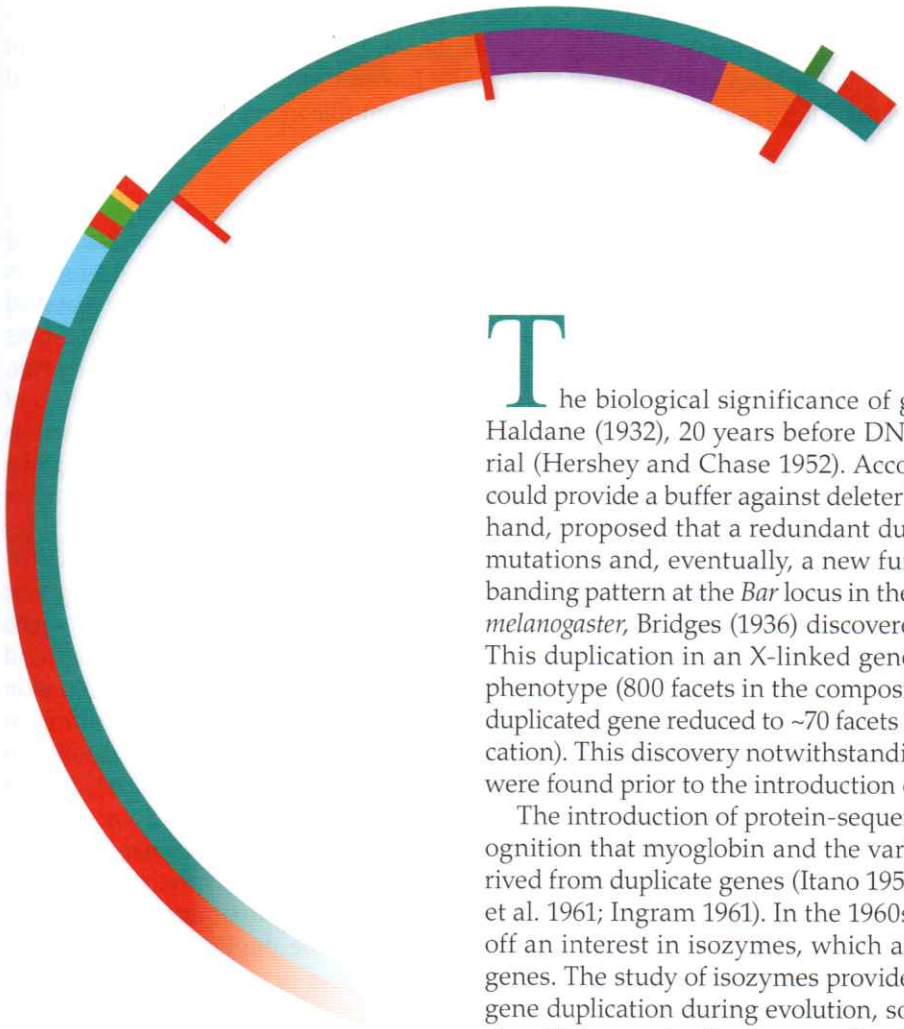
CHAPTER

7

Evolution by DNA Duplication



The biological significance of gene duplication was first recognized by Haldane (1932), 20 years before DNA was shown to be the hereditary material (Hershey and Chase 1952). According to Haldane (1932), duplicate genes could provide a buffer against deleterious mutations. Muller (1935), on the other hand, proposed that a redundant duplicate of a gene might acquire divergent mutations and, eventually, a new function. By microscopically examining the banding pattern at the *Bar* locus in the giant salivary chromosomes of *Drosophila melanogaster*, Bridges (1936) discovered the first example of a gene duplication. This duplication in an X-linked gene is responsible for the reduced-eye (*Bar*) phenotype (800 facets in the composite eye of females homozygous for the unduplicated gene reduced to ~70 facets in females homozygous for the gene duplication). This discovery notwithstanding, few other examples of duplicate genes were found prior to the introduction of molecular techniques.

The introduction of protein-sequencing methods in the 1950s led to the recognition that myoglobin and the various chains of hemoglobin have been derived from duplicate genes (Itano 1953, 1957; Rhinesmith et al. 1958; Braunitzer et al. 1961; Ingram 1961). In the 1960s, protein electrophoretic methodology set off an interest in isozymes, which are enzymes mostly encoded by duplicate genes. The study of isozymes provided evidence for the frequent occurrence of gene duplication during evolution, so much so that Ohno (1970) put forward a view that gene duplication was the only means by which a new function could arise: "natural selection merely modified, while redundancy created." Ohno's view was criticized initially, but it gained acceptance in the 1970s and has since stimulated much interest in the study of duplicate genes. Although other means of creating new functions are now known (Chapter 8), Ohno's view remains largely valid. With the advent of gene cloning and sequencing techniques in the 1980s, large quantities of duplicate-gene data started accumulating, culminating in a veritable deluge resulting from whole-genome sequencing.

In this chapter, we discuss the different types of duplication; the evolutionary fates of duplicated genes, including gene death and the acquisition of novel function; and the curious phenomenon of concerted evolution, i.e., the coordinated manner that characterizes the evolution of some repeated DNA sequences.

Types of DNA Duplication

An increase in the number of copies of a DNA segment can be brought about by several types of **DNA duplication** (Figure 7.1). These are usually classified according to the size or extent of the genomic region involved. The following types of duplication are recognized: (1) **internal** or **partial gene duplication**, (2) **complete gene duplication**, (3) **partial chromosomal duplication**, (4) **complete chromosomal duplication** (**polysomy**), and (5) **whole-genome duplication** (**polyploidy**). The terms **block duplication**, **regional duplication**, and **segmental duplication** have commonly been used to refer to duplications involving either two or more genes or lengthy genomic segments. Internal gene duplication will be discussed in Chapter 8.

Ohno (1970) argued that whole-genome duplication has generally been more important than regional duplication, because in the latter case only parts of the regulatory system of structural genes may be duplicated and such an imbalance may disrupt the normal function of the duplicated genes. However, as discussed later, regional duplications have apparently played a very important role in evolution.

Mechanisms of DNA Duplication

The mechanisms responsible for DNA duplication fall into two broad categories: (1) unequal crossing over, which is responsible for the generation of tandem repeats (Figure 7.2), and (2) transposition, which is responsible for the creation of dispersed sequences (Chapter 9). **Homologous unequal crossing over** refers to a crossing over that is initiated by the presence of highly similar sequences of substantial length, such as repetitive elements. **Nonhomologous crossing over** refers to a process of crossing over that is initiated by **microhomologies**, i.e., two or more short sequences that exhibit similarity to one another. (The term "microhomology" is something of a misnomer, since the similarity between the sequences is mostly due to chance rather than common descent.) Nonhomologous crossing over occurs at much lower rates than homologous events. Both types of crossing over result in two reciprocal chromosomal products: one will contain a duplication; the other, a deletion. It is important to remember that both these events are mutations, i.e., they affect genomic regions without regard to functional constraints and boundaries. The size of the duplicated region can vary from a few base pairs to tens or even hundreds of kilobases, and it can contain no genes, a portion of a gene, an entire gene, a few genes, or many genes.

An important feature associated with gene duplication is that as long as two or more copies of a gene exist in proximity to each other, the process of gene duplication by unequal

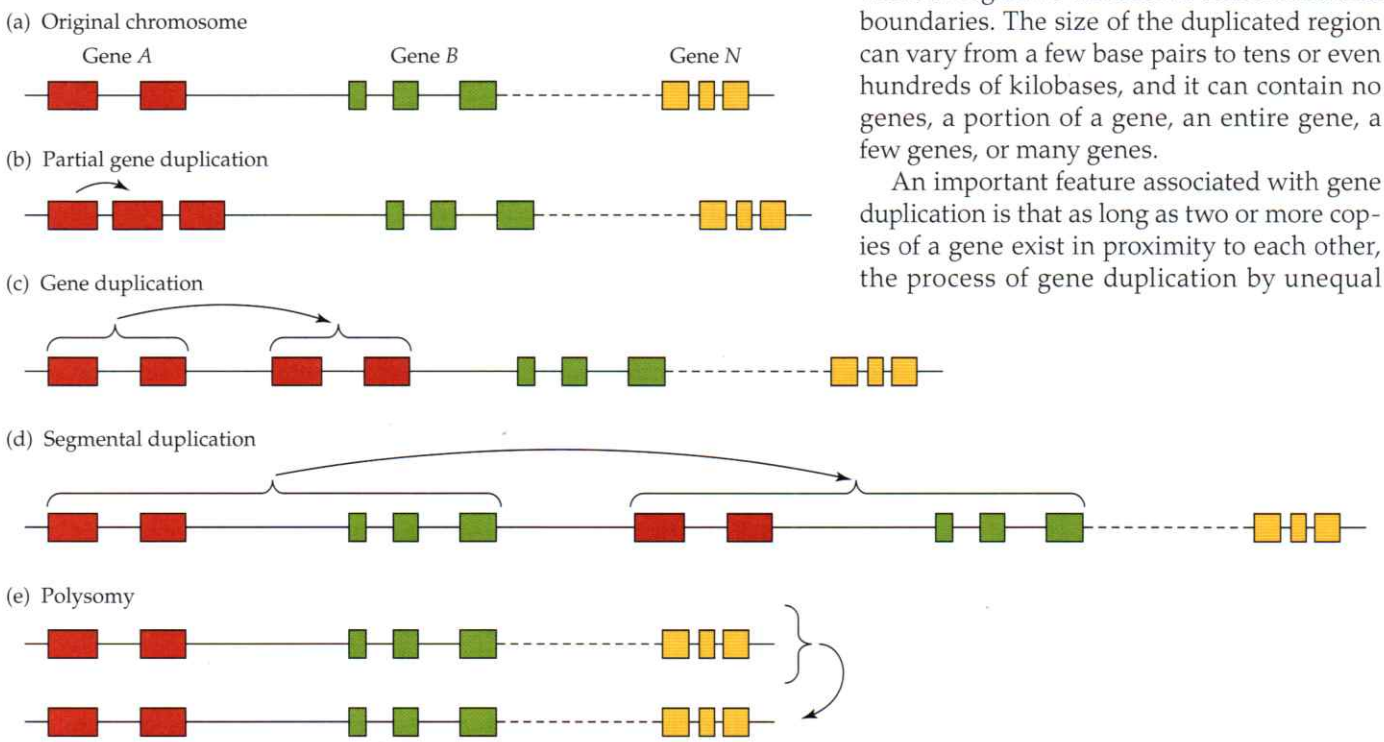


Figure 7.1 Types of DNA duplication. (a) A schematic chromosome. Three genes are shown, gene *A* with two exons, *B* with three exons, and *N* with three exons. The dashed line indicates that there are more than three genes on this chromosome. (b) Partial gene duplication results in the duplication of the first exon in gene *A*. (c) Complete duplication of gene *A*. (d) Segmental duplication of adjacent genes *A* and *B*. (e) Complete chromosomal duplication (polysomy).

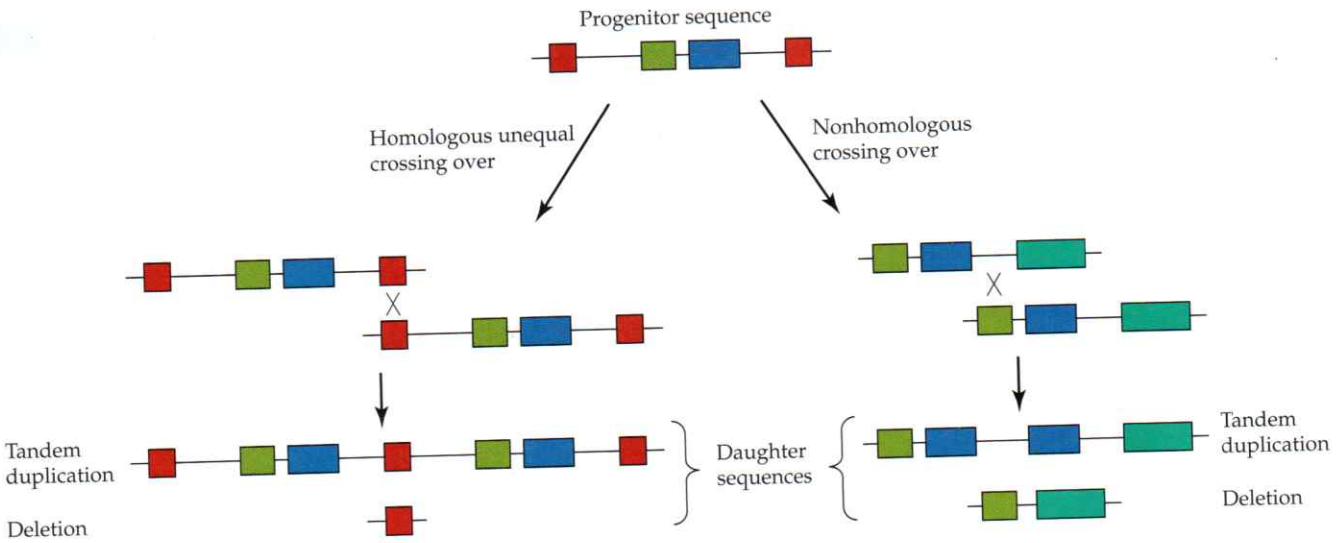


Figure 7.2 Mechanisms for sequence duplication and sequence deletion through crossing over. Single-copy functional sequences are denoted by blue and green rectangles. Repetitive sequences are denoted by red squares. If crossing over is initiated by the presence of similar repetitive sequences, it is referred to as homologous unequal crossing over. If crossing over is initiated by nonhomologous functional sequences, it is referred to as nonhomologous crossing over. In both cases, two reciprocal daughter sequences are produced: one sequence contains a tandem duplication; the other contains a deletion.

crossing over can be greatly accelerated in this region, and numerous copies may be produced.

Gene duplication can also occur by copy-and-paste transposition (Chapter 9). Copy-and-paste transposition can occur through either a DNA or an RNA intermediate. Both DNA-mediated transposition and RNA-mediated retrotransposition can result in the creation of dispersed duplicates.

Sequence homology due to gene duplication is called **paralogy**, in contradistinction with **orthology** and **xenology**, which are due to speciation and horizontal gene transfer, respectively (Chapters 5 and 9).

Dating Duplications

In Figure 7.3, genes *Y* and *Z* were derived from the duplication of an ancestral gene *X* and are therefore paralogous, while gene *Y* from species 1 and gene *Y* from species 2 are orthologous, as are genes *Z* from species 1 and gene *Z* from species 2. We can estimate the date of duplication, T_D, from sequence data if we know the rate of substitution in genes *Y* and *Z*. The rate of substitution can be estimated from the number of substitutions between the orthologous genes in conjunction with knowledge of the time of divergence, T_S, between species 1 and 2 (Figure 7.3).

For gene *Y*, let K_Y be the number of substitutions per site between the two species. Then, the rate of substitution in gene *Y*, r_Y, is estimated by

$$r_Y = \frac{K_Y}{2T_S} \qquad (7.1)$$

The rate of substitution in gene *Z*, r_Z, can be obtained in a similar manner. The average substitution rate for the two genes is given by

$$r = \frac{r_Y + r_Z}{2} \qquad (7.2)$$

To estimate T_D, we need to know the number of substitutions per site between genes *Y* and *Z* (K_{YZ}). This number can be estimated from four pairwise comparisons: (1) gene *Y* from species 1 and

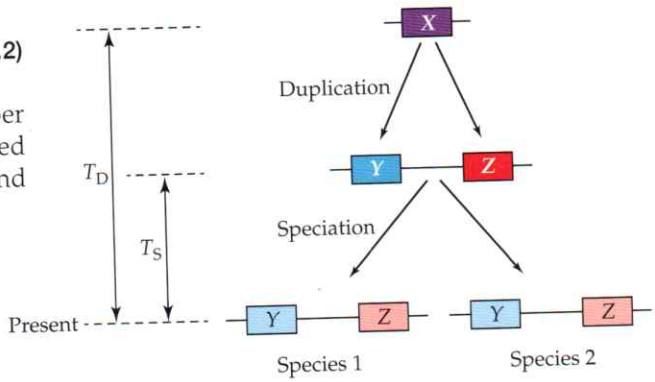


Figure 7.3 Model for estimating the time of a gene duplication event (T_D). Two genes, *Y* and *Z*, were derived from a duplication of gene *X* that occurred T_D time units ago in an ancestral species. The ancestral species split into two species (1 and 2) T_S time units ago.

gene Z from species 2, (2) gene Y from species 2 and gene Z from species 1, (3) gene Y and gene Z from species 1, and (4) gene Y and gene Z from species 2. From these four estimates we can compute the average value for K_{YZ}, i.e., $\overline{K}_{YZ}$, from which we can estimate T_D as

$$T_D = \frac{\overline{K}_{YZ}}{2r} \tag{7.3}$$

Note that, in the case of protein-coding genes, by using the numbers of synonymous and nonsynonymous substitutions separately, we can obtain two independent estimates of T_D. The average of these two estimates may be used as the final estimate of T_D. However, if the number of substitutions per synonymous site between genes Y and Z is large—say, larger than 1—then the number of synonymous substitutions cannot be estimated accurately, and synonymous substitutions may not provide a reliable estimate of T_D. In such cases, only the number of nonsynonymous substitutions should be used. Conversely, if the number of substitutions per nonsynonymous site between the paralogous genes is small, then the estimate of the number of nonsynonymous substitutions is subject to a large sampling error, and in such cases, only the number of synonymous substitutions should be used. Moreover, nonsynonymous sites may experience episodic selection such that their substitution rates may be less clocklike than those at synonymous sites.

In the above discussion, we have assumed rate constancy. Whether this assumption holds can be tested by the four pairwise comparisons mentioned above. If it does not hold, the T_D estimate may be erroneous. As will be discussed later, problems due to concerted evolutionary events may also arise, and these can complicate the estimation of T_D.

Another method for approximating the date of gene duplication events is to consider the phylogenetic distribution of genes in conjunction with paleontological data pertinent to the divergence date of the species in question. For example, all vertebrates with the exception of jawless fishes (hagfishes and lampreys) encode α- and β-globin chains. There are two possible explanations for this observation. One is that the duplication event producing the α- and β-globins occurred in the common ancestor of all vertebrates (Craniata), but the two jawless fish lineages (Myxini and Cephalaspidomorphi) have lost one of the two duplicates. This is possible but not very likely, because such a scenario would require the losses to occur independently in at least two evolutionary lineages. The other explanation is that the duplication event occurred after the divergence of jawless fishes from the ancestor of the jawed vertebrates (Gnathostomata), but before the radiation of the jawed vertebrates from each other. This latter explanation is thought to be more plausible, and the duplication date is commonly taken to be 450–500 million years ago.

Obviously, the above methods can only provide us with rough estimates of duplication dates, so all estimates should be taken with caution. Note that in estimating dates of divergence among species, one can use data from many genes belonging to many gene families. In comparison, in estimating dates of gene duplication, one must rely only on data from genes belonging to a single gene family. Because of the stringent limitations on the sequence data that can be used, estimates of gene duplication are often subject to very large standard errors.

Gene Duplication and Gene Families

The process of complete gene duplication produces paralogous repeats. Depending on their degree of differentiation, repeated genes can be divided into **invariant** and **variant repeats**. Invariant repeats are identical or nearly identical in sequence to one another. Variant repeats are paralogous copies of a gene that, although similar to each other, differ in their sequences to a lesser or greater extent. Surprisingly large numbers of homologous proteins that can perform markedly different functions have been revealed by sequencing projects (Todd et al. 2001). Several such examples are

Table 7.1

Similarity and dissimilarity among pairs of duplicate genes[a]

Gene pair products (organism)	Amino acid sequence similarity (%)	Estimated time of duplication (million years)	Regulation[b]	Chemical attributes[c]	Aggregation properties[d]	Place of expression[e]
Trypsin and chymotrypsin (human)	36	1,500	– – –	– –	+	+
Hemoglobin and myoglobin (human)	23	800	– – –	– –	– – –	– –
Lactate dehydrogenase M and H chains (human)	74	600	– –	– –	+	–
Hemoglobin α and β chains (human)	41	500	– – –	–	–	+
Immunogobulin H and L chains (human)	25	400	– – –	– –	–	+
Lactalbumin and lysozyme (human)	37	350	– – –	– – –	– –	– –
Growth hormone and prolactin (human)	25	330	– –	– –	+	–
Chymotrypsins A and B (human)	79	270	+	+	+	+
Carbonic anhydrases B and C (human)	60	180	–	–	+	+
Insulins I and II (rat)	96	30	–	+	+	+
Growth hormone and lactogen (human)	85	23	– –	–	+	– –
Alcohol dehydrogenase A and S chains (horse)	98	10	–	–	+	+

Source: Modified from Li (1983).

[a] + = similar; – = slightly different; – – = moderately different; – – – = markedly different.

[b] *Regulation* refers to differential expression over tissues or developmental stages, or to the rate of synthesis of the gene product if the two genes are expressed in the same tissue.

[c] *Chemical attributes* include catalytic properties and binding specificities to substrates, inhibitors, antigens, and so on.

[d] *Aggregation* refers to the number of subunits and the types of interactions between them.

[e] *Place of expression* refers to organs of the body or to types of differentiated cells.

listed in Table 7.1. They reveal that the rate with which functional differentiation can evolve can be quite variable. For example, lactalbumin, which is a subunit of the enzyme that catalyzes the synthesis of the sugar lactose, and lysozyme, which dissolves certain bacteria by cleaving the polysaccharide component of their cell walls, perform completely unrelated functions, are regulated differently, and are expressed in different tissues. Yet, the divergence time between lactalbumin and lysozyme is estimated to be only a third of that between trypsin and chymotrypsin, which perform very similar digestive functions and differ from each other merely in their recognition sites, with trypsin recognizing arginine and lysine, whereas chymotrypsin recognizes phenylalanine, tryptophan, and tyrosine (Barker and Dayhoff 1980).

Orthologous and paralogous genes, which share a common ancestral gene, are frequently referred to as a **gene family** or **multigene family**. The term **superfamily** was coined by Dayhoff (1978) to distinguish distantly related proteins from closely related ones. Subsequently, many rank terms referring to various degrees of relationships among homologous proteins, such as "subfamily" and "clan," have been suggested in the literature. As with many hierarchical biological terminologies, however, the terms are context-dependent and may sometimes seem quite erratic.

Gene family size, i.e., the number of genes within a gene family, was found to vary considerably among species and among gene families within a species (Li et al. 2001; Gu et al. 2002). Some families consist of **single-copy genes**. Other gene families consist of a small number of paralogous genes repeated within the genome. A few

gene families comprise hundreds of duplicated copies, in which case they are referred to as **highly repetitive genes**. The largest protein-coding gene family in *Drosophila melanogaster* is the trypsin family with 111 members, whereas the largest such family in humans is the olfactory receptor family with more than 1,000 members (Gilad et al. 2003; Zhang 2003). From sequence analyses of two yeast genomes (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), *Caenorhabditis elegans*, *D. melanogaster*, and *Escherichia coli*, Conant and Wagner (2002) inferred that ribosomal proteins and transcription factors generally form smaller gene families than do other protein-coding genes (e.g., those encoding proteins related to controlling cell cycles and metabolism).

The Prevalence of Gene Duplication

Gene duplications arise spontaneously at high rates in bacteria, bacteriophages, insects, and mammals and are generally viable (Fryxell 1996). Thus, the creation of duplications by mutation is not the rate-limiting step in the process of gene duplication and subsequent functional divergence. However, only a small fraction of all duplicated genes are retained, and an even smaller fraction evolve new functions. The reason is the much higher probability of nonfunctionalization in comparison with that of evolving a new function. We note, however, that in large populations, the probability of evolving a new function may be considerable (Walsh 1995; Nadeau and Sankoff 1997).

Genomic studies have shown that large proportions of genes have been generated by gene duplication in all three domains of life. For example, about 40% of all genes in such diverse organisms as *Homo sapiens*, the molicute bacterium *Mycoplasma pneumoniae*, and the fruit fly *Drosophila melanogaster* belong to multigene families. The fraction of genes belonging to multigene families is much higher in some organisms (e.g., 65% in *Arabidopsis thaliana*) and much lower in others (e.g., 17% in the Gram-negative pathogen *Helicobacter pylori*). These figures are almost certainly underestimates because many duplicated genes have diverged so much that virtually no sequence similarity is found.

Lynch and Conery (2000) estimated that in eukaryotes a new duplicate gene arises and is fixed in the population at an approximate rate of one duplication per gene per 100 million years. This rate is the **gene birth rate**, and it was derived from a study of recent duplications. Many fixed duplicated genes later become pseudogenes and may subsequently be deleted from the genome. The rate of duplication that gives rise to stably maintained genes can be calculated by multiplying the gene birth rate by the **retention rate**, which is expected to fluctuate widely with gene function (Zhang 2003).

Modes of Evolution of Multigene Families

Our notions concerning the evolution of multigene families have evolved considerably in the last 50 years. The archetype of evolution of multigene families before 1970 was the globin superfamily of genes (see below). The genes belonging to this family are related by descent and have diverged gradually from one another as the different copies have acquired new gene functions. This mode of evolution may be referred to as **divergent evolution** (Figure 7.4a). Evidence for the divergent mode of evolution accumulated rapidly as scientists realized that proteins performing unrelated biological functions may, in fact, be related by descent. Particularly surprising were the findings of Lazure et al. (1981), according to which functionally disparate proteins, such as tonin (a submaxillary angiotensinogen), γ-subunit nerve growth factor, epidermal growth factor-binding protein, and several serine proteinases, are evolutionarily related to one another.

Around 1970, however, a number of groups discovered that ribosomal RNAs in *Xenopus* are encoded by a large number of tandemly repeated genes and that the nucleotide sequences of the intergenic regions of the genes are more similar within a species than between two related species. These observations could not be explained by the model of divergent evolution, and a new model called **concerted evolution**

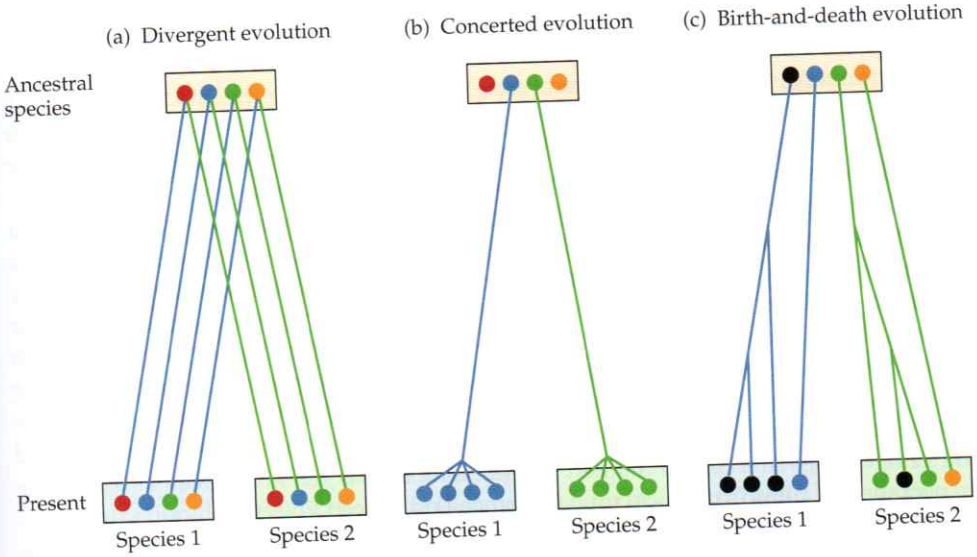


Figure 7.4 Schematic phylogenetic representation of (a) divergent, (b) concerted, and (c) birth-and-death evolution in multigene families. Colored circles represent functional genes; black circles signify pseudogenes.

was proposed (Figure 7.4b). In this model all the members of a gene family are assumed to evolve in a concerted manner rather than independently, and a mutation occurring in a repeat may spread through the entire group of member genes by repeated occurrences of unequal crossing over or gene conversion. It was subsequently discovered that the concerted evolution model may not be applicable to a substantial number of multigene families, and a third model, called **birth-and-death evolution**, was proposed. In this model, new genes are "born" by gene duplication, each being maintained in the genome for varying periods of time before "dying" through deletion or nonfunctionalization (Figure 7.4c).

Divergent Evolution of Duplicated Genes

Here we deal with the divergent evolution of duplicates created by complete gene duplication only. Divergent evolution refers to the fact that the duplicates evolve independently of one another. Usually such copies diverge in sequence, but in some cases, the copies remain similar to one another because of purifying selection due to functional constraints. How duplicated genes evolve varies from case to case (Figure 7.5). In this section, we discuss four general possibilities: (1) **nonfunctionalization**, (2) **retention of original function** or **gene conservation**, (3) **neofunctionalization**, and (4) four models of **subfunctionalization**. For a comprehensive discussion of these and other models pertaining to the evolution of function following gene duplication, see Innan and Kondrashov (2010).

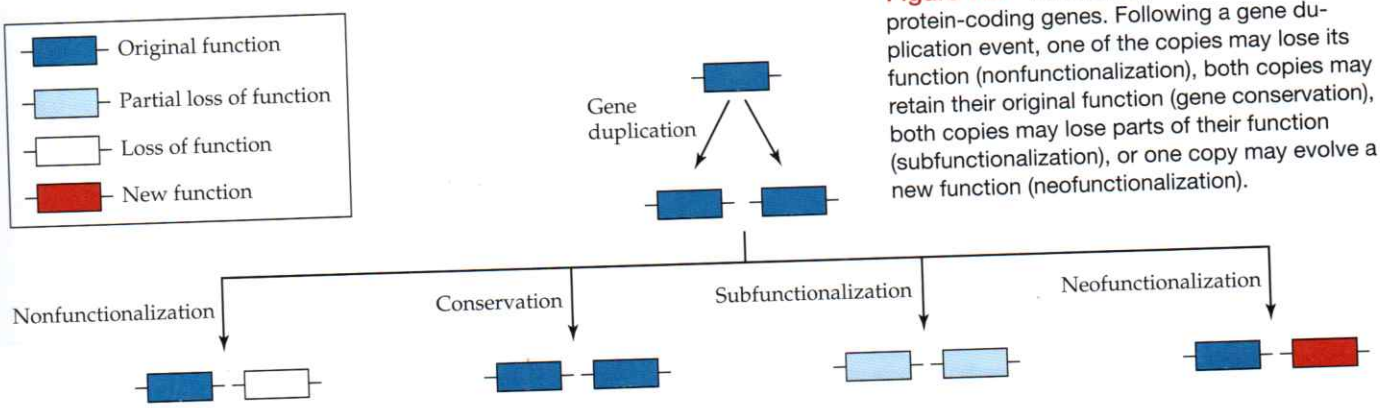


Figure 7.5 Evolutionary fates of duplicated protein-coding genes. Following a gene duplication event, one of the copies may lose its function (nonfunctionalization), both copies may retain their original function (gene conservation), both copies may lose parts of their function (subfunctionalization), or one copy may evolve a new function (neofunctionalization).

Nonfunctionalization and gene loss

The thousands of genetic diseases that have been documented in the medical literature (omim.org) and veterinary literature (omia.org) attest to the fact that mutations can easily destroy the function of a protein-coding gene. The vast majority of such mutations are deleterious and are either eliminated quickly from the population or maintained at very low frequencies by overdominant selection or genetic drift (Chapter 2). As noted by Haldane (1932), however, as long as there are other copies of a gene that function normally, a duplicate gene can accumulate deleterious mutations and become nonfunctional or be lost from the genome by deletion without adversely affecting the fitness of the organism. Indeed, because deleterious mutations occur far more often than advantageous ones, a redundant duplicate gene should be much more likely to become nonfunctional than to either retain its original function or evolve into a new gene (Ohno 1972). Interestingly, while nonfunctionalization is indeed the most likely fate of a redundant duplicate gene, the fraction of duplicated vertebrate genes that have remained functional in extant genomes greatly exceeds the expectation of Ohno's classical model (Prince and Pickett 2002).

The nonfunctionalization or **silencing** of a gene due to deleterious mutations produces a **nonprocessed** or **unprocessed pseudogene**, i.e., a pseudogene that has not gone through RNA processing (Chapter 9). The vast majority of nonprocessed pseudogenes are derived via the nonfunctionalization of duplicate copies of functional genes (duplicated pseudogenes). Some nonprocessed pseudogenes, such as $\psi\beta^x$ and $\psi\beta^z$ in the goat β-globin multigene family, have been derived from duplication of a preexisting pseudogene (Cleary et al. 1981), and a handful of nonprocessed pseudogenes have been derived from a functional gene without a prior duplication event (Chapter 8).

Table 7.2 lists the structural defects found in several globin pseudogenes. Most of these nonprocessed pseudogenes contain multiple defects such as frameshifts, premature stop codons, and obliteration of splicing sites or regulatory elements, so it has been difficult to identify the mutation that was the direct cause of gene silencing. In a few cases, identification of the "culprit" has been possible. For example, human $\psi\zeta$ contains only a single major defect—a nonsense mutation resulting in the premature termination of translation—that is probably the direct cause of its nonfunctionalization. In some cases, it has been possible to identify the mutation responsible for the nonfunctionalization of a gene through a phylogenetic analysis. For example, human pseudogene $\psi\beta$ in the β-globin family contains numerous defects, each of which could have been sufficient to silence it. The β-globin clusters in chimpanzee and gorilla, our closest relatives, were found to contain the same number of genes and pseudogenes

Table 7.2

Defects in nonprocessed globin pseudogenes[a]

Pseudogene	TATA box	Initiation codon	Frameshift	Premature stop codon	Essential amino acid	Splicing	Stop codon	Polyadenylation signal
Human $\psi\alpha 1$		+	+	+	+	+	+	+
Human $\psi\zeta 1$				+				
Mouse $\psi\alpha 3$	+		+	+		+		
Mouse $\psi\alpha 4$			+		+			
Goat $\psi\beta^x$	+		+	+	+	+	+	+
Goat $\psi\beta^z$	+		+	+	+	+	+	+
Rabbit $\psi\beta^2$			+	+	+	+		

Source: From Li (1983).

[a]A plus sign indicates the existence of a particular type of defect.

as in humans, indicating that the pseudogene was created and silenced before these three species diverged from one another. The three pseudogenes were found to have only three defects in common: either of two nonsynonymous substitutions in the initiation codon (ATG → GTA), a nonsense substitution in the tryptophan codon at position 15 (TGG → TGA), and a deletion in codon 20 resulting in a frameshift in the reading frame and a termination codon in the second exon (Chang and Slightom 1984). Thus, the "list of suspects" was reduced to three. Further studies showed that the same pseudogene exists in all primates and is, therefore, quite ancient. A comparison of the defects among all primate sequences showed that the initial mutation responsible for the nonfunctionalization of ψβ is the one in the initiation codon (Harris et al. 1984). It should be noted that most studies are biased by focusing on changes in the coding region; mutations in promoter regions that silence expression may go undetected.

Nonfunctionalization through missense mutations may not occur very often, because the production of a nonfunctional protein may be deleterious even if there are many functional copies of the gene that produce functional proteins. The reason is that a defective protein may be similar enough to the functional paralog to be incorporated into a biological structure but may interfere with the function of such structure. For example, there are dozens of chorion-coding genes in the genome of the silkworm *Bombyx mori*, yet if even one of them is rendered nonfunctional by a missense mutation, the entire eggshell becomes defective (e.g., Spoerel et al. 1989).

Because they are created by duplication, nonprocessed pseudogenes are usually found in the neighborhood of their paralogous functional genes. There are, exceptions, however; these are cases in which nonprocessed pseudogenes become dispersed as a consequence of genomic rearrangements (Chapter 11). For example, the α-globin cluster in the mouse is located on chromosome 11, and yet a nonprocessed α-globin pseudogene was found on chromosome 17 (Tan and Whitney 1993).

In genomes with large numbers of pseudogenes, nonprocessed pseudogenes are usually less frequent than processed pseudogenes (Chapter 9) but much more abundant than unitary pseudogenes (Chapter 8).

Nonfunctionalization time

The evolutionary history of a nonprocessed pseudogene is assumed to consist of two distinct periods. The first period starts with the gene duplication event and ends when the duplicate copy is rendered nonfunctional. During this period, the would-be pseudogene presumably retains its original function, and the rate of substitution is expected to remain roughly the same as it was before the duplication event. After the loss of function, the pseudogene is freed from all functional constraints and its rate of nucleotide substitution is expected to increase considerably. From the evolutionary point of view, it is interesting to estimate how long a redundant copy of a functional gene may remain functional after the duplication event. To estimate this **nonfunctionalization time**, the following method has been suggested (Li et al. 1981; Miyata and Yasunaga 1981).

Consider the phylogenetic tree in Figure 7.6. T denotes the divergence time between species 1 and 2, i.e., the time since the separation between the orthologous functional genes A and B; T_D denotes the time since duplication, i.e., the time of divergence between the functional gene A and its paralogous pseudogene ψA; and T_N denotes the time since the nonfunctionalization of pseudogene ψA. The numbers of nucleotide substitution per site at the ith position of codons (i = 1, 2, or 3) between ψA and A, between ψA and B, and between A and B are

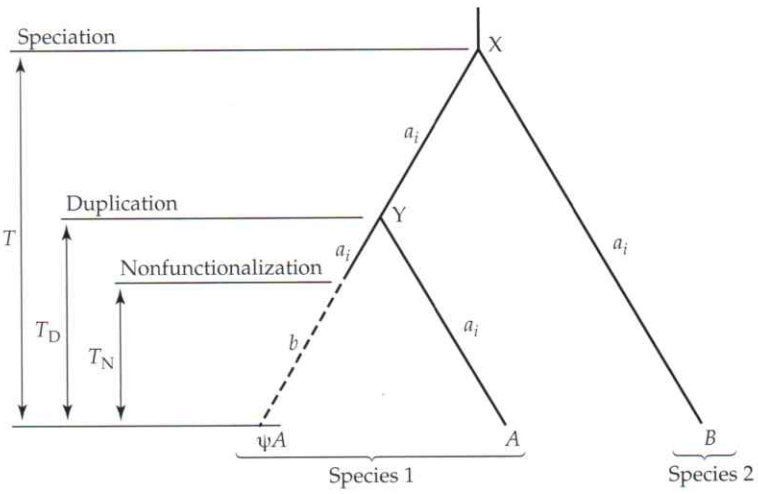


Figure 7.6 Schematic phylogenetic tree used to estimate the nonfunctionalization time of a nonprocessed pseudogene. T denotes the divergence time between species 1 and 2, T_D denotes the time since duplication of gene A, and T_N denotes the time since nonfunctionalization of pseudogene ψA. The term a_i denotes the rate of nucleotide substitution per site per year at the ith codon position in the functional genes, while b is the rate of substitution for the pseudogene. The node connecting the orthologous genes is denoted by X; the node connecting the paralogous genes is marked by Y. (Modified from Li et al. 1981.)

denoted as $d_{(\psi AA)i}$, $d_{(\psi AB)i}$, and $d_{(AB)i}$, respectively, and can be calculated directly from the sequence data (Chapter 3).

Let l_i, m_i, and n_i be the numbers of nucleotide substitutions per site at codon position i between points Y and ψA, Y and A, and Y and B, respectively. We then have

$$d_{(\psi AA)i} = l_i + m_i \tag{7.4a}$$

$$d_{(\psi AB)i} = l_i + n_i \tag{7.4b}$$

$$d_{(AB)i} = m_i + n_i \tag{7.4c}$$

Therefore, l_i, m_i, and n_i can be estimated by

$$l_i = \frac{d_{(\psi AA)i} + d_{(\psi AB)i} - d_{(AB)i}}{2} \tag{7.5a}$$

$$m_i = \frac{d_{(\psi AA)i} - d_{(\psi AB)i} + d_{(AB)i}}{2} \tag{7.5b}$$

$$n_i = \frac{-d_{(\psi AA)i} + d_{(\psi AB)i} + d_{(AB)i}}{2} \tag{7.5c}$$

In the following. we assume that the rates of substitution at a given codon position are equal in the functional genes A and B. We denote these rates by a_i, where the subscript i stands for the codon position. We also assume that once ψA became nonfunctional, i.e., once all functional constraints were obliterated, the rate of nucleotide substitution became the same for all three codon positions. We denote this rate by b. A reasonable expectation is that b would turn out to be much larger than a_1 and a_2, and possibly a little larger than a_3. From Figure 7.6, we obtain

$$d_{(\psi AA)i} = 2a_i T_D + (b - a_i) T_N \tag{7.6a}$$

$$d_{(\psi AB)i} = 2a_i T + (b - a_i) T_N \tag{7.6b}$$

$$d_{(AB)i} = 2a_i T \tag{7.6c}$$

If we know T, a_i can be estimated from Equation 7.6c by

$$a_i = \frac{d_{(AB)i}}{2T} \tag{7.7}$$

Let us denote the difference $d_{(\psi AB)i} - d_{(AB)i}$ as y_i. Note that

$$y_i = d_{(\psi AB)i} - d_{(AB)i} = bT_N - a_i T_N \tag{7.8}$$

Therefore, T_D can be estimated from Equations 7.6a and 7.8 by

$$T_D = \frac{\sum d_{(\psi AA)i} - \sum y_i}{2\sum a_i} \tag{7.9}$$

where Σ is the summation over i.

Two simple formulas for estimating T_N and b have been suggested by Li et al. (1981):

$$T_N = \frac{y_{12} - y_3}{a_3 - a_{12}} \tag{7.10}$$

$$b = \frac{a_3}{y_{12} - y_3} \tag{7.11}$$

where $y_{12} = (y_1 + y_2)/2$, and $a_{12} = (a_1 + a_2)/2$.

By setting the time of divergence, T, between mouse, rabbit, and human at about 80 million years ago, Li et al. (1981) estimated that the mouse globin pseudogene $\psi\alpha 3$ was created by gene duplication 27 ± 6 million years ago and became nonfunctional

23 ± 19 million years ago. Similarly, the human globin pseudogene ψα*1* was estimated to have been created by gene duplication 49 ± 8 million years ago and to have become nonfunctional 45 ± 37 million years ago. In both cases, it is statistically impossible to determine whether or not the nonfunctionalization time is different from 0. In general, it seems that those redundant duplicates that are ultimately destined to become pseudogenes retain their original function for only very short periods of time following the gene duplication event. Moreover, because gene duplication is a type of mutation that occurs with no regard for functional boundaries, a large fraction of duplicated genes may be "stillborn," i.e., they may be incomplete or may lack the proper regulatory elements necessary for expression, in which case the nonfunctionalization time should be 0.

Retention of original function following gene duplication

Rapid repeated rounds of gene duplication are typically referred to as **gene amplification**. The repeated unit of gene amplification is called an **amplicon**. Gene duplication and amplification produce identical copies of repeated units. With evolutionary time, however, these copies are expected to diverge from one another in sequence, structure, and function. Unexpectedly, in several cases, the copies maintain their sequence similarity to one another for long periods of time and are, therefore, invariant. In some cases, repetition of invariant sequences can be shown to be correlated with the synthesis of increased quantities of a gene product that is required for the normal function of the organism. Such repetitions are referred to as **dose** or **dosage repetitions**.

Selection for gene dosage can involve two different mechanisms. Selection for **increased dosage** involves a positive selection pressure to increase expression from a locus. The **dosage-compensation model**, on the other hand, invokes a negative selection pressure to retain the function and expression levels of both copies in order to preserve the correct stoichiometry—the appropriate amounts of the protein in relation to other proteins.

Selection for increased gene dosage is quite common when there is a metabolic need to produce large quantities of specific RNAs or proteins. For example, a duplication of the acid monophosphatase locus in yeast enables the carrier to produce twice the amount of enzyme, thus exploiting available phosphate more efficiently when phosphate is a limiting factor to growth (Hensche 1975). Perry et al. (2007) studied variation in the number of duplicates of human salivary amylase (*AMY1*) and found that human populations that consume starch-rich diets had on average more copies of *AMY1* per individual than populations relying less on starch in their diet. This translated into higher protein levels and an enhanced ability to break down starches. As these *AMY1* duplicates are very young and extremely similar or even identical in sequence, it is likely that they are maintained by selection, consistent with the dosage model.

There is now good evidence that an increase in gene number by gene amplification can occur quite rapidly under selection pressure for increased amounts of a gene product. For example, the genome of the wild-type strain of the peach-potato aphid (*Myzus persicae*) contains two genes encoding esterases E4 and FE4. The genes are very similar in sequence (98%), indicating that they have been duplicated recently (Field and Devonshire 1998). Following exposure to organophosphorous insecticides, which can be hydrolyzed and sequestered by esterase, resistant strains of *M. persicae* were found to contain multiple copies of *E4* and *FE4*. The gene duplications and subsequent increase in the frequency of the carriers of these duplications within the aphid population are likely to have occurred within the last 50 years, with the introduction of the selective agent. This is consistent with the finding that the individual copies of each duplicated gene, both within and between aphid clones, show no sequence divergence. Gene amplification has been a frequent strategy in the evolutionary response of insect populations to natural and synthetic pesticides (Mouches et al. 1986; Vontas et al. 2000; Li et al. 2007). A similar phenomenon has been observed