10

HOMOLOGY IN MOLECULAR BIOLOGY

David M. Hillis

Department of Zoology
The University of Texas
Austin, Texas 78712

I. INTRODUCTION

This book is evidence that concepts of homology are diverse among biological disciplines. These different concepts often lead to confusion when biologists of different flavors attempt to talk among themselves. Confusion, however, exists within specific research areas as well; molecular biologists

Homology: The Hierarchical Basis of Comparative Biology

inheritance from a common ancestor or evolutionarily independent acquisition), *homoplasy* (similarity that arises through evolutionary convergence, parallelism, or reversal), and *analogy* (superficial similarity that arises through functional convergence).

Although similarity at some level seems a necessary prerequisite to recognize homology (Patterson, 1988), some authors have taken the concept of homology to its logical conclusion: homologous structures include any "parts that arise from the same source" (Ghiselin, 1976, p. 138) or "trace back to a single genealogical precursor" (Goodman *et al.*, 1987, p. 146). Under such a definition, structures may have diverged into such dissimilar parts that they are no longer recognizably similar, although are still homologous because of their common ancestral origin. Thus, the word homology is now used in molecular biology to describe everything from simple similarity (whatever its cause) to common ancestry (no matter how dissimilar the structures). I fall at the end of the continuum that relates homology to common ancestry, and use the word similarity to describe the likeness of structures (including molecular sequences).

None of the discussion above requires any consideration of molecular biology. This introduction merely serves as a point of departure for considering the interesting, and at times complex, ramifications of homologous relationships among genes, parts of genes, and the immediate products of genes. For the remainder of this article, I will consider two biological molecules to be homologues if they are descended (via imperfect replication) from a common ancestor.

II. CLASSES OF MOLECULAR HOMOLOGY

If we accept that most genes are evolutionarily related, and that extant genomes were derived by duplication, modification, and recombination of a small number (perhaps one?) of original replicating sequences, then it is also the case that most genes are at some level homologous. It is thus necessary to constrain the concept of homology for certain applications to make it useful. In the context of inferred evolutionary relationships among genes, we typically are interested in the most recent relationship shared by two given genes. Moreover, different classes of homology have been constructed to address different processes of divergence that generate homologous genes. The most obvious of these processes are

speciation (the divergence of lineages of organisms)
gene duplication (the divergence of lineages of genes within an organismal lineage)
horizontal gene transfer (the divergence of lineages of genes by transfer across different organismal lineages)

Each of these processes is of interest to molecular evolutionary biologists, and each results in genes that "trace back to a single genealogical precursor." However, all three processes usually are not studied simultaneously, so it is necessary to distinguish among the various kinds of homology when only one of these processes is of interest in a given analysis. Fitch (1970) and Gray and Fitch (1983) proposed the following names for homologous macromolecules, based on the different generative processes:

orthologous genes (or their products) are homologues that diverged as a result of a speciation event
paralogous genes (or their products) are homologues that diverged as a result of a gene duplication event
xenologous genes (or their products) are homologues that diverged as a result of lateral gene transfer.

If one is interested in reconstructing the phylogenetic history of taxa by inferring relationships among genes contained in those taxa, then it is usually necessary to examine orthologous genes (but see Section III below). If the history of gene duplication is of interest, then the genes examined need include paralogues. Study of lateral gene transfer obviously requires examination of xenologs. Although these points are obvious corollaries of the definitions, they sometimes are not appreciated by practicing biologists. The important distinction is whether history of the taxa, or history of the genes is of primary concern. Confusion of orthology with paralogy and xenology is likely to result in misleading inferences about organismal evolution.

The example in Figure 1 illustrates some possible effects of confusing orthologous and paralogous genes. The diagram shows a simplified representation of the evolution of some globin genes in four species of vertebrates: a lamprey, a frog, a mammal, and a snake. All of these genes can be traced back to an ancestral sequence, so at some level all are homologous. After the split (speciation event) that led to the lamprey lineage on the one hand and the three tetrapod lineages on the other, there was a gene duplication event in the lineage that led to the tetrapods: this duplication event gave rise to the α- and β-hemoglobin gene families. Therefore, all three tetrapods have both types of globin genes, whereas the lamprey has only one. If we wished to infer relationships among these species, we could do so by analyzing sequences of either β-hemoglobins or α-hemoglobins from the three tetrapods, together with the unduplicated hemoglobin sequence from the lamprey (Fig. 1b). However, if we were unaware that the gene duplication had occurred, and analyzed a mixture of α- and β-hemoglobins from the various tetrapods, the evolutionary relationships that we inferred would not be of the taxa, but of the genes (Fig. 1c). In other words, to reconstruct the speciation events, we have to select genes that are orthologous. If the gene duplication events are of interest, then the paralogues need to be examined as well.

The example presented above is greatly simplified; there have been at least three additional gene duplication events in the α-hemoglobin family and six additional gene duplication events in the β-hemoglobin family (Doolittle, 1987; Goodman *et al.*, 1979, 1987). Some of these duplicated genes are expressed at different times in the ontogeny of various tetrapods (e.g., some are larval or fetal in expression, whereas others are expressed only in adults). There obviously exists a continuum of levels of duplication of genes, so that few genes are truly "single copy" (in the sense that there are no other close paralogues). In fact, much genic diversity of both eubacteria and eukaryotes is thought to have arisen through gene (or whole genome) duplication (Grime and Mowforth, 1982; Herdman, 1985; Rees and Jones, 1972; Sparrow and Nauman, 1976), so paralogy probably is the rule rather than the exception.

Orthologous, paralogous, and xenologous molecules are all likely to be detected by the similarity of their sequences. Paralogy may be distinguished from orthology by the test of conjunction: whether or not the two homologues are found in the same individual (Patterson, 1988). Xenology is inferred if

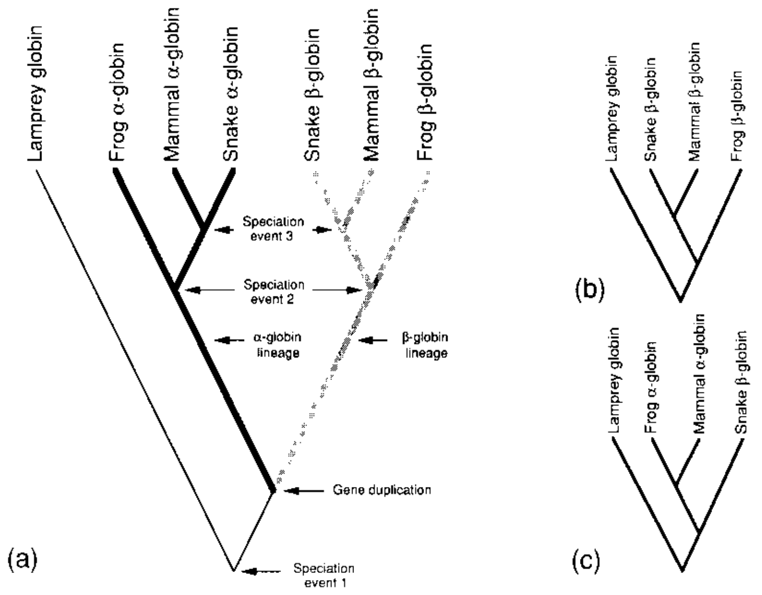


Fig. 1. (a) The origins of orthologous and paralogous globin genes in vertebrates. (b) The phylogeny of taxa can be inferred by analyzing either set of orthologous genes. (c) If a mixture of orthologous and paralogous genes is included in a phylogenetic analysis, the resulting tree will reflect gene relationships rather than relationships of taxa. (Simplified from Goodman *et al.*, 1987).

the homologues exist in distantly related species (i.e., information from the genes is highly incongruent with other information about phylogeny). Striking cases of xenology may result from retroviral transfer of genes, in which case the xenologous relationship may seem obvious. However, xenologous relationships (especially of alleles at a single locus) can also arise through hybridization, in which case xenology may be confused with convergence. In such cases, xenology of alleles may be inferred on the basis of the concentration of apparent convergence (across multiple loci) in particular branches of the inferred phylogenetic tree (see Buth, 1984; Duellman and Hillis, 1987).

III. CONCERTED EVOLUTION OF PARALOGOUS SEQUENCES

The above discussion assumes that duplicated genes will evolve independently following the duplication event. In fact, many duplicated genes continue to interact, so that evolution in the two (or more) duplicated sequences may not be independent. Sequences that are present in large numbers of tandem repeats rarely undergo independent evolution (see Arnheim, 1983; Dover, 1982, 1986; Ohta, 1980). Soon after sequences of tandem repeats were first studied, it was observed that the multiple copies of many repeated gene families were very similar within an individual and within a species, whereas the same families of repeated genes were often quite divergent among closely related species (Arnheim *et al.*, 1980; Brown *et al.*, 1972; Zimmer *et al.*, 1980). If the tandem duplications occurred in the ancestor of the two species, and the repeated sequences were evolving independently of one another, one would expect greater within-species than between-species divergence (Fig. 2). However, the observation was just the opposite: very little within-species, but great between-species divergence. It appeared that all the copies of the repeated sequences were evolving in concert; the phenomenon was termed *concerted evolution* (Zimmer *et al.*, 1980).

After concerted evolution of repeated DNA sequences was discovered, it was found to be nearly ubiquitous among mid- to highly repeated tandem gene families. In these families, concerted evolution often occurs so rapidly that relatively few differences can be detected among the copies (e.g., some nuclear ribosomal RNA genes; see Hillis and Dixon, 1991). The rate of concerted evolution among families of genes that are repeated at lower frequency is much more variable, from families that show few indications of concerted evolution (e.g., McElroy *et al.*, 1990; Shah *et al.*, 1983) to those in which the paralogous and orthologous relationships may be difficult to untangle (e.g., Doyle, 1991; Hughes and Nei, 1990; Irwin and Wilson, 1990). Furthermore, rates of concerted evolution may not be consistent within the same gene families among different taxa. In the case of ribosomal RNA genes, for instance, rates of concerted evolution may differ dramatically between tandemly repeated genes on a single chromosome versus sets

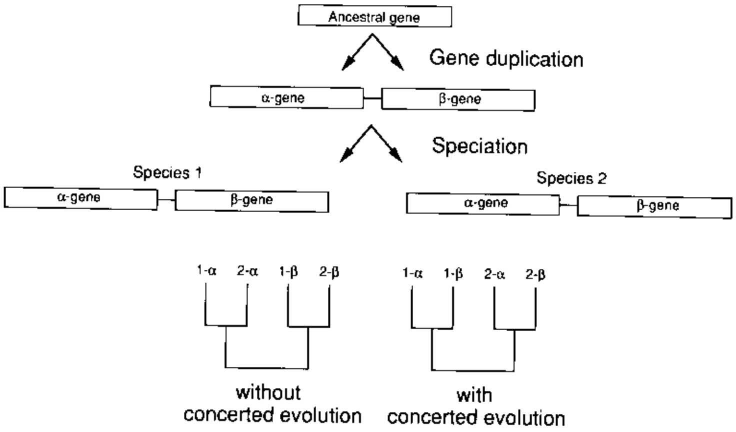


Fig. 2. The origin of a hypothetical set of paralogous sequences (α and β) through a gene duplication event, followed by a speciation event, producing two sets of orthologous relationships. In the absence of concerted evolution, the orthologues (α in species 1 and 2, or β in species 1 and 2) would appear to be more closely related to each other than to the paralogous sequences in the same species, because they share a more recent common ancestor (the original α- and β-genes, respectively). However, if the paralogous sequences are undergoing concerted evolution, then the paralogues are homogenized within each species, and the α-gene in species 1 will appear to be more closely related to the β-gene in species 1 than to either gene in species 2.

of repeats on different chromosomes. In some species, there may be clusters of ribosomal genes in which concerted evolution occurs, yet the different clusters may evolve independently (as in the clam genus *Corbicula*; D. M. Hillis, personal observation). In such cases, each cluster of ribosomal RNA repeats is evolving much like an orthologous locus, whereas relationships among the clusters appear paralogous.

Several mechanisms have been hypothesized to explain concerted evolution, of which two have received the most attention: unequal crossing-over (Coen *et al.*, 1982a; Ohta, 1980, 1983; Smith, 1974; Szostak and Wu, 1980) and gene conversion (Baltimore, 1981; Nagylaki, 1984; Nagylaki and Petes, 1982; Ohta, 1984; Ohta and Dover, 1983). Both mechanisms have received some empirical support as explanations for the concerted evolution of multigene families (Hillis *et al.*, 1991; Seperack *et al.*, 1988). Unequal crossing-over usually is considered a stochastic process, in which frequency of fixation of variants is directly related to the frequency at which the relevant mutations arise. Gene conversion, on the other hand, can be biased, such that a new variant sequence is favored over an ancestral sequence, and hence can spread rapidly through all copies of the repeated sequence. Biased gene conversion is thought to be the only mechanism that can adequately explain the rapid rate of homogenization seen in many families of genes that undergo concerted evolution (Coen *et al.*, 1982a,b).

If the rate of concerted evolution is high enough, an entire family of genes will evolve almost as if they were a single sequence present in many duplicate copies. Thus, paralogous sequences that have a high rate of concerted evolution behave like orthologous sequences: they show little divergence within species, but may evolve rapidly between species. For this reason, paralogous sequences evolving under high rates of concerted evolution can be used to infer phylogenetic relationships of taxa, without fear of reconstructing gene duplication events rather than speciation events (Sanderson and Doyle, 1992). Patterson (1988) suggested the term *plerology* to describe the relationship among paralogous sequences homogenized within taxa as a result of concerted evolution.

How high does the rate of concerted evolution have to be before the distinction between orthology and paralogy becomes blurred? Sanderson and Doyle (1992) simulated the effects of concerted evolution and found that when 70% of sites underwent concerted evolution between speciation events, the inferred trees always represented the correct (simulated) relationships among the taxa rather than the genes. In order to correctly infer gene trees among paralogues, concerted evolution had to involve fewer than 10% of the sites between speciation events in the simulations. At intermediate rates of concerted evolution, the inferred trees were likely to confound paralogous and orthologous relationships. Although these values are somewhat dependent on the details of the simulations, they represent a first-order approximation of relative rates of concerted evolution that are likely to confound the inferred relationships of paralogous and orthologous genes.

IV. PARTIAL HOMOLOGY OF MOLECULES: EXON SHUFFLING

Earlier in this chapter, I was critical of those who would say that two sequences are "50% homologous" when they mean the sequences share 50% of their aligned sites. However, that does not mean that homology must be an all-or-none condition, if the units of comparison are whole genes or their products. *Partial homology* is possible because some proteins have evolved through recombination of functional modules, which often correspond to the exons of genes (Fig. 3). For instance, the gene for tissue plasminogen activator is made up of exons that appear to have been captured from the genes for plasminogen, fibronectin, and epidermal growth factor (Patthy, 1985). Therefore, the corresponding functional modules are paralogous among the proteins, and the proteins (as well as their genes) are each partially paralogous. In addition to paralogy of modules among these proteins, are several cases of within-protein module paralogy (Fig. 3). The Kringle module present in two copies in tissue plasminogen activator is present in five tandem copies in plasminogen. Likewise, the finger module of fibronectin and the growth factor module of epidermal growth factor each are repeated many times (Fig. 3).

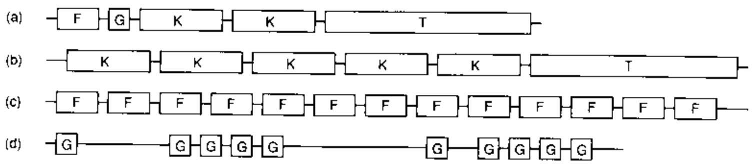


Fig. 3. Partial paralogy of some genes of proteins involved in blood coagulation and fibrinolysis. Functional modules are indicated by boxes: F, finger module; G, growth-factor module; K, Kringle module; T, trypsin-like module. (a) Tissue plasminogen activator protein. (b) plasminogen. (c) fibronectin. (d) epidermal growth factor. [Adapted from Patthy (1985) and Li and Graur (1991).]

Within-gene paralogy also can arise through slippage replication or unequal crossing-over of repeated units within genes (Hancock and Dover, 1988, 1990). Thus, genes may have regions that are internally paralogous. In the case of nuclear ribosomal RNA genes, two coevolved sets of paralogous regions may arise as a mechanism for maintaining secondary structure of the mature ribosomal RNA (Hancock and Dover, 1990).

V. POSITIONAL HOMOLOGY AND SEQUENCE ALIGNMENT

So far, I have been discussing homology of whole genes (or their products), or whole domains of genes. However, for most evolutionary analyses of sequence data, it is necessary to consider homology at a finer level: that of a single nucleotide site (or an amino acid site in the case of protein sequences). This is called *positional homology*. In a phylogenetic analysis of DNA sequences, for instance, the characters are nucleotide positions and the character states are the different nucleotides. When two orthologous genes are compared, the units of comparison typically are not the entire genes but the individual nucleotide positions. If all evolution has been through substitution, so that the two genes are exactly the same length, then it usually is a simple matter to align the homologous sites for analysis (Fig. 4a). However, when insertions or deletions have occurred in one or both sequences, gaps need to be inserted to preserve positional homology (Fig. 4b and c).

Alignment of sequences requires explicit and objective rules if inferences of positional homology are to be robust. If gaps are added without penalty to the alignment score, then unrelated (nonhomologous) sequences could be aligned without difficulty. However, if the penalty for gaps is too high, then the inferred positional homology is likely to be mistaken. Figure 5 shows a set of aligned ribosomal RNA genes in which the inferred positional homology differs depending on the weight assigned to gaps. If alternative alignments are equally good (or nearly so) for a given region, the ambiguous region should be excluded from analyses that assume accurate inference of positional homology, such as phylogenetic analyses (Swofford and Olsen, 1990).

```
    ATTCCGTAGCTGTTTCATCTTGATTGGTAACTG
(a) ||| ||||| ||||||||| ||||||||| |||
    ATTGCGTAGATGTTTCATCATGATTGGTATCTG

    ATTCCGTAGCTGTTTCATCTTGATTGGTAACTG
(b)   |       ||||||||| ||||||||| |||
     ATTGCGTAGTGTTTCATCATGATTGGTATCTG

    ATTCCGTAGCTGTTTCATCTTGATTGGTAACTG
(c) ||| ||||| ||||||||| ||||||||| |||
    ATTGCGTAG-TGTTTCATCATGATTGGTATCTG
```

Fig. 4. (a) Alignment of two homologous sequences is straightforward if all differences are the result of substitutions. (b) If two sequences are different in length, high apparent dissimilarity may exist if gaps are not allowed. (c) The same sequences as in (b), but with positional homology restored by the addition of a single gap (corresponding to an hypothesized deletion in the lower sequence).

Ideally, the weight of the penalty chosen for gaps should reflect the relative probability of insertion/deletion events relative to substitution events. The probability of insertion/deletion events varies greatly among different genomic regions. Such events are relatively rare in most protein-coding regions (because they usually lead to highly deleterious frameshift mutations), but are quite common within loop regions of ribosomal RNA genes and in many noncoding sequences. In regions where insertion/deletion events are common, accurate inferences of positional homology are highly unlikely in any but the most closely related sequences.

A number of secondary criteria (beyond sequence similarity) are often used to determine positional homology. In protein-coding genes, the translated amino acid sequence usually is more conserved than the encoding nucleotide sequence (because of the redundancy of the code), and often is of use for determining nucleotide alignment. Furthermore, because of the deleterious nature of frameshift mutations in protein-coding genes, it is reasonable to assign *a priori* heavy penalties (or disallow) gaps that do not correspond to multiples of codons (i.e., three nucleotides). If secondary structure is known for the gene product, alignments may be based on conserved secondary structure rather than conserved primary structure (i.e., sequence similarity).

```
                      22|00              22|20              22|40
Mus           GTCAGCCAGGACTCTCTACCCGCTCACGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rattus        GTCAGCCAGGACTCTCTACCCGCTCACGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Homo          GTCAGCCAGGACTCTCTACCCGCTCGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rhineura      GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Cacatua       GTCAGCCAGGATTCGCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Xenopus       GTCAGCCAGGATTCTCTACCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAGC
Rhyacotriton  GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Typhlonectes  GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Latimeria     GTCAGCCAGGATTCTCTACCCGCTTGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAGC
Cyprinella    GTCAGTCCAGGATTCCTACCCGCTGGCGGTCAAGCCTTCCCTCCGGCTACCGGAGGCAGC
                   * ** ** **   *     **   ** ** * *  **  *            *
```

```
                      22|00              22|20              22|40
Mus           GTCAG-CCAGGACTCTCTACCCGCTCACGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rattus        GTCAG-CCAGGACTCTCTACCCGCTCACGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Homo          GTCAG-CCAGGACTCTCTACCCGCTCGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rhineura      GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Cacatua       GTCAG-CCAGGATTCGCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Xenopus       GTCAG-CCAGGATTCTCTACCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAGC
Rhyacotriton  GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Typhlonectes  GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Latimeria     GTCAG-CCAGGATTCTCTACCCGCTTGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAGC
Cyprinella    GTCAGTCCAGGATTC-CTACCCGCTGGCGGTCAAGCCTTCCCT-CCGGCTACCGGAGGCAGC
                   *      *  *   *      **   *   *      *               *
```

Fig. 5. Alternative alignments of a segment of 28S ribosomal RNA genes in selected vertebrates. The upper set of sequences is aligned without gaps; there are 20 variable positions. Four gaps have been introduced into the lower set of sequences, thereby reducing the number of variable positions to 10 (including those with gaps). The alternative accepted depends on the weight of the penalty assigned to gaps. Sequences from Gonzales *et al.* (1985), Hadjiolov *et al.* (1984), Hassouna *et al.* (1984), Hillis and Dixon (1989), Hillis *et al.* (1990a), Larson and Wilson (1989), and Ware *et al.* (1983). (Adapted from Hillis *et al.* 1990b.)

Some authors claim that high similarity of aligned sequences never arises by convergence:

> in terms of nucleotide sequences, there seems to be no equivalent of convergence, or close similarity produced by evolution from different precursors (Goodman *et al.*, 1987, p. 147).

However, several cases of convergence of sequences are known, and more cases will undoubtedly come to light as molecular biologists begin to look for the phenomenon. For instance, the lysozymes of ruminants and colobine monkeys appear to have converged as a result of parallel development of foregut fermentation in these mammalian groups (Stewart and Wilson, 1987; Swanson *et al.*, 1991). Although these proteins are homologous as lysozymes, the high degree of sequence similarity clearly is convergent. (This convergence is analogous to the oft-cited case of convergence between bird and bat forelimbs: the structures are homologous as forelimbs, but convergent as wings.) Other cases of sequence convergence are likely to arise as a result of GC or AT biases (Aoki *et al.*, 1981; Hori and Osawa, 1986; Tamura, 1992; Wilson *et al.*, 1980), the presence of simple tandem repeats (e.g., Levinson *et al.*, 1985), or because of convergent mutational spectra (Muto and Osawa, 1987; Singer and Ames, 1970).

VI. HOMOLOGY IN INDIRECT (NONSEQUENCE) MOLECULAR TECHNIQUES

A. *DNA Hybridization*

Homology is an important concept for nonsequence molecular data as well. In DNA-DNA hybridization, low copy portions of the complete genomes of two taxa are annealed to form hybrid duplexes, and the average similarity of the genomes is measured as a function of melting temperature of the hybrid strands. Melting temperature (temperature at which 50% of the hybrid duplexes separate into single strands, or a similar standard of comparison; see Werman *et al.*, 1990) is directly related to the number and distribution of hydrogen bonds that form between complementary base pairs, so hybrid duplexes with few mismatches melt at a higher temperature than hybrid duplexes with many mismatches. Hybrid duplexes form between any two sequences with high levels of sequence similarity.

From the description above, it should be clear that DNA hybridization provides a measure of average sequence similarity of cross-hybridizing sequences, which undoubtedly include paralogous as well as orthologous genes. Moreover, sequence convergence at individual sites (homoplasy) cannot be directly distinguished from sequence similarity due to common ancestry (homology) by this technique (Bledsoe and Sheldon, 1990). The melting temperature thus is confounded by similarity due to orthology, paralogy, and homoplasy. Nonetheless, some proponents of the technique claim that DNA hybridization "solves the difficulty of determining homology," although the "solution" amounts to defining homology as similarity and ignoring the distinction between orthology and paralogy:

> Fortunately, DNA-DNA hybridization data are immune to convergence, because the conditions of the experiments preclude the formation of heteroduplexes between non-homologous sequences. To form a stable duplex DNA molecule at 60°C, 80 per cent of the bases in the two strands must be correctly paired, and *only homologous sequences* have this degree of complementarity. This solves the problem of homology and thereby eliminates the possibility of convergence. (Sibley and Ahlquist, 1987, p. 100; emphasis as in original)

Other proponents of this technique (e.g., Bledsoe and Sheldon, 1990; Werman *et al.*, 1990) have a more realistic view of the problems of determining homology in DNA-DNA hybridization studies.

B. Restriction Enzyme Analysis

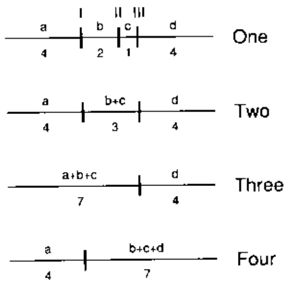
In restriction enzyme analyses, the targeted DNA may be entire genomes (e.g., Hillis *et al.*, 1992), an isolated organellar component (typically mitochondrial or chloroplast DNA; Moritz *et al.*, 1987; Palmer *et al.*, 1988), or a specific gene or gene region. In one common method of analyzing a specific gene, putatively homologous regions are examined by cross-hybridizing a cloned or otherwise isolated gene (called a probe) to genomic DNA that has been cleaved with a restriction enzyme, separated electrophoretically, and then attached to a support membrane (Southern, 1975). The hybrid probe-target heteroduplexes are visualized by autoradiography or its equivalent (Dowling *et al.*, 1990). As with DNA-DNA hybridization, this procedure relies on the low likelihood of hybridization between nonhomologous sequences. Hybridization between closely related paralogous sequences is commonly observed, but the presence of paralogous products usually can be detected on the basis of the presence of multiple regions of cross-hybridization. Target genes may be chosen to avoid complications of paralogy by process of elimination (Friedlander *et al.*, 1992).

Unfortunately, homology has taken on yet another meaning in the description of probes used in Southern blotting and similar procedures. A probe is often called "homologous" if it is used to study the same species from which it was derived, and "heterologous" if the probe was cloned from another species. This is somewhat confusing, since heterologous probes are used to examine presumably homologous genes in different species. To avoid confusion, the more appropriate terms *homospecific* and *heterospecific* can be used in place of *heterologous* and *homologous* to describe these relationships between probe and target.

An alternative approach for isolating a target sequence for restriction analysis is amplification by polymerase chain reaction or PCR (Kleppe *et al.*, 1971; Mullis and Faloona, 1987). This procedure involves hybridizing two short (typically 20 to 30 bases) oligonucleotide primers to opposite strands of a heat-denatured DNA target, replicating the DNA between the two primer sequences with a heat-stable DNA polymerase, denaturing the strands by heating, and repeating this cycle, usually about 30 times. The target DNA is doubled during each cycle (if all goes well), so that a large quantity of the target sequence is specifically amplified for analysis. Interpretation of the amplification products from PCR (using a given set of primers) as homologues is based on the assumption that only homologues will have the appropriate primer sequences separated by the appropriate distances in the genome; the size of the amplification product often is used to verify that the correct target was successfully amplified. However, this procedure does not distinguish orthologous from paralogous genes, both of which may retain the conserved primer sites and gene length. Worse, the amplification products may be recombinants of different alleles or even several paralogous loci — a phenomenon that produces "shuffled clones" (Saiki *et al.*, 1988; Scharf *et al.*, 1988a, b). This happens when a target sequence has not been fully replicated before the next round of denaturation and reannealing takes place. In the subsequent round, the partial product from one locus may rehybridize to a different locus to complete the product extension, thus producing a hybrid product that is partially derived from two different loci. Shuffled gene artifacts obviously are not limited to restriction analyses; any study that uses PCR should consider this process as a potential source of error that will confound assessments of homology.

Discussion so far has centered on determining homology of the portion of the genome targeted for restriction analysis. As with sequencing studies, an additional (finer) level of homology must be considered: homology of the individual characters. Data from restriction enzyme analyses typically are presented in one of two ways: restriction fragments or restriction sites (Fig. 6). Restriction fragments represent contiguous nucleotide sequences that fall between two recognition sites for a given restriction endonuclease. Such fragments may be coded as present or absent in a given individual.

Phylogenetic analysis of a presence/absence matrix of restriction fragments assumes that fragments of the same size represent homologous segments of the gene, whereas fragments of different sizes are not. There are three obvious sources of error with this assumption. The least of the three is that convergence of similarly sized but nonhomologous fragments is likely, especially if many fragments are examined and



Restriction sites

	I	II	III
One	+	+	+
Two	+	-	+
Three	-	-	+
Four	+	-	-

Restriction fragments

	1	2	3	4	7
One	+	+	-	+	-
Two	-	-	+	+	-
Three	-	-	-	+	+
Four	-	-	-	+	+

Fig. 6. Restriction site versus restriction fragment data. In this example, three restriction sites (I, II, and III) are variable in a linear segment of DNA among four individuals (One, Two, Three, and Four). The number under each restriction fragment indicates its respective length in arbitrary units. Two data matrices are shown: one of restriction sites and the other of restriction fragment size classes (in both, + indicates presence and - indicates absence). The three sites are independent characters, but the fragment size classes are not. In addition, note that the fragment patterns may be identical in species that share no restriction sites in common (e.g., individuals Three and Four).

the precision of measurement is not great. A more important problem is that insertion/deletion events will change the size of fragments and obscure homologies; this is serious because one insertion/deletion event may simultaneously affect many different but overlapping restriction fragments (each produced by a different restriction enzyme). Finally, gain of a restriction site will result in loss of one restriction fragment and gain of two smaller ones; obviously these are not independent events. The two smaller fragments are together homologous to the larger fragment. Because of all these sources of error, it is highly preferable to treat restriction sites (rather than restriction fragments) as characters in a restriction enzyme analysis (Fig. 6).

Restriction sites can be located in the target sequence by various mapping strategies (see Dowling *et al.*, 1990); the data matrix then indicates which sites are present versus absent in the studied individuals. The presence of a restriction site indicates that a specific base recognition sequence exists at a given location in the gene; a site that maps to the same position in two species is assumed to be homologous unless phylogenetic analysis suggests that the site is homoplastic. The primary source of error is related to imprecision of mapping: independent but adjacent restriction sites may be mapped to the same location. Thus, one individual may have two different restriction sites scored as one, or two individuals may have different restriction sites incorrectly assumed to be homologous. However, this source of error also affects restriction fragment data (because close sites produce small fragments that may be lost), so site data are highly preferable to fragment data when accurate assessments of homology are required.

C. *Random Amplified Polymorphic DNA (RAPD)*

As described above, the polymerase chain reaction can be used to amplify a specific target from a genome, based on *a priori* knowledge of the flanking regions. Welsh and McClelland (1990) and Williams *et al.* (1991) have described PCR-based methods for amplifying random polymorphic regions of the genome without any prior knowledge of specific flanking regions. The Williams *et al.* (1991) method, called RAPD analysis, utilizes a very short (typically 10 bases) oligonucleotide primer of an essentially arbitrary sequence (but with a constraint on GC content). Because of the small number of matches required for successful hybridization, the primer is likely to hybridize somewhere in the genome along opposite strands of DNA in the correct orientation (with the 3' ends facing each other), so that one or more fragments is amplified. This is especially likely at the location of inverted repeats. Some of these fragments will likely vary in length among individuals, and these variants are used as genetic markers in studies of populations and closely related species (e.g., Chapco et al., 1992; Crowhurst *et al.*, 1991; Goodwin and Annis, 1991; Hadrys *et al.*, 1992; Hunt and Page, 1992; Welsh *et al.*, 1992). Use of several different arbitrary primers can result in rapid collection of a large number of genetic markers.

Homology assignments of RAPD markers usually are based on considerations of fragment length or, less commonly, information on dominant or codominant segregation. Fragment length by itself is likely to be misleading about homology, for many of the same reasons discussed under Restriction Fragment Analysis (above). In addition, the same primer sites may exist in nonhomologous parts of the genome in two different species. Since segregation usually cannot be studied between species, assignments of homology of RAPD markers are highly tenuous in interspecific studies. If a primer site is lost (through substitution, for instance) but a second primer site exists nearby, two partially homologous fragments will be amplified that differ in length. Obviously, the presence of one such fragment is not independent of the presence of the other. Since many RAPD markers occur in regions of multiple repeats, the presence of multiple adjacent primer sites is not unlikely.

Although most of the above objections are theoretical, early empirical indications support the difficulties of using RAPD markers for phylogenetic analyses. Smith *et al.* (1993) found that loci that amplified in one strain of bacterium could be excluded from amplification in another strain because of competitive amplification of an unrelated locus. They also reported amplification of nonhomologous loci of indistinguishable size by the same primer, as well as amplification of multiple partially homologous fragments by a single primer. Smith *et al.* concluded that homology could not be reliably inferred from the RAPD fragment patterns. Furthermore, Kambhampati *et al.* (1991) found that RAPD-based inferences of phylogeny were incongruent with established, well-supported phylogenetic relationships of mosquitoes. This suggests that shared RAPD products were not orthologous. The difficulties of assigning homology to RAPD products suggests that the technique is not appropriate for applications that assume accurate assessments of homology, such as phylogenetic analyses. Production of nongenetic artifacts in RAPD studies (e.g., Ellsworth *et al.*, 1993) suggests caution for all applications of the technique.

D. *Allozyme Electrophoresis*

In enzyme electrophoresis, homology among genes is determined by several functional, structural, and expressional criteria (see Murphy *et al.*, 1990). Paralogous genes are explicitly recognized as distinct loci, and usually are assumed to be evolving independently. Multiple loci that code for enzymes with the same biochemical function are thought in most cases to be paralogues of each other. In many cases, the existence of multiple loci is clear, because they may be expressed together in the same tissue at the same time in the same individual. In other cases, paralogous loci are recognized on the basis of expression in different tissue types, at different times of development, or at different seasons. Paralogy also may be determined based on multimeric structure or differential cellular location of the enzyme (e.g., whether mitochondrial or cytosolic). Therefore, for a locus to be considered orthologous in two species in an allozyme study, the two relevant enzymes typically are expected to satisfy the following conditions: (1) they have the same catalytic function; (2) they have the same multimeric structure; (3) they are expressed at the same general cellular location; (4) they are expressed at the same time in development; and (5) they are expressed in the same tissues.

In addition to homology among genes, homology among electromorphs (putative alleles) is of concern in allozyme electrophoresis. Electromorphs are defined on the basis of their distance of movement in an electric field relative to each other or to an internal standard; two electromorphs are placed in the same class if their distances of migration are indistinguishable. Convergence in electrophoretic mobility is likely because many amino acid replacements produce the same changes in net charge; this is offset somewhat by the large number of potential states of overall charge for a protein. Homology among electromorphs is determined on the basis of congruence with other characters (usually other allozyme data) in a phylogenetic analysis. Electromorphs that are phylogenetically congruent are considered to be homologous, whereas incongruent electromorphs are considered to be homoplastic.

VII. SUMMARY

Some proponents of molecular techniques have claimed that molecular biology "solves the problem of homology" (Sibley and Ahlquist, 1987). Although a better understanding of molecular biology has provided many new insights into relationships among genes and their products, there is still ample room for mistaken inferences about homologous molecular relationships. Distinguishing orthology from paralogy is the greatest obstacle for most evolutionary applications of molecular techniques. Duplicated gene loci and paralogous pseudogenes appear to be the rule rather than the exception in the eukaryote nuclear genome, so care is required if orthologous products are to be compared among taxa.

Concerted evolution among paralogous sequences helps the reconstruction of phylogenetic relationships among taxa, but hinders reconstruction of gene relationships. Some knowledge of the degree of concerted evolution for a given gene family is required to assess whether inferred relationships are those of taxa or of genes. Rates of concerted evolution appear to be high enough in some repeated gene families that paralogy is unlikely to be a confounding factor in reconstructing relationships of taxa, unless branching points in the tree are separated by very small distances. However, additional empirical studies are needed to measure rates of concerted evolution to determine its effects on phylogenetic analyses.

Relationships of paralogy extend to parts of genes as well. Functional gene modules may be duplicated in different genes or in tandem within a gene. Slippage replication can also produce paralogy of small tandem repeats.

Convergence can occur in whole genes or their parts, leading to false hypotheses of homology. In the case of whole genes, convergence may be related to positive selection, or may be secondarily related to convergence in base composition or the presence of simple tandem repeats. At a finer level, convergence at individual nucleotide sites is common, and may be facilitated by mutational biases. Inferences of positional homology requires accurate alignment criteria, which in turn require some knowledge of the relative likelihood of substitutions relative to insertions and deletions.

DNA-DNA hybridization has considerable potential for conflating paralogy and orthology, as well as homology and homoplasy. Homology and homoplasy may be factored out in analysis, but the confusion of orthology and paralogy is potentially more difficult to correct. Restriction *fragment* analyses are also likely to suffer from inaccurate homology assignments, but restriction *site* analyses are much less sensitive to this problem. Analyses of randomly amplified polymorphic DNA regions probably often involve mistaken assignments of homology, and so are best restricted to applications in which accurate assignment of homology is not critical. In allozyme electrophoresis, orthologous and paralogous proteins are identified on the basis of catalytic function, multimeric structure, location of cellular expression, timing of expression, and tissue distribution; when these criteria are applied, confusion between paralogous and orthologous loci is unlikely. However, convergence of electromorphs may occur through amino acid replacements that produce convergent changes in net charge of the protein.

As we begin to discover more about the processes that produce molecular variation among species, relationships among genes and their products look increasingly complex. The difficulties of assigning homology to molecules parallel many of the difficulties of assigning homology to morphological structures (Patterson, 1988). This may be a partial explanation for the similar levels of homoplasy that are observed in molecular and morphological phylogenetic studies (Sanderson and Donoghue, 1989). Hopefully, then, a better understanding of homologous relationships at the molecular level can lead to a better understanding of homology in general.

ACKNOWLEDGMENTS

I thank Jim Bull, David Cannatella, Paul Chippindale, Mike Donoghue, Anna Goebel, Ken Halanych, Rodney Honeycutt, John Huelsenbeck, and Barbara Mable for useful comments on the manuscript, and Jim Smith for insights on RAPD analysis. This work was supported by grants from the National Science Foundation (DEB 9106746 and DEB 9221052).

REFERENCES

Aoki, K., Tateno, Y., and Takahata, N. (1981). Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary G+C content. *J. Mol. Evol.* **18**, 1-8.

Arnheim, N. (1983). Concerted evolution of multigene families. *In* "Evolution of Genes and Proteins" (M. Nei and R. K. Koehn, eds.), pp. 38-61. Sinauer Associates, Sunderland, MA.

Arnheim, N., Krystal, M., Schmickel, R., Wilson, G. Ryder, O., and Zimmer, E. (1980). Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7323-7327.

Baltimore, D. (1981). Gene conversion: Some implications for immunoglobulin genes. *Cell (Cambridge, Mass.)* **24**, 592-594.

Bledsoe, A. H., and Sheldon, F. H. (1990). Molecular homology and DNA hybridization. *J. Mol. Evol.* **30**, 425-433.

Brown, D. D., Wensink, P. C., and Jordan, E. (1972). A comparison of the ribosomal DNAs of *Xenopus laevis* and *Xenopus mulleri*: Evolution of tandem genes. *J. Mol. Biol.* **63**, 57-73.

Buth, D. G. (1984). Allozymes of the cyprinid fishes: Variation and application. *In* "Evolutionary Genetics of Fishes" (B. J. Turner, ed.), pp. 561-590. Plenum, New York.

Chapco, W., Ashton, N. W., Martel, R. K. B., Antonishyn, N., and Crosby, W. L. (1992). A feasibility study of the use of random amplified polymorphic DNA in the population genetics and systematics of grasshoppers. *Genome* **35**, 569-574.

Coen , E. S., Strachan, T., and Dover, G. A. (1982a). Dynamics of concerted evolution of rDNA and histone gene families in the *melanogaster* species subgroup of *Drosophila. J. Mol. Biol.* **158**, 17-35.

Coen , E. S., Thoday, J. M., and Dover, G. A. (1982b). Rate of turnover of structural variants in the rDNA gene family of *D. melanogaster. Nature (London)* **395**, 564-568.

Crowhurst, R. N., Hawthorn, B. T., Rikkerink, E. H. A., and Templeton, M. D. (1991). Differentiation of *Fusarium solani* f. sp. *cucurbitae* races 1 and 2 by random amplification of polymorphic DNA. *Curr. Genet.* **20**, 391-396.

Donoghue, M. J. (1992). Homology. *In* "Keywords in Evolutionary Biology" (E. F. Keller and E. A. Lloyd, eds.), pp. 170-179. Harvard Univ. Press, Cambridge, MA.

Doolittle, R. F. (1987). The evolution of the vertebrate plasma proteins. *Biol. Bull. (Woods Hole, Mass.)* **172**, 269-283.

Dover, G. A. (1982). Molecular drive: A cohesive mode of species evolution. *Nature (London)* **299**, 111-117.

Dover, G. A. (1986). Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. *Trends Gen.* **2**, 159-165.

Dowling, T. E., Moritz, C., and Palmer, J. D. (1990). Nucleic acids II: Restriction site analysis. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), pp. 250-317. Sinauer Associates, Sunderland, MA.

Doyle, J. J. (1991). Evolution of higher plant glutamine synthetase genes: Regulatory specificity as a criterion for predicting orthology. *Mol. Biol. Evol.* **8**, 366-377.

Duellman, W. E., and Hillis, D. M. (1987). Marsupial frogs (Anura: Hylidae: *Gastrotheca*) of the Ecuadorian Andes: Resolution of taxonomic problems and phylogenetic relationships. *Herpetologica* **43**, 141-173.

Ellsworth, D. L., Rittenhouse, K. D., and Honeycutt, R. L. (1993). Artifactual variation in randomly amplified polymorphic DNA banding patterns. *BioTechniques* **14**, 214-217.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113.

Friedlander, T. P., Regier, J. C., and Mitter, C. (1992). Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. *Syst. Biol.* **41**, 483-490.

Ghiselin, M. T. (1976). The nomenclature of correspondence: A new look at "homology" and "analogy." *In* "Evolution, Brain and Behavior: Persistent Problems" (R. B. Masterson, W. Hodos, and H. Jerison, eds.), pp. 129-142. Erlbaum, Hillsdale, NJ.

Gonzales, I. L., Gorski, J. L., Campden, T. J., Dorney, D. J., Erickson, J. M., Sylvester, J. E., and Schmickel, R. D. (1985). Variation among human 28S ribosomal RNA genes. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 7666-7670.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* **28**, 132-168.

Goodman, M., Miyamoto, M. M., and Czelusniak, J. (1987). Pattern and process in vertebrate phylogeny revealed by coevolution of molecules and morphologies. *In* "Molecules and Morphology in Evolution: Conflict or Compromise?" (C. Patterson, ed.), pp. 141-176. Cambridge Univ. Press, Cambridge.

Goodwin, P. H., and Annis, S. L. (1991). Rapid identification of genetic variation and pathotype of *Leptosphaeria maculans* by random amplified polymorphic DNA assay. *Appl. Environ. Microbiol.* **57**, 2482-2486.

Gray, G. S., and Fitch, W. M. (1983). Evolution of antibiotic resistance genes: The DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.* **1**, 57-66.

Grime, J. P., and Mowforth, M. A. (1982). Variation in genome size and ecological interpretation. *Nature (London)* **299**, 151-153.

Hadjiolov, A. A., Georgiev, O. I., Nosikov, V. V., and Yavachev, L. P. (1984). Primary and secondary structure of rat 28S ribosomal RNA. *Nucleic Acids Res.* **12**, 3677-3693.

Hadrys, H., Balick, M., and Schierwater, B. (1992). Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Mol. Ecol.* **1**, 55-63.

Hancock, J. M., and Dover, G. A. (1988). Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. *Mol. Biol. Evol.* **5**, 377-391.

Hancock, J. M., and Dover, G. A. (1990). 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res.* **18**, 5949-5954.

Hassouna, N., Michot, B., and Bachellerie, J.-P. (1984). The complete nucleotide sequence of mouse 28S rRNA gene.

Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res.* **12**, 3563-3583.

Herdman, M. (1985). The evolution of bacterial genomes. *In* "The Evolution of Genome Size" (T. Cavalier-Smith, ed.), pp. 37-68. Wiley, New York.

Hillis, D. M., and Dixon, M. T. (1989). Vertebrate phylogeny: Evidence from 28S ribosomal DNA sequences. *In* "The Hierarchy of Life" (B. Fernholm, K. Bremer, and H. Jörnvall, eds.), Proc. Nobel Symp. 70, pp. 355-367. Elsevier, Amsterdam.

Hillis, D. M., and Dixon, M. T. (1991). Ribosomal DNA: Molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66**, 411-453.

Hillis, D. M., Dixon, M. T., and Ammerman, L. K. (1990a). The relationships of the coelacanth *Latimeria: chalumnae:* Evidence from sequences of vertebrate 28S ribosomal RNA genes. *Environ. Biol. Fishes* **32**, 119-130.

Hillis, D. M., Larson, A., Davis, S. K., and Zimmer, E. A. (1990b). Nucleic acids III. Sequencing. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), pp. 318-370. Sinauer Associates, Sunderland, MA.

Hillis, D. M., Moritz, C., Porter, C. A., and Baker, R. J. (1991). Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* **251**, 308-310.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R., and Molineux, I. J. (1992). Experimental phylogenetics: Generation of a known phylogeny. *Science* **255**, 589-592.

Hori, H., and Osawa, S. (1986). Evolutionary change in 5S rRNA secondary structure and a phylogenetic tree of 352 rRNA species. *BioSystems* **19**, 163-172.

Hughes, A. L., and Nei, M. (1990). Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 958-962.

Hunt, G. J., and Page, R. E. (1992). Patterns of inheritance with RAPD molecular markers reveal novel types of polymorphisms in the honey bee. *Theor. Appl. Genet.* **85**, 15-20.

Irwin, D. M., and Wilson, A. C. (1990). Concerted evolution of ruminant stomach lysozymes. *J. Biol. Chem.* **265**, 4944-4952.

Kambhampati, S., Black, W. C., IV, and Rai, K. S. (1991). RAPD-PCR for identification and differentiation of mosquito species and populations: Techniques and statistical analysis. *J. Med. Entomol.* **29**, 939-945.

Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., and Khorana, H. G. (1971). Studies on polynucleotides XCVI. Repair replication of short synthetic DNAs as catalyzed by DNA polymerases. *J. Mol. Biol.* **56**, 341-361.

Lankester, E. R. (1870). On the use of the term homology in modern zoology. *Ann. Mag. Nat. Hist. Ser. [4]* **6**, 34-43.

Larson, A., and Wilson, A. C. (1989). Patterns of ribosomal RNA evolution in salamanders. *Mol. Biol. Evol.* **6**, 131-154.

Levinson, G., Marsh, J. L., Epplen, J. T., and Gutman, G. A. (1985). Cross-hybridizing snake satelite, *Drosophila*, and mouse DNA sequences may have arisen independently. *Mol. Biol. Evol.* **2**, 494-504.

Li, W.-H., and Graur, D. (1991). "Fundamentals of Molecular Evolution." Sinauer Associates, Sunderland, MA.

McElroy, D., Rothenberg, M., Reece, K. S., and Wu, R. (1990). Characterization of the rice actin gene family. *Plant Mol. Biol.* **15**, 257-268.

Moritz, C., Dowling, T. E., and Brown, W. M. (1987). Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. *Annu. Rev. Ecol. Syst.* **18**, 269-292.

Mullis, K. B., and Faloona, F. A. (1987). Specific synthesis of DNA *in vitro* via a polymerase catalyzed chain reaction. In "Methods in Enzymology" (S. Fleischer and B. Fleischer, eds.). Vol. **155**, pp. 335-350. Academic Press, FL.

Murphy, R. W., Sites, J. W., Jr., Buth, D. G., and Haufler, C. H. (1990). Proteins. I. Isozyme electrophoresis. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), pp. 45-126. Sinauer Associates, Sunderland, MA.

Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 166-169.

Nagylaki, T. (1984). The evolution of multigene families under intrachromosomal gene conversion. *Genetics* **106**, 529-548.

Nagylaki, T., and Petes, T. D. (1982). Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics* **100**, 315-337.

Ohta, T. (1980). "Evolution and Variation of Multigene Families." Springer-Verlag, Berlin.

Ohta, T. (1983). On the evolution of multigene families. *Theor. Popul. Biol.* **23**, 216-240.

Ohta, T. (1984). Some models of gene conversion for treating the evolution of multigene families. *Genetics* 106, 517-528.

Ohta, T., and Dover, G. A. (1983). Population genetics of multigene families that are dispersed into two or more chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 4079-4083.

Owen, R. (1843). "Lectures on the comparative anatomy and physiology of the invertebrate animals." Longman, Brown, Green, & Longmans, London.

Owen, R. (1848). "On the Archetype and Homologies of the Vertebrate Skeleton." Richard and John E. Taylor, London.

Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W., and Manhart, J. R. (1988). Chloroplast DNA and plant phylogeny. *Ann. Mo. Bot. Gard.* **75**, 1180-1206.

Patterson, C. (1988). Homology in classical and molecular biology. *Mol. Biol. Evol.* **5**, 603-625.

Patthy, L. (1985). Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. *Cell (Cambridge, Mass.)* **41**, 657-663.

Rees, H., and Jones, R. N. (1972). The origin of the wide species variation in nuclear DNA content. *Int. Rev. Cytol.* **32**, 53-92.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-491.

Sanderson, M. J., and Donoghue, M. J. (1989). Patterns of variation in levels of homoplasy. *Evolution (Lawrence, Kans.)* **43**, 1781-1795.

Sanderson, M. J., and Doyle, J. J. (1992). Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. *Syst. Biol.* **41**, 4-17.

Scharf, S. J., Long, C. M., and Erlich, H. A. (1988a). Sequence analysis of the HLA-DRβ and HLA-DQβ loci from three *Pemphigus vulgaris* patients. *Human Immunol.* **22**, 61-69.

Scharf, S. J., Friedmann, A., Brautbar, C., Szafer, F., Steinman, L., Horn, G., Gyllensten, U., and Erlich, H. A. (1988b). HLA class II allelic variation and susceptibility to *Pemphigus vulgaris*. *Proc. Natl. Acad. Sci. USA* **85**, 3504-3508.

Seperack, P., Slatkin, M., and Arnheim, N. (1988). Linkage disequilibrium in human ribosomal genes: Implications for multigene family evolution. *Genetics* **119**, 943-949.

Shah, D. M., Hightower, R. C., and Meagher, R. B. (1983). Genes encoding actins in higher plants: Intron positions are highly conserved but the coding sequences are not. *J. Mol. Appl. Genet.* **2**, 111-126.

Sibley, C. G., and Ahlquist, J. E. (1987). Avian phylogeny reconstructed from comparisons of the genetic material, DNA. *In* "Molecules and Morphology in Evolution: Conflict or Compromise?" (C. Patterson, ed.), pp. 95-121. Cambridge Univ. Press, Cambridge.

Singer, C. E., and Ames, B. N. (1970). Sunlight ultraviolet and bacterial DNA base ratios. *Science* **170**, 822-826.

Smith, G. P. (1974). Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 507-513.

Smith, J. J., Scott-Craig, J. S., Leadbetter, J. R., Bush, G. L., Roberts, D. L., and Fulbright, D. W. (1993). Characterization of random amplified polymorphic DNA (RAPD) products from *Xanthomonas campestris*: Implications for the use of RAPD products in phylogenetic analysis. *Mol. Phylog. Evol.* (in press).

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**, 503-517.

Sparrow, A. H., and Nauman, A. F. (1976). Evolution of genome size by DNA doublings. *Science* **192**, 524-529.

Stewart, C.-B., and Wilson, A. C. (1987). Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 891-899.

Swanson, K. W., Irwin, D. M., and Wilson, A. C. (1991). Stomach lysozyme gene of the langur monkey: Tests for convergence and positive selection. *J. Mol. Evol.* **33**, 418-425.

Swofford, D. L., and Olsen, G. J. (1990). Phylogeny reconstruction. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), pp. 411-501. Sinauer Associates, Sunderland, MA.

Szostak, J. W., and Wu, R. (1980). Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. *Nature (London)* **284**, 426-430.

Tamura, K. (1992). The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol. Biol. Evol.* **9**, 814-825.

Ware, V. C., Tague, B. W., Clark, C. G., Gourse, R. L., Brand, R. C., and Gerbi, S. A. (1983). Sequence analysis of 28S ribosomal DNA from the amphibian *Xenopus laevis*. *Nucleic Acids Res.* **11**, 7795-7817.

Wegnez, M. (1987). Letter to the editor. *Cell (Cambridge, Mass.)* **51**, 516.

Welsh, J., and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucl. Acids Res.* **18**, 7213-7219.

Welsh, J., Pretzman, C., Postic, D., Girons, I. S., Baranton, G., and McClelland, M. (1992). Genomic fingerprinting by arbitrarily primed polymerase chain reaction resolves *Borrelia burgdorferi* into three distinct phyletic groups. *Int. J. Syst. Bacteriol.* **42**, 370-377.

Werman, S. D., Springer, M. S., and Britten, R. J. (1990). Nucleic acids I: DNA-DNA hybridization. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), pp. 204-249. Sinauer Associates, Sunderland, MA.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1991). DNA polymorphisms amplified by arbitrary primers are useful genetic markers. *Nucleic Acids Res.* **18**, 6531-6535.

Wilson, J. T., Wilson, L. B., Reddy, V. B., Cavallesco, C., Ghosh, P. K., de Riel, J. K., Forget, B. G., and Weissman, S. M. (1980). Nucleotide sequence of the coding portion of human alpha globin messenger RNA. *J. Biol. Chem.* **255**, 2807-2815.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., and Wilson, A. C. (1980). Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2158-2162.