J Mol Evol (2004) 59:114–120
DOI: 10.1007/s00239-004-2616-9

JOURNAL OF MOLECULAR EVOLUTION

Recent Mammalian Gene Duplications: Robust Search for Functionally Divergent Gene Pairs

Austin L. Hughes, Robert Friedman

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Building, 700 Sumter Street, Columbia, SC 29208, USA

Received: 10 December 2003 / Accepted: 10 February 2004

Abstract. Comparison of 317 gene pairs in human and mouse that were duplicated after the most recent common ancestor of the two species was used to search for candidates that may have undergone functional differentiation. Even when corrected for multiple tests, Tajima's relative rate test showed significant rate differences in 36% of cases for which the test was applicable. However, a significant result in this case was increasingly likely as the sequence length increased; thus, a statistically significant result of a relative rate test may not be biologically meaningful. We used regression methods to provide more robust methods of testing for functionally differentiated gene pairs, which take into account the variation in the entire data set by examination of residuals from regression-identified gene pairs with unusually high nonsynonymous divergence from a reference sequence and from each other. This approach identified six duplicate gene pairs that appeared to be candidates for functional differentiation as a result of positive Darwinian selection.

Key words: Adaptive evolution — Gene duplications — Nucleotide substitution — Relative rate test

Introduction

Gene duplication, which can give rise to new genes encoding proteins with new functions, is believed to have played an important role in the evolutionary diversification of organisms (Nei 1969; Ohno 1970; Li 1982; Hughes 1994). There is evidence that gene duplication occurs continually over evolutionary time (Lynch and Conery 2000; Friedman and Hughes 2003). Only certain duplicate genes are actually retained in the genome, while the others are eventually lost. If a duplicate gene can assume a new function beneficial to the organism, it is more likely that it will be retained (Hughes 1994; Lynch et al. 2001). Thus, positive Darwinian selection may frequently be involved in the fixation of duplicate genes that have undergone beneficial mutations (Hughes 1999a).

It has often been difficult to obtain evidence that positive selection has acted on mutations leading to the functional differentiation of duplicate genes. A useful approach to testing for positive selection involves comparing the number of synonymous nucleotide substitutions per synonymous site (d_S) with the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) (Hughes and Nei 1988). In a number of cases, this approach has provided evidence of positive selection diversifying duplicate genes at the amino acid level (e.g., Hill and Hastie 1987; Tanaka and Nei 1989; Hughes 1999b, 2002; Hughes et al. 2000).

However, it seems unlikely that this approach will be able to detect positive selection in many cases

Correspondence to: Austin L. Hughes; *email:* austin@biol.sc.edu

involving multigene families. First, since positive selection is likely to be focused only on certain functional regions of the protein, this approach works best in cases in which structural and functional information is available (Hughes 1999a). Moreover, positive selection favoring specialization of duplicate genes may typically occur over a short time frame. Once the proteins encoded by two duplicate genes have become specialized for distinct functions, new amino acid changes may no longer be favored (Hughes 1999a). If so, purifying selection will again predominate, and eventually d_S will overtake d_N. There is evidence of such an evolutionary process in plots of d_N vs. d_S in pairwise comparisons among members of a variety of gene families. In such plots, d_N often exceeds d_S when d_S is low, and thus the genes compared have a recent common ancestor, but d_S exceeds d_N in more distant comparisons (e.g., Tanaka and Nei 1989; Hughes et al. 2000).

An additional correlate of functional divergence between two duplicated genes might be inequality (asymmetry) of the rates of nonsynonymous substitution in the two genes; such a pattern might indicate that one of the two genes has adopted a new function, whereas the other gene has retained a function closer to that of the ancestral gene. Some recent studies have made use of genomic data to survey for nonsynonymous rate asymmetry between duplicate genes. Kondrashov et al. (2002), in a study of 101 paralogs, found that significant nonsynonymous rate asymmetry occurred in only 5 cases. On the other hand, in analyses of *Saccharomyces cerevisiae, Schizosaccharomyces pombe*, and *Drosophila melanogaster*, Conant and Wagner (2003) found significant nonsynonymous rate asymmetry in 22 of 80 duplicate gene pairs. Furthermore, Zhang et al. (2003) found significant rate asymmetry at the amino acid level in 145 of 250 human duplicate gene pairs. Although all of these authors used different methods to test for rate asymmetry, none applied any correction for multiple tests. Thus, the true significance levels in these studies remain unclear.

Here we study pairs of paralogous genes in the human and mouse genomes that have arisen by gene duplication since the most recent common ancestor of the two species (estimated to have occurred about 110 million years ago; Kumar and Hedges 1998). We employ robust approaches to test for rate asymmetry between paralog, with an emphasis on methods that take into control for multiple testing. We apply simple regression-based methods that take into account the variability in the entire data set, with the goal of identifying gene pairs likely to have diversified functionally as a result of positive selection.

Methods

Sequence Data

The genomic data for human (version 16.33) and mouse (version 16.30) were obtained from Ensembl (http://www.ensembl.org). The version numbers refer to the database software (Ensembl version 16) and the assembly of the genomic sequences (NCBI versions 33 and 30). Genes predicted by Ensembl have been curated and verified by similarity with homologs discovered experimentally (Clamp et al. 2003). The numbers of annotated protein-coding genes were 32,035 for human and 32,911 for mouse. After removal of genes that were shorter than 100 bases and longer than 300,000 bases, the total gene sets were 29,606 in human and 32,296 in mouse. Further curation to remove overlapping loci resulted in a count of 20,387 genes in human and 23,222 genes in mouse.

Protein families were identified by homology and a single-linkage method employed by the BLASTCLUST software available in the Blast software package (Altschul et al 1997). Sequence homology was established by identifying matches using a conservative E-value of 10^{-6} with a minimum of 30% sequence identity across at least 50% of the length of two sequences. The single-linkage method assembles larger families by linking shared genes among families, thus ensuring that a given gene will be assigned to only one family. To identify recent duplicates, we chose families with exactly three members and at least one member from each of the two species. From these families, we selected those cases in which the number of synonymous substitutions per synonymous site (d_S) between the two conspecific genes was lower than that for either between-species comparison.

There is evidence that even recently duplicated genes can be chimeras as a result of exon shuffling (Katju and Lynch 2003). In order to rule out chimeric genes with marked differences between regions with respect to the extent of sequence divergence, we computed the proportion of amino acid difference in a window of 30 aligned amino acid residues along each pair of paralogs. One pair of paralogs showed a strong difference in sequence similarity between N-terminal and C-terminal regions, and this pair was found to correspond to a known chimeric gene (Paulding et al. 2003). Therefore, this gene family was excluded from the analysis. The resulting data set contained 316 families in which a gene duplication occurred in human (119 families) or in mouse (197 families) after the last common ancestor of human and mouse. The data are available from http://www.biol.sc.edu/~austin/.

Statistical Analyses

Homologous sequences were aligned at the amino acid level using the CLUSTAL W program (Thompson et al. 1994), and this alignment was imposed on the DNA sequences. The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated by a maximum likelihood method (Yang and Nielsen 2000) using the software package PAML (Yang 1997). We used Tajima's (1993) method to test the hypothesis that duplicate gene pairs evolved at equal rates at the amino acid, i.e., to test for rate asymmetry at the amino acid level. This test has an advantage over some other methods that have been used for such relative-rate tests because it is not model-dependent (Tajima 1993). However, the test statistic could not be computed in 32 of 317 families, either because the amino acid sequences were too similar or because they were too divergent. We also used approaches to identifying rate asymmetry based on linear regression, which have the advantage of taking into account stochastic error in the entire data set.

Table 1. Means (±SE) of variables comparing duplicate gene pairs without significant evidence of amino acid evolution rate asymmetry and those with significant evidence of rate asymmetry[a]

	No rate asymmetry ($N = 182$)	Rate asymmetry ($N = 102$)	p^b
No. codons	253.9 ± 11.6	354.5 ± 19.4	< 0.001
d_N between pair members	0.060 ± 0.007	0.118 ± 0.011	< 0.001
d_S between pair members	0.120 ± 0.014	0.207 ± 0.011	< 0.001
d_N–d_S	−0.060 ± 0.009	−0.089 ± 0.010	n.s.
d_N/d_S[c]	0.628 ± 0.029	0.624 ± 0.023	n.s.
Absolute value of standard residual from regression of d_{N1} vs. d_{N2} (comparison with reference sequence)	0400 ± 0.036	1.034 ± 0.113	< 0.001
Absolute value of standard residual from regression of d_N /d_{S1} vs. d_{N2} /d_{S2} (comparison with reference sequence)	0.558 ± 0.044	1.114 ± 0.079	< 0.001
Standard residual from regression of d_N vs. d_S (comparison between pair members)	−0.092 ± 0.067	0.296 ± 0.120	0.003

[a] Tajima's (1993) test ($p < 0.05$ with Bonferroni correction for multiple tests).
[b] t-test (two-tailed).
[c] Undefined in four cases.

Results

Relative Rate Tests

Tajima's (1993) relative rate test statistic could be computed for 284 duplicate gene pairs. Using a Bonferroni-corrected simultaneous significance level, 102 (35.9%) pairs showed a significant rate asymmetry in amino acid sequence evolution at the 5% level. The proportions of gene pairs showing rate asymmetry were similar for the two species: 41 of 109 (37.6%) in human and 60 of 175 (34.3%) in mouse. At the 1% level (Bonferroni-corrected), 81 of 284 (28.5%) gene pairs showed significant rate asymmetry, 35 of 109 (32.1%) in human and 46 of 175 (26.3%) in mouse. Thus despite the use of a conservative statistical test, our results showed a high frequency of rate asymmetry, comparable to other studies (Conant and Wagner 2003; Zhang et al. 2003).

In order to understand the factors contributing to a significant relative rate test, we compared duplicate gene pairs with statistically significant (at the 5% level) evidence of asymmetry in the rate of amino acid evolution with those showing no evidence of asymmetry. We found that gene pairs with significant asymmetry encoded significantly longer polypeptides on average (Table 1). The mean length of pairs with significant asymmetry was 354.5 residues (rang, 102–1134), while the mean length of pairs without significant asymmetry was 253.9 (range 52–962). A plausible explanation for the difference in mean length between the two groups is that the power of Tajima's (1993) test to detect a rate difference increases as the number of sites increases. Pairs with significant asymmetry also had greater mean values of both the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) (Table 1). These results also are best explained as reflecting the power of the test to detect rate differences, which is likely to increase as the number of differences between the sequences increases.

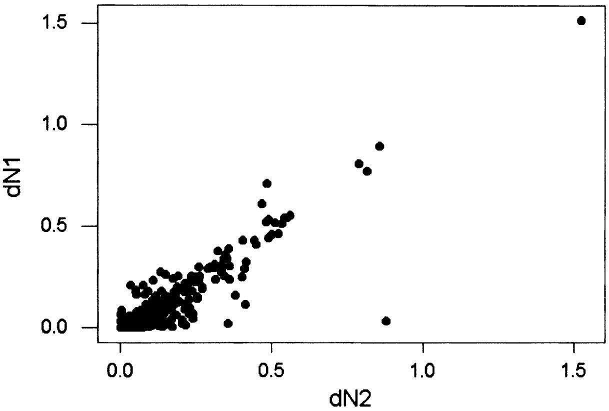

Fig. 1. Plot of d_{N1} (comparison of duplicate gene 1 vs. reference) vs. d_{N2} (comparison of duplicate gene 2 vs. reference). The linear regression line was $Y = -0.00134 + 0.887X$ ($R^2 = 0.789$, $p < 0.001$).

Thus, these results suggested that evidence of a significant rate asymmetry was largely a function of the statistical power to detect a rate difference. As a consequence, applying such a test to individual gene pairs may not be the optimal approach if the goal is to identify cases of exceptional rate asymmetry, which may be indicative of functional divergence

Table 2. Duplicate gene pairs with large (≥2.0) absolute values of the standard residual from the regression of d_{N1} vs. d_{N2}

Species	Protein function	d_{N1} ± SE	d_{N2} ± SE	st. resid.	Ensembl ID
Human	CDC10	1.516 ± 0.243	1.527 ± 0.245	2.64	ENSP00000317050, ENSP00000239730
	Serine protease inhibitor	0.612 ± 0.084	0.467 ± 0.066	2.83	ENSP00000274565, ENSP00000324870
	Coactosin-like	0.019 ± 0.009	0.356 ± 0.045	3.80	ENSP00000262428, ENSP00000329424
	dUTP pyrophosphatase[a]	0.116 ± 0.019	0.414 ± 0.042	31.9	ENSP00000249738, ENSP00000332786
	Unknown	0.186 ± 0.020	0.051 ± 0.014	2.06	ENSP00000328595, ENSP00000285718
	Vacuolar sorting-associated	0.182 ± 0.020	0.051 ± 0.010	2.00	ENSP00000332109, ENSP00000288304
	Unknown	0.011 ± 0.007	0.216 ± 0.032	2.23	ENSP00000261713, ENSP00000328254
	Unknown (homology to spindle pole body protein)[a]	0.209 ± 0.019	0.034 ± 0.007	2.58	ENSP00000311732, ENSP00000330408
	Phospholipase[a]	0.160 ± 0.014	0.379 ± 0.024	2.19	ENSP00000311732, ENSP00000330408
Mouse	*Drosophila* CG 14824 homolog	0.034 ± 0.009	0.882 ± 0.081	10.15	ENSPMUSP0000042926, ENSPMUSP0000053726
	LASP-1	0.209 ± 0.020	0.074 ± 0.011	2.09	ENSMUSP00000036100, ENSNfUSP00000052976
	Secreted and transmembrane protein 1[a]	0.708 ± 0.062	0.486 ± 0.044	3.90	ENSMUSP00000045748, ENSMUSP00000026162
	Unknown	0.273 ± 0.031	0.132 ± 0.020	2.27	ENSMUSP00000061176, ENSMUSP00000055682
	WW domain-binding protein[a]	0.044 ± 0.009	0.239 ± 0.022	2.07	ENSMUSP00000032340, ENSMUSP00000056886

[a] Indicates gene pairs with a high standard residual from the regression of d_N vs. d_S (see Table 4).

between duplicated genes. Instead, we used a number of approaches to identify duplicate gene pairs whose divergence at the amino acid level was unusually high in comparison to other gene pairs in the data set.

Regression Methods

The approach we chose involved conducting linear regression analyses and identifying outliers from the linear trend (as evidenced by high standardized residuals). First, we conducted a linear regression of d_N between one duplicate and to the reference sequence (the other species) against d_N between the other duplicate and the reference sequence. We refer to these values as d_{N1} and d_{N2} (Fig. 1). Because the order in which the two duplicates were compared to the reference was arbitrary, we used the absolute value of the standardized residual from this regression as an indicator of cases where the absolute difference between d_{N1} and d_{N2} was unusually large and thus there was asymmetry in the nonsynonymous rate. The mean absolute value of the standard residuals from the regression of d_{N1} vs. d_{N2} was significantly higher in cases where Tajima's test showed significant rate asymmetry than in cases where Tajima's test showed no rate asymmetry (Table 1). Thus, gene pairs showing high absolute values of the standard residuals from the regression of d_{N1} vs. d_{N2} seemed good candidates for functional divergence between duplicates. There were 14 cases (9 in human 5 in mouse) with standard residuals ≥2.0 in absolute value (Table 2).

In addition, we conducted regression of d_{N1}/d_{S1} vs. d_{N2}/d_{S2} and examined the absolute values of the standard residuals. In this analysis also, the mean absolute value of the standard residual was higher in cases where the Tajima test showed significant rate asymmetry than in cases where that test did not show significant rate asymmetry (Table 1).

In order to determine which factors were most associated with significant evidence of rate asymmetry in Tajima's (1993) test, we computed partial correlations between each of the variables whose means are summarized in Table 1 and the chi-square statistic for Tajima's test (Table 3). Only two variables showed significant partial correlations with the chi-square statistic when controlling simultaneously for all other variables: the number of codons in the sequence and the absolute value of the standard residual from the regression of d_{N1} vs. d_{N2} (Table 3). The significant partial correlation in the case of the latter variable showed that this variable is associated with a significant result of Tajima's (1993) test independent of the increase in power of that test as a function of increased sequence length and increased sequence divergence.

Synonymous and Nonsynonymous Substitutions

Figure 2 illustrates the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) in comparisons between duplicate pairs. Overall, mean d_S (0.139 ± 0.011 SE) was significantly greater than mean d_N (0.073 ± 0.006) (paired *t*-test; $p < 0.001$). However, in 36 (11.4%) of 316 gene pairs, d_N exceeded *ds*, while in 14 gene pairs (4.4%) d_N and d_S were equal. Only in two cases did d_N exceed d_S significantly (at the 5% level) by the widely used *z*-test. In both of these cases no synonymous substitutions were observed. The two gene pairs involved were the following: (1) a gene pair encoding

Table 3. Partial correlations between selected variables and the chi-square statistic for Tajima's (1993) relative rate test, simultaneously controlling for all other variables

	Partial correlation	p
No. codons	0.461	< 0.000001
d_N between pair members	0.018	n.s.
d_S between pair members	−0.020	n.s.
Absolute value of standard residual from regression of d_{N1} vs. d_{N2} (comparison with reference sequence)	0.312	< 0.000001
Absolute value of standard residual from regression of d_{N1}/d_{S1} vs. d_{N2}/d_{S2} (comparison with reference sequence)	0.129	n.s.
Standard residual from regression of d_N vs. d_S comparison between pair members)	0.017	n.s.

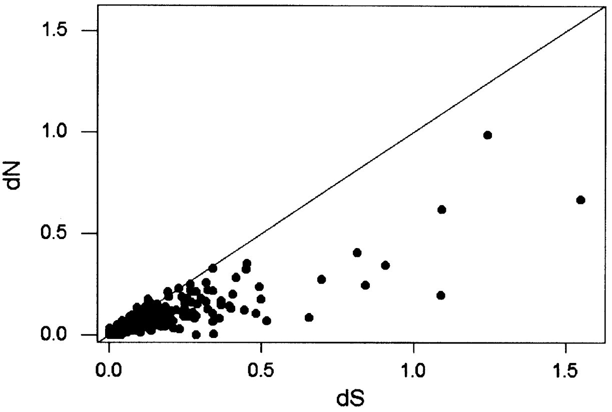

Fig. 2. Plot of d_N vs. $< d_S$ in the comparison of duplicate gene pairs. The line is a 45° line. The linear regression line was $Y = 0.0115 + 0.443X$ ($R^2 = 0.722, p < 0.001$).

proteins of unknown function (Ensembl I.D. ENSP00000295653 and ENSP00000303012) with $d_N = 0.0077 \pm 0.0031$ and (2) two genes encoding proteins related to pro-melanin-concentrating hormone (Ensembl I.D. ENSP00000323682 and ENSP00000295326) with $d_N = 0.0306 \pm 0.0126$. Because of the lack of synonymous substitutions, both of these cases evidently represented very recent duplicates. Also, it is worth pointing out that if the z-test results were corrected for multiple tests, the z-test would no longer be significant in either of these cases.

In order to identify other cases with unusually high d_N with respect to d_S, we searched for large positive standard residuals from the regression of d_N vs. d_S (Fig. 2). There were 10 gene pairs (5 from human, 5 from mouse) in which the standard residual was ≥2.0 (Table 4). In two of these pairs, d_N was greater than d_S although not significantly so by the z-test (Table 4). Furthermore, six of these gene pairs also showed unusually high absolute values of the standard residual from the regression of d_{N1} vs. d_{N2} (Tables 2 and 4). These six gene pairs thus show both unusually high d_N relative to d_S, as indicated by high standard residuals (Table 4), and strong nonsynonymous rate asymmetry. As a consequence, these six gene pairs seem to be good initial candidates for duplicates that have diverged functionally.

Discussion

Relative rate tests (Wu and Li 1985; Tajima 1993; Takezaki et al. 1995) have been widely used to test the hypothesis of equality of the rate of molecular evolution between sequences (or groups of sequences) by comparison to an outgroup or reference. Such tests typically rely on the assumption that all sites evolve independently. As a consequence, it is expected that, as the number of sites examined increases, the power of the test to detect rate asymmetry will increase. However, such small rate differences may actually results from stochastic error and thus may not be biologically meaningful. Consistent with the theoretical prediction that the power of these tests increases as the number of sites examined increases, we found that the number of sites examined was a good predictor of statistical significance in Tajima's (1993) test, even when a very conservative correction for multiple testing was applied (Tables 1 and 3).

In studies whose goal is to identify cases where duplicate gene pairs may have diverged as a result of adaptation to distinct functions, it seems preferable to use approaches that take into account the variance across gene pairs. In the present paper, we analyzed data on recently duplicated gene pairs in mammals using robust approaches based on linear regression. We show that these approaches can be used to search for gene pairs whose divergence at the amino acid level is unusually high in comparison to others in the data set.

One of these approaches is based on the absolute values of the standard residuals from the regression of d_N in the comparison of one duplicate gene with the reference sequence against d_N in the comparison

Table 4. Duplicate gene pairs with large (≥2.0) values of the standard residual from the regression d_N vs. d_S

Species	Protein function	d_N ± SE	d_S ± SE	St.resid.	Ensembl ID
Human	TNF receptor superfamily	0.351 ± 0.027	0.452 ± 0.059	2.72	ENSP00000221132, ENSP00000310263
	dUTP pyrophosphatase[a]	0323 ± 0.035	0.450 ± 0.076	2.19	ENSP00000249783, ENSP00000332786
	Rieske iron–sulfur polypeptide	0.255 ± 0.027	0.317 ± 0.054	2.02	ENSP00000306397, ENSP00000332578
	Unknown (homology to spindle pole body protein)[a]	0.213 ± 0.019	0.194 ± 0.028	2.25	ENSP00000311732, ENSP00000330408
	Phospholipase[a]	0.247 ± 0.018	0.265 ± 0.036	2.31	ENSP00000322900, ENSP00000332937
Mouse	CD59	0.174 ± 0.39	0.126 ± 0.056	2.09	ENSMUSP00000048041, ENMUSP00000052773
	GTPase effector[a]	0.986 ± 0.096	1.246 ± 0.261	8.73	ENSMUSP00000042926, ENSMUSP00000053726
	Unknown	0.323 ± 0.049	0.340 ± 0.092	3.24	ENSMUSP0000033146, ENSMUSP00000033147
	Secreted transmembrane protein 1[a]	0.618 ± 0.054	1.095 ± 0.025	2.47	ENSMUSP00000052018, ENSMUSP00000038020
	WW domain binding protein[a]	0.227 ± 0.021	0.228 ± 0.032	2.23	ENSMUSP00000032340, ENSMUSP00000056886

[a]Indicates gene pairs with a high standard residual from the regression of d_{N1} vs. d_{N2}.

of the other duplicate gene with the reference sequence. Partial correlation analysis showed that the absolute value of the standard residuals from this regression was significantly correlated with the chi-square statistic in Tajima's (1993) test, independent of the effect of sequence length (Table 3). An additional approach was based on the standard residuals from the regression of d_N vs. d_S between the two duplicated genes. Interestingly, five gene pairs were identified by both of these methods (Tables 2 and 4). These gene pairs seem the best candidates in our data set for functional diverged duplicate gene pairs.

Computation of d_S and d_N over the entire coding region of a gene can rarely provide a meaningful test of the hypothesis of positive Darwinian selection, because such selection typically acts only on a limited region involved in the function that is under selection (Hughes 1999a). In the present data set, only two gene pairs showed a significantly greater value of d_N than d_S for the entire gene by the commonly used *z*-test However, in both of these cases, no synonymous substitutions were observed, and d_N was quite low. These cases may represent positive selection that occurred soon after gene duplication. On the other hand, the difference between d_N than d_S may be due to stochastic error. Additional information on the structure of the proteins encoded by these gene will be needed to definitively rule out the latter possibility.

Of the 10 cases in which the standard residuals from the regression of d_N against d_S were unusually large, d_N exceeded d_S in only 2, and in neither of these cases was the difference significant by the *z*-test (Table 4). On the other hand, these 10 cases were identified by a method that takes into account the variance in d_S and d_N over the entire data set. Such cases may actually be at least as plausible candidates for positive selection as the two cases in which d_N of exceeded d_S significantly by the *z*-test. In searching for cases of adaptive evolution at the molecular level, it may be preferable to employ a two-step procedure: (1) using regression of d_N vs. d_S, identify cases with an unusually high d_N for a given d_S; (2) using structural information, identify functionally important regions of these molecules and compute d_N and d_S separately in each region. In the present data set, this approach identified six duplicate gene pairs as good candidates for functional divergence. Detailed structural information was lacking for these six duplicate pairs, but further application of this approach may uncover candidates with known structure and may eventually inspire structural studies on genes whose duplication and functional divergence may have played an important role in the evolution of biological processes.

Acknowledgments. This research was supported by NIH Grant GM66710. We are grateful for comments on the manuscript by Vaishali Katju and two anonymous reviewers.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI- BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E. (2003) Ensembl 2002: Accommodating comparative genomics. Nucleic Acids Res 31:38–42

Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome Res 13:2052–2058

Friedman R, Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol 20:154–161

Hill RE, Hastie ND (1987) Accelerated evolution in the reactive center regions of serine protease inhibitors. Nature 326:96–99

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B 256:119–124

Hughes AL (1999a) Adaptive evolution of genes and genomes. Oxford University Press, New York

Hughes AL (1999b) Evolutionary diversification of the mammalian defensins. Cell Mol Life Sci 56:94–103

Hughes AL (2002) Evolution of the human killer cell inhibitory receptor family. Mol Phylogenet Evol 25:330–340

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at MHC class I genes reveals overdominant selection. Nature 335:167–170

Hughes AL, Green JA, Garbayo JM, Roberts RM (2000) Adaptive diversification within a large family of recently duplicated, placentally expressed genes. Proc Natl Acad Sci USA 97:3319–3327

Katju V, Lynch M (2003) The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165:1793–1803

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:Research 0009.1–0008.9

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Li W-H (1982) Evolutionary change of duplicate genes. Isozymes 6:55–92

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804

Nei M (1969) Gene duplication and nucleotide substitution in evolution. Nature 221:40–42

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin

Paulding CA, Ruvolo M, Haber DA (2003) The *Tre2 (USP6)* oncogene is a hominid-specific gene. Proc Natl Acad Sci USA 100:2507–2511

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:559–607

Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823–833

Tanaka T, Nei M (1989) Positive selection observed at the variable-region genes of immunoglobulin. Mol Biol Evol 6:447–459

Thompson JD, Higgins DG, Gibson T (1994) CLUSTALW: 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

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43

Zhang P, Gu Z, Li W-H (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56