Significantly Different Patterns of Amino Acid Replacement After Gene Duplication as Compared to After Speciation

Cathal Seoighe, Catrióna R. Johnston, and Denis C. Shields
Clinical Pharmacology Department, Royal College of Surgeons in Ireland, Dublin, Ireland

We have performed a large-scale analysis of amino acid sequence evolution after gene duplication by comparing evolution after gene duplication with evolution after speciation in over 1,800 phylogenetic trees constructed from manually curated alignments of protein domains downloaded from the PFAM database. The site-specific rate of evolution is significantly altered by gene duplication. A significant increase in the proportion of amino acid substitutions at constrained (slowly evolving) sites after duplication was observed. An increase in the proportion of replacements at normally constrained amino acid sites could result from relaxation of purifying selective pressure. However, the proportion of amino acid replacements involving radical changes in amino acid properties after duplication does not appear to be significantly increased by relaxed selective pressure. The increased proportion of replacements at constrained sites was observed over a relatively large range of protein change (up to 25% amino acid replacements per site). These findings have implications for our understanding of the nature of evolution after duplication and may help to shed light on the evolution of novel protein functions through gene duplication.

Introduction

The duplication of protein-coding genes can increase the number of different proteins encoded within a genome and has been proposed as a major source of evolutionary novelty and increased complexity (Ohno 1970; Hughes 1994; Li 1997). After gene duplication, one copy may be silenced by a deleterious mutation, and it will then degenerate and ultimately disappear through further mutations, in the absence of any selective constraint maintaining it within the genome. Alternatively, both copies may acquire mutually complementary degenerative mutations that require the maintenance of both copies (Force et al. 1999). A third possibility is that a novel function may evolve in one copy that increases the adaptive fitness of the genome relative to the original genome that only had one copy of the gene. This latter process is of most interest from the perspective of understanding how gene duplication contributes to the phenotypic evolution of novel functions. A survey of completed eukaryotic genomes by Lynch and Conery (2000) suggests that detectable gene duplication is fairly frequent in many organisms (0.01 per gene per Myr), but after gene duplication, there is a high rate of silencing (half-life of 4 Myr, after about 0.05 changes per silent site). Kondrashov et al. (2002) also confirmed the expectation that recently duplicated genes go through a period of low selective constraint on the protein sequence, since the ratio of replacement to silent DNA changes in recently duplicated paralogous pairs is greater than the equivalent ratio for orthologs. However, the authors pointed out that such data could be biased by multiple substitutions per site. Patterns of DNA change are therefore only likely to be informative about the period immediately after gene duplication. Since most recently diverged duplicates are likely to be lost, they are not representative of the duplicated genes that are maintained for long periods within the genome and which may well confer phenotypic differences and functional advantages. In this study, we investigated the pattern of protein sequence change after gene duplication over a wider range of evolutionary distance and contrasted it with the pattern in evolutionary branches that do not follow gene duplication. Previous specific studies of lysozyme (Stewart and Wilson 1987; Messier and Stewart 1997), haemoglobin (Braunitzer and Hiebl 1988), and integrins (Hughes 1992) are consistent with the theory that adaptation to novel functions gives rise to changes at otherwise constrained (slower-evolving) residues. Our large survey over many protein families supports this theory, suggesting that altered site-specific replacement rates are a general phenomenon after duplication.

Data and Methods

Data were derived from “seed” alignments of PFAM protein domain families (Bateman et al. 2000), downloaded from the Pfam Web site (http://www.sanger.ac.uk/Software/Pfam/) on August 17, 2000. Only “seed” alignments were used to maximize alignment quality. After a number of families for which it was not possible to perform ancestral reconstructions in a reasonable amount of time were omitted, the remaining 1,821 families were analyzed.

Phylogenetic Tree and Ancestral Node Reconstruction

We performed a large-scale comparison at the amino acid level between evolutionary patterns after duplication and speciation events. Phylogenetic trees were constructed automatically from 1,821 Pfam protein domain “seed” alignments using ClustalW (Thompson et al. 1997) software, which created neighbor-joining trees, allowing for multiple amino acid replacements. Trees were rooted using Retree from the phylip package (Felsenstein 1993). Because the trees were constructed and rooted automatically, some errors in topology and direction of evolution are expected. Whereas such errors introduce noise into the data set, they should not be biased with regard to definitions of speciation and duplication branches. Ancestral states of sequences were determined for each internal
node of each tree using software from the Paml (Yang 1997) package.

Defining Postduplication Branches

If you were to start out with a tree with only proteins from one organism, it would only have duplication nodes and branches. If you now introduce additional species, the effect would be like “randomly” introducing additional speciation nodes along these branches. Evolution after the speciation nodes should be the same as evolution at random points along the original duplication branches. The speciation nodes may be taken to reflect random samples from somewhere along the original duplication branches.

Each branch of the tree was classified as being after gene duplication or after speciation, and the nature of the evolutionary changes between the node sequences at the beginning and end of the branch was analyzed (fig. 1). Duplication nodes, and the subsequent “duplication branches,” were defined as all those nodes whose two descendent branches were ancestral to members of the same species. All other nodes were considered to be nonduplication, or “speciation” nodes. When the topology is correct, nodes arising through gene duplication should be identified correctly, although nodes may sometimes be incorrectly identified as speciation nodes when paralogs descended from them have not been identified or have been lost through deletion. Misclassification reduces the differences observed between postduplication and postspeciation branches, so that the true differences are probably greater than the observed difference. A total of 35,500 speciation and 14,377 duplication branches were assessed.

Radical Amino Acid Changes

Amino acid changes for which there were three or more property changes among 10 physicochemical properties (Zvelebil et al. 1987) were classified as radical. The ratio of radical to nonradical amino acid changes, rad, was measured and compared for duplication and speciation branches at different ranges of branch lengths.

Constrained Sites

Sequence positions for which there was evidence (at the 90% significance level) that the position was evolving at a slower rate than the sequence as a whole (i.e., \( r_i < 1 \) [see below]) were classified as constrained sites (on average 23% of sequence positions). Let \( d_c \) be the inferred number of replacements per constrained site and \( d \), the number of replacements per site. \( d_c \) and \( d \) are corrected separately for multiple hits. The ratio \( b = d_c / d \) gives an indication of the proportion of replacements at constrained sites.

The determination of whether a residue was significantly slower evolving than the sequence as a whole was based on the following approach. The probability of a given amino acid substitution at a given residue position in a branch is described by the PAM matrix (Dayhoff 1972) equivalent to \( L r_i \), where \( L \) is the actual PAM length of the branch and \( r_i \) is the relative rate of evolution at position \( i \) (PAM = 10 is equivalent to 10 replacements per 100 sites). The accuracy of the estimate of \( r_i \) depends on the number of branches, on the lengths of the branches, and on the number of residues in the protein. We generated a confidence interval for the rate \( r_i \) allowing for different amounts of information concerning \( r_i \) in branches of different lengths. From Bayes’s Theorem,

\[
P(r_i = r \mid B) \propto P(B \mid r_i = r)P(r_i = r),
\]

where \( P(B \mid r_i = r) \) is the likelihood of generating the observed amino acid substitutions at position \( i \) in the set of branches \( B \) if \( r_i = r \) and \( P(r_i = r) \) is the prior probability of the rate \( r \). The prior probability distribution describes the variation in the rate of evolution among sites. We use the gamma distribution (with \( a = b \) (Yang 1993; Durbin et al. 1998)) and use the reciprocal of the variance of the number of amino acid replacements at each position across the complete alignment for the parameter \( a \). Integrating this probability distribution provides estimates of the upper and lower bounds of \( r_i \), lying at the 10% and 90% points on the cumulative distribution. If a replacement has occurred at position \( i \), then the upper bound of \( r_i \) is used to produce a conservative estimate of the likelihood of the mutation. If no mutation has occurred at position \( i \), then the lower bound of \( r_i \) is used.

Branches with Unusually High Numbers of Constrained Site Changes

Assuming a binomial distribution for replacements at constrained and unconstrained sites on a branch, the probability of the observed or a greater number of replacements at constrained sites occurring on a branch was calculated as

\[
P = \sum_{i=0}^{n} x^i (1-x)^{n-i} \frac{n!}{i!(n-i)!},
\]

where \( x \) is the proportion of all replacements at all branches in the alignment that occur at constrained sites,
\( n \) is the number of replacements occurring in the branch, and \( r \) is the number of replacements occurring at constrained sites in the branch. Small values of this probability indicate that the rate of evolution at constrained sites exceeds expectation. The expression should be approximately correct, provided \( s \) and \( s_x \) are large compared with \( n \), where \( s \) is the total number of sites. Otherwise the above expression will tend to overestimate \( P \) and thus provide a conservative statistic.

Error bars for the figures were calculated by the bootstrap method (Efron 1979): the data sets were sampled randomly, with replacement, 1,000 times and 2.5% tails of the distribution were used to produce the error bars for figure 2.

**Results**

We tested whether the kinds of changes occurring after duplication tended more towards “radical” amino acid replacements between residues of very different physicochemical properties. The strongest effect was seen for branches of PAM 6 to 10, with a very slight excess of radical mutations in duplication branches (\( \text{rad} = 0.225 \) [see Data and Methods]) compared with postspeciation branches (\( \text{rad} = 0.219 \)), but this was not significant. Therefore, the occurrence of changes (see below), rather than the nature of the amino acid changes, may be the most important distinguishing feature of protein evolution after duplication.

We tested whether there was a greater tendency for replacements at constrained sites after duplication. “Constrained sites” were defined as those with a significantly slower than average rate of evolution. For a given branch, \( b \) is defined as the ratio of the number of replacements per constrained site to the number of replacements per site. Figure 2 compares the values of \( b \) for branches after gene duplication and for branches after speciation. Over a range of evolutionary distances, there are consistent and significant excesses of constrained site change after duplication. For example, looking at branches of PAM 6 to 10, there is approximately a 60% increase in \( b \) for postduplication branches. Thus, constrained sites are subject to greater change after gene duplication than after speciation nodes. Within vertebrate lineages alone, in the range \( 5 < \text{PAM} < 31 \), the value of \( b \) after duplication nodes was 0.23, significantly higher (\( P = 0.001 \)) than after nonduplication nodes (0.17). Thus, vertebrate protein evolution shows a similar excess of constrained site change after duplication.

![Figure 2](image-url)
On average, nodes preceding gene duplications occur earlier in the phylogenetic trees than speciation nodes. To test that this did not introduce bias, we compared postduplication and postspeciation branches in bins of evolutionary depth. Evolutionary depth was approximated as the average PAM length between the node and its descendant terminal nodes. Whereas the ratio of replacements at constrained sites increases with evolutionary depth, the difference between postduplication branches and postspeciation branches remained significant (fig. 3). The increase in constrained site change with evolutionary depth could reflect long-term selection for diversity of protein function or gradual changes in the set of residues that is constrained (Miyamoto and Fitch 1995).

The 50 protein families containing the postduplication branches with the lowest \( P \)-values (see Data and Methods) are shown in figure 4. A wide range of protein functions is represented in the 50 postduplication branches for which there was the greatest excess of constrained site changes, as indicated by the lowest probabilities (fig. 4a). The functional categories of the unusual vertebrate postduplication branches were not significantly different from those of postspeciation branches \( (P = 0.08, \text{ Fisher’s exact test}) \). Neither was there a significant difference between postduplication and postspeciation categories for the overall data set \( (P = 0.09) \) (fig. 4a). Thus, excess change at constrained sites after duplication is a general phenomenon, distributed across a broad functional range of proteins.

There are a number of technical limitations with the approach we adopted that future studies may redress with sets of larger, but reliable, alignments. Although alignment error is a source of noise, in our analysis the fact that we only compared branches of the same length probably prevents bias. It will be of interest to apply the method used here within a data set of DNA alignments, to allow direct comparison of synonymous and nonsynonymous rates of substitution for gene families for which large numbers of closely related sequences are available (Liberles et al. 2001).

**Discussion**

Gene duplication is a widespread phenomenon that occurs with high frequency in the genomes of most organisms (Lynch and Conery 2000). After gene duplication, functional redundancy may be resolved quickly through the rapid loss of one member of the duplicated pair. Alternatively, redundancy may be resolved through the evolution of novel functionality or specificity. It is not clear how rapidly the expression or function of duplicated genes diverges after duplication so that there is selective pressure to maintain both copies. Gene loss after
duplication typically occurs rapidly (half-life = 4 Myr [Lynch and Conery 2000]). Therefore, if novel functionality determines the fate of duplicated genes, the evolution of novel function must take place soon after duplication.

It is also possible that functional redundancy is maintained for a long time after duplication for a subset of duplicated genes (Nowak et al. 1997; Taverna and Goldstein 2000; Wagner 2000). Because these genes are maintained in duplicate over much longer times than most gene pairs, the opportunity for evolution of novel function in these genes is much greater than for most duplicated gene pairs. If this is the case, the establishment of long-term redundancy or near-redundancy may be frequent after gene duplication, and the raw material for the evolution of novel function may be long-term functionally redundant gene pairs rather than the more common short-lived duplicated gene pairs. Changes in amino acid sequences with divergence times far in excess of the half-life that has been estimated for duplicate genes (Lynch and Conery 2000) are thus likely to be relevant to the fate of duplicated genes.

Previous investigations (Lynch and Conery 2000; Kondrashov et al. 2002) have revealed an increase in the ratio of nonsynonymous nucleotide substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS) after gene duplication. Although many examples of the evolution of functional novelty after gene duplication have been found (Rosenberg et al. 1995; Zhang, Rosenberg, and Nei 1998; Zhang, Kimberly, and Rosenberg 2000), it is not clear whether the increase in the rate of nonsynonymous substitution more often reflects reduction in purifying selection, allowing an increased number of slightly deleterious mutations to be fixed. If the increase in dN reflects a reduction in the efficiency of purifying selection during a prolonged period in which functional redundancy is maintained, then the number of slightly deleterious replacements should increase. Maintenance of functional redundancy will require that replacements that completely alter or destroy the protein function continue to be selected against. As a result, we might not expect the increase in slightly deleterious replacements to be accompanied by an increase in replacements at highly constrained sites and the proportion of potentially function-altering replacements should not increase after duplication. We have found evidence that there is an increase in the proportion of amino acid replacements at amino acid sites that are normally constrained after duplication. We argue that amino acid replacements at constrained sites are indicative of changes in protein function and suggest that an increased level of functional change after gene duplication is a general phenomenon that completely alter or destroy the protein function should not increase after duplication. As a result, we might not expect the increase in slightly deleterious replacements to be accompanied by an increase in replacements at highly constrained sites and the proportion of potentially function-altering replacements should not increase after duplication. We have found evidence that there is an increase in the proportion of amino acid replacements at amino acid sites that are normally constrained after duplication. We argue that amino acid replacements at constrained sites are indicative of changes in protein function and suggest that an increased level of functional change after gene duplication is a general phenomenon that can be observed as an increased relative rate of replacement at constrained sites.

Constrained sites reflect the action of purifying selection against amino acid changes, which can alter the properties of a protein in such a way as to prevent it from carrying out its function effectively. We have interpreted the increase in the rate of replacement at constrained sites as evidence of greater functional change after duplication than after speciation. This interpretation is reasonable, provided that change in protein function is used in a broad sense to include the nonfunctioning, impaired function or reduced functioning of one or both of the gene duplicates as well as the evolution of a novel function. Whereas a change in the functioning of a duplicated gene is not in itself sufficient to prove that a novel protein function has evolved, over large timescales, the probability that both copies of the gene will be maintained will be increased for proteins that have evolved novel functions or specificities. The function change could also be temporary. For example, a change that reduces the fitness of one duplicate could allow a protein to move between distinct fitness peaks during a period of relaxed selection after duplication, followed by increased selection.

A change in protein function may, in some cases, be caused by a set of amino acid replacements at normally constrained sites. Shortly after duplication, the preduplication site-specific rates can be restored. This has been referred to as type II functional divergence (Gu 1999). Alternatively, a change in protein function may change the site-specific rates of replacement permanently, so that the constrained sites after the duplication are different from
the constrained sites before duplication. This kind of functional divergence has been termed type I (Gu 1999). Although the method we have used to analyze functional divergence after duplication is more suited to the detection of type II divergence, it should also be possible to detect divergence of type I in many cases. We have identified constrained amino acid sites based on the set of replacements that have been observed across an entire protein family. Type II functional divergence is associated with an increase in the proportion of amino acid replacements at sites that are constrained throughout the whole family. Type I divergence may alter the set of sites that are constrained. Provided the total length of branches after the duplication is significantly less than the remaining length of the tree, the constrained sites will be determined largely by remaining branches of the tree. An altered rate of replacement after duplication may then cause an elevation in the proportion of replacements at constrained sites.

Previous codon-based comparisons of non synonymous changes per non synonymous site (dN) with synonymous changes per synonymous site (dS) point to a period of relaxed selection immediately after gene duplication (Lynch and Conery 2000; Kondrashov et al. 2002). Our results show no increase in the proportion of replacements at constrained sites on short branches immediately after gene duplication. There are several possible explanations for this. The accuracy with which phylogenetic relationships involving short branches can be inferred is lower than for longer branches. This will impact on the accuracy with which we have labeled nodes as being derived either from duplication or speciation as well as the amino acid replacements associated with the branch. It is also the case that protein pairs immediately after duplication are more likely to be functionally redundant. During a period in which complete functional redundancy is maintained, the number of slightly deleterious mutations may increase without a concomitant increase in replacements at highly constrained sites (as argued above). If these shorter branches include large numbers of protein pairs that are remaining functionally redundant as well as proteins undergoing functional change, then the effect of the redundant pairs may mask any functional adaptation along these branches. The change in the proportion of replacements at constrained sites reported here is distinct from the relaxed selection that has previously been reported on two grounds. First, the timescale on which we have observed our result (five to 25 amino acid replacements per 100 sites) is greater than the timescale of most of the comparisons of dN with dS after duplication (Lynch and Conery 2000; Kondrashov et al. 2002). Second, our method highlights amino acid replacements at a subset of residues and cannot be brought about by an increase in the overall proportion of non synonymous change alone.

Comparisons of sequence divergence after gene duplication and speciation should be approached with caution. In several published studies, an excess of dN over dS immediately after duplication has been observed followed by a gradual decrease in the dN/dS with time (Hughes 1999; Lynch and Conery 2000). These results have often been interpreted as evidence that there is a period of relaxed selection after duplication with the normal mode of selection being resumed after some time. The need for caution arises from the fact that nonsynonymous sequence positions, unlike synonymous positions, are under a wide range of selective pressures. If, for example, a sequence has a small proportion of nonsynonymous positions that are normally evolving under diversifying selection (e.g., an antigenic region of a pathogen), dN/dS may be greater than 1 for comparisons of closely related sequences, regardless of whether the sequences have diverged through duplication or speciation. By comparing amino acid sequences and grouping branches in bins according to length, our comparison of postduplication and postspeciation evolution avoids this problem. To infer that positive selection has been brought about by the duplication event, sequence divergence at a specific time point after duplication should be compared with divergence at the same time point after speciation for the same gene. The same caution needs to be applied to inference of selection from the ratio of radical and nonradical amino acid changes.

Unusual changes at constrained sites provide important insights into the evolution of functional differences (Dermitzakis and Clark 2001; Knudsen and Miyamoto 2001) Identifying clusters of such changes within a protein pinpoints regions of proteins that may confer functional specificity (Casari, Sander, and Valencia 1995; Lichtarge, Bourne, and Cohen 1996; Gu 1999; Caffrey, O’Neill, and Shields 2000; Gu 2001a, 2001b, 2001c), complementing other approaches of identifying adaptive change, such as the identification of excesses of nonsynonymous over synonymous DNA substitutions (Iwabe, Kuma, and Miyata 1996; Hughes 1999; Libersle et al. 2001). For example, this could define target regions of proteins in the design of therapeutic drugs that are specific for the protein. The findings of this study provide a strong rationale for seeking to identify residues conferring functional specificity and indicate that in the majority of protein families, such searches should be focused on evolutionary changes that have occurred over a reasonably wide evolutionary time frame after duplication.

Acknowledgments

This work is a publication from the Biopharmaceutical Sciences Network, supported by the Higher Education Authority (Ireland). We thank Daniel Caffrey, Ken Wolfe, Karen Crum, and Gearoid Tuohy for discussion.

Literature Cited


William Taylor, Associate Editor

Accepted November 6, 2002