Copyright © 2001 by the Genetics Society of America

Disparity Index: A Simple Statistic to Measure and Test the Homogeneity of Substitution Patterns Between Molecular Sequences

Sudhir Kumar and Sudhindra R. Gadagkar

Department of Biology, Arizona State University, Tempe, Arizona 85287-1501

Manuscript received February 8, 2001
Accepted for publication April 6, 2001

ABSTRACT

A common assumption in comparative sequence analysis is that the sequences have evolved with the same pattern of nucleotide substitution (homogeneity of the evolutionary process). Violation of this assumption is known to adversely impact the accuracy of phylogenetic inference and tests of evolutionary hypotheses. Here we propose a disparity index, I_D, which measures the observed difference in evolutionary patterns for a pair of sequences. On the basis of this index, we have developed a Monte Carlo procedure to test the homogeneity of the observed patterns. This test does not require *a priori* knowledge of the pattern of substitutions, extent of rate heterogeneity among sites, or the evolutionary relationship among sequences. Computer simulations show that the I_D-test is more powerful than the commonly used χ^2-test under a variety of biologically realistic models of sequence evolution. An application of this test in an analysis of 3789 pairs of orthologous human and mouse protein-coding genes reveals that the observed evolutionary patterns in neutral sites are not homogeneous in 41% of the genes, apparently due to shifts in G + C content. Thus, the proposed test can be used as a diagnostic tool to identify genes and lineages that have evolved with substantially different evolutionary processes as reflected in the observed patterns of change. Identification of such genes and lineages is an important early step in comparative genomics and molecular phylogenetic studies to discover evolutionary processes that have shaped organismal genomes.

MOLECULAR sequences are routinely used to reconstruct phylogenetic histories of species and multigene families and to detect nonneutral evolution at the molecular level. Most of these methods assume that the sequences analyzed have evolved with the same process of nucleotide substitution in their evolutionary history (homogeneity assumption). If this assumption is not satisfied, the inferred phylogenetic trees may have erroneous branching patterns and tests of neutral evolutionary hypotheses may become unreliable (Hasegawa *et al.* 1993; Steel *et al.* 1993; Funk *et al.* 1995; Galtier and Gouy 1998; Naylor and Brown 1998; Rodriguez-Trelles *et al.* 2000; Tarrio *et al.* 2000). Therefore, it is important to test this assumption for a given set of sequences *prior to* molecular evolutionary analysis. Knowledge of the violation of the homogeneity assumption would allow the investigators to choose advanced methods of phylogenetic reconstruction (*e.g.*, Lockhart *et al.* 1994; Galtier and Gouy 1998) or to conduct phylogenetic analyses with the offending sequences removed, if possible. Identification of genes and species with atypical patterns of change is also useful for elucidating the evolutionary mechanisms responsible for the observed differences.

In general, sequences that have evolved with the same substitution process (that is, where the relative probability of change from one state to another is the same in the lineages being compared) are expected to have similar nucleotide (and amino acid) compositions. Therefore, differences in the substitution process among lineages can be detected by comparing the observed patterns of nucleotide frequencies in the extant sequences. In the following, we propose a simple measure, disparity index (I_D), to quantify the difference in observed patterns and use it to develop a statistical test. We examine the performance of this test under biologically realistic conditions and compare it to other tests by means of computer simulation as well as empirical data analysis.

DISPARITY INDEX TO MEASURE SUBSTITUTION PATTERN DIVERGENCE

Let X and Y be two DNA sequences of length L each. Let x_i be the count of the ith type of nucleotide (i = A, T, C, or G) in sequence X and let y_i be the corresponding count in sequence Y. The composition distance between these two sequences can then be defined as

$$D_C = \frac{1}{2}\sum_i (x_i - y_i)^2, \quad \text{where } i = \text{A, T, C, or G.} \qquad (1)$$

The expected value of D_C can be obtained in the following way. Let us represent sequences X and Y as

$$\begin{pmatrix} a_1\ a_2\ a_3\ \ldots\ a_L \\ b_1\ b_2\ b_3\ \ldots\ b_L \end{pmatrix}.$$

For a given nucleotide type i at a given site k, we define

Corresponding author: Sudhir Kumar, Life Sciences A 371, Department of Biology, Arizona State University, Tempe, AZ 85287-1501. E-mail: s.kumar@asu.edu

$$\delta_i^k = \begin{cases} +1 & \text{for } a_k = i \text{ and } b_k \neq i \\ -1 & \text{for } a_k \neq i \text{ and } b_k = i \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Using Equation 2, we can write (1) as

$$D_C = \frac{1}{2}\sum_i \left(\sum_{k=1}^{L} \delta_i^k \right)^2. \quad (3)$$

The expected value is given by

$$E(D_C) = \frac{1}{2} E\left(\sum_i \left(\sum_{k=1}^{L} \delta_i^k \right)^2 \right). \quad (4)$$

Assuming independence among sites, we get

$$E(D_C) = \frac{1}{2} E\left(\sum_i \sum_k (\delta_i^k)^2 \right) + \frac{1}{2} E\left(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'} \right). \quad (5)$$

The first term on the right-hand side is simply the expected number of nucleotides different between the two sequences (N_d), which is determined by the extent of sequence divergence, pattern of evolutionary change, and the extent of evolutionary rate heterogeneity among sites. That is,

$$\frac{1}{2} E\left(\sum_i \sum_k (\delta_i^k)^2 \right) = E(N_d). \quad (6)$$

The second term on the right-hand side in (5) can be written as follows, because the summations are over independent sites:

$$E\left(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'} \right) = \sum_i \left(E\left[\sum_k \delta_i^k \right] E\left[\sum_{k' \neq k} \delta_i^{k'} \right] \right). \quad (7)$$

When the underlying substitution process is homogeneous, then for a given nucleotide pair (i, j), $E(n_{ij}) = E(n_{ji})$, where $n_{i\cdot} = \Sigma_{j \neq i} n_{ij}$, $n_{\cdot i} = \Sigma_{j \neq i} n_{ji}$, and n_{ij} is the number of sites showing nucleotide i in sequence X and j in sequence Y. Thus,

$$E\left(\sum_k \delta_i^k \right) = E(n_{i\cdot}) - E(n_{\cdot i}) = 0. \quad (8)$$

Therefore,

$$E\left[\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'} \right] = 0. \quad (9)$$

Substituting (6) and (9) into (5), we get

$$E(D_C) = E(N_d), \quad (10)$$

where N_d is the number of sites with different nucleotides in sequences X and Y. This proof works for any number of states.

Equation 10 shows that the expected number of differences between two sequences is simply half the sum over all states of the squared differences of the corresponding base (amino acid) frequencies in the sequences compared. Cornish-Bowden (1977) first presented the statistic given in Equation 1. However, the proof of Equation 10 presented in that work implicitly assumed homogeneity of the evolutionary process and also that the counts x_i's and y_i's are independent. The second assumption is clearly invalid because these counts are correlated due to the common ancestry of sequences X and Y. Our proof (Equations 1–10) does not require this assumption and holds irrespective of the complexity of the nucleotide (or amino acid) substitution model to be applied to the observed pattern and the extent of among-site rate heterogeneity among sites. Computer simulations reaffirm this fact over a variety of conditions (Figure 1).

When the two sequences compared do not exhibit the same substitution pattern (heterogeneity scenario), the composition distance obtained using Equation 1 is expected to be larger than that obtained under the homogeneity case. This is because the observed difference in frequency of the same state in two sequences $(x_i - y_i)$ will then be larger. To show this, we conducted computer simulations for amino acid sequences. In this simulation, the probability of change from one amino acid residue to another was made the same for all residues in the sequence evolution in both lineages (homogeneity case, open circles in Figure 2). In the heterogeneity scenario, the transition probability to a given residue is made increasingly larger in a preselected lineage to effect a larger deviation in the substitution pattern [pattern deviation factor (pdf)], with all other transition probabilities kept equal. pdf is a factor by which the probability of change to a prespecified amino acid (or nucleotide) differs from that expected under the homogeneity case. For s states ($s = 4$ for nucleotides and $s = 20$ for amino acids), the probability of substitution to a given state is $1/(s - 1)$ when all changes are equally likely. A pdf equal to f means that the probability of substitution from any state to this prespecified state is $f/(s - 1)$; $f = 1$ corresponds to the homogeneous process. The probability of change to any other state is equal and is given by $(1 - f/[s - 1])/(s - 2)$. A higher value of pdf indicates greater heterogeneity in the patterns of substitution.

Figure 2 shows the results of computer simulations for the homogeneity and heterogeneity cases. It is clear that D_C is higher when the evolutionary process is heterogeneous. This disparity increases with increasing heterogeneity; we call this difference the disparity index (I_D). I_D increases when the number of substitutions increases with pdf kept constant (Figure 3A) and when the pdf increases with the number of substitutions kept constant (Figure 3B). The relationship in both cases is explained approximately by a second order power curve, as the frequency difference is squared in the D_C formula.

In empirical data analysis, we obtain I_D for a given pair of sequences using the equation

$$I_D = \frac{1}{2}\sum_i (x_i - y_i)^2 - N_d, \quad (11)$$

where x_i and y_i are the counts of ith type of nucleotide (or amino acid) in sequences X and Y, respectively, and

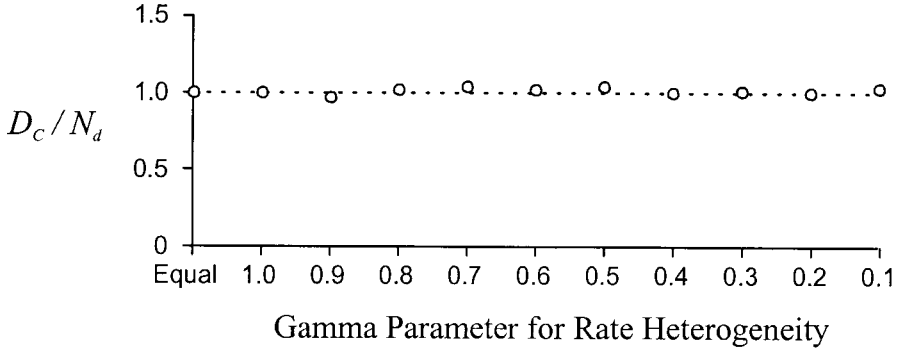

FIGURE 1.—Relationship of the composition distance (D_C) between two sequences and its expected value (the number of differences, N_d) when the two sequences are evolving with the same evolutionary process, under simple and complex patterns of nucleotide substitutions and equal and heterogeneous rates among sites. Each point is an average obtained from 5000 computer simulation replicates with sequence length of 500 nucleotides and 100 substitutions (50 in each lineage). Results presented come from diverse simulation conditions to show the robustness of the relationship between D_C and N_d under different biologically realistic scenarios. For the 11 points shown, the simulation conditions (g_A, g_T, g_C, g_G, a, α/β) were as follows, in order from left to right: (1) 0.25, 0.25, 0.25, 0.25, ∞, 1; (2) 0.05, 0.45, 0.05, 0.45, 0.1, 5; (3) 0.20, 0.30, 0.20, 0.30, 0.2, 2; (4) 0.15, 0.35, 0.15, 0.35, 0.3, 3; (5) 0.10, 0.40, 0.10, 0.40, 0.4, 4; (6) 0.05, 0.45, 0.05, 0.45, 0.5, 5; (7) 0.20, 0.30, 0.20, 0.30, 0.6, 2; (8) 0.15, 0.35, 0.15, 0.35, 0.7, 3; (9) 0.10, 0.40, 0.10, 0.40, 0.8, 4; (10) 0.05, 0.45, 0.05, 0.45, 0.9, 5; (11) 0.05, 0.45, 0.05, 0.45, 1.0, 5. g_A, g_T, g_C, and g_C refer to the respective equilibrium frequencies of the four nucleotides, a is the value of the gamma parameter quantifying the extent of rate heterogeneity among sites, and α/β is the transition/transversion rate ratio. Similar results were obtained for a general reversible model with and without rate heterogeneity among sites.

the observed N_d is used as an estimator of the D_C expected under homogeneity. When the homogeneity assumption is satisfied, $E(I_D) = 0$, because the expected value for the first term is the same as that for the second term (Equation 10).

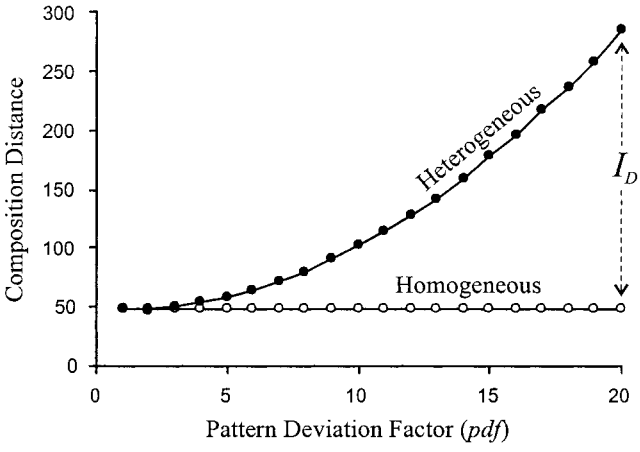

FIGURE 2.—Composition distances when sequences are evolving with homogeneous and heterogeneous patterns of substitution in computer simulations. I_D shows the difference in D_C for the heterogeneous and the homogeneous conditions. Each value is an average obtained from 5000 computer simulation replicates with a pair of protein sequences of length 500 and with 50 substitutions. In each simulation replicate, an ancestral sequence of the specified length was generated assuming equal amino acid frequencies. This sequence was then subjected to substitution in two lineages to produce sequences 1 and 2. In lineage 1, the probability of change from one amino acid to another was the same for all amino acids. For the homogeneity case, the same evolutionary process was used in lineage 2. In the heterogeneity scenario, the evolutionary process in lineage 2 differed from lineage 1 in that the probability of change to a prespecified amino acid was modified by a multiplication factor (pdf) higher than in lineage 1.

MONTE CARLO METHOD FOR TESTING THE HOMOGENEITY ASSUMPTION

We test the homogeneity assumption by calculating the probability of observing a composition distance (D_{CO}) greater than that expected under the null hypothesis of homogeneity, *i.e.*, $I_D > 0$. Because the actual distribution of D_C under homogeneity for the given base frequencies and number of differences is not known *a priori*, we derive it using a Monte Carlo approach. In each replicate of the Monte Carlo method, we start with a random sequence of length L; the expected frequencies are made equal to the average base frequencies computed using the given pair of sequences. Two descendent sequences are then generated by introducing substitutions randomly until the number of differences between the descendent sequences becomes equal to N_d for the original pair of sequences. This is done to obtain D_C under the homogeneity assumption from the observed data, given the average base frequencies for the original pair of sequences. For effecting a substitution, we randomly select one of the two descendent sequences and then choose a site in this sequence at random. We replace the nucleotide at this site (irrespective of its current base) with another chosen randomly on the basis of the average observed frequencies obtained above. Therefore, the resulting sequences are expected to have the same base frequencies, as the substitutions occur with the same evolutionary process in both lineages. (This scheme is chosen because there is no *a priori* information on the null pattern of substitution and evolutionary rate heterogeneity among sites or between lineages.) Using the two sequences generated in the current replicate (say *b*), we compute $D_{C,b}$. This process is repeated a desired number of times and the proportion of replicates in which D_{CO} is higher than the $D_{C,b}$ ($I_D > 0$) is computed. If this proportion is >95%,

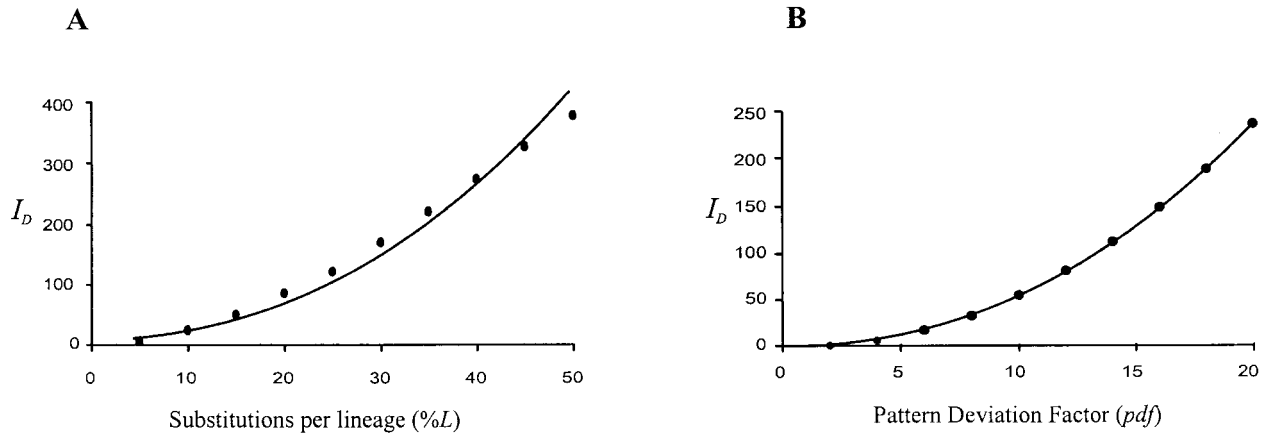

FIGURE 3.—The value of the disparity index (I_D) increases with (A) increasing numbers of substitutions per lineage keeping pdf constant (= 2); (B) increasing pdf with the percentage substitutions per lineage held constant at 5%. I_D estimates are averages obtained from 5000 computer simulation replicates with protein sequences of length 500 amino acids. See Figure 2 legend for details on computer simulation.

we can reject the null hypothesis at the 5% level. As an example, we show the distribution of D_C for the amino acid sequence of human and mouse myeloid differentiation primary response proteins in Figure 4. For this pair, $D_{CO} = 93$, $N_d = 56$, and therefore $I_D = 37$. This I_D is >0 at the 5% level as D_{CO} is located on the right of the 95% cutoff point ($D_C = 92$) in the D_C distribution.

POWER OF THE I_D-TEST

To assess the power of the Monte Carlo test in detecting differences in the evolutionary patterns, we conducted computer simulations under biologically diverse conditions. Figure 5A shows the type I error of the I_D-test at the 5% significance level when the pattern of substitution is homogeneous for three sets of conditions: (1) the Jukes-Cantor (JC; JUKES and CANTOR 1969) model with the same evolutionary rate among sites; (2) the Hasegawa-Kishino-Yano (HKY + G; HASEGAWA *et al.* 1985) model with biased base composition, transition/transversion rate bias, and extreme rate heterogeneity among sites; and (3) the general time-reversible model with rate heterogeneity among sites (GTR + G; reviewed in NEI and KUMAR 2000). Results in Figure 5A clearly show that the type I error at the 5% significance level is ~5%, and thus the test is not conservative. Similar results were obtained in simulations involving unequal rates of evolution between lineages and for protein sequences (results not shown). Given that the type I error could be >5% in some cases (Figure 5A), we recommend that a 1% significance level may be more appropriate.

Figure 5, B and C, shows the power of the I_D-test in rejecting a false null hypothesis when the sequences compared have actually evolved with different evolutionary processes. The statistical power of the I_D-test in rejecting the null hypothesis increases with the number of substitutions and sequence length (Figure 5B). For a given sequence length and number of substitutions, its power increases quickly with even small deviations in the evolutionary pattern between sequences (pdf = 2; Figure 5C). Similar results are found when the sequence evolution followed HKY, HKY + G, GTR, and GTR + G models.

Relative power of the I_D-test: The χ^2-test is often employed to examine if the base frequencies are similar between sequences. In this case,

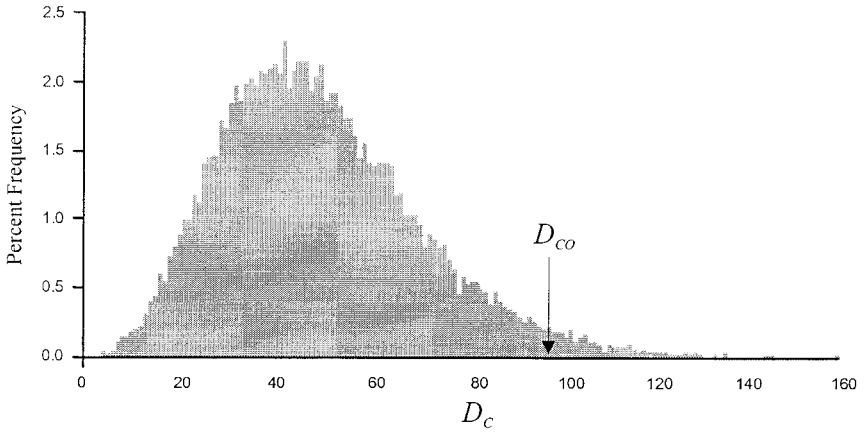

FIGURE 4.—Distribution of the composition distance under the homogeneity assumption (D_C) obtained by the Monte Carlo method (30,000 replicates) for the amino acid sequence of the myeloid differentiation primary response protein in human and mouse (GenBank accession nos. U70451 and U84409). The location of the observed composition distance (D_{CO}) in this distribution is shown with an arrow.

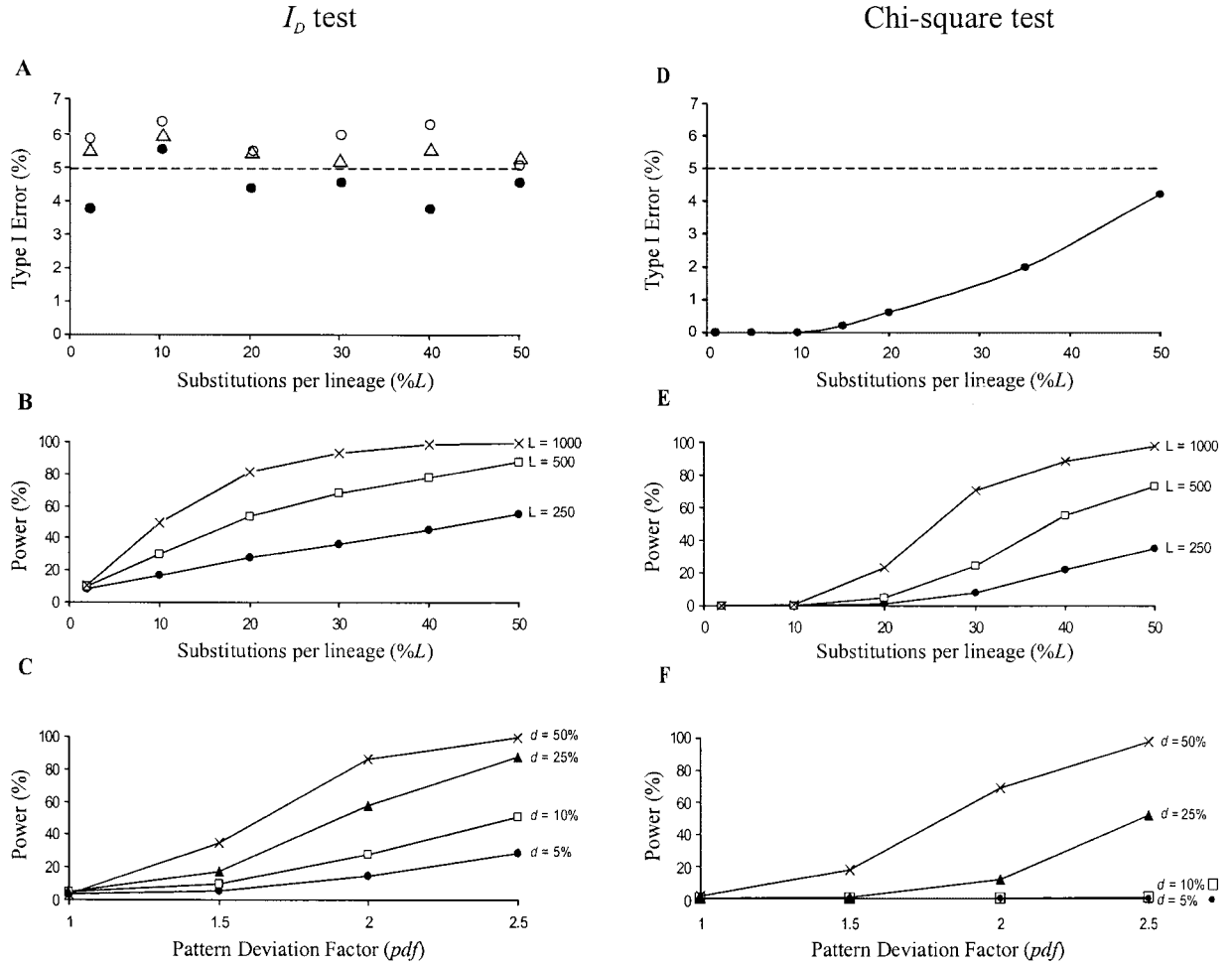

FIGURE 5.—Performance of the I_D-test for nucleotide sequences evolving under the same and different patterns of nucleotide substitutions. (A) Type I error of the I_D-test at a 5% significance level for nucleotide sequences of length 250 when the compared sequences evolved with the same JUKES and CANTOR (1969) model (solid circles) and HASEGAWA *et al.* (1985) model (open circles) with the parameters $g_A = 0.10$, $g_T = 0.40$, $g_C = 0.10$, $g_G = 0.40$, $a = 0.10$, $\alpha/\beta = 5$; and general-reversible model (open triangles) with parameters $g_A = 0.10$, $g_T = 0.40$, $g_C = 0.10$, $g_G = 0.40$, $a = 0.10$, A ↔ C:A ↔ G:A ↔ T:C ↔ G:C ↔ T:G ↔ T::5:1:2:1:5:2, respectively, where g_A, g_T, g_C, and g_G refer to the respective frequencies of the four nucleotides, a is the value of the gamma parameter quantifying the extent of rate heterogeneity among sites, and α/β is the transition/transversion rate ratio. The dashed line indicates the 5% error level. (B) The power of the I_D-test for increasing number of substitutions, when the pattern of nucleotide substitution differs (pdf = 2). (C) The proportion of replicates in which the I_D-test rejects a false null hypothesis (power) for increasing pdf is shown, for sequence length = 500 nucleotides. d gives the number of substitutions per site. The numbers of simulation replicates were 5000 for A and 1000 for B and C. (D–F) Type I error and power of the χ^2-tests in computer simulations under the same conditions as for A–C. Similar results were obtained for all cases described above when the general reversible model with rate heterogeneity among sites was used.

$$\chi^2 = \sum_i (f_{1i} - f_{2i})^2/(f_{1i} + f_{2i}) \qquad (12)$$

is used, where f_{1i} and f_{2i} are the respective counts of the ith state in sequences 1 and 2. Type I errors of the χ^2-test at the 5% significance level obtained in computer simulations under homogeneity assumption are given in Figure 5D. The χ^2-test is clearly a conservative test. This conservative nature is also manifested in the power curves for the χ^2-test when the null hypothesis is false (Figure 5, E and F). In all our simulations, the I_D-test was more powerful than the χ^2-test (Figure 5; other results not shown).

The reason for the conservative nature of the classical χ^2-test is the underlying assumption that the counts are independent. This is not so because the frequencies obtained from homologous sequences are not independent due to the shared evolutionary history. This nonindependence inflates the denominator in the χ^2-test formula as it incorporates information from all sites, even including those that have not undergone any substitutions. Inclusion of these invariant positions in the denominator makes the χ^2-value too low, whereas their contribution in the numerator automatically cancels out. This effect is more severe for closely related sequences than for distantly related sequences, as a larger fraction of sites are identical by descent and thus invariant in the former. We conducted computer simulation studies to examine the type I error of the χ^2-test on the

TABLE 1

Proportion of 3789 human and mouse genes for which the homogeneity assumption is rejected

Data	Average length	I_D-test[a] α[b] = 0.05 (%)	I_D-test[a] α = 0.01 (%)	χ^2-test α = 0.05 (%)	χ^2-test α = 0.01 (%)
Fourfold degenerate sites	219	40.5	27.6	22.9	14.4
Zerofold degenerate sites	871	12.4	5.8	0.3	0.2
Protein sequences	481	12.9	5.4	0.2	0.1

[a] Monte Carlo test with 500 replications.
[b] Significance level.

basis of only those sites that had undergone change in one or both lineages (we refer to this as the V^2-test). In this case, the test became liberal with type I error almost two times the significance level when the null hypothesis is true. One might consider constructing a null distribution for the V^2-test using the Monte Carlo approach, but it is unclear what the expected V^2 is under homogeneity. In any case, D_C and V^2 are quite similar in form and the expected distribution of D_C under homogeneity can be easily constructed. Furthermore, the V^2 statistic has no clear-cut biological interpretation, unlike the D_C statistic.

The problem of observing the same base at a site due to factors such as identity by descent was also considered by Rzhetsky and Nei (1995). They developed a rigorous statistical test of equality of nucleotide (amino acid) frequencies among multiple sequences. Our computer simulations (not shown) in conditions equivalent to those in Figure 5, however, show that the Rzhetsky-Nei test is also a liberal test, which may be due to the violation of some of the assumptions made in their test.

TESTING THE HOMOGENEITY OF MOLECULAR EVOLUTIONARY PATTERNS IN HUMAN AND MOUSE GENES

Human and mouse genome sequencing projects provide DNA sequences of a large number of genes, which gives us an opportunity to examine the homogeneity of patterns of substitution for different genes in human and mouse lineages on a genome-wide scale. We assembled a data set consisting of cDNA sequences of 3789 human genes and their mouse orthologs using the July 1999 release of the HOVERGEN database (Duret *et al.* 1994). Sequence orthology in each case was determined using multigene family trees constructed using the neighbor-joining method (Saitou and Nei 1987) from protein sequence alignments and the homogeneity assumption examined for neutral substitutions in individual genes in human and mouse lineages. For this purpose, we use the fourfold degenerate sites, which are known to best reflect the neutral evolutionary patterns, as no nucleotide change in fourfold degenerate sites will alter the amino acid encoded by the codon. We took a stringent approach in identifying fourfold degenerate sites by choosing sites that have potentially remained fourfold degenerate throughout the evolutionary history of human and mouse. This was accomplished by considering a site to be fourfold degenerate only if it was so in both human and mouse codons. With this definition only ~15% of all sites in a gene were fourfold degenerate, with the average number being ~220.

We tested the null hypothesis of similarity of the evolutionary process in human and mouse lineages (homogeneity assumption) for each gene by the I_D-test. Results show that the null hypothesis can be rejected in 41% of the genes at the 5% significance level (Table 1). This indicates that the neutral evolutionary sites are potentially evolving with significantly different substitution patterns between human and mouse lineages. Homogeneity-rejected genes are not necessarily evolving faster than other genes because the average proportion of sites different in the two cases was similar (0.36 and 0.32, respectively). As expected, the χ^2-test was conservative as it rejected the null hypothesis in only 23% of genes at the 5% level and only 14.4% at the 1% level (Table 1). Therefore the χ^2-test is only one-half as powerful as the I_D-test for these data.

Mammalian genomes are mosaics of regions of homogeneous base compositions (see review in Bernardi 2000). These isochores are characterized by their G + C content, which is also reflected in the third codon positions (and fourfold degenerate sites). If homogeneity is rejected in many genes due to shifts in G + C content either in the human sequence or the mouse sequence for a gene, then the average G + C content difference between human and mouse sequences at fourfold degenerate sites ($|\Delta GC4|$) is expected to be higher for homogeneity-rejected genes as compared to the other genes. This was indeed the case, as the average $|\Delta GC4|$ over homogeneity-rejected genes was 12.9%, which is almost three times that observed in all other genes (4.6%). In fact, the I_D-test for G + C content difference almost always rejects the same genes (40.8% at the 5% level).

Significant differences in G + C content between

genes could arise if the G + C content of one of the two genomes has experienced an overall change. This does not appear to be the case as the percentages of G + C content averaged over all genes in fourfold degenerate sites are 59.7 and 58.3%, respectively, for human and mouse genomes. Another possibility is chromosomal rearrangement. Mammalian genomes are also known to rearrange at a high rate (KUMAR *et al.* 2001), which can relocate genes from one isochore to another with substantially different G + C content. Such an event may lead to a directional change in the G + C content of the fourfold degenerate sites in the transposed gene such that it becomes similar to that of its surroundings. As the genetic maps of human and mouse genomes become available from completed gene sequencing projects, we plan to examine the contribution of gene relocation in the observed shifts in patterns of neutral evolution.

We also conducted tests of the homogeneity assumption for zerofold degenerate sites, which are under strong purifying selection because all changes at these nucleotide sites produce a change in the amino acid encoded. The χ^2-test rejected the null hypothesis in only 0.3% of the cases, which is much lower than that expected by chance alone at the 5% significance level. The I_D-test rejects the null hypothesis in 12.4% of the genes (Table 1). A similar result was seen in the analysis of protein sequences, in which the null hypothesis was rejected in 12.9% of the cases. These results indicate that protein sequences have evolved with a more homogeneous process than evolutionarily neutral sites because 41% of the genes were rejected in the latter case.

Thus, we have shown the usefulness of the I_D statistic as a diagnostic tool to identify pairs of sequences that are evolving with significantly different substitution patterns. In molecular phylogenetics, the ability to identify such sequence pairs prior to evolutionary tree reconstruction using the I_D-test is potentially useful for deciding on whether or not to use phylogenetic reconstruction methods that relax the homogeneity assumption (*e.g.*, LOCKHART *et al.* 1994; GALTIER and GOUY 1998). However, it is worth noting that the parameter richness of such sophisticated methods can be a hindrance in obtaining results with high statistical confidence (reviewed in NEI and KUMAR 2000). Alternatively, investigators may choose to remove sequences that do not satisfy the homogeneity assumption using the I_D-test and use simpler models to make more robust phylogenetic estimations. In general, the I_D-test will be useful as a diagnostic tool to screen for lineages and genes evolving with atypical patterns of change.

We thank S. Blair Hedges, Michael Douglas, Marlis Douglas, Tom Dowling, Mark Miller, and Philip Hedrick for comments on an earlier draft of this article; Sankar Subramanian for help with cDNA sequence alignments; and Michael Rosenberg for invaluable help with the simulation study. We also thank two anonymous reviewers and Dr. Marcy Uyenoyama for many insightful comments and making the derivation of Equation 10 more concise. This work was supported by research grants to S.K. from the National Institutes of Health (HG02096), National Science Foundation (DBI-9983133), and Burroughs-Wellcome Fund (BWF-1001311). Methods described in this work are available in the computer software *MEGA2* (http://www.megasoftware.net).

LITERATURE CITED

BERNARDI, G., 2000 Isochores and the evolutionary genomics of vertebrates. Gene **241:** 3–17.

CORNISH-BOWDEN, A., 1977 Assessment of protein sequence identity from amino acid composition data. J. Theor. Biol. **65:** 735–742.

DURET, L., D. MOUCHIROUD and M. GOUY, 1994 HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. **22:** 2360–2365.

FUNK, D. J., D. J. FUTUYMA, G. ORTI and A. MEYER, 1995 Mitochondrial DNA sequences and multiple data sets: a phylogenetic study of phytophagous beetles (Chrysomelidae: Ophraella). Mol. Biol. Evol. **12:** 627–640.

GALTIER, N., and M. GOUY, 1998 Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15:** 871–879.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

HASEGAWA, M., T. HASHIMOTO, J. ADACHI, N. IWABE and T. MIYATA, 1993 Early branchings in the evolution of eukaryotes: ancient divergences of entamoeba that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. **36:** 380–388.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KUMAR, S., S. R. GADAGKAR, A. FILIPSKI and X. GU, 2001 Determination of the number of conserved chromosomal segments between species. Genetics **157:** 1387–1395.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY and D. PENNY, 1994 Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11:** 605–612.

NAYLOR, G. J. P., and W. M. BROWN, 1998 Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. **47:** 61–76.

NEI, M., and S. KUMAR, 2000 *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

RODRIGUEZ-TRELLES, F., R. TARRIO and F. J. AYALA, 2000 Evidence for a high ancestral GC content in Drosophila. Mol. Biol. Evol. **17:** 1710–1717.

RZHETSKY, A., and M. NEI, 1995 Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. **12:** 131–151.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

STEEL, M. A., P. J. LOCKHART and D. PENNY, 1993 Confidence in evolutionary trees from biological sequence data. Nature **364:** 440–442.

TARRIO, R., F. RODRIGUEZ-TRELLES and F. J. AYALA, 2000 Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. Mol. Phylogenet. Evol. **16:** 344–349.

Communicating editor: M. K. UYENOYAMA