Evaluation of Whether Accelerated Protein Evolution in Chordates Has Occurred before, after, or Simultaneously with Gene Duplication

Catrióna R. Johnston, Colm O'Dushlaine,† David A. Fitzpatrick,* Richard J. Edwards,* and Denis C. Shields**

*UCD Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland; and †Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland

Gene duplication and loss are predicted to be at least of the order of the substitution rate and are key contributors to the development of novel gene function and overall genome evolution. Although it has been established that proteins evolve more rapidly after gene duplication, we were interested in testing to what extent this reflects causation or association. Therefore, we investigated the rate of evolution prior to gene duplication in chordates. Two patterns emerged; firstly, branches, which are both preceded by a duplication and followed by a duplication, display an elevated rate of amino acid replacement. This is reflected in the ratio of nonsynonymous to synonymous substitution (mean nonsynonymous to synonymous nucleotide substitution rate ratio [Ka:Ks]) of 0.44 compared with branches preceded by and followed by a speciation (mean Ka:Ks of 0.23). The observed patterns suggest that there can be simultaneous alteration in the selection pressures on both gene duplication and amino acid replacement, which may be consistent with cooccurring increases in positive selection, or alternatively with concurrent relaxation of purifying selection. The pattern is largely, but perhaps not completely, explained by the existence of certain families that have elevated rates of both gene duplication and amino acid replacement. Secondly, we observed accelerated amino acid replacement prior to duplication (mean Ka:Ks for postspeciation preduplication branches was 0.27). In some cases, this could reflect adaptive changes in protein function precipitating a gene duplication event. In conclusion, the circumstances surrounding the birth of new proteins may frequently involve a simultaneous change in selection pressures on both gene-copy number and amino acid replacement. More precise modeling of the relative importance of preduplication, postduplication, and simultaneous amino acid replacement will require larger and denser genomic data sets from multiple species, allowing simultaneous estimation of lineage-specific fluctuations in mutation rates and adaptive constraints.

Introduction

Gene duplication is instrumental in generating new genes encoding proteins with novel or altered function(s) (Ohno 1970). The rate of gene duplication and loss is high and may occur continually over evolutionary time (Lynch and Conery 2000). The maintenance of duplicated copies of genes relies on both copies contributing to fitness, and many initial duplicates are lost (Lynch and Conery 2000). Postduplication, the original genes' functions can be subdivided between the duplicated gene copies. This subfunctionalization (Serebrovsky 1938; Jensen 1976; Wistow and Piatigorsky 1987; Hughes 1994; Force et al. 1999; Stoltzfus 1999; Ward and Durrett 2004; He and Zhang 2005) does not require positive selection (generally indicated by a Ka:Ks > 1, where Ka is rate of nonsynonymous amino acid change and Ks is rate of synonymous change) to acquire subdivided roles and is more likely to occur when the original gene contains divisible functions or multiple *cis* motifs (He and Zhang 2006). Alternatively, 1 postduplication copy of the gene may maintain the original function and the other develop novel gene function (neofunctionalization) or indeed one copy may be lost (pseudogenization) (Eketjall et al. 2004; Suyama et al. 2006). Because evolving proteins are capable of gaining new functions while maintaining their original function (Aharoni et al. 2005), a period of selective pressure for duality of roles can be followed by gene duplication that then subdivides the different roles into the separate protein copies. This hypothesis implies there would be accelerated protein evolution prior to some gene duplication events along with the more generally acknowledged acceleration of protein evolution postduplication. An alternative scenario is that periodic increases in selection pressure may act at the same time on both gene-copy number and on amino acid replacement rates. Here, we characterize the pattern of pre- and postduplication change in chordate families in The Adaptive Evolution Database (TAED), to determine the evidence for either of these hypotheses. We chose the TAED chordate database as it represents a large collection of 15,452 gene families with multiple information held for each of these families, including multiple sequence alignments, phylogenetic trees, ancestral sequence data, as well as nonsynonymous to synonymous nucleotide substitution rate ratio (Ka:Ks) for each branch, all of which are important in order to attempt to investigate the events surrounding single-gene duplication events. In addition, we investigate a subset of this data set, human–rodent duplication events, in an attempt to homogenize the taxonomic groupings and evolutionary time lines to see if any trend is apparent surrounding rodent duplication events.

Materials and Methods

The TAED Database

We analyzed the chordate database from the TAED, from http://www.bioinfo.no/tools/TAED/files/taed_chordata_138.tar.gz (Liberles et al. 2001; Roth et al. 2005). The database includes over 15,452 chordate protein alignments. The methods involved in generating this database have been presented elsewhere (Liberles et al. 2001; Roth et al. 2005) but are summarized below for convenience:

GenBank 138 protein-encoding gene sequences greater than 10 amino acids (excluding annotated pseudogenes) were grouped into families, based on a Blast (Altschul et al.

Key words: genome evolution, gene duplication, positive selection.

E-mail: denis.shields@ucd.ie.

Mol. Biol. Evol. 24(1):315–323. 2007
doi:10.1093/molbev/msl162
Advance Access publication October 25, 2006

1990) *E*-value cutoff of 1.0 and global point accepted mutation (PAM) distances (Gonnet et al. 2000) of 100 or less. Multiple sequence alignments were calculated using partial order graphs (POA) (Grasso and Lee 2004) with the Blosum 80 substitution matrix, requiring each family member to be similar to all others (PAM $\leq$ 70) over at least 85% of their sequence. Majority-rule consensus protein phylogenies were estimated with MrBayes (Huelsenbeck and Ronquist 2001). Trees were simultaneously rooted and mapped onto the National Center for Biotechnology Information (NCBI) taxonomy using a soft parsimony approach (Steffansson 2004). Nodes with low posterior probabilities of less than 0.7 that conflicted with the NCBI taxonomy were corrected according to the NCBI taxonomy (Roth et al. 2005). NCBI taxonomy is held to be good for higher eukaryotes where the taxonomy is largely known, and the adaptive evolution database (TAED) data set in use in this paper consists of Chordates. Nodes that remained nonbinary were resolved using unweighted pair group method with arithmatic mean (UPGMA). For clades where all members come from the same species, only the most recent and complete (as defined by coding sequence [CDS]) sequence was kept, thus effectively pruning in-paralogues from the trees. The TAED trees used in this study are therefore drawn from trees with a broad range of multiple combinations of chordate species, where no specific species taxonomy is being included or excluded. Ancestral sequence predictions, Ka (rate of nonsynonymous change) and Ks (rate of synonymous change), were calculated for branches on an evolutionary tree between nodes using the methods described in these papers (Li et al. 1985; Li 1993; Pamilo and Bianchi 1993), with full treatment of probabilistic ancestral sequences (Benner et al. 1998), and normalized to exclude high Ka:Ks values where the high Ka:Ks is due to a single change. Ka:Ks values for each branch were calculated by maximum parsimony methodology (Fitch 1971). This method of Ka:Ks calculation excludes instances where positive selection on relatively few sites is overwhelmed by purifying selection on the rest of the sites in a given branch. In order to identify these sites, it would be necessary to use an alternative method that calculates the Ka:Ks values for individual sites—this work was considered outside the scope of this general study. For further information on TAED construction refer to the TAED papers (Liberles et al. 2001; Roth et al. 2005).

Branch Classification

In this study, we labeled all internal nodes in all phylogenies in the TAED chordate database as duplication or speciation nodes, based on the species classification of the terminal sequence downstream of the node. The branch support for all trees in the data set was of the order of $\geq$0.7 (Roth et al. 2005), and any instances where the taxonomy deviated from the NCBI taxonomy were corrected. We therefore feel our node assignments are reasonably accurate, given the current available sequence data; errors are likely to mainly reflect scenarios where paralogues are incorrectly identified, owing to incomplete sampling of sequences from a species, gene deletion, or incorrect annotation. Internal branches were classified as after speciation–before speciation (S–S), after speciation–before duplication (S–D), after duplication–before speciation (D–S), and after duplication–before duplication (D–D), depending on the given branch's flanking node labels. (Note: In our nomenclature, the branch in question is represented by a "–," and its flanking nodes are represented by a S or D for speciation and duplication, respectively; hence, D–S is a branch following a duplication and preceding a speciation.) Branches leading to terminal nodes were not considered in these classifications as they are more likely to have given rise to pseudogenes.

Branch Category Ka:Ks Comparison and Statistics

Taking the evolutionary branches within a range of synonymous site change from 0.0 to 0.155 (the cutoff of 0.155 was chosen as this represented a natural split in the data [data not shown]). The data was analyzed for the frequency branches at different values of Ks. It was observed that there was a separation of the data into 4 peaks with sufficient numbers for analysis. The peaks for the 2 higher Ks values (above 0.155) were broader than those for the lower 2, thus including branches that were too long for this analysis; hence, the range of the first 2 makes up our data set. The fraction of each branch category at incremental Ka:Ks was plotted (fig. 1*a*) as well as cumulative Ka:Ks plotted at incremental Ka:Ks (fig. 1*b*) for all 4 branch categories. The number of branches in each category are as follows: S–S has 5695 branches, S–D has 498 branches, D–S has 1147 branches, and D–D has 774 branches. Given the nonnormal distribution of Ka:Ks values, the Mann–Whitney test was used to assess the statistical significance of the difference in Ka:Ks values among each grouping compared with the S–S group (table 1). As an overall test of the difference in Ka:Ks distributions, the Kolmogorov–Smirnov test was applied (table 1). All statistical tests were performed using the Stata 8.0 software package (StataCorp, Station Road, Texas).

Family-specific weighted mean Ka:Ks values were calculated for each branch category (weighted on the inverse of the variance of the Ka:Ks), and their respective S–D, D–S, and D–D values were plotted against their family-specific S–S counterparts. Families that lack S–S branches were excluded from this analysis (fig. 2).

Preduplication Analysis

We took the 117 branches in 92 protein families, where $\mathrm{Ka:Ks}_{(\mathrm{S-D})} > 0.4$, and we categorized residues in the preduplication branch as those that had been changed and those that had not. We chose 0.4 as our cutoff because in figure 1*a* this is the approximate point where we observe a shift of S–D Ka:Ks away from that of S–S. We examined what happened to these residues at other branches after the duplication event and also in the rest of the tree (excluding the preduplication branch from the analysis [supplementary fig. 1a, Supplementary Material online]). Preduplication-changed sites were only used where both TAED (Liberles et al. 2001; Roth et al. 2005) and Gapped Ancestral Sequence Prediction for proteins (Edwards and Shields 2004) unambiguously agreed on the ancestral state and the ancestral changes. The simplest approach to assessing

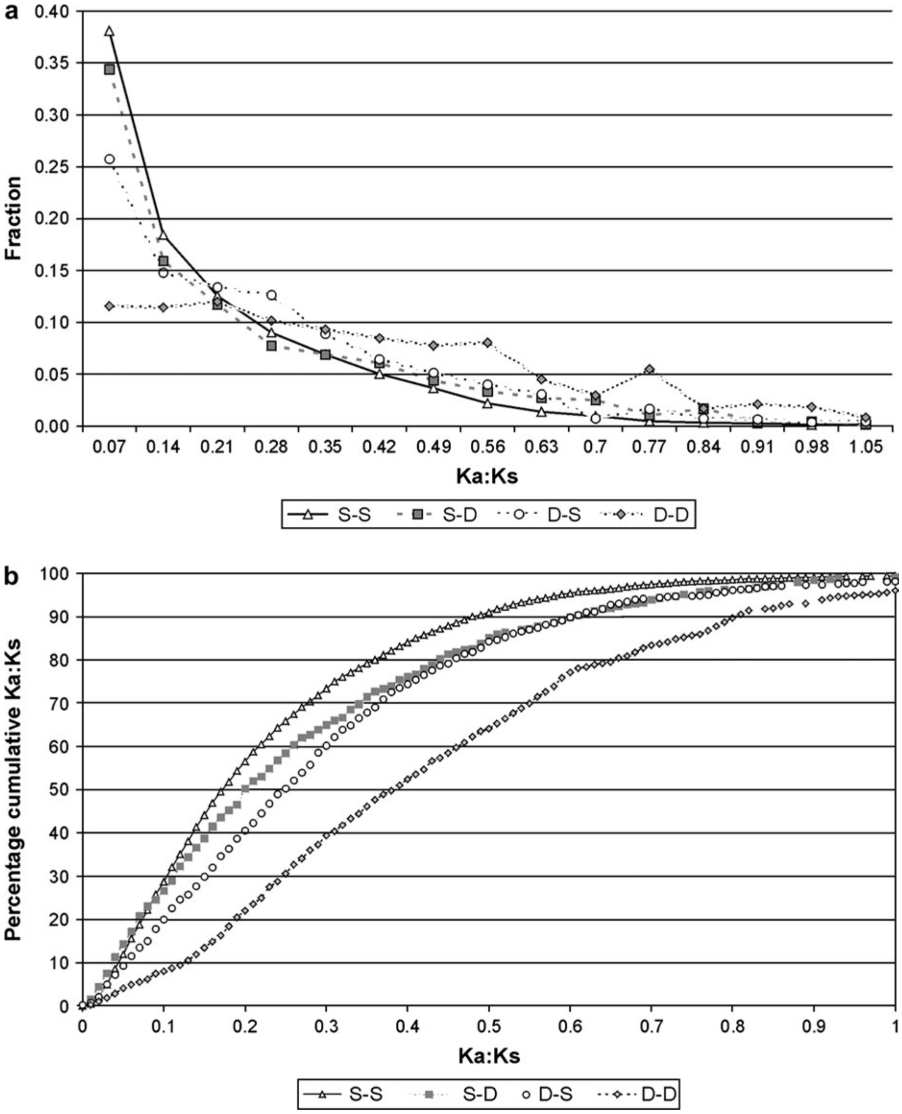

FIG. 1.—Acceleration of amino acid change prior to gene duplication. (*a*) Ka:Ks Frequency plot of after speciation–before speciation (S–S), after speciation–before duplication (S–D), after duplication–before speciation (D–S), and after duplication–before duplication (D–D) branches within a Ks range of 0.0–0.155 (significance of difference see table 1). Ka: rate of nonsynonymous (amino acid altering) DNA change. Ks: rate of synonymous DNA change. (*b*) Percentage cumulative Ka:Ks plot for after speciation–before speciation (S–S), after speciation–before duplication (S–D), after duplication–before speciation (D–S), and after duplication–before duplication (D–D) branches within a Ks range of 0.0–0.155.

statistical significance was to evaluate the odds ratio (OR) by logistic regression. The OR was calculated from the ratio of postduplication changes in all subsequent descendent branches out of all changes at those residues across the rest of the tree, with the ratio seen at unchanged residues (supplementary fig. 1a, Supplementary Material online). Statistical significance of the OR was estimated using Stata 8.0 statistical software (StataCorp, Station Road, Texas). To visualize the data, we plotted the $\log_{10}$ of the change ratio versus the $\log_{10}$ of the no-change ratio as well as the fraction of the significant families to nonsignificant families at varying ORs (supplementary fig. 1b, Supplementary Material online and fig. 3).

Phylogenies for the 92 well-aligned protein families that contained a preduplication branch (S–D) with a Ka:Ks $>$ 0.4 were further investigated. Bayesian phylogenies derived from their TAED alignment were inferred using MrBayes (Huelsenbeck and Ronquist 2001). Using the Jones amino acid matrix, 4 chains were calculated for 1 million generations. After a burn-in of 250,000 generations, a majority-rule consensus tree was calculated from trees sampled every 100 generations. To assess whether differences in topology between the soft parsimony approach and Bayesian trees are no greater than expected by random chance, we performed Shimodaira–Hasegawa tests (Shimodaira and Hasegawa 1999) using Tree-Puzzle 5.1 (Schmidt et al. 2002). If the Bayesian phylogeny was found to fit the underlying alignment significantly better than the soft parsimony tree, we removed the protein family from our analysis. Further, families were removed where there were in-paralogues (caused by species name ambiguity in GenBank, e.g., *Mus* sp. etc.), the preduplication branch

Table 1
Significance Tests of Differences in Ka:Ks Distribution Compared with Branches Flanked by Speciation Nodes (S–S)

	Significance *P* Value		
Branch category	D–D	D–S	S–D
Mann–Whitney full data set	<0.00009	<0.00009	0.0006
Kolmogorov–Smirnov full data set	<0.0009	<0.0009	<0.0009
Mann–Whitney Ka:Ks > 0.4	<0.00009	0.0004	0.0117
Kolmogorov–Smirnov Ka:Ks > 0.4	<0.0009	0.001	0.022

NOTE.—D–D = after duplication–before duplication branches, D–S = after duplication–before speciation branches, and S–D = after speciation–before duplication branches.

was near the root of the tree, and the Ka:Ks for the preduplication branch was greater than the mean of the rest of the branches in the tree (i.e., the preduplication effect is not explicable by general high Ka:Ks values for the protein family in question). We analyzed the remaining 9 protein families for site changes prior to duplication (table 2).

Human–Rodent Taxonomy Analysis

We performed a subanalysis on the 31 trees containing an outgroup followed by a human speciation node, followed by a rodent duplication prior to the speciation of rat and mouse (with duplicated paralogues observed in both rat and mouse, or only observed in 1 of the species) (fig. 4). The point of human–rodent common ancestor and all branches down from the human–rodent common ancestor were then analyzed. Each branch was categorized by its nodal definitions as described above; this time terminal branches were included but segregated into 2 types, terminal human branches (S human) and terminal rodent branches (S rodent), whose Ka:Ks values were compared with the internal D–S and S–D branches. The data set is too small for reliable statistical testing, and we simply visualized the data (supplementary fig. 2, Supplementary Material online).

Results and Discussion

Evolutionary Change Both before and after Duplication

To detect effects occurring close to the time of the duplication, we restricted analysis to evolutionary branches with a modest degree of synonymous site change (removing branches where Ks > 0.155). This also avoids problems with saturation for Ks seen over more extended evolutionary periods (see Materials and Methods for branch data set details). Branch classes were compared with branches flanked by speciation nodes (S–S). These S–S branches are isolated from duplication events and are therefore the best estimate of the evolutionary rate in the absence of duplication events. The node classification process may include some error (e.g., where a long branch may include a duplication event that was subsequently lost and thus be defined as a S–S branch). However, TAED phylogenies are based on the NCBI taxonomy, and thus the majority of phylogenies will not include taxonomic errors, and the large number of phylogenies in this analysis should occlude those with inaccuracies giving a reasonable portrayal of the overall pattern. Biases may arise because of potential differences in sampling of different branch categories. For example, branches bounded by duplication events would tend to be shorter. To check the scale of any such effects, we compared the difference in mean Ks among categories (for those with Ks < 0.155). The difference between the mean D–D Ks and the mean S–S Ks was 0.0122. The same differences between D–S and S–S and S–D and S–S branches were 0.011 and 0.001, respectively. This indicates that there is not a major difference in branch lengths between the different branch categories.

In branches following a duplication event and preceding a speciation event (D–S) as well as branches both preceding and following a duplication event (D–D), we observed, using the Mann–Whitney (testing for differences between nonnormally distributed data sets) and the Kolmogorov–Smirnov (testing for random fluctuations explaining the difference between the data sets) statistical tests, a significant excess of change compared with speciation-flanked branches (S–S). This significance was true for both the entire data set as well as for branches in these categories, with a Ka:Ks greater than 0.4 (fig. 1 and table 1). Figure 1*b* shows smooth curves representing the percentage cumulative Ka:Ks values for both D–S and D–D. If the excess of change were due to single families having large shifts in Ka:Ks, this curve would be less smooth.

In addition, we observed a significant excess of change in branches preceding a duplication event (S–D) compared with speciation branches (S–S) (fig. 1 and table 1). This is a smaller but significant difference. In this study, the preduplication mean $\mathrm{Ka{:}Ks}_{(S-D)}$ was 0.27 compared with the $\mathrm{Ka{:}Ks}_{(S-S)}$ prespeciation mean of 0.23. In our data set, $\mathrm{Ka{:}Ks}_{(D-S)}$ postduplication mean was 0.30 ($P < 0.00009$). The scale of this postduplication effect (D–S) is smaller than the previously observed Ka acceleration after gene duplication (Lynch and Conery 2000; Kondrashov et al. 2002; Conant and Wagner 2003; Zhang et al. 2003). Evolutionary depth is unlikely to be confounding this effect as the same trend is seen at varying nodal depths (supplementary fig. 3, Supplementary Material online). However, the most striking acceleration in amino acid replacement is seen among those branches, which are preceded by and followed by a gene duplication, with mean $\mathrm{Ka{:}Ks}_{D-D}$ being 0.44. We repeated the analysis excluding very short branches (removing branches where Ks < 0.055) because many short-lived duplications can be lost as pseudogenes and have much less impact on long-term evolution of function (Lynch and Conery 2000). The exclusion of the very short branches had little effect on the results (e.g., the least significant difference was for the Mann–Whitney test for greater Ka:Ks in S–D branches vs. S–S branches, $P = 0.005$).

The very high frequency of change in branches that are both preceded and followed by gene duplication (D–D) (fig. 1) suggests a very strong association of gene duplication and amino acid replacement in genes with a reasonable half-life (Ks > 0.055). Mean $\mathrm{Ka{:}Ks}_{D-D}$ is much higher (0.44) than that observed for postduplication branches preceded by speciation (0.3) ($P < 0.0000$). This association may

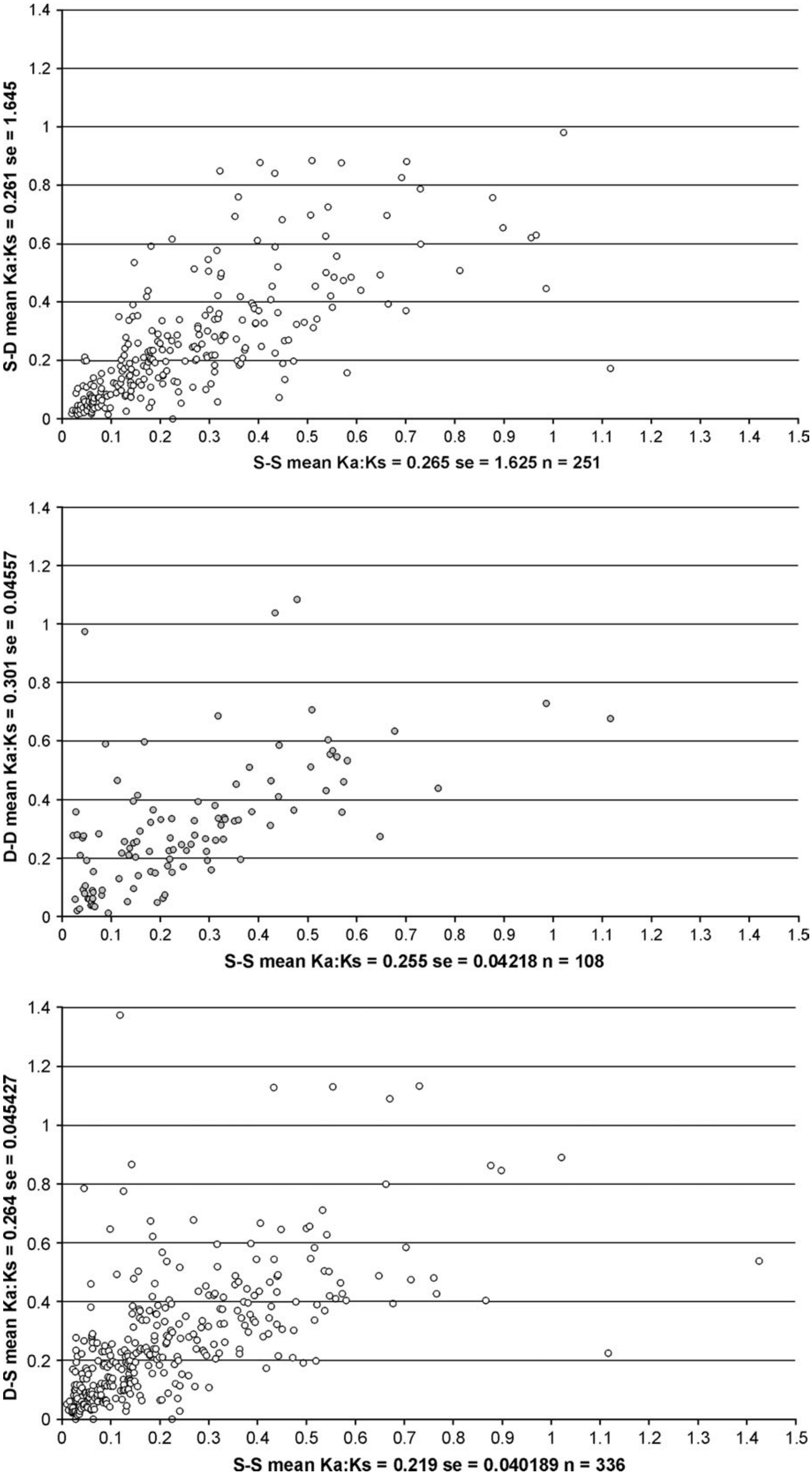

FIG. 2.—Individual family comparisons of Ka:Ks for S–S versus S–D, D–S, and D–D. The weighted mean Ka:Ks for each branch category in a given protein family was calculated for all families in the data set. Some families may be missing branch categories. For a given family the weighted mean Ka:Ks for the S–S branches were plotted on the *x* axis and those for the given families other branch categories were plotted on the *y* axis (individual plots for each category). The weighted means were calculated weighting on the inverse of the variance of each Ka:Ks. The number of families (n), the mean overall Ka:Ks across families and its standard error (se) for each axis in each plot is given beside the plot. Families in the given categories where there were no S–S branches were excluded from these plots.

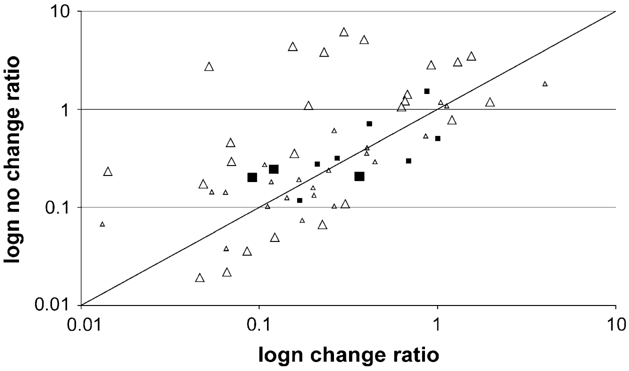

FIG. 3.—Change ratio versus no-change ratio for each of the 92 protein families. OR significant families are represented by larger squares or triangles and nonsignificant families are represented by small squares or triangles. Black squares represent the 9 instances of preduplication shown in table 2, white triangles represent other families. The line represents the neutral expectation: families above the line have more conservation of sites changed prior to duplication, whereas families below the line contain more change at these sites. (GO category information on this 92 protein family data set can be seen in supplementary table 1, Supplementary Material online).

be specific to certain evolutionary time points or may be a more general feature of particular protein families, that is, proteins which accumulate more duplications may have a higher rate of amino acid replacement. To explore this, we directly compared branch categories in D–D–containing families and found that a surprisingly large proportion of D–D branch–containing families lack S–S branches (~75%). Figure 2 shows that the mean D–D Ka:Ks is 0.046 greater than that of S–S branches. This is a very similar increase to that seen for D–S compared with S–S branches (0.045). We conclude that the initially observed high Ka:Ks in D–D branches largely reflects an increase in certain families of both duplications and accelerated protein evolution.

Evolutionary Branches with an Excess of Change prior to Gene Duplication

To identify the families contributing to the observed significant preduplication trend (fig. 1), we investigated residues that had been replaced prior to the duplication, at other branches in the tree in the 92 protein families where $Ka{:}Ks_{(S–D)} > 0.4$ for the preduplication branch. This revealed 23 protein families with significantly ($P < 0.05$) less change postduplication, consistent with a model of evolution of novel secondary function prior to duplication that is then maintained as an advantageous subfunctionalization (fig. 3). A smaller number of families showed a significant excess of change postduplication, suggesting that these sites were released of constraints or were undergoing subsequent positive selection (fig. 3). Although periodic fluctuations in the rate of gene evolution (Philippe et al. 2003) can contribute to departures from expectation, these patterns are suggestive of 2 alternative modes of evolutionary constraint.

A preduplication subset of 9 protein families with well-supported phylogenies and Ka:Ks greater than the mean of the rest of the tree (see Materials and Methods) is shown in table 2. A fucosyltransferase showed a number of instances of amino acid reversion to the state seen prior to the changes incurred in the preduplication branch in 1 descendant lineage (highlighted in table 2); there was no evidence for a gene conversion mechanism to explain this because the changes were distributed along the alignment, rather than in distinct clusters. Thus, 1 lineage has apparently reverted to an optimal ancestral function, whereas the other maintains the novel function acquired preduplication. In addition, both cathepsin and fertilin protein families had a significant excess of sites with no further change postduplication (highlighted in table 2), perhaps reflecting a maintenance of new functionality.

Amino acid changes for which there were 3 or more property changes among 10 physicochemical properties

Table 2
Selected Families with Apparent Excess of Change in Preduplication Branches

				Preduplication Changed Sites, Site Type		Proportion of Radical to Nonradical Changes		
TAED Family ID	Function	Preduplication Nonchanged Sites	Preduplication Changed Sites	No Further Change[a]	Descendant Reversion[b]	Pre–S–D[c]	S–D[d]	Post–S–D[e]
42003	Pancreatic ribonuclease	182	11	5	1	1.55	1.33	1.08
230007	Olfactory receptor	494	10	9	0	0.93	0.13	0.00
230120	Olfactory receptor	372	26	14	1	0.64	0.35	0.09
785007	Cathepsin	305	34[f]	1	2	0.70	1.80	1.18
977011	Interferon alpha 12	170	24	7	0	3.00	0.57	0.61
2474003	Myeloid antimicrobial peptide	279	14	5	0	2.71	0.50	3.00
2748002	Fertilin alpha	940	57[f]	7	7	0.59	4.00	1.89
3220009	Fucosyltransferase	419	37	19	9	2.08	0.65	3.63
13478000	Leukemia inhibitory factor	183	21[g]	1	0	0.08	0.15	0.63

NOTE.—The 3 highlighted families show an excess of radical changes following the duplication node, radical change being changes for which there were 3 or more property changes among 10 physicochemical properties (Zvelebil et al. 1987).

[a] Sites that are completely conserved after duplication.

[b] Sites where a descendant branch reverts to the ancestral state of the site prior to the preduplication change.

[c] All other branches in the phylogeny excluding S–D and post–S–D. Radical changes: 3 or more changes in the physiochemical properties of the amino acid.

[d] Branch preceding duplication, showing high Ka:Ks.

[e] All of the phylogenetic tree descendent from the duplication node (D) following the S–D branch.

[f] Significant deficit of change postduplication (see fig. 4).

[g] Significant excess of change postduplication (see fig. 4).

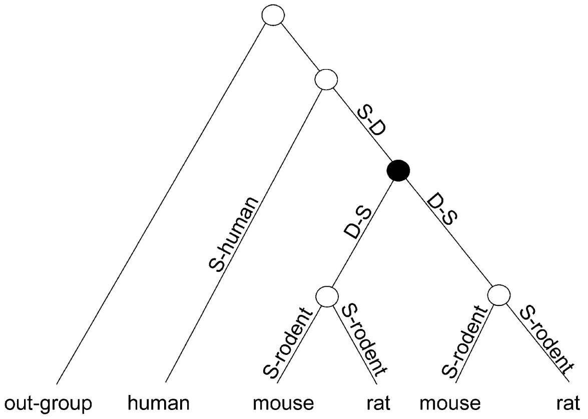

FIG. 4.—Example tree for human–rodent taxonomy analysis.

(Zvelebil et al. 1987) were classified as radical changes. We compared the radical and nonradical changes at and after the S–D branch, as well as in the rest of the tree, for each of the final 9 protein families suspected of preduplication activity (table 2). The main observation is that the ratio of radical to nonradical change appears to vary considerably in all 9 families, comparing the S–D branch to branches before and after, but that the particular pattern of variation is relatively specific to each family and presumably reflects separate evolutionary constraints on functional change at different stages. We observed an excess of radical to nonradical changes at the preduplication branch in 3 protein families (table 2: S–D column). This is consistent with cases such as immune response genes, which favor less conservative changes during positive selection (Hughes et al. 1990). However, in 6 of the 9 families, we also observe an excess of nonradical changes, consistent with the suggestion that some genes (male reproductive genes) can favor conservative changes during positive selection (Wyckoff et al. 2000). Because we observe a degree of both radical and nonradical change in all families, we cannot rule out that this observation could be due to either positive selection or a relaxation in selection. The fucosyltransferase, cathepsin, and fertilin protein families had an increase in the proportion of radical to nonradical changes after the S–D branch compared with the rest of the tree (exclusive of the S–D branch) (highlighted in table 2), which is consistent with further selection for functional change after the duplication. This is not a contradiction of the above data because these increased radical changes postduplication occur at sites other than the sites of no further change observed above for the cathepsin and fertilin families.

Sensitivity of Preduplication Analysis to Phylogenetic Subsets

Chordates are a deep clade, where nucleotide substitution rates can vary substantially among lineages. Branches with the same Ks in different parts of the tree could correspond to very different time frames. Because the species in different trees varies, there is the potential that the broad survey presented here is biased by overrepresentation of certain clades in the sampling of different types of branches. The overall analysis was carried out across a variety of different phylogenetic contexts, and we were interested to see what patterns emerge when a smaller, more specific, subset is used.

To investigate the effect of restricting the analysis to particular subclades, we looked at duplication events in rodents because of their divergence from humans (fig. 4), where the phylogeny appears reliable. The relevant comparison of pre- and postduplication branch Ka:Ks values are with the postspeciation branches in the tree, which lead to either a human or rodent terminal node (S human or S rodent, respectively). Although preduplication branches showed a modest excess (Ka:Ks = 0.145) compared with 0.131 for S human and 0.138 for S rodent (fig. 4), this difference was not statistically significant ($P = 0.4$, Mann–Whitney). Thus, the overall trend seen in figure 1*a* is suggested within a more controlled set of phylogenies, but a larger sample of trees will be needed to robustly confirm it. In addition, more accurate modeling of the rate of synonymous change is required for each evolutionary interval before any very strong evolutionary inference can be drawn. This will require much denser genome sequence sampling contributing to reconstruction of ancestral sequences flanking duplication events because rate accelerations in different lineages and subregions of trees can introduce systematic biases into the observed patterns. It could be argued that the complexity of the data would lead to the conclusion that the broad analysis we have performed across many phylogenetic scenarios is too contaminated with various biases to be reasonably interpretable. However, it is reasonable to assume that such biases may average out over a large number of trees and give a reasonable indication of the overall pattern. Therefore, it appears useful to present the overall patterns and highlight the main features (fig. 1) but to interpret these with caution. If the general trend seen for overall Ka:Ks acceleration prior to duplication is indeed borne out by other studies, it is likely there may be a large number of instances of particular sites undergoing selective change prior to duplication, which are not reflected in a more general change in Ka:Ks.

Conclusion

The existing interpretation of the association between gene duplication and acceleration in amino acid change focuses on possible increases in positive selection and reductions in purifying selection following duplication (duplication first model; fig. 5). However, in figure 1*a* and *b* rate acceleration in branches occurring both before and after duplication suggests an alternative model, where selection pressures for novel functions act on both the amino acid frequency and the gene-copy number simultaneously, making certain protein families more prone to accumulate duplicates (simultaneous model; fig. 5). Such a model of simultaneous pressures on both gene duplication and site mutations is perhaps not surprising, given that the mutation rate toward duplication and toward amino acid replacement are of the same order of magnitude (Lynch and Conery 2000). Related transient or persistent positive selection pressures may relate to sub- or neofunctionalization, or simply to gene dosage (Francino 2005). Maintenance of both

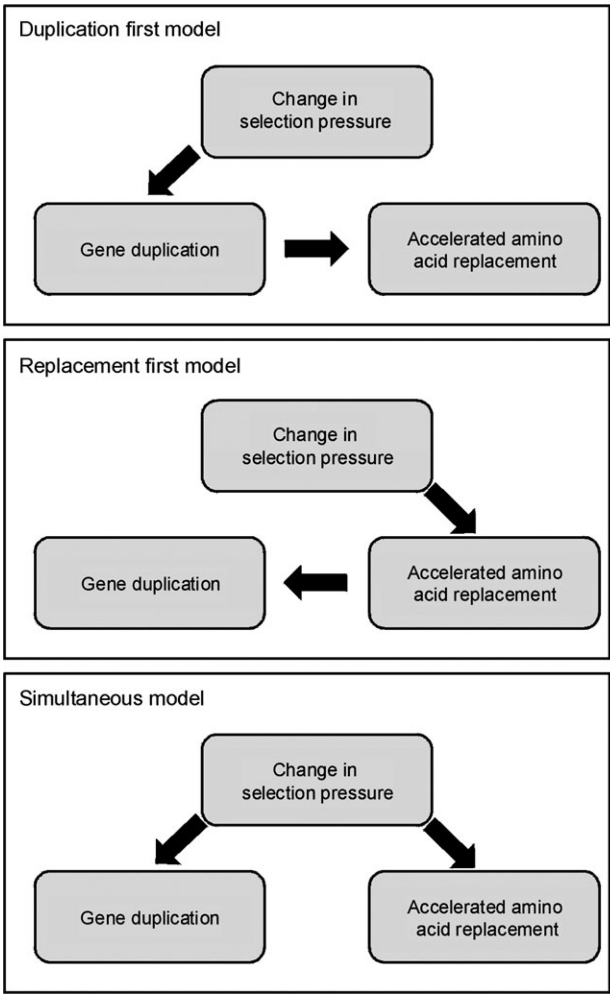
Duplication first model
Change in selection pressure
Gene duplication
Accelerated amino acid replacement
Replacement first model
Change in selection pressure
Gene duplication
Accelerated amino acid replacement
Simultaneous model
Change in selection pressure
Gene duplication
Accelerated amino acid replacement

FIG. 5.—Three alternative models relating observed patterns of duplication and amino acid replacement to underlying selection pressures.

copies over long evolutionary periods is likely to require subfunctionalization, but this may or may not be coupled with the transient selection pressure.

Against the background of such systematic correlations of Ka:Ks and gene duplication rates (fig. 5), we have observed a distinct but less common mode of evolution; in a subset of proteins (in 9 [table 2] out of 92 preduplication-containing families), accelerated change is seen prior to duplication. One model to explain this is that a multifunctionalizing selection pressure acts initially on a protein sequence, giving rise to accelerated change, which is then resolved to a state of higher fitness by the novel and ancestral functions segregating into separate duplicates (replacement first model; fig. 5). This tentatively supports the hypothesis put forward by Aharoni et al. (2005) from their analysis of experimental evolution of novel functions that gene duplication may be preceded by the acquisition of multiple functions.

The enrichment of mammalian gene histories with increasing genome sequences will allow a more formal quantification of the roles of altered selection pressures on amino acid rates arising from gene duplication (duplication first model; fig. 5), of altered selection pressures on amino acid sequences leading to subsequent gene duplication (replacement first model), and of transient selection pressures that may act on both independently for a period of time (simultaneous model).

Supplementary Material

Supplementary figures 1–3 and table 1 are available at http://bioinformatics.ucd.ie/shields/kjohnston/ and at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Science Foundation Ireland. We are grateful to M. Betts and D. Liberles for information and data on the TAED database and to M. Fares and A. McLysaght for comments.

Literature Cited

Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS. 2005. The 'evolvability' of promiscuous protein functions. Nat Genet. 37:73–76.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Benner SA, Trabesinger N, Schreiber D. 1998. Post-genomic science: converting primary structure into physiological function. Adv Enzyme Regul. 38:155–180.

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

Edwards RJ, Shields DC. 2004. GASP: gapped ancestral sequence prediction for proteins. BMC Bioinformatics. 5:123.

Eketjall S, Jornvall H, Lonnerberg P, Kobayashi S, Ibanez CF. 2004. Recent evolutionary origin within the primate lineage of two pseudogenes with similarity to members of the transforming growth factor-beta superfamily. Cell Mol Life Sci. 61:488–496.

Fitch W. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool. 20:406–416.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531–1545.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. Nat Genet. 37:573–577.

Gonnet GH, Hallett MT, Korostensky C, Bernardin L. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics. 16:101–103.

Grasso C, Lee C. 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 20:1546–1556.

He X, Zhang J. 2006. Higher duplicability of less important genes in yeast genomes. Mol Biol Evol. 23:144–151.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 169:1157–1164.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 256:119–124.

Hughes AL, Ota T, Nei M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol. 7:515–524.

Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol. 30:409–425.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biol. 3:RESEARCH0008.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2:150–174.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. 2001. The adaptive evolution database (TAED). Genome Biol. 2:RESEARCH0028.

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

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer Verlag.

Pamilo P, Bianchi NO. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol. 10:271–281.

Philippe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life. 55:257–265.

Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA. 2005. The adaptive evolution database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res. 33:D495–D497.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.

Serebrovsky A. 1938. Genes scute and achaete in *Drosophila melanogaster* and a hypothesis of gene divergency. C R Acad Sci URSS. 19:77–81.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

Steffansson P. 2004. Building consensus trees using gene sequences—a phylogenetic approach [master's thesis]. Sweden: Stockholm University.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. J Mol Evol. 49:169–181.

Suyama M, Harrington E, Bork P, Torrents D. 2006. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. PLoS Comput Biol. 2:e76.

Ward R, Durrett R. 2004. Subfunctionalization: how often does it occur? How long does it take? Theor Popul Biol. 66:93–100.

Wistow G, Piatigorsky J. 1987. Recruitment of enzymes as lens structural proteins. Science. 236:1554–1556.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature. 403:304–309.

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

Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol. 195:957–961.

Dr Billie Swalla, Associate Editor

Accepted October 17, 2006