Evolution by gene duplication: an update

Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan, 3003 Nat. Sci. Bldg, 830 N. University Ave, Ann Arbor, MI 48109, USA

The importance of gene duplication in supplying raw genetic material to biological evolution has been recognized since the 1930s. Recent genomic sequence data provide substantial evidence for the abundance of duplicated genes in all organisms surveyed. But how do newly duplicated genes survive and acquire novel functions, and what role does gene duplication play in the evolution of genomes and organisms? Detailed molecular characterization of individual gene families, computational analysis of genomic sequences and population genetic modeling can all be used to help us uncover the mechanisms behind the evolution by gene duplication.

In 1936, Bridges reported one of the earliest observations of gene duplication from the doubling of a chromosomal band in a mutant of the fruit fly *Drosophila melanogaster*, which exhibited extreme reduction in eye size [1]. The potential role of gene duplication in evolution was subsequently suggested and possible scenarios of duplicate gene evolution were proposed [2–4]. Ohno's seminal book in 1970, *Evolution by Gene Duplication* [5], further popularized this idea among biologists. It was, however, not until the late 1990s, when many genome sequences were determined and analyzed, that the prevalence and importance of gene duplication was clearly demonstrated. Through genomic sequence analysis, population genetic modeling and molecular experimentation, rapid progress has also been made in disclosing the mechanisms by which duplicate genes diverge in function and contribute to evolution. Here, I review current understandings of these mechanisms. I do not discuss genome duplication, as there have been several recent reviews of this topic [6–8].

Prevalence of gene duplication in all three domains of life

Table 1 lists the estimated numbers of duplicated genes in completely or nearly completely sequenced genomes of representative bacteria, archaebacteria and eukaryotes. One finds that, in all three domains of life, large proportions of genes were generated by gene duplication. It is almost certain that these proportions are underestimates, because many duplicated genes have diverged so much that virtually no sequence similarity is found.

Lynch and Conery estimated that gene duplication arises (and is fixed in populations) at an approximate rate of 1 gene^{-1} 100 million years (MY)$^{-1}$ in eukaryotes such as *Homo sapiens*, *Mus musculus, D. melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Saccharomyces cerevisiae* [9]. This rate is comparable to that of nucleotide substitution, which is 0.1–0.5 site^{-1} 100 MY^{-1} in nuclear genomes of vertebrates [10]. The above duplication rate is the gene-birth rate, which was derived from recent duplications. Many fixed duplicated genes later become PSEUDOGENES (see Glossary) and are deleted from the genome. The rate of duplication that gives rise to stably maintained genes is the birth rate multiplied by the retention rate, which is expected to fluctuate with gene function, among other things.

Duplicated genes are often referred to as paralogous genes, which form gene families. Several authors have tabulated the distribution of gene family size for a few completely sequenced genomes [11,12] and this varies substantially among species and gene families [13]; for instance, the biggest gene family in *D. melanogaster* is the

Table 1. Prevalence of gene duplication in all three domains of life[a]

	Total number of genes	Number of duplicate genes (% of duplicate genes)	Refs
Bacteria			
Mycoplasma pneumoniae	677	298 (44)	[65]
Helicobacter pylori	1590	266 (17)	[66]
Haemophilus influenzae	1709	284 (17)	[67]
Archaea			
Archaeoglobus fulgidus	2436	719 (30)	[68]
Eukarya			
Saccharomyces cerevisiae	6241	1858 (30)	[67]
Caenorhabditis elegans	18 424	8971 (49)	[67]
Drosophila melanogaster	13 601	5536 (41)	[67]
Arabidopsis thaliana	25 498	16 574 (65)	[69]
Homo sapiens	40 580[b]	15 343 (38)	[11]

[a] Use of different computational methods or criteria results in slightly different estimates of the number of duplicated genes [12].

[b] The most recent estimate is ~30 000 [61].

Glossary

Concerted evolution: a mode of gene family evolution in which members of a family remain similar in sequence and function because of frequent gene conversion and/or unequal crossing over.
Gene conversion: a recombination process that nonreciprocally homogenizes gene sequences.
Nonsynonymous (nucleotide substitution): a nucleotide substitution in the coding region of a gene that changes the protein sequence.
Positive (darwinian) selection: natural selection that promotes the fixation of advantageous alleles.
Pseudogene: a DNA sequence derived from a functional gene but has been rendered nonfunctional by mutations.
Purifying selection: natural selection that prevents the fixation of deleterious alleles.
Operon: a unit of gene expression and regulation, including structural genes and control elements.
Synonymous (nucleotide substitution): a nucleotide substitution in the coding region of a gene that does not change the protein sequence.

Corresponding author: Jianzhi Zhang (jianzhi@umich.edu).

http://tree.trends.com 0169-5347/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved. doi:10.1016/S0169-5347(03)00033-8

trypsin family [12], with 111 members, whereas the biggest family in mammals is the olfactory receptor family, with ~1000 members [14,15]. From a genomic sequence analysis of the bacterium *Escherichia coli*, two yeasts, *C. elegans* and *D. melanogaster*, Conant and Wagner found that ribosomal proteins and transcription factors generally form smaller gene families than do other proteins, such as those controlling cell cycles and metabolism [16].

Generation of duplicate genes

Gene duplication can result from unequal crossing over (Fig. 1a), retroposition (Fig. 1b), or chromosomal (or genome) duplication, the outcomes of which are quite different. Unequal crossing over usually generates tandem gene duplication; that is, duplicated genes are linked in a chromosome (Fig. 1a). Depending on the position of crossing over, the duplicated region can contain part of a gene, an entire gene, or several genes. In the latter two cases, introns, if present in the original genes, will also be present in the duplicated genes. This is in sharp contrast to the result from retroposition (Fig. 1b). Retroposition occurs when a message RNA (mRNA) is retrotranscribed to complementary DNA (cDNA) and then inserted into the genome. As expected from this process, there are several molecular features of retroposition: loss of introns and regulatory sequences, presence of poly A tracts, and presence of flanking short direct repeats, although deviations from these common patterns do occasionally occur [17]. Another major difference from unequal crossing over is that a duplicated gene generated by retroposition is usually unlinked to the original gene, because the insertion of cDNA into the genome is more or less random. It is also impossible to have blocks of genes duplicated together by retroposition unless the genes involved are all in an OPERON. Only those genes that are expressed in the germ line are subject to heritable retroposition. Because promoter and regulatory sequences of a gene are not transcribed and hence not duplicated by retroposition, the resulting duplicate often lacks necessary elements for transcription and thus immediately becomes a pseudogene. Nevertheless, several retroposition-mediated duplicate genes are expressed, probably because of the chance insertion of cDNA into a genomic location that is downstream of a promoter sequence [17]. Chromosomal or genome duplication occurs probably by a lack of disjunction among daughter chromosomes after DNA replication. Substantial evidence shows that these large-scale duplications occurred frequently in plants but infrequently in animals [10]. Recent human genome analysis reveals another type of large-scale duplication, segmental duplication, which often involves 1000 to >200 000 nucleotides [18]. That most segmental duplications do not generate tandem repeats suggests that unequal crossing over is probably not responsible, although the exact duplication mechanism is unclear [18].

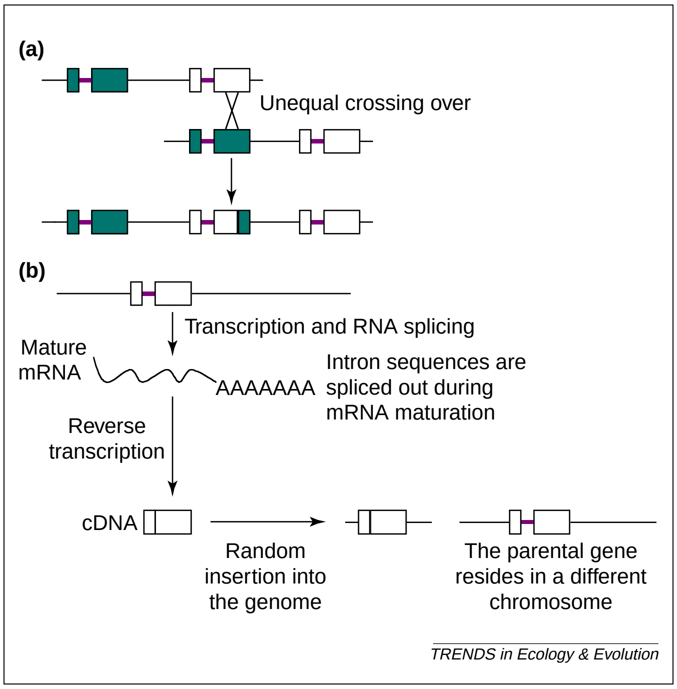

(a)
Unequal crossing over
(b)
Transcription and RNA splicing
Mature mRNA
AAAAAAA
Intron sequences are spliced out during mRNA maturation
Reverse transcription
cDNA
Random insertion into the genome
The parental gene resides in a different chromosome
TRENDS in Ecology & Evolution

Fig. 1. Two common modes of gene duplication. (a) Unequal crossing over, which results in a recombination event in which the two recombining sites lie at nonidentical locations in the two parental DNA molecules. (b) Retroposition, which occurs when a message RNA (mRNA) is retrotranscribed to complementary DNA (cDNA) and then inserted into the genome. Squares represent exons and bold lines represent introns.

Evolutionary fate of duplicate genes

Duplication occurs in an individual, and can be fixed or lost in the population, similar to a point mutation. If a new allele comprising duplicate genes is selectively neutral, compared with pre-existing alleles, it only has a small probability, $1/2N$, of being fixed in a diploid population [19], where N is the effective population size. This suggests that many duplicated genes will be lost. For those that do become fixed, fixation is time consuming, because it takes, on average, $4N$ generations for a neutral allele to become fixed [19]. Upon fixation, the long-term evolutionary fate of duplication will still be determined by functions of the duplicate genes. The birth and death of genes is a common theme in gene family and genome evolution [20,21], with those genes involved in the physiologies that vary greatly among species (e.g. immunity, reproduction and sensory systems) probably having high rates of gene birth and death.

Pseudogenization

Gene duplication generates functional redundancy, as it is often not advantageous to have two identical genes. In other words, mutations disrupting the structure and function of one of the two genes are not deleterious and are not removed by selection. Gradually, the mutation-containing gene becomes a pseudogene, which is either unexpressed or functionless, an evolutionary fate that has been shown by population genetic modeling [22,23] as well as by genomic analysis [9,24]. After a long time evolutionarily speaking, pseudogenes will either be deleted from the genome or become so diverged from the parental genes that they are no longer identifiable. Relatively young pseudogenes are recognizable because of sequence similarity. For example, genomic analyses have identified 2168 pseudogenes in *C. elegans*, or about one pseudogene for every eight functional genes [25]. More pseudogenes exist in humans, with about one pseudogene for every two functional genes in the two completely sequenced chromosomes [24]. Pseudogenization, the process by which a functional gene becomes a pseudogene, usually occurs in the first few million years after duplication if the duplicated gene is not under any selection [9]. Nevertheless, some duplicated genes had been maintained in the genome for a long time for specific functions, before recently becoming pseudogenes because of the relaxation of functional constraints. For example, the size of the olfactory receptor gene family (~1000) is similar in humans and mice, but the percentage of pseudogenes is >60% in humans and only 20% in mice. Many olfactory receptor genes have become pseudogenes since the origin of hominoids [26]. This is probably related to the reduced use of olfaction in hominoids, which can be compensated for by other sensory mechanisms, such as better vision.

Pseudogenes do occasionally serve some function. In chickens, there is only one functional gene (*VH1*) encoding the heavy chain variable region of immunoglobulins, and immunoglobulin diversity is generated by GENE CONVERSION of the *VH1* gene by the many duplicated variable region pseudogenes that occur on its 5′ side [27]. Although unlikely, pseudogenes can also be revived. In cows, the pancreatic ribonuclease gene has a paralogous gene called the seminal ribonuclease gene, which is expressed in semen. These two genes are the result of gene duplication that occurred before the radiation of ruminants at least 35 MY ago. In all other ruminants, the seminal ribonuclease gene either contains deleterious mutations or is not expressed [28–30], which suggests that the seminal ribonuclease gene had been a pseudogene for much of its history, but was revived recently in the cow. How this could have happened is unclear.

In my view, there have not been sufficient studies of pseudogenization probably because pseudogenes are regarded to be uninteresting. In fact, lineage-specific pseudogenization, such as the aforementioned example of olfactory receptor genes of hominoids, provides rich information about organismal evolution.

Conservation of gene function

The presence of duplicate genes is sometimes beneficial simply because extra amounts of protein or RNA products are provided. This applies mainly to strongly expressed genes the products of which are in high demand, such as rRNAs and histones. How can two paralogous genes maintain the same function after duplication? One way is by gene conversion. Under frequent gene conversion, two paralogous genes will have very similar sequences and functions, and this mode of evolution is often referred to as CONCERTED EVOLUTION [10]. Alternatively, strong PURIFYING SELECTION against mutations that modify gene function can also prevent duplicated genes from diverging. Purifying selection can be distinguished from gene conversion by an examination of synonymous (or silent) nucleotide differences among duplicated genes. Synonymous differences are more or less immune to selection and cannot be reduced by purifying selection. But they can be removed by gene conversion, because gene conversion homogenizes DNA sequences regardless of whether the differences are synonymous or nonsynonymous (amino-acid-altering). Using this strategy, Nei and his associates re-examined several large gene families that were previously thought to be under concerted evolution. Their results suggest that purifying selection is much more important than is gene conversion in maintaining common functions of these duplicated genes [21,31]. A recent population genetic analysis also suggested that the conditions for gene conversion to be favored selectively are relatively restrictive [32].

Subfunctionalization

Unless the presence of an extra amount of gene product is advantageous, two genes with identical functions are unlikely to be stably maintained in the genome [33]. Theoretical population genetics predicates that both duplicates can be stably maintained when they differ in some aspects of their functions [33], which can occur by subfunctionalization, in which each daughter gene adopts part of the functions of their parental gene [34–36]. One form of subfunctionalization that is potentially important in the evolution of development is division of gene expression after duplication ([37], Fig. 2). Several duplicate genes have been demonstrated to evolve following this model of subfunctionalization [37]. For example, zebrafish *engrailed-1* and *engrailed-1b* are a pair of transcription factor genes generated by a chromosomal segmental duplication that occurred in the lineage of ray-finned fish. Zebrafish *engrailed-1* is expressed in the pectoral appendage bud, whereas *engrailed-1b* is expressed in a specific set of neurons in the hindbrain/spinal cord [37]. The sole *engrailed-1* gene of the mouse, orthologous to both genes of the zebrafish, is expressed in both pectoral appendage bud and hindbrain/spinal cord. Changes of gene expression after gene duplication appear to be a general rule rather than exception [38,39] and these changes often occur quickly after gene duplication [39].

Subfunctionalization can also occur at the protein function level and can lead to functional specialization when one of the duplicate genes becomes better at performing one of the original functions of the progenitor gene [40]. A recent study illustrates how a specialized digestive

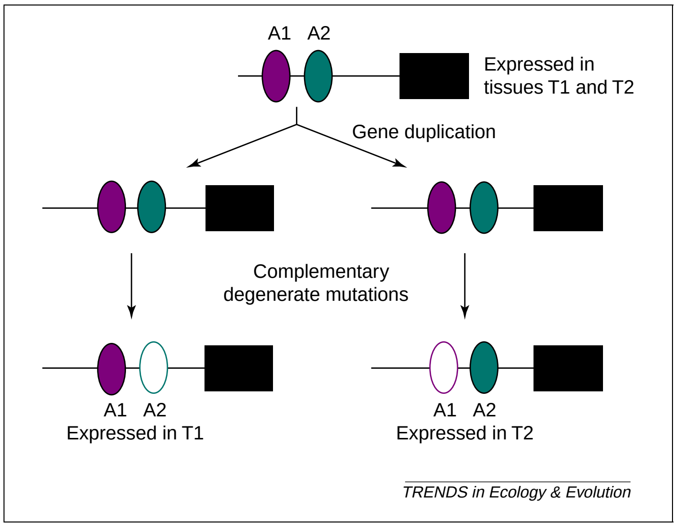

A1 A2
Expressed in tissues T1 and T2
Gene duplication
Complementary degenerate mutations
A1 A2
Expressed in T1
A1 A2
Expressed in T2
TRENDS in Ecology & Evolution

Fig. 2. Division of expression after gene duplication. Squares represent genes, closed ovals represent *cis*-acting elements that regulate gene transcription, and open ovals represent deactivated *cis*-elements. Consider a gene that is expressed in tissues T1 and T2, with a *cis*-acting regulatory element A1 controlling the expression in T1 and A2 controlling the expression in T2. Following gene duplication, one daughter gene might lose the A1 element whereas the other gene might lose A2, so that each is expressed in only one of the two tissues. Under such conditions, both genes are necessary and therefore will be maintained in the genome.

enzyme gene emerged following the duplication of a bifunctional gene in the leaf-eating monkey douc langur ([41] Box 1). It is unclear, however, what percentage of duplicate genes evolved by subfunctionalization.

Neofunctionalization

One of the most important outcomes of gene duplication is the origin of novel function. Although it seems improbable that an entirely new function could emerge in a duplicate gene, there are several examples. For instance, the eosinophil-derived neurotoxin (*EDN*) and eosinophil cationic protein (*ECP*) genes of humans were generated in the lineage of hominoids and Old World monkeys via gene duplication [42]. Both genes belong to the RNase A gene superfamily. After duplication, a novel antibacterial activity emerged in *ECP*. This activity is absent in human *EDN* and the *EDN* of New World monkeys, which represents the progenitor gene before duplication. More surprisingly, the antibacterial activity of *ECP* does not depend on the ribonuclease activity [43]. Molecular evolutionary analysis suggests that the new function is probably conferred by a large number of arginine substitutions that occurred in a short period after duplication [42]. *ECP* is toxic to bacteria because it makes their cell membranes porous; the positively charged arginine residues might be important for establishing tight contact between the *ECP* and negatively charged bacterial cell membranes in the pore-formation process [42].

In many cases, however, a related function, rather than an entirely new function, evolves after gene duplication. One good example is the red- and green-sensitive opsin genes of humans, which were generated by gene duplication in hominoids and Old World monkeys [44]. After duplication, the two opsins have diverged in function, resulting in a 30-nm difference in the maximum absorption wavelength. This confers the sensitivity to a wide range of colors that humans and related primates have.

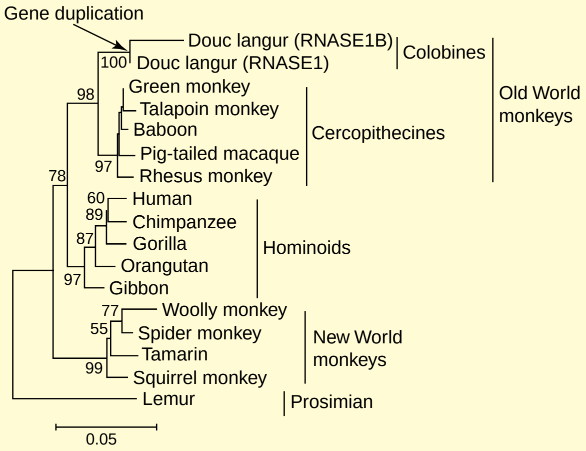
Neofunctionalization of duplicated genes requires varying numbers of amino acid substitutions. The functional change in *ECP* probably required many substitutions [42], but the functional difference between the red and green opsins is largely attributable to two substitutions [45]. In the case of neofunctionalization by a large number of genetic changes, an intriguing question is what function the protein has during the many steps of the genetic changes. This could be answered by the reconstruction of ancestral proteins and functional analysis of these proteins.

Box 1. Digestive RNases of a leaf-eating monkey: a case study of duplicate gene evolution

Colobine monkeys are unique among primates in using leaves rather than fruits and insects as their primary food source; the leaves are fermented in the foregut by symbiotic bacteria [70]. Similar to ruminants, colobines recover nutrients by breaking and digesting the bacteria with the use of various enzymes, including RNase1, which is secreted from the pancreas and transported into the small intestine to degrade RNAs [71,72]. Initial studies revealed a substantially greater amount of RNases in the pancreas of colobines and ruminants than in other mammals [71,72]. This is probably because rapidly growing bacteria have the highest RNA-nitrogen:total nitrogen ratio of all cells, and high concentrations of RNases are needed to break down bacterial RNAs so that nitrogen can be recycled efficiently [72]. Zhang *et al.* identified two copies (*RNase1* and *RNase1B*) of the RNase gene in douc langur *Pygathrix nemaeus*, an Asian colobine, but only one copy (*RNase1*) in each of the 15 noncolobine primates examined [41]. Phylogenetic analyses suggested that the gene duplication occurred after colobines diverged from other monkeys (Fig. I). After duplication, *RNase1B* evolved much more rapidly than did *RNase1* (Fig. I). In fact, the rate of nucleotide substitution in *RNase1B* since duplication was significantly higher at nonsynonymous sites than at synonymous and noncoding sites, providing evidence for the action of positive selection on *RNase1B*.

Because the pH in the small intestine of colobines is 6–7, whereas that in humans and other monkeys is 7.4–8, Zhang *et al.* hypothesized that RNase1B has been under selection for a high catalytic efficiency in an acidic environment. To test this hypothesis, recombinant proteins from the douc langur *RNase1B* gene as well as the *RNase1* genes of humans, rhesus monkeys and douc langur were prepared and their ribonucleolytic activities were quantified at different pHs. The catalytic optimal pH of douc langur RNase1B was found to be 6.3, whereas that of *RNase1* was 7.4 for all three species examined. At pH 6.3, RNase1B is approximately six times more efficient in degrading RNA than is RNase1 [41].

Why has the douc langur *RNase1* been conserved after duplication and why does its optimal catalytic pH remain at 7.4? Human *RNase1* is expressed in many tissues other than the pancreas and has a second enzyme activity (EA_{dsRNA}) in degrading double-stranded RNA, an activity that might be related to defense against viral infection but unrelated to digestion. Zhang *et al.* found similar EA_{dsRNA} among the RNase1 proteins of the human, rhesus monkey and douc langur, although that of the douc langur RNase1B was reduced by >300-fold [41]. Apparently, RNase1B can lose EA_{dsRNA} and become specialized in digesting bacterial RNAs because the paralogous RNase1 retains EA_{dsRNA}. There are nine amino acid differences between douc langur RNase1 and RNase1B. Using site-directed mutagenesis, Zhang *et al.* made protein mutants and showed that each of the nine substitutions in RNase1B was detrimental to EA_{dsRNA}. Thus, if the EA_{dsRNA} function had not been relaxed in RNase1B, none of the adaptive substitutions that shifted the optimal pH could have occurred [41]. This detailed molecular evolutionary study demonstrated complementary roles of positive selection and relaxation of purifying selection in functional divergence of duplicate genes and provides a vivid example of the contribution of gene duplication toward organismal adaptation to changing environments.



TRENDS in Ecology & Evolution

Fig. I. Phylogenetic relationships of *RNase1* and *RNase1B* genes of primates. Douc langur has both genes, whereas other species have only the *RNase1* gene. Bootstrap percentages >50 are shown. Branch lengths are drawn to scale, indicating the number of nucleotide substitutions per site. This neighbor-joining tree was reconstructed with the use of Kimura's two-parameter distances. Reprinted, with permission, from [41].

Evolutionary forces behind functional divergence of duplicate genes

In the case of division of expression, such as the *engrailed-1* and *engrailed-1b* genes of zebrafish, it is likely that random fixations of complementary degenerate mutations under relaxed functional constraints are the main cause [37]. In other words, it is a result of neutral evolution, without the involvement of POSITIVE SELECTION. In the case of functional specialization and neofunctionalization, two models have been widely cited. The first model is known as the Dykhuizen–Hartl effect, which does not require positive selection [19,42,46,47]. In this model, after gene duplication, random mutations are fixed in one daughter gene under relaxed purifying selection, which occurs by reduced functional constraint provided by genetic redundancy. These fixed mutations later induce a change in gene function when the environment or the genetic background is altered.

The second model requires positive selection and involves two scenarios. In the first scenario, after gene duplication, a few neutral or nearly neutral substitutions create a new, but only weakly active function in one daughter gene, and positive selection then accelerates the fixation of advantageous mutations that enhance the activity of the novel function [42]. In the second scenario, the ancestral gene already has dual functions. Gene duplication provides the opportunity for each daughter gene to adopt one ancestral function, and further substitutions under positive selection can refine the functions [40]. Acceleration of protein sequence evolution following gene duplication is often observed [9,48,49]; however, this can be explained by either model. Under such circumstances, one often assumes the null hypothesis of neutral evolution with relaxed purifying selection. A significantly higher rate of NONSYNONYMOUS than SYNONYMOUS NUCLEOTIDE SUBSTITUTION can be used to reject the null hypothesis and to establish the action of positive selection. Indeed, several cases of positive selection after gene duplication have been reported, including immunoglobulins, conotoxins, ribonucleases, pregnancy-associated glycoproteins, triosephosphate isomerase and the *ECP* gene [40–42,50–54]. A cautionary note is that relaxation of purifying selection is often treated as the null hypothesis and is thus accepted even without direct evidence. Because of the relatively low power of statistical methods for detecting positive selection, actions of positive selection have probably been overlooked and relaxation of purifying selection incorrectly invoked. Box 1 illustrates how molecular experimentation can be used to test the hypothesis of relaxation of purifying selection. Both positive selection and relaxation of purifying selection are necessary in the functional divergence of duplicate genes [41] (Box 1). This might be particularly so when the functional change involves multiple amino acid substitutions.

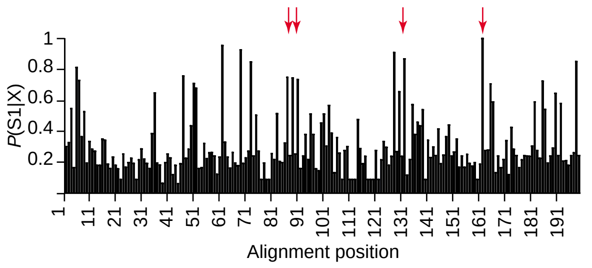
When specialization or neofunctionalization is completed, duplicate genes are likely to be maintained under different functional constraints and show different substitution patterns. Several statistical methods have been developed to identify amino acid sites evolving with altered substitution rates and to test the rate difference [55–58]. Box 2 shows an application of such statistical methods in finding candidate sites responsible for functional differences between two subfamilies of the proteases involved in apoptosis. These computational results can guide molecular experimentation, leading to the identification of functionally important amino acid substitutions, which might help us to reveal the molecular basis of functional divergence as well as provide crucial information for drug design and other biomedical applications.

Box 2. Identifying amino acid changes behind the functional differences among caspases

When two duplicated genes A and B have different functions, the amino acid substitution rate in A might differ from that in B at certain sites. Let X be a sequence data set that usually contains multiple orthologous sequences of protein A and B, and let S1 be the event that the substitution rate in A and B are different at a given site. Gu has developed a method to compute $P(S1|X)$, the posterior probability that a site has different substitution rates in A and B, given the data X [55]. He has also invented a statistic θ as a measure of the overall site-specific rate difference between proteins A and B [55]. The computer program for these methods [73] is available at http://xgu1.zoo1.iastate.edu.

Wang and Gu [74] applied this method to cysteine aspartyl proteases (caspases), which are key components in apoptosis. The authors were interested in caspases because, in vertebrates, these can be divided into two subfamilies (CED-3 and ICE), which have different functions. CED-3 type caspases are essential for most apoptotic pathways whereas the major function of ICE-type caspases is to mediate immune response, although some members are also involved in apoptosis under some circumstances. The two subfamilies also show structural differences.

The authors found $\theta = 0.29 \pm 0.05$, suggesting significant site-specific rate differences between the CED-3 and ICE subfamilies (Fig. I). The significantly positive value of θ was found to be due to 21 sites, as θ becomes virtually 0 when these sites are removed. In fact, experimental evidence is available supporting the functional importance of four of the 21 sites. It would be interesting to examine experimentally whether the rest of the sites are also functionally important. Such experiments will also help computational biologists to improve predictive methods.



TRENDS in Ecology & Evolution

Fig. I. The site-specific profile for predicting crucial amino acid residues responsible for the functional divergence between CED-3 and the ICE subfamilies, measured by the posterior probability $P(S_1|X)$. The arrows point to four amino acid residues at which functional divergence between two subfamilies has been verified by experiments. Reprinted, with permission, from [74].

Contributions of gene duplication to genomic and organismal evolution

The most obvious contribution of gene duplication to evolution is providing new genetic material for mutation, drift and selection to act upon, the result of which is specialized or new gene functions. Without gene duplication,

Box 3. Outstanding questions about the evolution of duplicate genes

1. What is the relative importance of positive darwinian selection and relaxation of purifying selection in functional divergence of duplicated genes? In spite of many case studies, a general pattern is still lacking. This is likely to be a long-standing question, because sequence analysis, even at the genomic level, only provides a partial answer because of the inefficiency of current statistical methods in detecting selection.
2. How does an entirely new function originate after gene duplication? More detailed molecular studies of model gene families are needed to look into the emergence of novel gene function.
3. What roles does gene duplication play in the establishment of complex gene expression networks and protein–protein interaction networks, which are key characteristics of biological systems? A few studies have been conducted using genomic data from model organisms [59,75,76], but more comprehensive studies of gene duplication in the evolution of gene/protein networks are needed.
4. How does genetic buffering function? Biological systems are quite robust and deletions of some genes often do not show severe phenotypes. Is this because of the existence of duplicate genes that share functional similarity with the deleted genes or because of redundant metabolic networks? A recent study using mammalian gene sequences suggested that gene duplication is as important as redundant metabolic networks [77]. Another study using yeast gene knockout experiments showed that, on average, deletions of genes that have closely related duplicates exhibit less severe phenotypes than do deletions of genes with distantly related duplicates or without duplicates [78]. These studies provide the first evidence that gene duplication plays a key role in making biological systems robust against genetic turbulence. Related to this issue, one wonders whether functional redundancy of duplicated genes is adaptive or simply a result of structural constraints brought by a common origin.
5. How important is gene duplication to the origin of species-specific features and speciation? Answering this question requires empirical data from closely related species, particularly the molecular details of the genes involved in reproductive isolation.

the plasticity of a genome or species in adapting to changing environments would be severely limited, because no more than two variants (alleles) exist at any locus within a (diploid) individual. It seems difficult to imagine, for instance, how the vertebrate adaptive immune system (with dozens of duplicated immunoglobulin genes) could have evolved without gene duplication.

Gene duplication has probably also contributed to the evolution of gene networks in such a way that sophisticated expression regulations can be established [59]. An interesting case is *eyeless* (also known as *Pax6*), the master control gene of eye development in metazoans. This gene was duplicated in *Drosophila*, and its paralog, *twin of eyeless*, now regulates the expression of *eyeless* [60].

Species-specific gene duplication can also lead to species-specific gene functions, which might facilitate species-specific adaptation, as exemplified in [41] (Box 1). In other words, gene duplication contributes to species divergence and origins of species-specific features. For example, if the gene duplication rate of 1 gene^{-1} 100 MY^{-1} [9] is used, I estimate that there have been $1 \times 30\,000 \times 6/100 = 1800$ gene duplications in the human genome since humans diverged from chimpanzees. Here, 30 000 is the estimated total number of human genes [61] and 6 is the approximate time in MY since the human–chimpanzee split. A more conservative estimate of $\sim$720 gene duplications can be obtained using the recent result from human segmental duplications [62,63]. Although many of these duplicated genes might have become pseudogenes, it is possible that some acquired new functions. Identification of human-specific gene duplications might help pinpoint the genetic basis of human-unique features. With the human genome sequence now available, such studies should be feasible. In fact, several potential cases of human-specific gene duplication are known [64]. It is also possible, as Lynch has suggested [63], that differential gene duplication and pseudogenization in geographically isolated populations causes reproductive isolation and speciation, although this intriguing hypothesis awaits empirical evidence. Box 3 lists several outstanding questions on the evolution of gene duplication that I believe are important. It can be expected that, with an explosive increase in genomic data and rapid advances in molecular genetic technology, the manifold and fundamental roles of gene duplication will become even more evident and the once imaginative idea of evolution by gene duplication will be established as one of the cornerstones of evolutionary biology.

Acknowledgements

I thank Lizhi Gao, Xun Gu, David Mindell, David Webb and three anonymous reviewers for valuable comments. This work was supported by a startup fund and a Rackham research fellowship of the University of Michigan and an NIH grant (R01-GM67030) to J.Z.

References

1 Bridges, C.B. (1936) The Bar ‘gene’ a duplication. *Science* 83, 210–211
2 Stephens, S.G. (1951) Possible significance of duplication in evolution. *Adv. Genet.* 4, 247–265
3 Ohno, S. (1967) *Sex Chromosomes and Sex-Linked Genes*, Springer
4 Nei, M. (1969) Gene duplication and nucleotide substitution in evolution. *Nature* 221, 40–42
5 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer
6 Wolfe, K. (2001) Yesterday’s polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341
7 Friedman, R. and Hughes, A.L. (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res.* 11, 1842–1847
8 Spring, J. (2002) Genome duplication strikes back. *Nat. Genet.* 31, 128–129
9 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
10 Li, W.H. (1997) *Molecular Evolution*, Sinauer
11 Li, W.H. *et al.* (2001) Evolutionary analyses of the human genome. *Nature* 409, 847–849
12 Gu, Z. *et al.* (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* 19, 256–262
13 Lespinet, O. *et al.* (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059
14 Mombaerts, P. (2001) The human repertoire of odorant receptor genes and pseudogenes. *Annu. Rev. Genomics Hum. Genet.* 2, 493–510
15 Zhang, X. and Firestein, S. (2002) The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci.* 5, 124–133
16 Conant, G.C. and Wagner, A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30, 3378–3386
17 Long, M. (2001) Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11, 673–680
18 Samonte, R.V. and Eichler, E.E. (2002) Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* 3, 65–72
19 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press

20 Hughes, A.L. and Nei, M. (1989) Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Mol. Biol. Evol.* 6, 559–579

21 Nei, M. *et al.* (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10866–10871

22 Walsh, J.B. (1995) How often do duplicated genes evolve new functions? *Genetics* 139, 421–428

23 Lynch, M. *et al.* (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159, 1789–1804

24 Harisson, P.M. *et al.* (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 12, 272–280

25 Harrison, P.M. *et al.* (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* 29, 818–830

26 Rouquier, S. *et al.* (2000) The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2870–2874

27 Ota, T. and Nei, M. (1995) Evolution of immunoglobulin VH pseudogenes in chickens. *Mol. Biol. Evol.* 12, 94–102

28 Trabesinger-Ruef, N. *et al.* (1996) Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? *FEBS Lett.* 382, 319–322

29 Breukelman, H.J. *et al.* (1998) Secretory ribonuclease genes and pseudogenes in true ruminants. *Gene* 212, 259–268

30 Kleineidam, R.G. *et al.* (1999) Seminal-type ribonuclease genes in ruminants, sequence conservation without protein expression? *Gene* 231, 147–153

31 Piontkivska, H. *et al.* (2002) Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol. Biol. Evol.* 19, 689–697

32 Hurst, L.D. and Smith, N.G.C. (1998) The evolution of concerted evolution. *Proc. R. Soc. Lond. Ser. B* 265, 121–127

33 Nowak, M.A. *et al.* (1997) Evolution of genetic redundancy. *Nature* 388, 167–171

34 Jensen, R.A. (1976) Enzyme recruitment in the evolution of new function. *Annu. Rev. Microbiol.* 30, 409–425

35 Orgel, L.E. (1977) Gene duplication and the origin of proteins with novel functions. *J. Theor. Biol.* 67, 773

36 Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. Ser. B* 256, 119–124

37 Force, A. *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545

38 Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6579–6584

39 Gu, Z. *et al.* (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18, 609–613

40 Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*, Oxford University Press

41 Zhang, J. *et al.* (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* 30, 411–415

42 Zhang, J. *et al.* (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3708–3713

43 Rosenberg, H.F. (1995) Recombinant human eosinophil cationic protein. Ribonuclease activity is not essential for cytotoxicity. *J. Biol. Chem.* 270, 7876–7881

44 Yokoyama, S. and Yokoyama, R. (1989) Molecular evolution of human visual pigment genes. *Mol. Biol. Evol.* 6, 186–197

45 Asenjo, A.B. *et al.* (1994) Molecular determinants of human red/green color discrimination. *Neuron* 12, 1131–1138

46 Dykhuizen, D. and Hartl, D.L. (1980) Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* 96, 801–817

47 Li, W.H. (1983) Evolution of duplicate genes and pseudogenes. In *Evolution of Genes and Proteins* (Nei, M. and Koehn, R.K., eds) pp. 14–37, Sinauer

48 Ohta, T. (1994) Further examples of evolution by gene duplication revealed through DNA sequence comparisons. *Genetics* 138, 1331–1337

49 Van de Peer, Y. *et al.* (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.* 53, 436–446

50 Tanaka, T. and Nei, M. (1989) Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol. Biol. Evol.* 6, 447–459

51 Duda, T.F. Jr and Palumbi, S.R. (1999) Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6820–6823

52 Zhang, J. *et al.* (2000) Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4701–4706

53 Hughes, A.L. *et al.* (2000) Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3319–3323

54 Merritt, T.J. and Quattro, J.M. (2001) Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. *Genetics* 159, 689–697

55 Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674

56 Dermitzakis, E.T. and Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* 18, 557–562

57 Knudsen, B. and Miyamoto, M.M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14512–14517

58 Gaucher, E.A. *et al.* (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* 27, 315–321

59 Wagner, A. (1994) Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. U. S. A.* 91, 4387–4391

60 Czerny, T. *et al.* (1999) Twin of eyeless, a second *Pax-6* gene of *Drosophila*, acts upstream of eyeless in the control of eye development. *Mol. Cell* 3, 297–307

61 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562

62 Bailey, J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007

63 Gagneux, P. and Varki, A. (2001) Genetic differences between humans and great apes. *Mol. Phylogenet. Evol.* 18, 2–13

64 Lynch, M. (2002) Gene duplication and evolution. *Science* 297, 945–947

65 Himmelreich, R. *et al.* (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24, 4420–4449

66 Tomb, J.F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacterpylori*. *Nature* 388, 539–547

67 Rubin, G.M. *et al.* (2000) Comparative genomics of the eukaryotes. *Science* 287, 2204–2215

68 Klenk, H.P. *et al.* (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364–370

69 The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815

70 Kay, R.N.B. and Davies, A.G. (1994) Digestive physiology. In *Colobine Monkeys: Their Ecology Behaviour and Evolution* (Davies, A.G. and Oates, J.F., eds) pp. 229–250, Cambridge University Press

71 Barnard, E.A. (1969) Biological function of pancreatic ribonuclease. *Nature* 221, 340–344

72 Beintema, J.J. (1990) The primary structure of langur (*Presbytis entellus*) pancreatic ribonuclease adaptive features in digestive enzymes in mammals. *Mol. Biol. Evol.* 7, 470–477

73 Gu, X. and Vander Velden, K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* 18, 500–501

74 Wang, Y. and Gu, X. (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics* 158, 1311–1320

75 Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18, 1283–1292

76 Bhan, C. *et al.* (2002) A duplication growth model of gene expression networks. *Bioinformatics* 18, 1486–1493

77 Kitami, T. and Nadeau, J.H. (2002) Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nat. Genet.* 32, 191–194

78 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66