The K_A/K_S Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study

Anton Nekrutenko, Kateryna D. Makova and Wen-Hsiung Li

Genome Res. 2002 12: 198-202
Access the most recent version at doi:10.1101/gr.200901

The K_A/K_S Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study

Anton Nekrutenko, Kateryna D. Makova, and Wen-Hsiung Li[1]

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

Comparative genomics is a simple, powerful way to increase the accuracy of gene prediction. In this study, we show the utility of a simple test for the identification of protein-coding exons using human/mouse sequence comparisons. The test takes advantage of the fact that in the vast majority of coding regions, synonymous substitutions (K_S) occur much more frequently than nonsynonymous ones (K_A) and uses the K_A/K_S ratio as the criterion. We show the following: (1) most of the human and mouse exons are sufficiently long and have a suitable degree of sequence divergence for the test to perform reliably; (2) the test is suited for the identification of long exons and single exon genes, which are difficult to predict by current methods; (3) the test has a false-negative rate, lower than most of current gene prediction methods and a false-positive rate lower than all current methods; (4) the test has been automated and can be used in combination with other existing gene-prediction methods.

Although computational gene prediction has made much progress over the last decade, its reliability remains a challenging problem. A recent survey by Rogic et al. (2001) revealed that the most widely used gene prediction tools have such common problems as high false-positive and false-negative rates and decreased accuracy toward long exons and single exon genes. Some of these difficulties stem from the fact that current gene prediction tools use a similar set of criteria and try to infer all necessary information from one sequence. This one-sequence approach reflects the history of genomics: the amount of genomic sequence data for comparative analysis was limited just a couple of years ago. Today, with the human genome near completion and the complete mouse genome expected in the near future, gene prediction algorithms should make extensive use of cross-species comparison. At present, this area remains largely unexplored with the exception of a few recent reports (Batzoglou et al. 2000; Kent and Zahler 2000).

The utility of cross-species comparisons in gene prediction is based on two key assumptions. First, nondistantly-related species, such as human and mouse, share the vast majority of their genes. In fact, it has been estimated that only 1% of human genes have no homology with other animals (The International Human Genome Sequencing Consortium 2001). Second, most genes are subject to purifying selection with stronger selective constraints for nonsynonymous changes than for synonymous ones (Li 1997; Makalowski and Boguski 1998). This provides a basis for the identification of genes by comparing genomic sequences from two species. One may first search for regions conserved between the two genomes and then assess the protein-coding potential of each conserved region. For the later step, one can use the nonsynonymous-to-synonymous substitution ratio test (the K_A/K_S test), which is described below. Nucleotide substitutions in protein-coding regions are divided into two classes, ones that change amino acid (nonsynonymous) and those that do not (silent or synonymous). From these two numbers and the numbers of synonymous and nonsynonymous sites, one can calculate two normalized values, K_A, the number of nonsynonymous substitutions per nonsynonymous site, and K_S, the number of synonymous substitutions per synonymous site. Because, on average, nonsynonymous substitutions occur less frequently than synonymous substitutions (e.g., $K_A < K_S$), the ratio K_A/K_S has been found to be significantly smaller than one for most protein-coding regions (Makalowski and Boguski 1998).

From the above reasoning, we propose the following approach to gene prediction. If a genomic region is very similar between two species and contains a reading frame that satisfies the requirement $K_A/K_S < 1$, then this region is likely to represent a protein-coding exon. However, before we can apply this method to unannotated genomic regions, we need to test the reliability of the K_A/K_S test on a set of biologically validated protein-coding exons. The goal of this report is to study the reliability of the K_A/K_S test in exon prediction by use of human and mouse exons. In addition, we conducted a simulation study to evaluate its false-positive rate of prediction. We asked the following questions: (1) is a typical mammalian exon long enough for the test to perform reliably; (2) is the degree of sequence divergence between human and mouse exons suitable for this approach; (3) how does the test perform for long exons or single exon genes, which are difficult to predict by current methods; and (4) how often does the test misidentify a noncoding region as a coding region?

RESULTS

Set of Orthologous Exons

Our dataset included 153 genes representing a total of 1244 exons. This included 25 single exon genes. The average number of exons in multi-exon genes is 17, with the majority of genes (~72%) containing 10 or fewer exons. The maximum

[1]**Corresponding author.**
E-MAIL whli@uchicago.edu; FAX (773) 702-9740.
Article published on-line before print in December 2001: *Genome Res.*, 10.1101/gr.200902.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.200901.

12:198–202 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

number of exons (64) was in the human lamilin α 2 chain gene (GenPept accession no. AAB18388). Some descriptive statistics of exons are given in Table 1. We divided all exons into three categories according to their location in the coding region, initial (first coding exon of a gene), internal, and terminal (last coding exon of a gene). This was done to examine the possibility that the initial and terminal exons of a gene may behave differently from internal exons. As initial and terminal exons frequently contain untranslated regions, only the coding parts (excluding stop-codons) were included in our analyses. On average, internal exons tend to be shorter than initial and terminal exons and have a nearly normal length distribution with the mean of 132 bp (Table 1). These numbers agree with the data reported by the International Human Genome Sequencing Consortium (2001). Our data show that human exons have an average nucleotide identity of 86.5% (SD = 8.0) with orthologous mouse sequences, close to the estimates reported by Makalowski and Boguski (1998) and Batzoglou et al. (2000).

Synonymous and Nonsynonymous Rates

Synonymous (K_S) and nonsynonymous (K_A) distances were estimated by use of the maximum likelihood method implemented in the codeml program (Yang 1999) under the F3 × 4 model of codon substitution (Goldman and Yang 1994). This model accounts for both transition/transversion and codon usage biases. The results of this analysis are summarized in Table 2. The upper limit of synonymous rate is unusually high due to the presence of several outliers with a small number of synonymous sites. The mean K_A value is one order of magnitude smaller than the mean K_S value and this is also the case when the medians instead of the means are considered. There were only 4 cases (of 1244) in which the K_A/K_S ratio was higher than 1 (1.061, 1.598, 2.024, and 2.138). However, none of them was significantly greater than 1 because these exons were either very short (36–63 bp) or too divergent (<70% nucleotide identity).

The K_A/K_S Ratio Test

Among the 1244 exons tested, only 118 (9.5%) failed to produce a K_A/K_S ratio significantly smaller than 1 (at the 5% significance level). The proportion of exons failing the test of $K_A/K_S < 1$ ($P < 0.05$) was different for the three exon types, 16% ($N = 114$, 95% CI: 9.3%–22.7%) of first exons, 14% ($N = 114$, 95% CI: 7.6%–20.4%) of last exons, and only 8% ($N = 991$, 95% CI: 6.3%–9.7%) of internal exons failed the test. Thus, although this difference was not statistically significant, internal exons tend to be recognized best by this approach. The power of the test depends on the exon length and on the degree of divergence between sequences. To study the effects of exon length on the performance of the test, we classified the exons into six length classes. Figure 1A shows the total number of exons as well as the number of exons that failed the test in each of the six length classes. As expected, the longer the exon is, the better the test performs. The mean exon length in our dataset is 132 bp (Table 1). As shown in Figure 1A, the majority of internal exons (~70%) is concentrated in the two relatively short length classes, 100 and 150 bp. However, only a small proportion of them (43 of 692, or ~6%) failed the test. Thus, the test performs well on most typical mammalian exons.

Let us now consider the effect of human–mouse sequence divergence on the outcome of the K_A/K_S test (Fig. 1B). The exons in our dataset were divided into five bins (categories) on the basis of the percent divergence between human and mouse sequences. For each category, we counted exons that did not meet the criterion of $K_A/K_S < 1$ ($P < 0.05$) and divided the obtained number by the total number of exons in this category. Figure 1B shows that the majority of exons (~85%) fall into three categories (5%–10%, 10%–15%, and 15%–20% bins) for which the proportion of exons failing the test is smaller than 8% (7.3%, 4.1%, and 6.6%, respectively). Therefore, the divergence observed between human and mouse exons is suitable for our purpose.

Characteristics of Exons that Failed the Test

A total of 118 exons (from 57 genes) failed the test (false negatives). The average length of these exons is 89 bp (SD = 49; median = 78 bp), which is significantly smaller than the rest of the exons in our dataset (*t*-test, $P < 0.0001$). There was no case in which all exons of a gene failed the test. Of the 57 genes containing exons failing the test, 28 (49%), 14 (25%), and 15 (26%) contained 1, 2, and >2 failing exons, respectively. In these three classes of genes, the number of exons per gene had the following ranges: 2–62, 6–33, and 7–35, respectively. The average proportion of failing exons in a gene was 17% (median = 9%). The largest number of exons failing the test in a gene was 6. This was the aldehyde oxidase gene (human accession no. AAB83966, mouse accession no. NM_009676), which contains a total of 35 protein-coding ex-

Table 1. Descriptive Statistics of Human Protein-Coding Exons

Exons	No. of exons	Length[a] range	Length[a] mean	Length[a] median	GC% overall	GC% third pos.	% Identity with mouse
All[c]	1219	30–1842	147	123	52.1 ± 9.6	59.4 ± 19.3	86.5 ± 8.0
Initial[d]	114	30–1842	222	190	59.6 ± 10.2	72.7 ± 17.0	
Internal	991	30–624	132	123	51.0 ± 9.1	57.5 ± 19.1	
Terminal[d]	114	30–1167	203	190	54.5 ± 10.2	63.9 ± 18.7	
Single-exon genes	25	327–4737	1376	1228	56.1 ± 9.6	68.2 ± 19.0	

[a]Comparisons of human and mouse sequences often involve insertion/deletion events. Therefore, the alignable length of exons in this study is sometimes smaller than their actual length. However, in the majority of cases (1104 of 1244, or 88%) alignable length is equal to the actual lengths of human and mouse sequences.
[b]Minimal alignable length was set to 30 bp.
[c]Excluding single-exon genes.
[d]Only coding parts of the initial and terminal exons were included.

Table 2. Summary Statistics for K_a and K_s

	K_a			K_s			K_a/K_s		
Exons	mean (SD)	median	range	mean (S)	median	range	mean(SD)	median	range
First	0.083 (0.111)	0.041	0–0.615	0.920 (1.134)	0.620	0–9.802	0.139 (0.237)	0.057	0.001–2.024
Internal	0.062 (0.125)	0.034	0–1.863	0.820 (0.812)	0.602	0–7.904	0.100 (0.150)	0.054	0.001–2.138
Last	0.082 (0.152)	0.045	0–1.340	0.869 (0.747)	0.693	0–4.053	0.122 (0.207)	0.061	0.001–1.600
Single-exon	0.055 (0.051)	0.045	0–0.209	0.665 (0.273)	0.646	0–1.335	0.088 (0.080)	0.071	0.005–0.383
All	0.065 (0.125)	0.036	0–0.863	0.830 (0.835)	0.608	0–9.802	0.105 (0.164)	0.054	0.001–2.138

ons. Therefore, every gene in our dataset had a small proportion of exons that did not pass the K_A/K_S test.

The False-Positive Rate of the Test

How frequently does the test show K_A/K_S <1 (P <0.05) by chance alone? In other words, how frequently does the test misidentify a noncoding region as a coding region? To answer this question, we did a computer simulation by generating pairs of random sequences and performed the K_A/K_S test. Sequences were generated in 24 discrete length/divergence classes (Table 3). Each class contained 1000 replicates. From Table 3, we see that the proportion of sequences passing the K_A/K_S test (false positives) ranges from a maximum of 5% (length = 51 bp, mean divergence = 15%) to a minimum of 1.1% (length = 1002 bp, mean divergence = 20%). The proportion of false positives averaged across all length/divergence classes is 2.6% (95% CI: 2.40–2.78). The proportion of false positives tends to decrease as the length of the sequence increases. Several outliers, such as 3.5% of false positives among sequences with a length of 300 bp and mean divergence of 15% are likely the result of statistical fluctuations, as only 1000 sequence pairs were considered.

DISCUSSION

With the recent publication of the human genome and anticipated completion of the mouse genome in the near future, comparative analysis becomes a powerful tool for understanding genomic data. A number of such studies have already been published (Makalowski and Boguski 1998; Batzoglou et al. 2000; Dubchak et al. 2000; Jareborg et al. 1999; Venter et al. 2001). Because mammals, such as human and mouse, share the majority of their genes, we may be able to identify most of them by searching for blocks of high similarity between human and mouse genomic sequences. Once such blocks are identified, their protein-coding capacity can be assessed with the simple K_A/K_S test. In this study, we showed the suitability of this approach by applying it to a set of orthologous exon sequences from human and mouse, and we showed that it has very low false-positive and negative rates.

The reliability of the K_A/K_S test depends on the number of sites compared. The first objective of this study was to determine whether a typical mammalian exon contains sufficient information for the test. Figure 1A shows that the proportion of exons failing the test is small for the most common length classes, suggesting that the test will perform well on the genome-wide scale. On the other hand, the test performance also depends on the degree of sequence divergence between compared species. If the divergence is too small, then the number of informative sites is small and the test has no power. Similarly, when the two sequences are too distant, saturation of nucleotide substitutions at synonymous sites may compromise the test. Figure 1B shows that the test is expected to work well for the majority of exons in human–mouse comparisons.

Currently, all gene prediction algorithms experience a significant drop in accuracy for exons that are longer than 300 bp (Rogic et al. 2001). In contrast, in this study, all exons ≥300 bp (N = 87 including 25 single exon genes) display K_A/K_S significantly smaller than 1. This has important implications for reliable gene prediction because many genes, such as

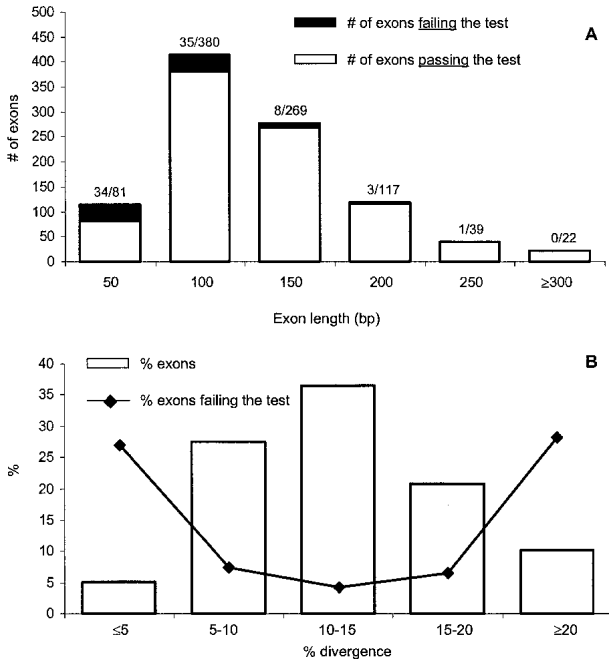

Figure 1 (*A*) Distribution of exon lengths. Exons were stratified into six length classes. For example, the 100-bp class contains exons with lengths ranging from 75 to 125 bp. The white area of each bar represents the number of exons that show K_A/K_S significantly smaller than 1 (passing the test), whereas the shaded area corresponds to the number of exons that have K_A/K_S not statistically different from 1 (failing the test). The numbers above each bar indicates the ratio of the number of exons in the shaded area to the number of exons in the white area. For example, the group with the mean length 50 bp contains 81 exons; 34 of them did not pass the K_A/K_S test. (*B*) Relationship between human–mouse sequence divergence and the number of false negatives (exons that fail the test). Bars represent the proportions of exons in each of the five divergence classes. Points on the curve indicate the proportion of false negatives within each identity class. For example, ~35% of exons in our dataset belong to a class in which divergence ranges from 10% to 15%. Within this class ,~4% exons did not pass the K_A/K_S test.

Table 3. Proportion of Random Sequences with K_A/K_S Significantly Smaller than 1 ($P < 0.05$)[a]

	Length (codons/nucleotides)					
% Divergence	17/5	34/102	50/150	67/201	100/300	334/1002
5	3.0 (2.0–4.1)	2.2 (1.3–3.1)	2.5 (1.5–3.5)	2.5 (1.5–3.5)	2.0 (1.1–2.9)	2.2 (1.3–3.1)
10	2.9 (1.9–3.9)	2.8 (1.8–3.8)	2.6 (1.6–3.6)	1.8 (1.0–2.6)	2.2 (1.3–3.1)	1.9 (1.1–2.7)
15	5.0 (3.6–6.4)	3.3 (2.2–4.4)	1.7 (0.9–2.5)	2.1 (1.2–3.0)	3.5 (2.4–4.6)	1.9 (1.1–2.7)
20	4.8 (3.5–6.1)	2.9 (1.5–3.5)	2.9 (1.9–3.9)	2.5 (1.5–3.5)	2.5 (1.5–3.5)	1.1 (0.5–1.7)

[a]The values in parentheses indicate the 95% confidence interval.

genes for G-protein-coupled receptors, are intronless and possess unusually long exons.

In this study, the test missed <10% of the exons. These exons were on average shorter ($P < 0.0001$) than the rest of the data set. The missed exons all belonged to multiexon genes (there was no case in which a single exon gene was missed). In a multiexon gene, the missed exons typically constituted only a small proportion of the total exon number. Therefore, the problem of missed exons is not likely to cause difficulties in recovering complete gene structures, especially when multiple lines of evidence are used (such as the OTTO system; Venter et al. 2001). A way to recover these exons is to add a third species to the comparison, which will increase the number of informative sites. A preliminary analysis by our group (unpubl.) has shown that this approach worked well even for short exons (the test became significant in all cases when a third species was added). However, at present, this idea cannot be applied on the genome-wide level because only a very limited amount of sequence data is available for mammalian species other than human and mouse. Genomic sequences from a third mammal (neither a primate nor a rodent) should be the most useful for the K_A/K_S analysis. For example, dog or cow would be a good candidate, because these species have a suitable degree of divergence relative to human and mouse. Unfortunately, these two species are poorly represented in sequence databases. For example, GenBank (as of May 2001) contained 4,668,765 human and 2,871,750 mouse entries, whereas the total number of sequences from cow and dog combined was only 179,878! This situation further emphasizes the importance of pursuing genome sequencing projects for other nonprimate and nonrodent mammals.

The proportion of exons missed by the test can be viewed as its false-negative rate. Thus, we can say that when the test is used for the exon identification the false-negative rate is 9.5%. This is comparable with that of the GenScan algorithm (proportion of missed exons is 8%) and lower than most other gene prediction tools (range, 12%–32%; from Rogic et al. 2001). It is also important to know the false-positive rate of the test. In other words, when the test is applied to genomic regions conserved between human and mouse, how many of them will have $K_A/K_S < 1$ ($P < 0.05$) by chance? To determine this, we performed a simulation by applying the test to a set of randomly generated sequences. This experiment showed that the test is expected to have the false-positive rate of only 2%–3% for typical human–mouse exon pairs (length 100–150 bp, divergence 15%; Table 3). This is significantly better than most of the currently used gene prediction tools. For example, according to Rogic et al. (2001) the proportion of wrong exons (false positives, do not overlap with any known exon) ranges from 7% to 28% (9% for GenScan). Note that both the false-negative and false-positive rates in our test can be reduced to virtually 0 by including a third or more genomes.

In our empirical study of the K_A/K_S test, we selected human and mouse orthologous genes from the homol_seq_pairs table in LocuLink. Some of the pairs in this dataset might be paralogous (derived from gene duplication) rather than orthologous (derived from speciation). This possibility does not pose a problem because even for a paralogous gene pair, nonsynonymous substitutions are usually expected to occur less frequently than synonymous substitutions. However, in some cases, duplicate genes may have gone through relaxation of functional constraints or advantageous divergence at the protein level, so that the K_A/K_S ratio may not be smaller than 1. Therefore, orthologous genes are preferred over paralogous genes. In practice, however, it may be too tedious to make this distinction.

The K_A/K_S test alone cannot be used for gene prediction. In particular, it does not have the capability to correctly predict exon/intron boundaries. However, our group is developing an integrated system that will combine the test with existing methods for identification of exon/intron boundaries, promoters, and polyadenylation signals. This system will allow prediction of complete gene structures using human/mouse genome comparisons.

Although the amount of sequence data generated by genome initiatives is overwhelming, we are still struggling with the same old question, how many genes are in a genome? Comparative analysis will help us to answer these questions by providing an alternative estimate of the genome's protein-coding capacity. This study illustrates how a simple evolutionary approach can become a powerful tool in making sense of genomic sequences.

METHODS

Selection of Genes

We chose the Locus Link database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/LocusLink/) because it has been manually curated to ensure high-data quality. Locus Link data were downloaded locally and converted into MySQL tables (MySQL is a relational database system available at no charge from http://www.mysql.com). We selected only those human genes that satisfied the following two conditions: (1) they have known exon-intron structure and (2) they were reviewed by NCBI staff (i.e., have the REVIEWED status). To find mouse orthologs, we used homol_seq_pairs table, which is a part of the standard Locus Link distribution. The resulting dataset contained 153 orthologous pairs of human and mouse genes and is available from (http://pondside.uchicago.edu/~lilab/ka_ks_data.html).

Automated Data Analysis System

To analyze genes selected for the analysis, we designed and implemented an automated system that contains six core modules written in PERL. The system is designed to be used under the Linux environment, but can be used with any other UNIX clone. The system performs the following steps: (1) extract coding sequences (CDS), exon coordinates and CDS translations (protein sequences) from a set of GenBank records representing pairs of orthologous genes; (2) align orthologous protein sequences using CLUSTALW1.81 (Thompson et al. 1994); (3) use protein alignments as a guide to align corresponding CDSs. This eliminates frameshifts that may occur due to incorrect placement of gaps; (4) cut resulting CDS alignments into exons and trim them if necessary to preserve reading frames. At this point, the number of human/mouse orthologs with known exon-intron structure is very limited, so we assumed that the homologous mouse gene has the same exon-intron structure as its human counterpart. The results of Batzoglou et al. (2000) suggest that this assumption is justifiable because sizes of orthologous exons in human and mouse are identical in the majority (73%) of cases; (5) run the codeml program from PAML package (Yang 1999) on each exon alignment twice, first, with the K_A/K_S ratio fixed at 1 and second, with the K_A/K_S ratio as a free parameter. Collect maximum likelihood values ML_1 and ML_2 from the two runs and calculate the likelihood ratio as $LR = 2(lnML_1 - lnML_2)$. LR is then compared against the χ^2 distribution with one degree of freedom to test whether K_A/K_S is significantly different from 1 (see Yang and Bielawski 2000 for explanation of the test); (6) the results of the test are uploaded directly to MySQL and analyzed. Resultant tables can be imported into any spreadsheet software (such as Excel). The results of the analysis described here can be downloaded from http://pondside.uchicago.edu/~lilab/ka_ks_analysis.tar.gz

Computer Simulation

The goal was to generate two random sequences with a given length and divergence. First, a single sequence of specified length is generated by randomly choosing codons. Each of the 61 sense codons has the same probability of being selected. Next, we generate the second sequence by introducing random nucleotide substitutions to the first sequence regardless of codon position until the desired degree of divergence is attained. If a nucleotide substitution creates a stop codon, then this site is mutated again until a nonstop codon different from the original one is created. Third, we perform the K_A/K_S test on each pair of generated sequences using the automated data analysis system developed in this study.

ACKNOWLEDGMENTS

We thank Rick Blocker for the UNIX/Linux system maintenance and Meng-Hsin Hsiao for help. This study was supported by NIH grants GM30998, GM55759, and HD38287.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., and Frazer, K.A. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10:** 1304–1306.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–919.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Kent, W.J. and Zahler, A.M. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res.* **10:** 1115–1125.

Li, W.-H. 1997. Rates and patterns of nucleotide substitutions. In, *Molecular Evolution* (ed. W.-H. Li), pp. 177–214. Sinauer Associates, Sunderland, MA.

Makalowski, W. and Boguski, M.S. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci.* **95:** 9407–9412.

Rogic, S., Mackworth, A.K., and Ouellette, F.B.F. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Venter, J.C., Adams, M.D., Myers, E.W., Li, W.P., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Yang, Z. 1999. *Phylogenetic analysis by maximum likelihood (PAML), Version 2.* University College London, England.

Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15:** 496–503.

Received June 14, 2001; accepted in revised form October 19, 2001.