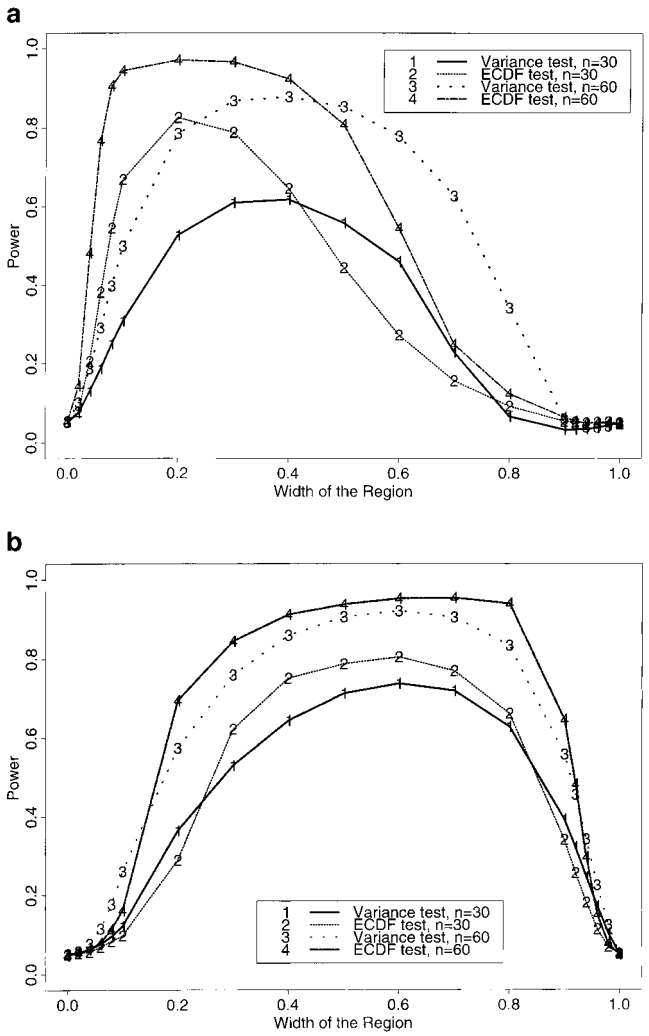

Figure 4.—Comparison of the power of the Goss and Lewontin variance test and the ECDF test. The *x*-axis shows the width, *w*, of the hot or cold spot located at the center of the sequence. (a) Hot spot, $r = 5{:}1$; (b) cold spot, $r = 1{:}5$.

ditional information. Holding *r* at the 5:1 level and *w* at 10% of the entire sequence, the power is only 0.27 for $n = 10$, but increases to 0.67 for $n = 30$, and reaches 0.99 when $n = 100$.

By the depth of the differential region we mean the ratio, *r*, of substitution rate between the differential and background regions. An increase in the deviation in substitution rate in the differential region raises the power. For 60 events and holding *w* at 10% of the total length, the power is 0.17 when the probability in the differential region is twice as high as that of the constant region, but it increases to 0.94 when *r* is 5:1, *ceteris paribus*.

Figure 3, a and b, shows that for both alternative hypotheses A and B, the power of the test peaks when the differential region is of moderate width. Note that when the differential region is narrow, a hot spot is easier to detect than a cold spot. But when the differential region becomes wide, a cold spot is easier to detect. This is not surprising. If several events take place in a relatively narrow hot spot, the spacings will be significantly shorter. This will produce a highly positive ΔG in that region, and hence high power. But if the narrow differential region is a cold spot, the best that can happen is that no event takes place in this region. Thus, the cold region is completely covered by one spacing, and all the events occur in the constant region with equal probability. If this one spacing covering the cold spot is not long enough to produce a significant ΔG, the test is most likely to be insignificant.

In comparison with Goss and Lewontin's variance test, Figure 4 indicates that the two methods have comparable power, and neither dominates the other over the range of the width of the differential region.

In the above simulations, we always centered the differential region to make the two flanking regions equal in length. The power is not affected significantly if this differential region is slightly off center. But when it moves to the extremities of the sequence, the power is affected due to an edge effect, which arises because the first and the last spacings in an observed sequence are necessarily shorter than the expected length, n/N. Consider a situation in which one of the flanking regions in alternative hypothesis B is degenerate. The cold spot is located at one end of the sequence, but the first spacing at this end is still likely to be shorter than expected due to the edge effect. The cold spot is then less likely to be detected, and the power of the test decreases. The same reasoning argues that the power should increase if the head/tail region is in fact a hot spot. A remedy for the edge effect is to circularize the two ends of the sequence and then cut it at the first event. In other words, we can cut off the first spacing and append it at the end. The new sites of the events are $X'_k = X_{k+1} - X_1$, for $k = 1, \ldots, n-1$. Note that we have one less spacing because the first and the last spacings in the original sequence have merged into one. The power of the test against all alternative hypotheses then becomes independent of the location of the alternative region (data not shown). Finally, even without circularizing the sequence, the performance of the test is not affected much as soon as the flanking regions on both sides reach 5% of the total length of the sequence.

Changing the number of differential regions: Intuitively, we expect that one hot spot is easier to detect than two hot spots, each half in size. It is also generally true that one hot spot is easier to detect than two isolated hot spots of the same size, unless the two hot spots are extremely close to each other. In that case they behave like one hot spot and, as seen in Figure 4, the power is not proportional to the width of the differential region. We show these two properties by two experiments.

First, we examine the change in power when one differential region is split into two or more smaller regions. We start with alternative hypothesis A, with one hot spot of width 0.3. We then divide this region into two or more smaller regions, keeping the total width of all these regions at 0.3. Between each pair of two consecutive smaller hot spots is a segment of constant region of width 0.1. Each of these alternative distributions is arranged symmetrically with respect to the center of the sequence. Figure 5 shows the change in the power to reject the null hypothesis when the original region is split into 2–6 regions. Note that the power decreases less sharply for larger sample size. Nonetheless, it decreases in each case. For this reason, our method of hypothesis testing should not be used to detect very fine-grained rate variation. It will not, for example, detect a higher substitution rate of the third position in each codon.

In the second experiment, we compare the change

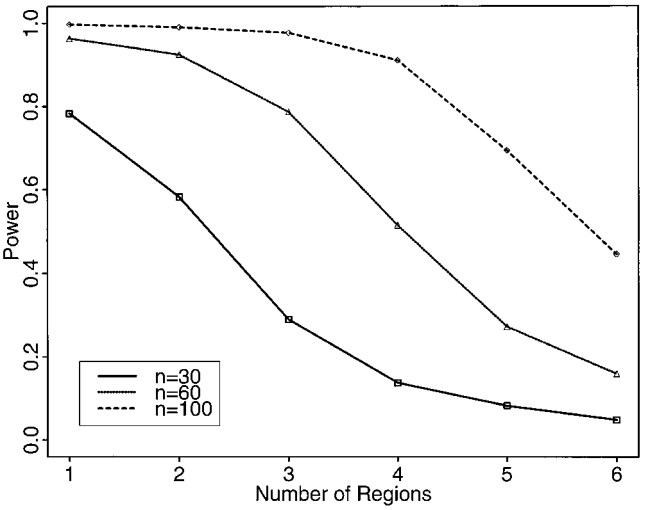

Figure 5.—Change in power when a hot spot, $r = 5{:}1$, is divided into two or more smaller regions. The x-axis shows the number of smaller regions. The summed width of the regions is 0.3 in all cases.

in power when there is more than one differential region of the same size. We begin with alternative hypothesis A, with a hot spot of width 0.1. We then add the second, third, fourth, and fifth hot spots, all with length 0.1. Again, each pair of neighboring hot spots is separated by a constant region of width 0.1, and each of these alternative hypotheses is arranged symmetrically with respect to the center of the sequence. The results are almost identical to the picture given in Figure 5. As expected, a decrease in power is observed for all sample sizes.

Although we have only presented results when the alternative distribution contains multiple hot spots, the same argument applies to the case of multiple cold spots. Denote the power of rejecting an alternative hypothesis with m similar differential regions each of length j as $P(m,j)$. Using the results in Figure 5, we conclude that in most cases,

$$P(m, j) \leq \min(P(1, j), P(1, mj)).$$

This gives an upper bound of the power when the alternative distribution contains more than one region. The presence of a mixture of hot spots and cold spots in effect increases the ratio of substitution rate (r) and results in an increase in power compared to the cases with only hot spots or only cold spots (data not shown).

ESTIMATION OF THE DIFFERENTIAL REGION

We now turn to the problem of estimating the differential region. We restrict ourselves to the simple cases where the underlying distributions are in the shapes of alternative hypotheses A or B. Our goal is to estimate the location of the central region. As the distribution

becomes more complex, it is not at all clear what we identify as the "differential" region.

Sequential estimation method: Using the hypothesis testing method with smoothing presented in the last section, the estimation procedure and the testing procedure go hand in hand.

1. Test for heterogeneity in the sequence. If the test is not significant we conclude that substitution rate is uniform and there is no region to be estimated.
2. If $|T| > T^*(\alpha, n)$ we estimate the differential region as *the interval where G changes almost monotonically and over which the absolute magnitude of* ΔG *is maximized.*
3. Remove the estimated region from the sequence, and repeat steps 1 and 2. Continue this iterated process until no significant heterogeneity is found.

In such a recursive procedure the most deviant observations are being successively removed while the sample size is being reduced progressively. Thus, the probability of a type I error will be greater than the nominal value of α in tests after the first one, while n and N are being successively reduced. Such a recursive procedure will then be both conservative and of lower power for each successive test.

Unlike point estimates, there is no established measure of accuracy for an interval estimate. Intuitively, we want our estimate to coincide with the true central region as closely as possible. To this end we look at the distributions of two proportions: the proportion (P) of the estimated region that falls in the true differential region, and the proportion (Q) of the true differential region that is covered by the estimated region. Of these two proportions, the former is more important than the latter. If we simply estimated the region by including the entire sequence, we would be bound to cover the true differential region completely every time (if one is present). But such an estimate is meaningless. At the same time, we hope that our estimated region includes as much of the true differential region as possible. An estimated region of 1 bp long is not very informative. Alternatively, we can treat the endpoints of the estimated interval as two point estimates and examine their marginal and joint distributions.

Once the boundary of the central region is identified, we can further estimate the relative rate of substitution by comparing the ratio of the number of substitutions in each region to the width of that region.

Results: Figure 6, a and b, shows single simulated examples of plots of G when the sequence contains a hot spot or a cold spot, respectively. As seen in Figure 3b from the previous section, the power of detecting a cold spot is low when the true region is very narrow. Therefore, we use examples of cold spots of width 0.2 in this section.

Figure 7 shows the joint distribution of P and Q under

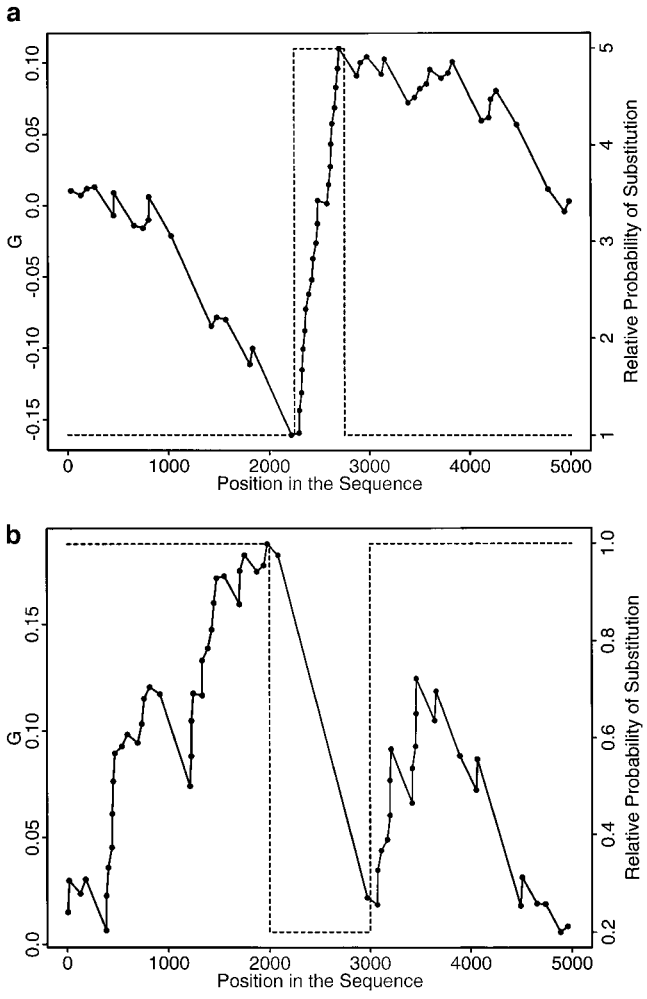

Figure 6.—Plots of G under alternative hypotheses, n = 60. (a) Alternative hypothesis A, shown in dotted lines. The estimated differential region is [2220–2669] and $\Delta G = 0.2694$. (b) Alternative hypothesis B, shown in dotted lines. The estimated differential region is [1946–3073] and $\Delta G = 0.1688$.

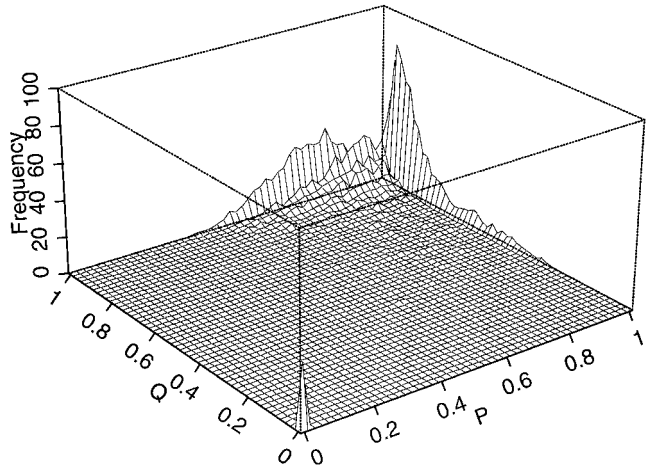

Figure 7.—Joint distribution of P and Q under alternative hypothesis A, with n = 60, r = 5:1, and w = 0.1.

TABLE 2

Accuracy of the estimated region under the alternative hypothesis A ($w = 0.1$), a hot spot, and B ($w = 0.2$), a cold spot

	Hypothesis A		Hypothesis B	
n	Mean (P)	Mean (Q)	Mean (P)	Mean (Q)
30	0.79	0.89	0.54	0.99
60	0.85	0.91	0.78	0.98
100	0.87	0.93	0.84	0.94

alternative hypothesis A and $n = 60$. The mean length of the estimated region is 554 sites long, while the true central region spans 500 sites. The distribution shows that in most cases the proportion, P, of the estimated region that overlaps the true differential region is >60%. Figure 7 also shows the distribution of Q, the proportion of the truly differential region that is included within the estimated region under the alternative hypothesis. In almost all cases, 70% of the differential region is covered. Of course, the estimates do not coincide exactly with the true differential region. The plot of joint distribution of P and Q in Figure 7 shows that the mode occurs where 99% of the estimated region overlaps with the true region and 94–96% of the true region is identified. The mean proportions of P and Q under alternative hypotheses A and B with different numbers of events are tabulated in Table 2. Counterintuitively, the mean of Q for a cold spot decreases as n increases. Such an apparent trend, however, should not be interpreted to mean that a larger sample size somehow lowers the accuracy of the estimation. The length of the true differential region is constant for all three cases ($w = 0.2$). When n is small, the estimated regions are much larger (low average of P) than the true one, so they cover most of the true region; as n increases, the estimated regions are narrower, and therefore are less likely to cover the entire true differential region. Although Q decreases slightly as we accumulate more events, data with larger n still give a more informative estimate.

As an extension, we may look at the quantity $R = PQ$:

$$E\{R\} = \mathrm{Cov}(P, Q) + E\{P\} * E\{Q\}.$$

The advantage of R is that it exacts a penalty if the estimated region is either too small or too large. When the estimated region coincides exactly with the true differential region, R is 1.0. R becomes small if either P or Q is small. Figure 7 shows that most of the time R is quite high.

Alternatively, we can look at the distribution of each endpoint. The distributions of both endpoints are shown in Figure 8, a and b. The sharpness of the peaks

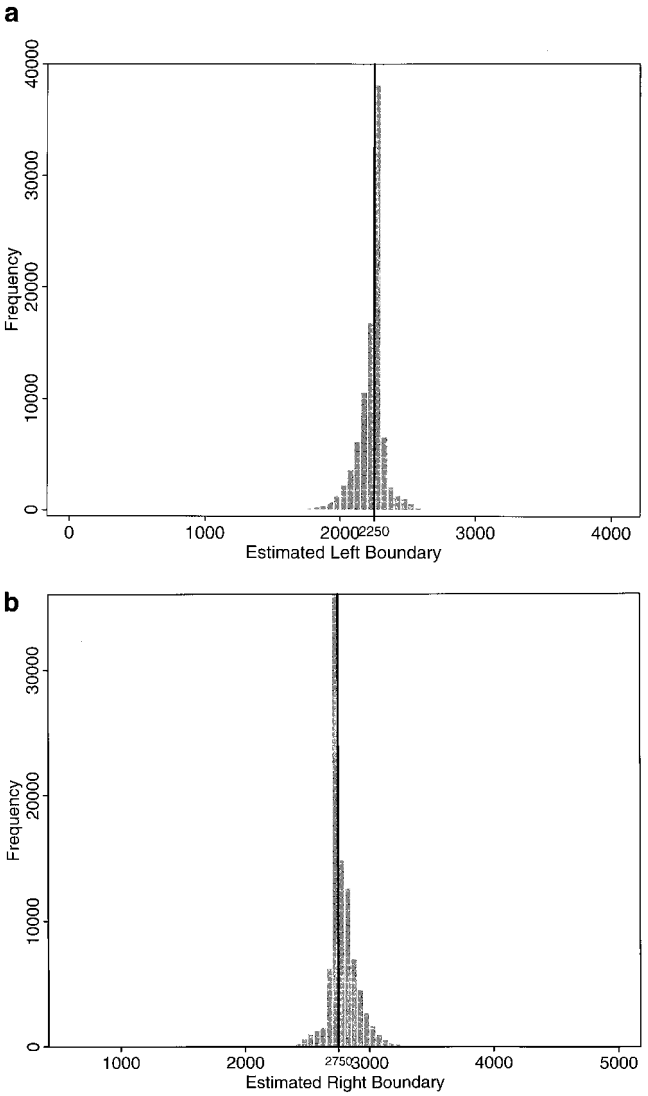

Figure 8.—(a) Distribution of left endpoint of the estimated region. Estimated mean = 2219; true value = 2250. (b) Distribution of right endpoint of the estimated region. Estimated mean = 2788; true value = 2750.

in these distributions is another indication of the accuracy of the estimates.

Table 3 lists the means of the estimated endpoints. In each case, the true left and right endpoints are 2250 and 2750, respectively. It seems that the estimated left

TABLE 3

The means of the estimated endpoints with different sample sizes

n	Mean (left)	Mean (right)
30	2169.341	2835.692
60	2219.216	2788.282
100	2232.847	2779.878
True endpoints	2250	2750

TABLE 4

Power of the test and the accuracy of the estimates with varying smoothing parameters

			Left estimate		Right estimate	
Δ	Critical value	Power	Median	SD	Median	SD
0	0.22213	0.6697	2256	351.9	2742	337.6
0.001	0.22440	0.6739	2255	343.4	2743	340.7
0.005	0.22767	0.6773	2254	320.4	2748	331.8
0.01	0.23987	0.6602	2246	327.5	2770	326.7
∞	0.27213	0.6078	2145	465.0	2853	468.7

The true distribution contains a hot spot between positions 2250 and 2750. $n = 30$, $w = 0.1$, $r = 5{:}1$, $N = 5000$.

endpoint tends to be too small, and that of the right endpoint too large. But the bias decreases as the sample size increases. One potential source for the bias is that each estimate is always a site where a substitution actually occurs. For an estimate to fall exactly on the true endpoint, a necessary condition is that an event occurs there. The probability of such an occurrence is low, especially when n is small. If there is no event taking place at the true change point, the spacing that covers the true change point is likely to be shorter (if the differential region is a hot spot) than the expected length of spacing. So the left estimate is often the site of substitution right before the differential region starts. By the same token, the right estimates are often the site of substitution immediately after the differential region ends. If this is true we may correct the bias by taking some inner portion of the current estimate.

The smoothing parameter: As noted earlier, we have somewhat arbitrarily neglected an excursion of $\Delta<0.005$ as "noise." For a sequence of 5000 bp long, this allows a spacing in a hot spot to be as much as 25 bp longer than average even if it should be shorter than the average length under the null hypothesis. Is this relaxation enough, or is it too much? This depends on the resolution of the study. It should be clear that as long as the same smoothing scheme is used to simulate the critical values under the null hypothesis, the type I error will not be affected. Table 4 compares the 0.05-level critical value, the power of the test, the median, and the standard deviation of the left and right estimates of the differential region to varying Δ. The alternative region is a centered hot spot with ratio 5:1. It shows that the power fluctuates as Δ varies, while the estimated regions grow wider as Δ increases. In the limiting cases, when $\Delta = 0$ there is no smoothing, and when $\Delta \to \infty$ the statistic is the same as the V statistic, suggested by Kuiper (1960), which uses the difference between the maximum and minimum of G over the entire sequence. For this particular alternative hypothesis, a smoothing parameter between 0.005 and 0.01 optimizes the estimates in the sense that the medians are closest to the true endpoints and the spread of the estimates, measured by the standard deviations, are smallest. Nonetheless, Table 4 indicates that the power and the estimates are reasonably stable for a wide range of Δ; hence the method we present here is quite robust even if the "optimal" smoothing parameter is not used.

HOT SPOTS *vs.* COLD SPOTS

One shortcoming in many previously suggested tests is that one cannot distinguish a hot spot from a cold spot. For example, the shortest and longest interval tests require prior knowledge of the probability of the differential region (Goss and Lewontin 1996). After all, there is a shortest interval and a longest interval in every sequence, and testing with both methods runs into the problem of multiple comparison and an increased probability of type I error. The variance test provides a unified test for both cold and hot spots, but the test statistic is always positive, and it does not provide insight on the nature of the differential region (Goss and Lewontin 1996). The method presented here requires no prior assumption of either cold or hot spot. The test is two-tailed when we have no knowledge about the alternative distribution. An extremely low (negative) value of T suggests the existence of a cold spot, while a highly positive T value indicates the presence of a hot spot.

A careful look at Figure 7 reveals that there are cases in which the estimated region completely misses the true differential region (P, Q, and R are all 0). In most of these cases the T statistic is highly negative. What happens in these cases is that, instead of identifying the central region as a hot spot, we have estimated the location of one of the flanking regions as a cold spot. This reveals a weakness of our definition of differential region. In this study we have arbitrarily defined the central region as the differential region. As discussed above, because of the lack of a baseline probability of substitution, alternative hypothesis A can be considered as a hot spot or, equivalently, as two cold spots. When the central region becomes very wide, the test is more likely to indicate one of the cold spots rather than the

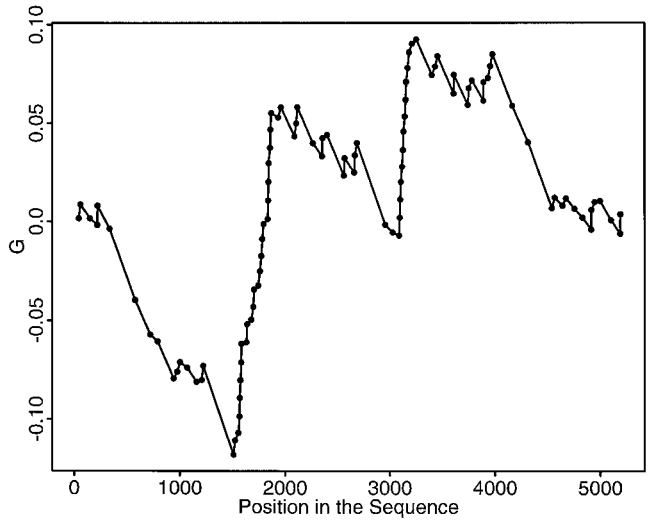

Figure 9.—Plot of G for Hin region of the *dpp* gene in *Drosophila melanogaster*. N = 5208 bp, n = 99. Positions of variant sites: 44, 60, 149, 219, 221, 334, 574, 718, 789, 940, 975, 1002, 1069, 1159, 1207, 1222, 1509, 1524, 1557, 1566, 1570, 1576, 1581, 1584, 1633, 1638, 1678, 1697, 1704, 1746, 1761, 1774, 1782, 1795, 1834, 1838, 1841, 1844, 1856, 1860, 1870, 1934, 1960, 2089, 2107, 2117, 2265, 2352, 2357, 2401, 2561, 2567, 2658, 2665, 2685, 2954, 3027, 3088, 3093, 3098, 3104, 3116, 3125, 3128, 3141, 3150, 3155, 3171, 3182, 3212, 3253, 3400, 3430, 3455, 3607, 3610, 3742, 3751, 3782, 3889, 3892, 3934, 3955, 3976, 4165, 4314, 4541, 4566, 4640, 4673, 4753, 4829, 4913, 4914, 4946, 4995, 5099, 5187, 5188. T^* = 0.10966. Null hypothesis rejected at the 0.05 level. Estimated differential region is a hot spot between base pairs 1509 and 1960.

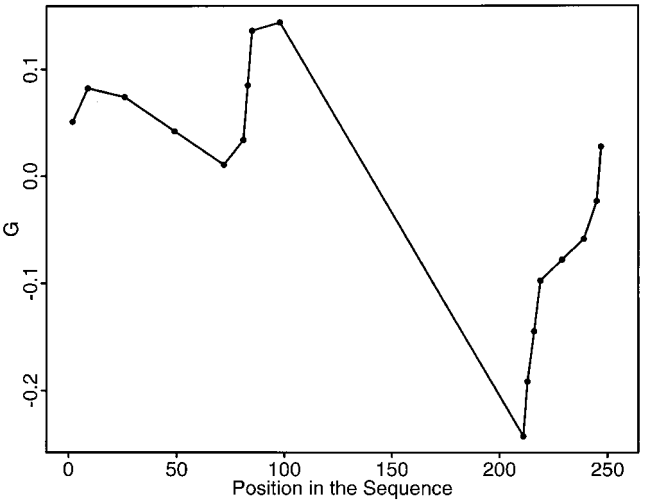

Figure 10.—Plot of G for the Adh protein in *D. melanogaster*. N = 254 amino acids; n = 17. T^* = 0.30477. Divergent sites at 2, 9, 26, 49, 72, 81, 83, 85, 98, 211, 213, 216, 219, 229, 239, 245, 247. The null hypothesis is rejected at the 0.05 level. Estimated differential region is a cold spot between amino acids 98 and 211.

hot spot. If we have prior knowledge that the differential region is a hot spot, we can perform a one-tailed test and only look for intervals where ΔG is positive. If we could detect both cold spots and leave only the central region, the estimate would be just as useful.

NUMERICAL EXAMPLES

As a first example, consider the Hin region of the *dpp* gene on chromosome 2 of the *D. melanogaster* genome. Each sequence consists of 5208 bp. Within 18 genomes sequenced, 99 loci are found to be polymorphic (Richter *et al.* 1997). Positions of polymorphism in the sequence are given in the legend of Figure 9, which shows the plot of G at each position of polymorphism. The test statistic T is 0.1688. The null hypothesis is rejected with a P value of 0.00054. Further, Figure 9 shows that the differential region contains a hot spot between sites 1509 and 1960. This corresponds to the apparent cluster at the 5′ end of intron 2. If we cut off this segment and anneal the remaining two pieces, the resulting sequence is 4757 bp long, with 73 sites of polymorphism. Applying the test to this second sequence, the null hypothesis is again rejected at the 0.05 level (T = 0.1442 and P = 0.01338). This time, another hot spot is identified between base pairs 3027 to 3253. This corresponds to the cluster at the 3′ end of intron 2. If we delete this hot spot, anneal the remaining two pieces, and test for heterogeneity in substitution again, the test is not significant. Finally, if we anneal all of the exons into one sequence and examine the distribution of substitutions, there is no evidence of heterogeneity. This result is consistent with the Goss and Lewontin variance test (Richter *et al.* 1997).

The second example concerns the Drosophila *Adh* protein, which consists of 245 amino acids (sequences in GenBank; URL: http://www.ncbi.nlm.nih.gov/). The data are analyzed as a numerical example by Goss and Lewontin (1996). Figure 10 lists the 17 sites of fixed divergence among eight species of the *melanogaster* subgroup. The test statistic T is -0.386, giving a P value of 0.0058. This result enables us to reject the null hypothesis, and is consistent with that in Goss and Lewontin (1996). Figure 10 indicates a cold spot between the amino acids 98 and 211.

We gratefully acknowledge Herman Chernoff, Peter Goss, and Steven Scott for their helpful discussions and references. A wholehearted thank you goes out to Susan Holmes for her 2 years of moral support of the senior author. Finally, we express our gratitude to Steven Scott, Daniel Larson, Richard Lung, and all those not already mentioned who have read the manuscript.

LITERATURE CITED

Feltz, C. J., and G. A. Goldin, 1992 Generalization of the Kolmogorov-Smirnov goodness-of-fit test, using group invariance. Nonparametric Stat. **1:** 357–370.

Goss, P. J. E., and R. C. Lewontin, 1996 Detecting heterogeneity of substitution along DNA and protein sequences. Genetics **143:** 589–602.

Hinkley, D. V., 1971 Inference about the change-point from cumulative sum tests. Biometrika **58:** 509–523.

Kuiper, N. H., 1960 Tests concerning random points on a circle. Ned. Acad. Wetensch. Proc. Ser. A **63:** 38–47.

Richter, B., M. Long, R. C. Lewontin and E. Nitasaka, 1997 Nucleotide variation and conservation at the dpp locus, a gene controlling early development in Drosophila. Genetics **145:** 211–323.

Simonoff, J. S., 1996 *Smoothing Methods in Statistics.* Springer, New York.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*, Ed. 3. W. H. Freeman and Company, New York.

Stephens, M. A., 1986a Tests based on ECDF statistics, pp. 97–193 in *Goodness of Fit Techniques*, edited by R. B. D'Agostino and M. A. Stephens. Marcel Dekker, London.

Stephens, M. A., 1986b Tests for the uniform distribution, pp. 331–366 in *Goodness of Fit Techniques*, edited by R. B. D'Agostino and M. A. Stephens. Marcel Dekker, London.

Communicating editor: A. G. Clark