matic activity. Of course, *RNASE1B* can afford to loose its dsRNA activity, because its paralog, *RNASE1*, retained it. We thus end up with a case that can be described by a mixed evolutionary model of neofunctionalization and subfunctionalization.

Rates of Evolution in Duplicated Genes

Following a gene duplication event, the resultant duplicates will start to diverge, although occasionally they may be partially or wholly homogenized by gene conversion (p. 302). Since duplication creates redundancy, a commonly held view is that the rate of evolution will be accelerated by relaxation of functional constraint. Lynch and Conery (2000) conducted the first genome-wide analysis of duplicate genes, using data from five animals, two plants, and yeast, and found an elevated rate of nonsynonymous substitution in a large proportion of duplicated genes. This observation was confirmed by subsequent studies (Nembaware et al. 2002; Jordan et al. 2004; Scannell and Wolfe 2008) and has been interpreted to mean that duplicated genes are subject to weaker purifying selection than single-copy genes (Kondrashov et al. 2002).

It has also been observed that duplicated genes often exhibit dissimilar evolutionary rates, i.e., one copy evolves more slowly than the other (Van de Peer et al. 2001; Conant and Wagner 2003; Zhang et al. 2003; Brunet et al. 2006; Scannell and Wolfe 2008). This asymmetry has been taken as evidence supporting Ohno's (1970) model of neofunctionalization, whereby one duplicate (the slow copy) maintains the ancestral functional role and evolves at a rate similar to that of the ancestral gene, while the other copy evolves a novel beneficial function and undergoes rate acceleration due to positive selection (Kellis et al. 2004).

In yeast, Scannell and Wolfe (2008) discovered a sharp rate increase in protein sequence evolution affecting both copies. However, the symmetrical rate acceleration lasted for only a very short time following the duplication event; subsequently a rate-asymmetry pattern was commonly established, with one copy continuing to evolve at an elevated rate while the other reverted to the slower preduplication rate. The generality of these observations, however, remains to be tested in other taxa.

Rates and patterns of expression divergence between duplicated genes

Expression divergence between duplicated genes has long been considered an important step in the functional differentiation between duplicated genes (Markert 1964; Ohno 1970; Ferris and Whitt 1979). Indeed, investigations on tissue expression divergence among enzymes encoded by duplicated genes (isozymes) started soon after the introduction of protein electrophoresis (see Markert 1964). Such studies provided examples of differential tissue expression of duplicated genes and suggested a role for expression divergence in functional refinement and diversification. The availability of sequenced genomes and gene expression data has stimulated numerous studies, and general patterns in the relationship between sequence divergence and expression divergence are beginning to emerge.

Is expression divergence correlated with coding-sequence divergence? Dealing with this question requires two variables: a measure for sequence divergence, and a measure for expression divergence. As far as sequence divergence is concerned, one may use the number of substitutions per site between the two sequences (Chapter 3). For a measure of **expression dissimilarity** or **expression divergence**, one may use $1 - r$, where r is the Pearson correlation coefficient between the expression levels of two genes over different tissues or experimental conditions (Gu et al. 2002b).

In all organisms so far studied, a significant negative correlation was found between the two measures of divergence, indicating that expression divergence increases with sequence divergence and, hence, with time. Studies of duplicate genes in yeast, human, and *Arabidopsis* have revealed a rapid phase of initial expression divergence between duplicates followed by a plateau during which the measure of expression divergence remains relatively unchanged (Gu et al. 2002b; Makova and Li 2003; Blanc and Wolfe

2004; Casneuf et al. 2005; Gu et al. 2005; Yang et al. 2005; Ganko et al. 2008). It is not clear why a rapid divergence in expression between duplicates frequently occurs soon after the duplication event. One possible reason is incomplete duplication of *cis*-regulatory elements in one of the two genes (Katju and Lynch 2003). However, although this may explain the phenomenon as far as dispersed duplications or short tandem duplications are concerned, it is less likely to follow large segmental duplications, because a large duplicated segment would likely include all *cis* regulatory elements. Haberer et al. (2004) noted that in *Arabidopsis*, tandem and segmental duplicate gene pairs have divergent expression even when they share many similar *cis*-regulatory sequences, suggesting that changes in a small fraction of *cis* elements may be sufficient for expression divergence. Most intriguingly, despite the fact that whole-genome duplication should duplicate all regulatory elements, studies have shown that some genes are silenced immediately after allopolyploidization (see Adams and Wendel 2005). One reason why significant expression changes occur so soon after the formation of the polyploid is that transposable elements and other mobile sequences from one genome may invade the other genome in an allopolyploid (Wang et al. 2004).

There are three possible patterns of tissue-specific expression divergence following gene duplication: (1) both copies retain the ancestral pattern of expression, thereby creating redundancy; (2) the two copies diverge asymmetrically, whereby one gene is expressed in only a small subset of the tissues or conditions while its duplicate retains the ancestral pattern of expression; and (3) the two copies diverge complementarily, thus at least one descendant duplicate is expressed in each of the tissues in which the ancestral gene was expressed (Figure 7.14).

Data on the pattern of expression divergence in *Arabidopsis thaliana* indicate a preponderance of expression asymmetry over the other two possible patterns of expression divergence (Zhao et al. 2003; Casneuf et al. 2006; Duarte et al. 2006; Ganko et al. 2008). The asymmetric pattern was found in over 70% of gene pairs, whereas the complementary expression pattern was observed in only ~10% of the pairs. Asymmetric expression divergence was observed in many organisms, including yeast, *Drosophila*, rice, and whitefish. Thus, asymmetry in expression divergence between duplicate genes seems to be a universal phenomenon.

One possible reason for asymmetry is that once a *cis*-regulatory element is lost in one copy, leading to a lower expression level, then a similar mutation in the other copy would be deleterious and selected against and, hence, unobservable in natural populations. Papp et al. (2003b) found that the number of *cis*-regulatory elements shared between duplicated genes decreases over evolutionary time, though the total number of element types in the two genes remains relatively constant.

Conant and Wagner (2002) and Kim and Yi (2006) found that the K_A/K_S ratio was related to the degree of expression asymmetry between duplicated genes in yeast. However, the relative roles of positive selection, purifying selection, and relaxation of selective constraints in determining rates and patterns of expression divergence following gene duplication remain to be elucidated.

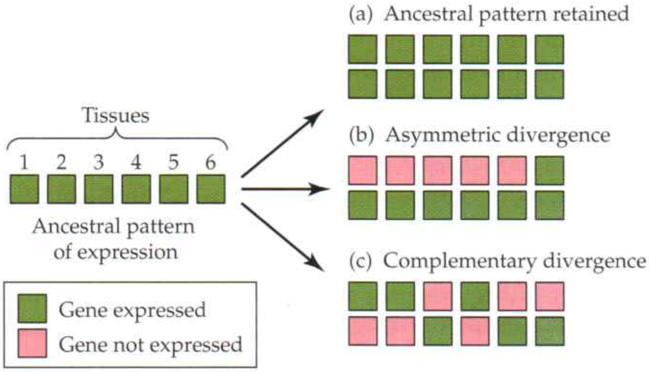
(a) Ancestral pattern retained
Tissues
1 2 3 4 5 6
(b) Asymmetric divergence
Ancestral pattern of expression
(c) Complementary divergence
Gene expressed
Gene not expressed

Figure 7.14 Three patterns of tissue-specific expression divergence following duplication of a gene that is expressed in six different tissues. (a) The two copies both retain the ancestral pattern of gene expression. (b) The two genes diverge asymmetrically, whereby one gene is expressed in only a small subset of the tissues, while its duplicate remains expressed in the original six tissues. (c) The two genes diverge complementarily, so no tissue-specific expression is lost. (Modified from Caseneuf et al. 2006.)

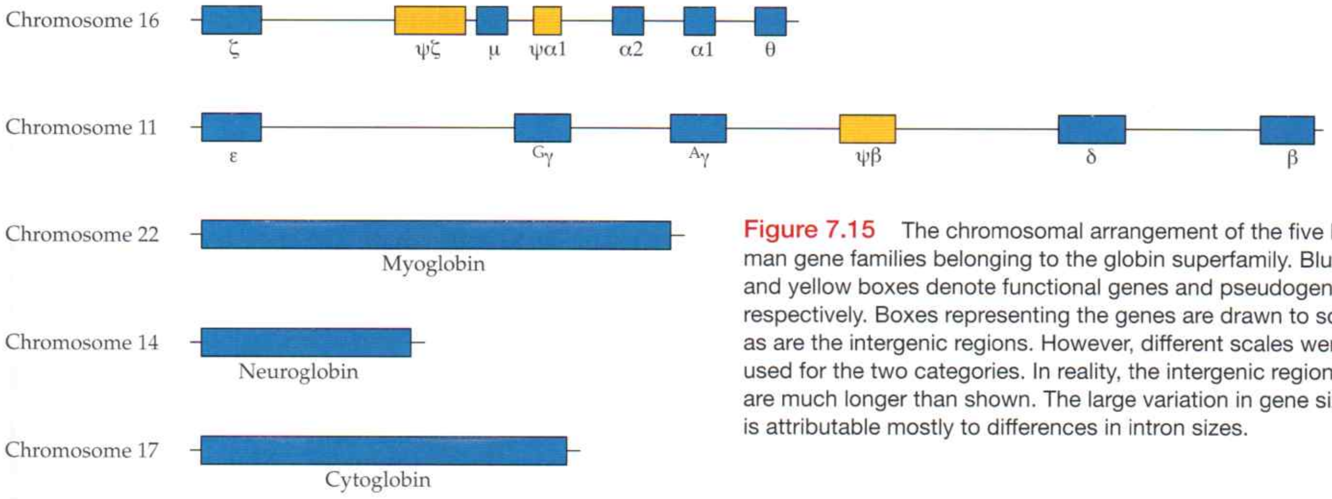

Figure 7.15 The chromosomal arrangement of the five human gene families belonging to the globin superfamily. Blue and yellow boxes denote functional genes and pseudogenes, respectively. Boxes representing the genes are drawn to scale, as are the intergenic regions. However, different scales were used for the two categories. In reality, the intergenic regions are much longer than shown. The large variation in gene sizes is attributable mostly to differences in intron sizes.

Human Globins

Here, we use the evolution of the human globin superfamily of genes as an illustration of the manner in which gene duplication creates working biological systems through the divergent evolutionary pathways described above.

In humans, the globin superfamily of genes consists of five chromosomal clusters, each with at least one functional member: the α-globin family on chromosome 16, the β-globin family on chromosome 11, and three single-gene families: myoglobin on chromosome 22, neuroglobin on chromosome 14, and cytoglobin on chromosome 17 (**Figure 7.15**).

Judging by the fact that globin and globin-like genes exist in both prokaryotes and eukaryotes, the globin superfamily must be very ancient in origin (Lecomte et al. 2005; Vinogradov et al. 2006). Phylogenetic studies indicate that the globin genes represented in the human genome originated more than 800 million years ago, long before mammals existed, probably even preceding the emergence of annelid worms (**Figure 7.16**). The first divergence event gave rise to neuroglobin, which in mammals is predominantly expressed in nerve cells. Neuroglobin is a monomer that reversibly binds oxygen with an affinity higher than that of hemoglobin. It is thought to protect neurons under hypoxic or ischemic conditions, potentially limiting brain

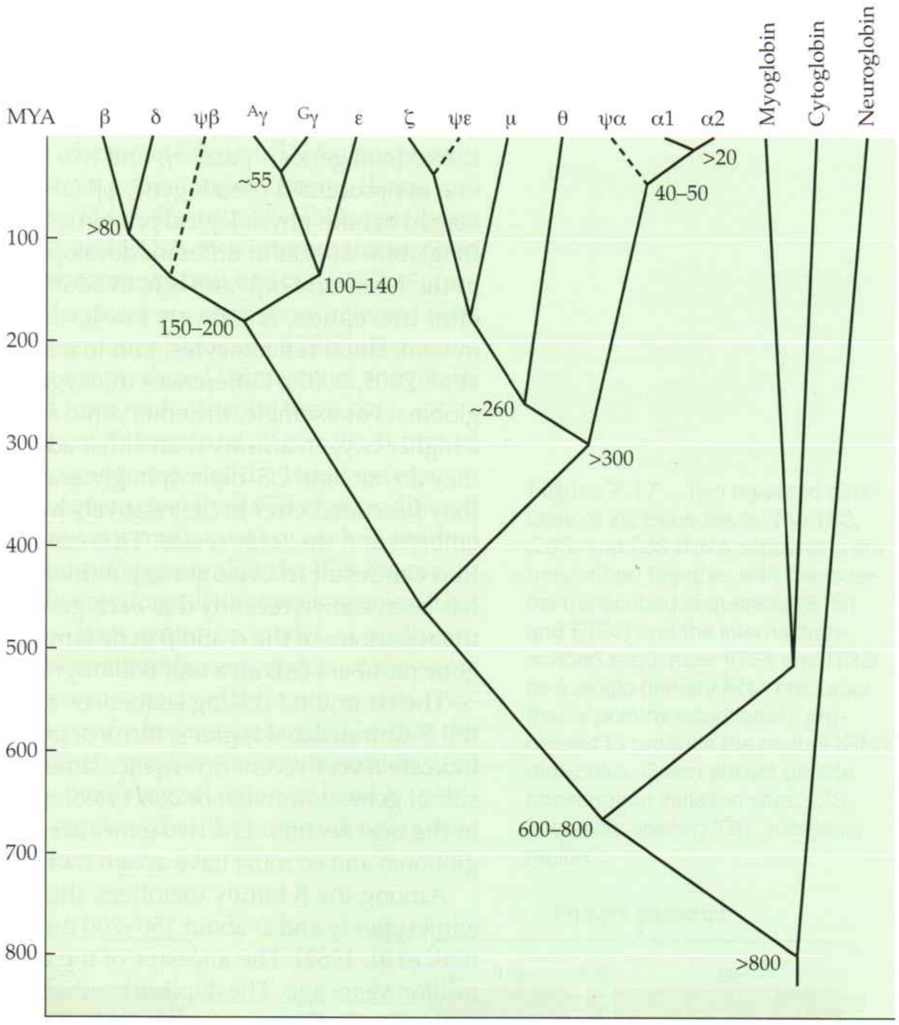

Figure 7.16 Evolutionary history of human globin genes. The broken lines denote pseudogene lineages.

damage. Evolutionarily, it is closely related to the nerve globins of invertebrates (Pesce et al. 2002).

The next divergence event gave rise to the ancestor of myoglobin and cytoglobin, on the one hand, and the ancestor of the α and β gene families on the other. This event occurred in the range of 600 to 800 million years ago. Myoglobin and cytoglobin diverged from each other quite early in vertebrate evolution, and both retained the plesiomorphic monomeric state. Cytoglobin is expressed in many different tissues. Its physiological roles are not completely understood. Although supplying cells with O_2 is a likely function, it is also possible that it acts as an O_2-consuming enzyme, an O_2 sensor, or a scavenger of nitric oxide. Myoglobin is the oxygen-storage protein in muscles, and its affinity for oxygen is higher than that of hemoglobin.

The α and β families diverged following a gene duplication 450–500 million years ago (Figure 7.16). Jawless fishes contain only one type of monomeric hemoglobin, while the vast majority of vertebrates have hemoglobin made up of two types of chains, one encoded by an α family member, the other by a β family member. Thus, the duplication that gave rise to the α- and β-globin families most probably also ushered in the era of tetrameric hemoglobin, the oxygen carrier in blood. This heteromeric structure allowed hemoglobin to acquire several novel capabilities that are absent in the monomeric globins. Among these are (1) binding of four oxygen molecules cooperatively, (2) responding to the acidity and carbon dioxide concentration inside red blood cells (the Bohr effect), and (3) regulating its own oxygen affinity through the level of organic phosphate in the blood.

In humans, the α-globin gene cluster is located on chromosome 16. It spans about 30 Kb and includes seven loci, for five functional genes and two pseudogenes: ζ, ψζ, μ, ψ*α1*, *α2*, *α1*, and θ. The β cluster consists of five functional genes: the embryonic gene ε; two fetal genes, $^{G}\gamma$ and $^{A}\gamma$; and two adult genes, β and δ. The family also contains one nonprocessed pseudogene, ψβ (also referred to as ψη). The two families have diverged in both physiological properties and ontological regulation. In fact, distinct hemoglobins appear at different developmental stages: $\zeta_2\varepsilon_2$ and $\alpha_2\varepsilon_2$ in the embryo, $\alpha_2\gamma_2$ in the fetus, and $\alpha_2\beta_2$ and $\alpha_2\delta_2$ in adults. The θ gene is mainly transcribed 5–8 weeks after conception, but at very low levels. The μ gene is transcribed quite abundantly in cord-blood reticulocytes, and to a lesser extent in adult-blood reticulocytes (Goh et al. 2005, 2007). Differences in oxygen-binding affinity have evolved among these globins. For example, the embryonic and fetal hemoglobins ($\zeta_2\varepsilon_2$, $\alpha_2\varepsilon_2$, and $\alpha_2\gamma_2$) have a higher oxygen affinity than either adult hemoglobin ($\alpha_2\beta_2$ and $\alpha_2\delta_2$), mainly because they do not bind 2,3-diphosphoglycerate as strongly as the adult forms. Consequently, they function better in the relatively hypoxic (low oxygen) environment in which the embryo and the fetus reside. This example illustrates once again how gene duplication can result in evolutionary refinements of physiological systems. In addition, it has been shown recently that each gene may produce more than one transcript; thus, the repertoire of the α and β gene families may be even greater than implied by mere gene numbers (Alvarez and Ballantyne 2009).

The *α1* and *α2* coding sequences are identical, but there are some differences in the 5′ untranslated regions, introns, and 3′ untranslated regions. This would seem to indicate a very recent divergence time. However, the similarity could also be the result of gene conversion or concerted evolution, a phenomenon that will be discussed in the next section. The two genes are present in humans and all the apes (including gibbons) and so must have arisen more than 20 million years ago.

Among the β family members, the adult types (β and δ) diverged from the nonadult types (γ and ε) about 150–200 million years ago (Efstratiadis et al. 1980; Czelusniak et al. 1982). The ancestor of the two γ genes diverged from the ε gene 100–140 million years ago. The duplication that created $^{G}\gamma$ and $^{A}\gamma$ occurred after the separation of the simian lineage (Anthropoidea) from the prosimians about 55 million years ago (Hayasaka et al. 1992). Soon after that, the ancestor of the two γ genes, which was originally an embryonically expressed gene, became a fetal gene. The change in temporal ontogenetic expression was brought about by sequence changes within a 4 Kb region surrounding the γ gene (TomHon et al. 1997). The divergence between the δ and β genes is estimated to have occurred before the eutherian radiation, more than 80 million years ago (Goodman et al. 1984; Hardison and Margot 1984).

We emphasize that the description above only applies to the evolution of globins currently found in the human genome. The evolution of the globin superfamily of genes in vertebrates constitutes a wonderful, but as yet only partially deciphered, narrative of gene and genome duplications, gene birth-and-death processes, convergence and divergence, and horizontal movements of globin genes from one locus to another (see Goh et al. 2005; Hardison 2007; Opazo et al. 2008a, 2008b; Patel et al. 2008; Hoffman et al. 2010; Patel and Deakin 2010).

Concerted Evolution

From the mid-1960s to the mid-1970s, a large number of DNA reannealing and hybridization studies were conducted to explore the structure and organization of eukaryotic genomes. These studies revealed that the genome of multicellular organisms is composed of highly and moderately repeated sequences as well as single-copy sequences (Chapter 11). They also revealed an unexpected evolutionary phenomenon, namely that the members of a repeated-sequence gene family are sometimes very similar to each other within one species, although members of the family from even fairly closely related species may differ greatly from each other. A fine illustration of this phenomenon was first provided by Brown et al. (1972) in a comparison of the ribosomal RNA genes from the African frogs *Xenopus laevis* and *X. borealis* (the latter being misidentified at the time as *X. mulleri*).

In these two *Xenopus* species, as well as in the vast majority of multicellular eukaryotes, the genes specifying the 18S and 28S ribosomal RNAs (rRNAs) are present in hundreds of copies and are arranged in one or a few tandem arrays (Long and Dawid 1980). Each repeated unit consists of the 18S, 5.8S, and 28S RNA-specifying genes, two **external transcribed spacers** (ETS1 and ETS2), two **internal transcribed spacers** (ITS1 and ITS2), and an **intergenic nontranscribed spacer** (IGS) (Figure 7.17). The IGS and the two external transcribed spacers are sometimes referred to as the **intergenic region** (IGR). The transcribed segment produces a 45S RNA precursor from which the functional ribosomal RNAs are produced by means of enzymatic cleavage. The transcribed repeats are separated from each other by the IGS.

In a comparison of the ribosomal RNA genes of *X. laevis* and *X. borealis*, Brown et al. (1972) found that, while the rRNA-specifying sequences of the two species were virtually identical to each other, the nucleotide sequences of their IGS regions differed by about 10%. In contrast, the IGS regions were very similar within each individual and among individuals within each species. Thus, it appears that the IGS regions in each species evolved together, although they diverged rapidly between species. This observation could not be explained by the divergent evolution model, according to which the differences in nucleotide sequence between different repeats of the same species are expected to be as large as those between repeats of different species (Figure 7.18). One simple explanation for the intraspecific homogeneity may be that the function of the repeats depends strongly upon their specific nucleotide sequence, so new mutations have been either eliminated by purifying selection or fixed by positive selection (Figure 7.19a). This explanation requires that the same advantageous

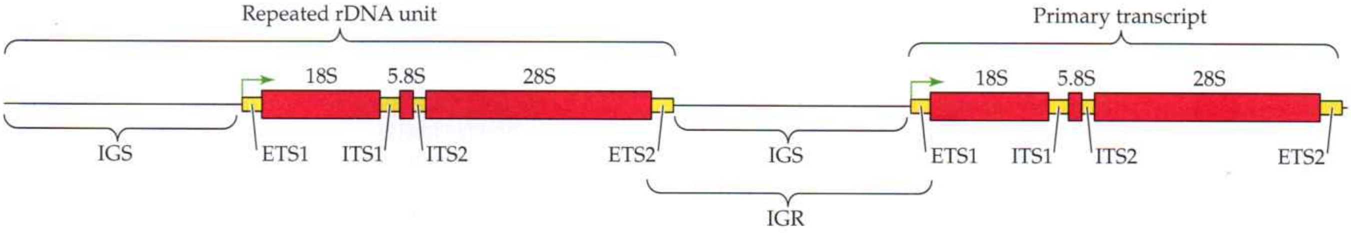

Figure 7.17 Two repeated rDNA units of *Xenopus laevis*. The 18S, 5.8S, and 28S rRNA sequences are transcribed together with the external transcribed sequences (ETS1 and ETS2) and the internal transcribed sequences (ITS1 and ITS2) as a single primary RNA precursor that is posttranscriptionally processed to produce the mature rRNA molecules. Green arrows denote transcription initiation sites. IGS, intergenic spacer; IGR, intergenic region.

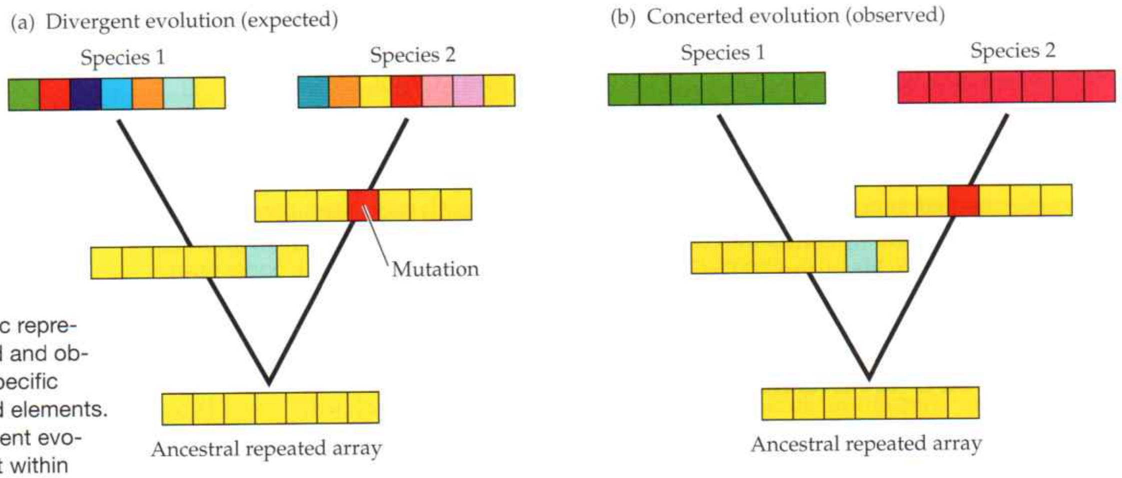

Figure 7.18 Schematic representation of the expected and observed patterns of intraspecific variation among repeated elements. (a) Under classical divergent evolution, each repeated unit within an array evolves independently. As a consequence, the similarity between any two randomly chosen units within a species is expected to be the same as that between two units randomly chosen from different species. (b) The observed patterns of intraspecific variation reveal a high degree of within-species homogeneity among repeated units. Differences from the ancestral sequence are shown in different colors.

mutation occur repeatedly in all copies of the gene. Moreover, the IGS regions have no known function and do not appear to be subject to stringent selective constraints, and yet are conserved within the species. Another simple explanation may be that the family has arisen from a recent amplification of a single unit (Figure 7.19b). In this case, the homogeneity would simply reflect the fact that there has not been enough time for the members of the gene family to diverge from one another. If this is the case, it is expected that the homogeneity of the family would gradually decrease because over evolutionary time mutations would accumulate in the family members through genetic drift, particularly in regions that are not subject to stringent structural constraints. Under this model of independent evolution, the degree of intraspecific variation among the repeated elements is expected to be approximately equal to the degree of interspecific variation.

The empirical data from *Xenopus* supports neither model. Rather, the intraspecific homogeneity among the repeated units seems to be maintained by a mechanism through which mutations can spread horizontally to all members in a multigene family. Brown et al. (1972) concluded that a "correction" mechanism must have operated to spread a mutation from one sequence to another faster than new changes can arise by mutation in these sequences (Figure 7.19c). They called this phenomenon, which

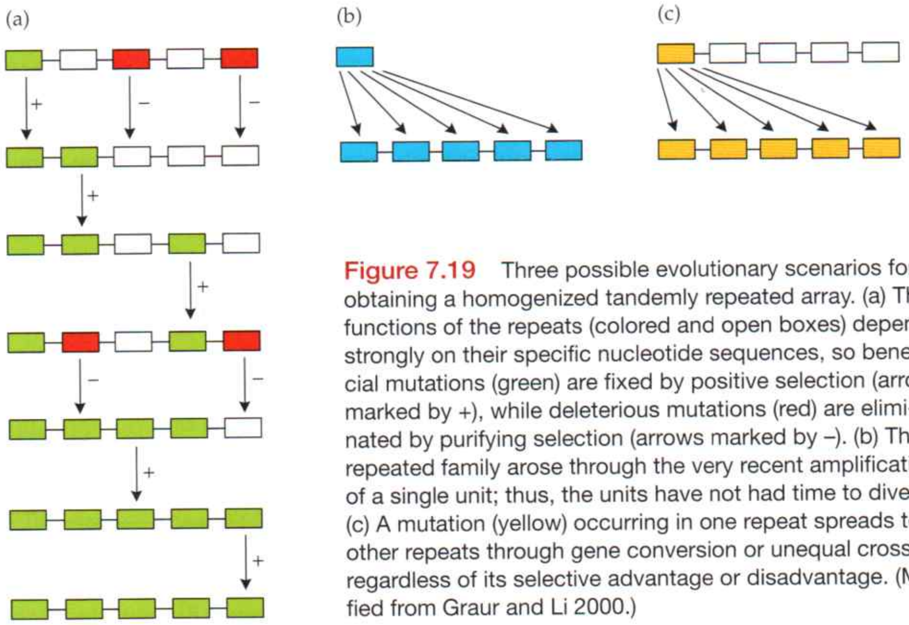

Figure 7.19 Three possible evolutionary scenarios for obtaining a homogenized tandemly repeated array. (a) The functions of the repeats (colored and open boxes) depend strongly on their specific nucleotide sequences, so beneficial mutations (green) are fixed by positive selection (arrows marked by +), while deleterious mutations (red) are eliminated by purifying selection (arrows marked by –). (b) The repeated family arose through the very recent amplification of a single unit; thus, the units have not had time to diverge. (c) A mutation (yellow) occurring in one repeat spreads to all other repeats through gene conversion or unequal crossover regardless of its selective advantage or disadvantage. (Modified from Graur and Li 2000.)

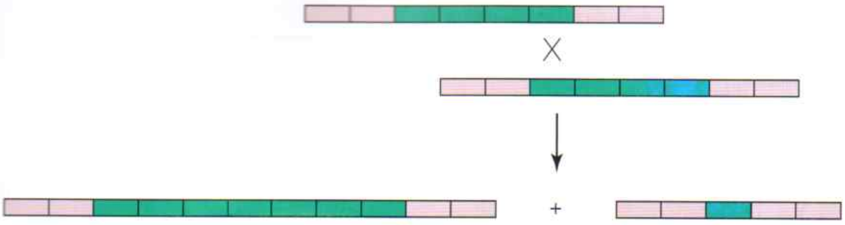

Figure 7.20 Model of the homogenizing process of repeated units in an array through unequal crossing over (×).The two repeat types have equal frequencies (50%) in the parental arrays. As a result of unequal crossing over, the repeat makeups of the two daughter arrays are more homogeneous than those of the parents (64% and 80%).

must originate in a single individual and later become fixed in the population, horizontal evolution, in contrast to vertical evolution, which refers to the spread of a mutation in a breeding population from one generation to the next. Many other terms have since been suggested in the literature; the term most commonly used nowadays is **concerted evolution** (Zimmer et al. 1980).

The term "concerted evolution" implies that an individual member of a gene family does not evolve independently of the other members of the family. Rather, repeats in a family exchange sequence information with each other, either reciprocally or nonreciprocally, so a high degree of intrafamilial sequence homogeneity is maintained. Through genetic interactions among its members, a multigene family evolves as a unit in a concerted fashion. The result of concerted evolution is a homogenized array of nonallelic homologous sequences. It is important to emphasize that concerted evolution requires not only the horizontal transfer of mutations among the members of the family (homogenization), but also the spread of the homogenized array of repeats in the population (fixation).

Gene conversion and unequal crossover are considered to be the two most important mechanisms responsible for concerted evolution. Although these two mechanisms have received the most extensive quantitative coverage in the literature and will be explained in some detail below, there are other mechanisms, such as slipped-strand mispairing and duplicative transposition, that can result in the creation of homogeneous families of repeated sequences.

Unequal crossing over

Unequal crossing over may occur either between the two sister chromatids of a chromosome during mitosis in a germline cell, or between two homologous chromosomes at meiosis. Unequal crossing over is a reciprocal recombination process that creates a sequence duplication in one chromatid or chromosome and a corresponding deletion in the other (Chapter 1). A hypothetical example in which an unequal crossover has led to the duplication of three repeats in one daughter chromosome and the deletion of three repeats in the other is shown in Figure 7.20. As a result of this exchange, the repeated arrays in both daughter chromosomes have become more homogeneous than those in the parental chromosomes. If this process is repeated, the numbers of each variant repeat on a chromosome will fluctuate with time, and eventually one type will become dominant in the array. Figure 7.21 illustrates how one type of repeat may spread throughout a gene family through repeated rounds of unequal crossing over. The process of concerted evolution by unequal crossing over has been investigated mathematically in detail and has received considerable experimental support (Smith 1976; Ohta 1984; Nei and Rooney 2005).

In the example of the rRNA genes in *Xenopus laevis* and *X. borealis*, the family size is large and it is conceivable that the number of genes may fluctuate with time without adverse consequences. Thus, unequal crossing over, rather than gene conversion, was apparently the driving force in the homogenization of the repeats.

Like unequal crossing over, slipped-strand mispairing is an expansion-contraction process that leads to the homogenization of the members of a tandem

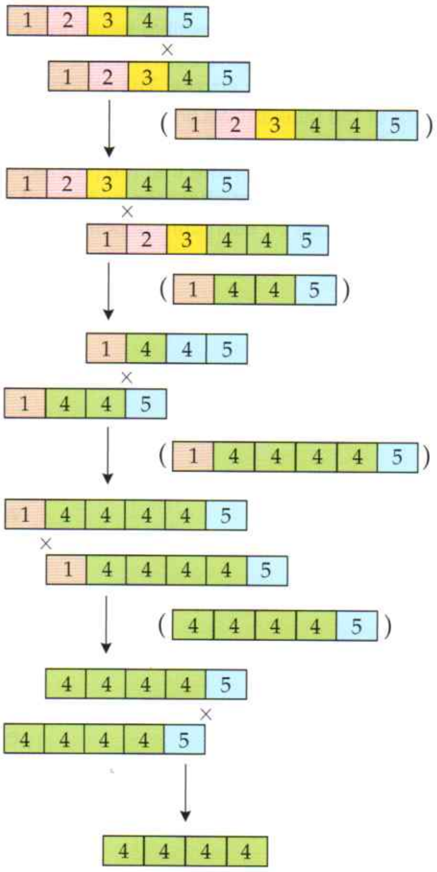

Figure 7.21 Concerted evolution by unequal crossing over (×). Repeated cycles of unequal crossover events cause the duplicated sequences on each chromosome to become progressively more homogenized. Different sequences are numbered and differently colored. The sequences in parentheses on the right are the ones selected for the next round of unequal crossover. The end result is a takeover by the "green" repeat (sequence 4). Note that unequal crossing over affects the number of repeated sequences on each chromosome. (Modified from Ohta 1980.)

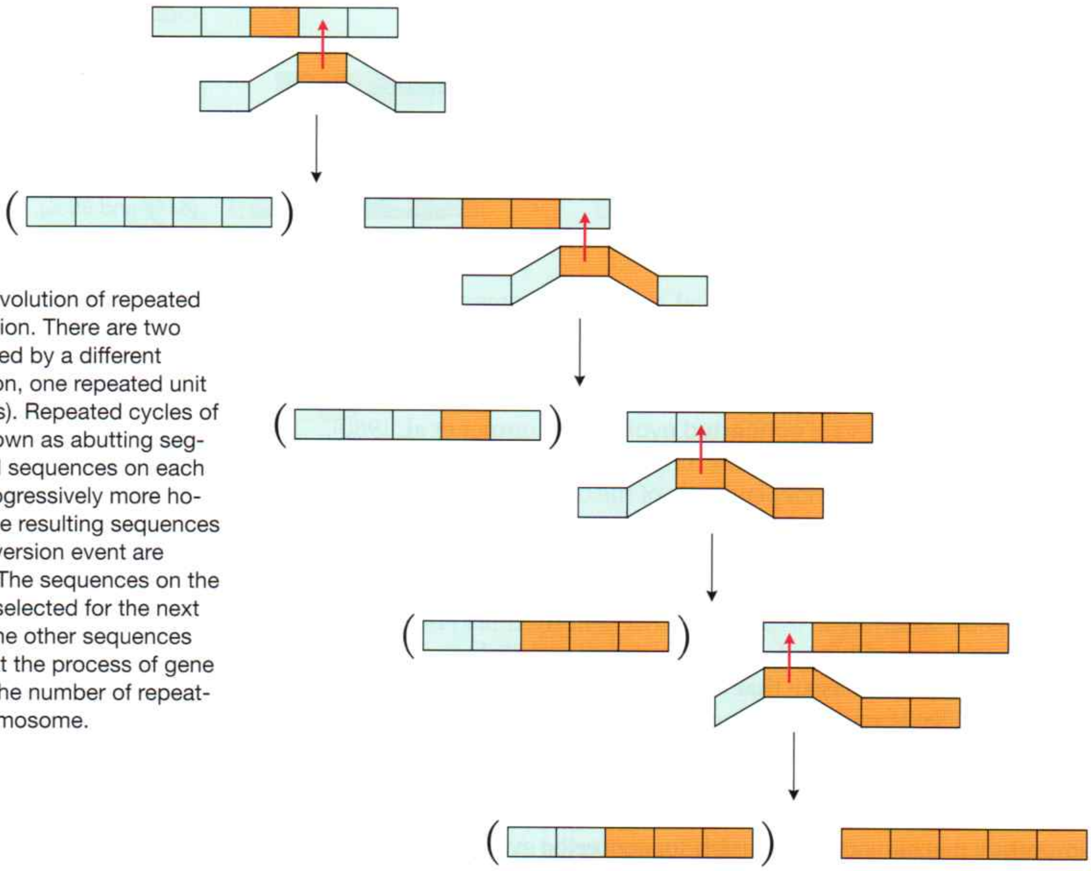

Figure 7.22 Concerted evolution of repeated sequences by gene conversion. There are two types of repeats, each marked by a different color. During gene conversion, one repeated unit converts another (red arrows). Repeated cycles of gene conversion events (shown as abutting segments) cause the duplicated sequences on each chromosome to become progressively more homogenized. The two possible resulting sequences from an unbiased gene conversion event are shown beneath the arrows. The sequences on the right of the blue arrows are selected for the next round of gene conversion; the other sequences are in parentheses. Note that the process of gene conversion does not affect the number of repeated sequences on each chromosome.

repeat family. As will be addressed in Chapter 11, while unequal crossover usually affects large tracts of DNA, slipped-strand mispairing is involved in the generation of tandem arrays of short repeats.

Gene conversion

Gene conversion is a nonreciprocal recombination process in which two sequences interact in such a way that one is converted by the other (Chapter 1). Although the mechanisms of gene conversion have not yet entirely been elucidated, several models of gene conversion have gathered support in the literature (see Chen et al. 2007). Theoretical studies (e.g., Ohta 1990) have shown that gene conversion can produce concerted evolution (Figure 7.22).

Based on the chromatids involved in the process, gene conversion can be divided into several types (Figure 7.23). When the exchange occurs between two paralogous sequences on the same chromatid, the process is called **intrachromatid conversion**. An exchange between two paralogous sequences from complementary chromatids is called **sister chromatid conversion**. **Classical conversion** involves exchanges between two alleles at the same locus. **Semiclassical conversion** involves an exchange between two paralogous genes from two homologous chromosomes. If the exchange occurs between paralogous sequences located on two nonhomologous chromosomes, the process is called **ectopic conversion**. From the viewpoint of the evolution of duplicated genes, the most important types of gene conversion are the **nonallelic conversions** (i.e., conversions between genes located at different loci and not between allelic forms).

Gene conversion may be biased or unbiased. **Unbiased gene conversion** means that sequence A has as much chance of converting sequence B as sequence B has of converting sequence A. **Biased gene conversion** means that the two possible direc-

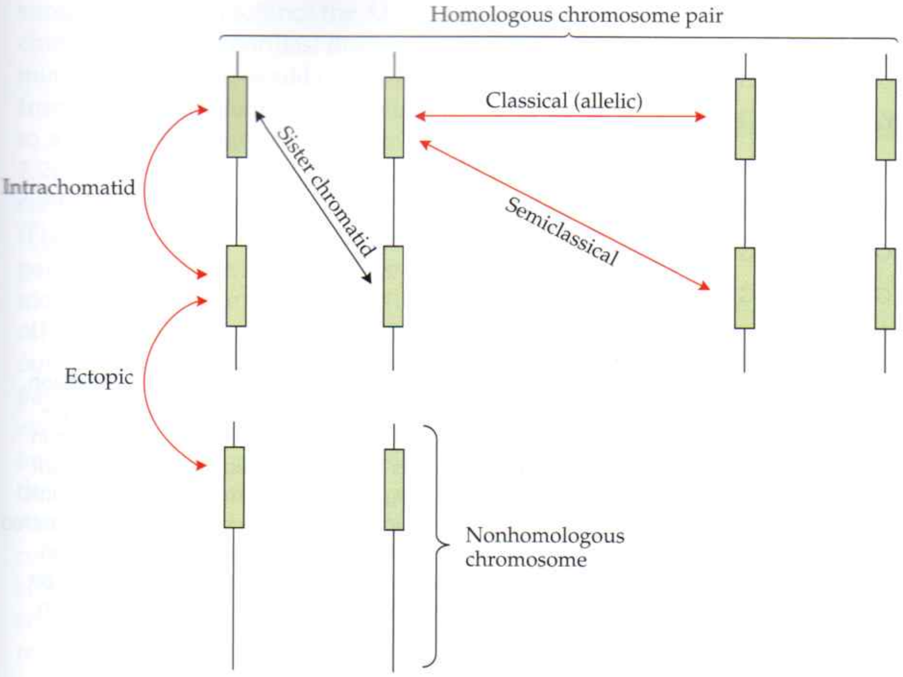

Figure 7.23 Types of gene conversion (red double-headed arrows) between repeated sequences (green boxes). The two chromatids of a chromosome are shown as neighboring pairs. Homologous double-stranded chromosomes are shown at the top; a nonhomologous chromosome is shown at the bottom. (Modified from Graur and Li 2000.)

tions of gene conversion occur with unequal probabilities. If deviation from parity occurs, we may speak of **conversional advantage** or **disadvantage** of one sequence over the other. Available data indicate that biased gene conversion is more common than the unbiased type. One particularly important type of biased gene conversion is **GC-biased gene conversion**, in which GC-rich sequences are favored over AT-rich ones (p. 304).

The amount of DNA involved in a gene conversion event varies from a few base pairs to a few thousand base pairs. Finally, the rate and the probability of occurrence of gene conversion vary with location; some locations are more prone to conversions than others.

DETECTING GENE CONVERSION EVENTS Many procedures for detecting putative gene conversion events have been put forward in the literature (see Drouin et al. 1999). In this section, we will briefly describe the method of Sawyer (1999), which asks whether or not a run of consecutive identical sites from among all polymorphic sites between two paralogous sequences is longer than would be expected by chance.

Consider the set of six aligned DNA sequences in Figure 7.24a. The length of the alignment is 50. We first identify the longest run of matched sites between any two sequences in the alignment. In our case, we find that sequences 2 and 4 have a run of 27 consecutive sites in common. Because we aim to identify identities due to gene conversion rather than extreme conservation due to stringent selective constraint or insufficient time for divergence subsequent to the duplication event, we next discard the 26 sites that are monomorphic in the six sequences. The condensed alignment is shown in Figure 7.24b. Now, the longest run of matched sites is 14 in an alignment of size 24. One way to calculate the statistical significance of this run is to use a runs test, which is a statistical procedure that examines whether or not a string of data occurs randomly given a specific distribution. The number of differences between sequences 2 and 4 in the condensed alignment is 6 out of 24, i.e., $p = 0.25$. Let us now use the runs test to calculate the probability of obtaining a run of matched sites of size 14 or longer starting at position 3. This probability is $0.25 \times (1 - 0.25)^{14} = 0.0045$. We note that there are 11 possible starting points for runs of length 14 or more in an alignment of length 24. Thus, the corrected probability of obtaining a run of length 14 or more is $0.0045 \times 11 = 0.05$. We now need to concern ourselves with the fact that

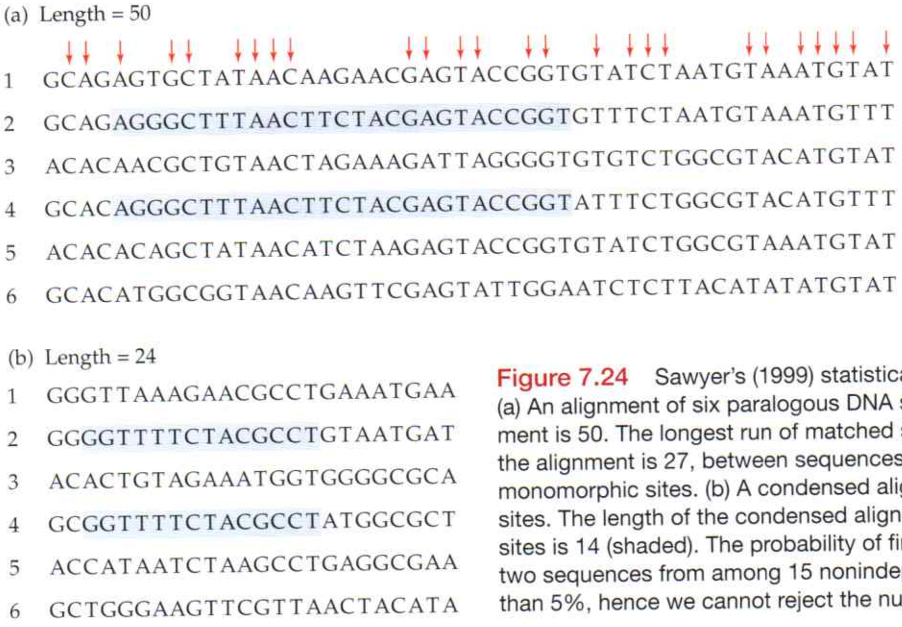
(a) Length = 50
1 GCAGAGTGCTATAACAAGAACGAGTACCGGTGTATCTAATGTAAATGTAT
2 GCAGAGGGCTTTAACTTCTACGAGTACCGGTGTTTCTAATGTAAATGTTT
3 ACACAACGCTGTAACTAGAAAGATTAGGGGTGTGTCTGGCGTACATGTAT
4 GCACAGGGCTTTAACTTCTACGAGTACCGGTATTTCTGGCGTACATGTTT
5 ACACACAGCTATAACATCTAAGAGTACCGGTGTATCTGGCGTAAATGTAT
6 GCACATGGCGGTAACAAGTTCGAGTATTGGAATCTCTTACATATATGTAT
(b) Length = 24
1 GGGTTAAAGAACGCCTGAAATGAA
2 GGGGTTTTCTACGCCTGTAATGAT
3 ACACTGTAGAAATGGTGGGGCGCA
4 GCGGTTTTCTACGCCTATGGCGCT
5 ACCATAATCTAAGCCTGAGGCGAA
6 GCTGGGAAGTTCGTTAACTACATA

Figure 7.24 Sawyer's (1999) statistical test for detecting gene conversion. (a) An alignment of six paralogous DNA sequences. The length of the alignment is 50. The longest run of matched sites between any two sequences in the alignment is 27, between sequences 2 and 4 (shaded). Red arrows mark monomorphic sites. (b) A condensed alignment after removing monomorphic sites. The length of the condensed alignment is 24. The longest run of matched sites is 14 (shaded). The probability of finding a run of 14 matches between two sequences from among 15 nonindependent comparisons is much larger than 5%, hence we cannot reject the null hypothesis of no gene conversion.

with six sequences there are 15 sequence-pair nonindependent comparisons. Thus, the true probability of obtaining such a run of matched sites by chance becomes larger than 5%. Thus, despite the fact that sequences 2 and 4 display a large run of sequence identity, we cannot rule out chance in explaining this identity, and there is no support for the gene conversion hypothesis.

PREVALENCE OF GENE CONVERSION How often do duplicate genes undergo gene conversion? Ezawa et al. (2006) examined 2,641 gene pairs in the mouse and rat genomes that were duplicated after the human-rodent split but before the mouse-rat split. They found strong evidence of gene conversion in 488 pairs (18%), the vast majority of which (407/488 = 83%) were linked on the same chromosome. Since detection of gene conversion has a low power, this represents an underestimate. Thus, gene conversion seems to occur with a high frequency between young duplicate genes.

GC-BIASED GENE CONVERSION There is evidence that in the vast majority of eukaryotes, gene conversion is biased, in that G and C nucleotides are favored over A and T, a phenomenon called **GC-biased gene conversion** (Galtier et al. 2001). GC-biased gene conversion is expected to increase the GC content of recombining DNA segments over evolutionary time, and it is considered a major contributor to the variation in GC content within and between genomes. GC-biased gene conversion has major methodological implications, for example, in detecting positive selection. Interestingly, GC-biased gene conversion can often be an antiadaptive force, by favoring GC-rich alleles (or paralogs) even if the AT-rich ones happen to be superior in fitness (Galtier et al. 2009).

Examples of gene conversion

In the examples described below, only two duplicate genes are involved, so concerted evolution must have occurred through gene conversion alone. In general, when the family size is small, unequal crossover events resulting in a number of repeats below a critical threshold may have a low fitness, so unequal crossover most probably does not play an important role in the concerted evolution of such gene families.

THE $^{A}\gamma$- AND $^{G}\gamma$-GLOBIN GENES IN THE GREAT APES An interesting case of concerted evolution by gene conversion involves the $^{A}\gamma$- and $^{G}\gamma$-globin genes, which were created by a duplication approximately 55 million years ago, after the divergence between prosimians and simians. Since the African great apes (humans, chimpanzees, and gorillas) diverged from one another at a much later date, we would expect the ${}^{G}\gamma$ orthologous genes from apes to be much more similar to one another than to any of the ${}^{A}\gamma$ paralogs. However, as shown in **Figure 7.25a**, this is only true for the 3′ part of the gene, which contains exon 3. The 5′ part, which contains exons 1 and 2 (Figure 7.25b), exhibits a different phylogenetic pattern, i.e., paralogous exons within each species resemble each other more than they resemble their orthologous counterparts in other apes (Slightom et al. 1985). This discrepancy is obvious when counting the nucleotide differences between two paralogous genes from the same species. In humans, for example, the 5′ parts of ${}^{A}\gamma$ and ${}^{G}\gamma$ differ from each other at only 7 out of 1,550 nucleotide positions (0.5%). In contrast, the 3′ part shows a difference that is 20 times larger, 145 out of 1,550 nucleotides (9.4%). Assuming that the 5′ and 3′ parts are subject to similar functional constraints, we may conclude that the 5′ end of the gene underwent gene conversion. This conclusion is strengthened by the fact that the second intron in both genes in all apes contains a stretch of the simple repeated DNA sequence $(TG)_n$ that can serve as a hotspot for the recombination events involved in the process of gene conversion.

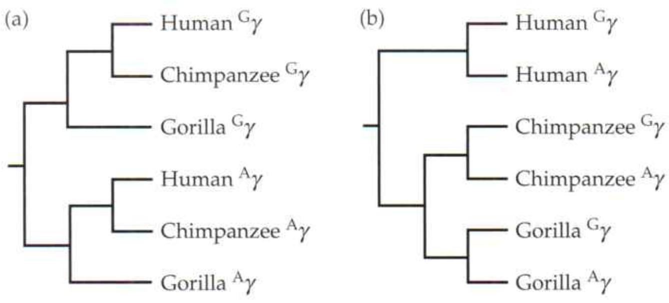

Figure 7.25 Phylogenetic trees for (a) exon 3 and (b) exons 1 and 2 of the ${}^{G}\gamma$- and ${}^{A}\gamma$-globin genes from human, chimpanzee, and gorilla. (Modified from Graur and Li 2000.)

Assuming that both the converted and unconverted parts of the genes evolve at equal rates, it is possible to date the last gene conversion event by using the degrees of similarity between the two sequences in conjunction with the date for the gene duplication event. The last conversion event that has been fixed in the human lineage occurred about 1–2 million years ago, i.e., after the divergence between human and chimpanzee, which indicates that the conversions in the chimpanzee and gorilla lineages occurred independently. Indeed, there are indications that gene conversion events between the 5′ parts of ${}^{A}\gamma$ and ${}^{G}\gamma$ genes occur quite frequently in human populations (e.g., Kalamaras et al. 2008).

GENE CONVERSION OF GENES AND PSEUDOGENES: WHEN DEATH IS NOT FINAL, LIFE IS PRECARIOUS, AND DISTINGUISHING BETWEEN THE TWO IS DIFFICULT

Pancreatic ribonuclease is a ubiquitous protein secreted by the pancreas in all vertebrates. In mammals, this protein is usually encoded by a single-copy gene. In the ancestor of true ruminants (suborder Pecora), the gene underwent two rounds of duplication, from which emerged three paralogous genes encoding pancreatic, seminal, and cerebral ribonucleases. Interestingly, a functional gene for seminal ribonuclease was only found in the closely related bovine species *Bos taurus* (cattle), *Bubalus bubalis* (Asian water buffalo), and *Syncerus caffer* (Cape buffalo), whereas in all other pecorans, such as deer and giraffe, the orthologous sequence was found to be a pseudogene. Even in *Tragelaphus imberbis* (the lesser kudu), which belongs to the same subfamily (Bovinae) as *Bos*, *Bubalus*, and *Syncerus*, the orthologous sequence is a pseudogene (Confalone et al. 1995; Breukelman et al. 1998).

One explanation for the data is that the gene was nonfunctionalized repeatedly and independently in numerous ruminant lineages within families Bovidae and Giraffidae (Sassi et al. 2007). The other most parsimonious explanation is that the original seminal ribonuclease gene in the ancestor of the true ruminants was a pseudogene that was subsequently "resurrected" in one lineage and became expressed in the seminal fluid, while in the other lineages, it stayed "dead" (**Figure 7.26**). Because of the taxonomic distribution of the seminal ribonuclease genes and pseudogenes, it is possible to date this resurrection at between 5 and 10 million years ago, after the divergence of the lesser kudu, but before the divergence of the Asian water buffalo.

A detailed analysis of the ribonuclease sequences indicated that the resurrection may have involved a gene conversion event, i.e., a transfer of information from the gene for the pancreatic enzyme to that for the seminal ribonuclease (Trabesinger-Ruef