CHAPTER 5

Molecular Systematics

One of the most exciting developments in the past decade has been the application of nucleic acid data to problems in systematics. The term **molecular systematics** is used to mean macromolecular systematics—the use of DNA and RNA to infer relationships among organisms. Although technically, isozyme methods and flavonoid data are also molecular, they are usually discussed separately. In this chapter we outline the major types of nucleic acid data currently available, the molecules and genomes that have been most commonly used, and some aspects of data analysis that are unique to molecular data.

Molecular data have revolutionized our view of phylogenetic relationships, although not for the reasons initially suggested. Early proponents of molecular systematics claimed that molecular data were more likely to reflect the true phylogeny than morphological data, ostensibly because they reflected gene-level changes, which were thought to be less subject to convergence and parallelism than were morphological traits. This early assurance now appears to be wrong, and molecular data are in fact subject to most of the same problems that morphological data are. The big difference is that there are simply many more molecular characters available, and their interpretation is generally easier—an adenine is an adenine, whereas compound leaves, for example, can form in quite different ways in different plants. As a result, molecular data are now widely used for generating phylogenetic hypotheses.

In many cases, molecular data have supported the monophyly of groups that were recognized on morphological grounds (e.g., Poaceae, Fabaceae, Rosaceae). More importantly, molecular data often have allowed systematists to choose among competing hypotheses of relationships (e.g., to decide what group is the sister group of the Asteraceae, or the Poaceae). In other cases, molecular data have allowed the placement of taxa whose relationships were known to be problematic. For example, although the Hydrangeaceae were traditionally placed in or near the Saxifragaceae, it was clear that the two were unrelated. Only with molecular data, however, was there a strong alternative hypothesis for the placement of the Hydrangeaceae: in the order Cornales. It has been fairly rare for molecular data to suggest something completely novel, although there have been a few dramatic cases, such as the monophyly of the glucosino-

late clade and the placement therein of the Limnanthaceae, the inclusion of the Vochysiaceae in the Myrtales, and the documentation of introgression between species that were apparently intersterile.

TABLE 5.1 ***Comparison of the three genomes in a plant cell.***

	Genome size (kbp)	Inheritance
Chloroplast	135–160	Generally maternal (from the seed parent)
Mitochondrion	200–2500	Generally maternal (from the seed parent)
Nucleus	1.1×10^6 to 110×10^9	Biparental

Plant Genomes

The plant cell contains three different genomes: those of the chloroplast, the mitochondrion, and the nucleus (Table 5.1). Systematists have used data from all three. The two organelles are generally inherited uniparentally (usually maternally in angiosperms); the nucleus is biparental. The three genomes differ dramatically in size, with the nucleus being by far the largest—measured in megabases. The mitochondrial genome includes several hundred kilobase pairs (kbp) of DNA (200–2500 kbp), which is small relative to the nuclear genome, but quite large relative to the mitochondrial genomes of animals (which tend to be about 16 kbp). The chloroplast genome is the smallest of the three plant genomes, in most plants ranging from 135 to 160 kbp.

Like the eubacteria from which they are derived, mitochondria and chloroplasts have circular genomes. The order of genes in the mitochondrion is variable, and they are separated by large regions of noncoding DNA. The mitochondrial genome rearranges itself frequently, so that many rearranged forms can occur in the same cell. This means that rearrangements of the genome occur so often within individual plants that they do not characterize or differentiate species or groups of species; therefore they are not especially useful for inferring relationships.

The chloroplast, in contrast, is stable, both within cells and within species. The most obvious feature of the chloroplast genome is the presence of two regions that encode the same genes, but in opposite directions; these are known as the **inverted repeats**. Between them are a small single-copy region and a large single-copy region (Figure 5.1). Rearrangements of the chloroplast genome are rare enough in evolution that they can be used to demarcate major groups. Also, gains and losses of genes, or their introns, are common enough to be worth looking for, but rare enough to be a stable marker of evolutionary change.

The order of genes in the nuclear genome is also presumed to be stable, at least within species, and may be stable across groups of species as well. Some information on gene order has been revealed by classic techniques of cytogenetics, but much more detailed information is now being provided by genome mapping (see below). In the coming years, this could become an important source of systematic information.

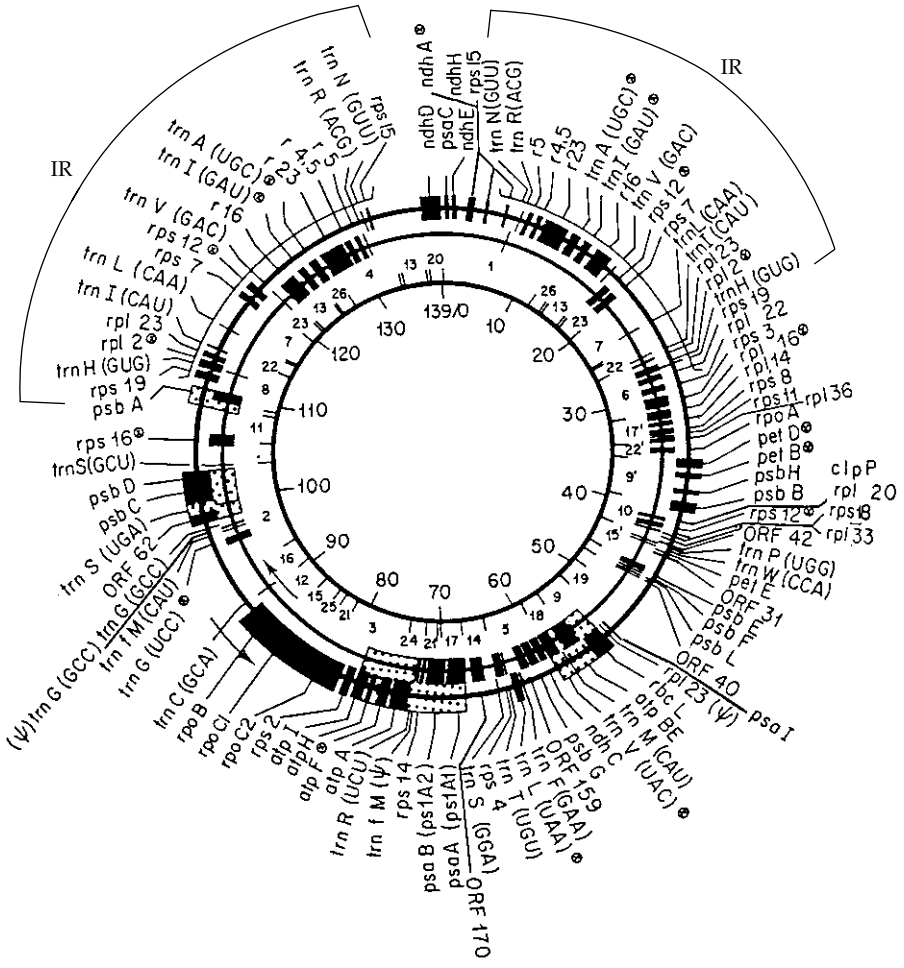

Figure 5.1 Diagram of the chloroplast genome of maize, showing the locations of some major genes and the inverted repeat regions. (From S. Rodermel, Maize Newsletter 1993.)

DNA sequences change at a rate different from the rate of genomic rearrangement. Chloroplast genes tend to accumulate mutations more rapidly than do mitochondrial genes in plants. It is harder to generalize about nuclear genes, which is hardly surprising because there are so many of them, and because we know less about them.

Generating Molecular Data

Molecular systematics has been and remains technique-driven; as new molecular methods become available, they expand the kinds and amounts of systematic data that can be extracted from nucleic acids. A number of good textbooks describing the techniques for generating and analyzing molecular data are available (Hillis et al. 1996; Soltis et al. 1997, 1998; Crawford 1990; Miyamoto and Cracraft 1991).

If useful comparisons are to be made across many taxa, the technique applied has to be fast and easy. This meant that molecular systematics was barely possible until the invention of recombinant DNA, became easier as sequencing techniques were improved, and took another leap forward with the invention of the polymerase chain reaction (PCR) technique. The current advances in automated sequencing will continue to add speed to the mechanical task of data collection, which is already being done by robots in some large sequencing facilities.

GENE MAPPING

Most molecular systematic studies were initially done using **restriction site analysis**. This technique can be used to generate maps of individual genes or entire genomes. Much of what we now know about chloroplast and mitochondrial genome structure comes from such studies.

In this procedure, DNA is extracted from a plant, and is then cut with restriction enzymes—enzymes that cut DNA at a particular sequence. The enzyme known as *Bam*HI, for example, cuts DNA everywhere it finds the sequence GGATCC, and *Eco*RI cuts at GAATTC. (The names of restriction enzymes are acronyms based on the first letter of the genus and first two letters of the species of the bacterium from which the enzyme was isolated. So *Bam*HI is from *Bacillus amylofaciens*, *Eco*RI from *Escherichia coli*.) A map is constructed by cutting the DNA with one enzyme and examining the resulting pattern of fragment sizes, then cutting it with a second enzyme, and finally cutting it with both enzymes together. This creates a sort of puzzle from which the order of the restriction sites can be constructed.

Such studies were initially done by the laborious process of separating organelles from plant tissue, isolating their DNA, cutting it with a particular restriction enzyme, and then measuring the fragment sizes on an ethidium-stained gel. Once a couple of cloned organellar genomes became widely available, it proved easier to use them as probes. In this process, the DNA of interest is cut and then transferred to a nylon membrane. A cloned piece of chloroplast DNA (the probe) is then labeled with radioactive phosphorus and denatured to produce single-stranded DNA. This single-stranded DNA is then allowed to bind to the DNA on the membrane; it will bind only to matching (chloroplast) sequences. The membrane is then placed next to a piece of X-ray film. The bands where the probe has bound appear as dark lines on the film. (This technique is known as Southern blotting, named after E. M. Southern, who invented it.)

Restriction site analysis is less widely used today than it was initially, but it remains common for studying variation in the chloroplast genome and in ribosomal RNA spacers, particularly among congeneric species and sometimes within species as well. Restriction site studies are also sometimes used for assessing variation among PCR fragments, although with recent advances in gene sequencing techniques, this is less of a shortcut than it once was.

Nuclear genome mapping remains a large commitment of time and effort. It requires that two plants be crossed, and their F_1 offspring self-pollinated to produce a large number (100 or more) of F_2 plants. Then the genotypes of both parents and offspring are determined using restriction site, RAPD, or AFLP markers (see below); these must be polymorphic between the parents. The parents and offspring are then scored for whatever morphological characters are of interest, and these are mapped according to their linkage to the molecular markers. This technique generally requires the use of sophisticated statistical programs to infer linkage relationships.

GENE SEQUENCING

DNA sequencing of genes, parts of genes, or noncoding regions is becoming more and more common and is now widely used in systematics. Sequencing determines the precise order of nucleotides—adenine (A), cytosine (C), guanine (G), or thymine (T)—in a stretch of DNA. The central difficulty of sequencing has always been getting enough DNA to work with. This was done initially by cloning genes into bacteria and allowing the bacteria to replicate the genes along with their own genomes. Genes were taken from genomic libraries, which were made by cutting all the DNA of an organism with a restriction enzyme and then cloning all the resulting fragments into an appropriate plasmid, bacteriophage, or other vector. This method is quite slow, but is reliable and avoids some of the possible artifacts of more efficient methods. It is also the only method available if only a few sequences of the gene are known.

This laborious approach was later replaced by the polymerase chain reaction technique (**PCR**), in which DNA is replicated enzymatically, allowing the cloning step to be omitted (Figure 5.2). PCR requires some knowledge of the sequence to be studied. Small pieces of

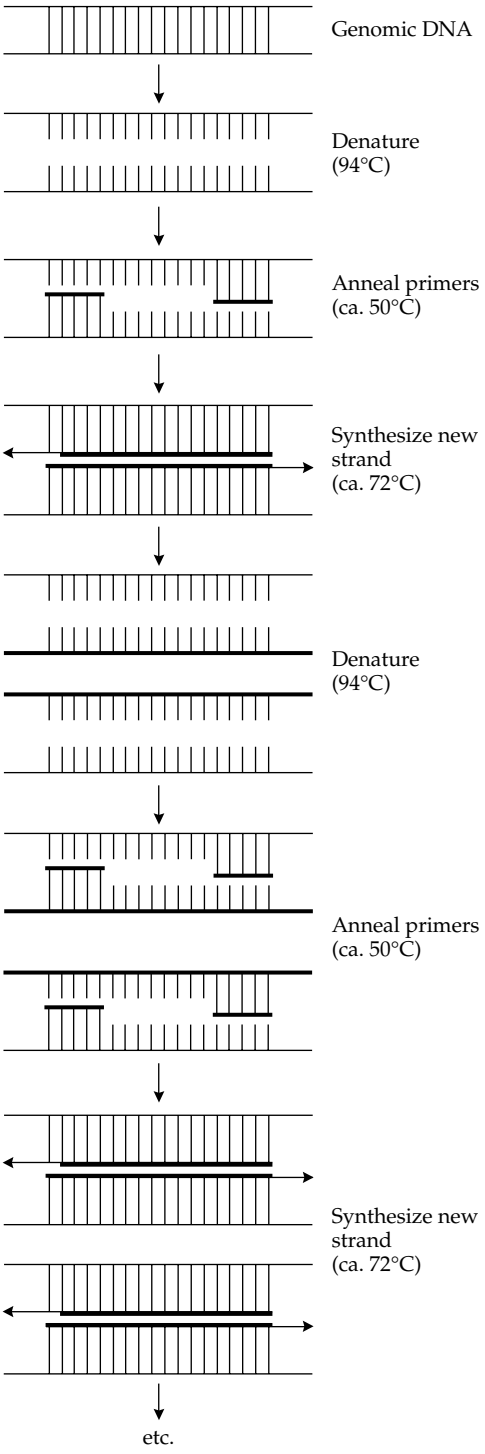

Figure 5.2 The polymerase chain reaction.

single-stranded DNA (primers) are produced to match the DNA sequences at either end of the region of interest. These primers are placed in a tube with DNA from the organism, a DNA polymerase, and free nucleotides. The mix is then subjected to repeated heating and cooling. As it heats, the double-stranded organismal DNA denatures and becomes single-stranded. Then, as it cools, the primers bind to their complementary sequences at either end of the target region. The temperature is then raised to the point at which the polymerase becomes active. It binds to the DNA + primer complex and begins synthesizing a complementary strand using the free nucleotides in the solution. Then the temperature is raised to denature the DNA, and the cycle repeats. The DNA in the region between the primers is thus copied, and the amount increases exponentially. The PCR product can be sequenced directly or can be cloned and then sequenced.

This rapid method has allowed systematists to study the same region in many species of a particular group. One disadvantage of PCR is that the polymerase itself introduces occasional errors, which could in theory affect an estimate of phylogeny, particularly if the sequences being compared are extremely similar. In practice, this may be not be a serious problem, in that the errors are not likely to be biased in favor of any one grouping.

One way to reduce potential error in sequencing is to sequence both strands of the molecule. This is required by some journals before the results will be published, but is not universal practice. The decision of how accurately to sequence depends on the relative costs of an error versus the costs of repeatedly sequencing the same region. Systematists must often make a choice between highly accurate sequences from fewer taxa or less accurate sequences from more taxa.

Direct sequencing of the PCR product will not generally reveal minor variants of the sequence if they are present. This is sometimes a problem with highly repetitive genes such as ribosomal genes, for which the many copies often are not identical. Direct sequencing also cannot distinguish between alleles of the same gene. Imagine that two alleles differ from each other at two positions, such that one allele has an A at the first position and a T at the second, whereas the other has a T at the first position and an A at the second. Both positions will appear as A/T polymorphisms on a sequencing gel, and it is impossible to tell which allele has which base at which position. The latter problem can be overcome by cloning the PCR products (which remains easier than screening gene libraries but can create its own sampling bias).

As the major genome sequencing projects progress, the technology for gene sequencing is improving and becoming increasingly automated. It seems likely that systematics laboratories will make increasing use of commercial sequencing facilities in the future, so that the attention of systematists can be focused on the more intellectually demanding work of analyzing sequences.

Types of Molecular Data

Virtually all molecular phylogenetics is now done using either genome rearrangements or sequences of DNA as characters. The former are studied using restriction sites as genomic markers, and the latter are either sequenced completely or sampled with restriction enzymes.

GENOME REARRANGEMENTS

Studies of the chloroplast and mitochondrial genomes proceed by constructing genomic maps, which reveal the order of genes in the genome. One of the early successes of molecular systematics was the identification of the earliest diverging members of the Compositae by Jansen and Palmer (1987). Using restriction site mapping, they found that almost all members of the family had a unique order of genes in the large single-copy region of the chloroplast genome. This order could be explained by a single inversion of the DNA. All other angiosperms lacked the inversion. The few composites that had the ancestral arrangement of the genome were members of the subtribe Barnedisiinae, a South American group with bilabiate corollas. This finding strongly suggested that the Barnedisiinae (now treated as a subfamily, the Barnedisioideae) is the sister group to the rest of the enormous sunflower family, and that the latter is monophyletic.

The context of this discovery is important. Many previous researchers had speculated on what the "ancestral" composite might have looked like, and had suggested several extant groups that might represent the earliest lineages. The Barnedisiinae was one of several possibilities, and had been supported by cladistic analyses of morphological data. In this case, then, the molecular data did not completely change our view of evolution, but they did resolve the ambiguity of the morphological data.

Other mapping data have supported a number of major groups. The grass family (Poaceae), for example, has three inversions in the chloroplast genome. One of these is unique to the Poaceae, one is shared with the Joinvilleaceae, and a third is shared with the Joinvilleaceae and Restionaceae. The inversion that is unique to the Poaceae is hardly a surprise—the family is unquestionably monophyletic, a result that can be confirmed by almost any sort of data. The inversions shared with the Joinvilleaceae and Restionaceae again helped to clarify the morphological data, which suggested that either might be the sister group to the grasses.

Loss of genes from the chloroplast is also common. Some groups, for instance, have lost one of the inverted repeats. This has occurred among the angiosperms (e.g., in a group of papilionioid legumes), in all conifers, and in *Euglena* (a flagellated photosynthetic eukaryote unrelated to green plants). In general, however, losses of a smaller piece of DNA, such as an intron or an entire gene, are more common than major rearrangements, and may occur multiple times in evolution. For example, most angiosperms have an intron in the chloroplast gene *rpo*C1, but this intron has been lost in grasses, one subfamily of cacti (Cactoideae), at least two members of Goodeniaceae, some Aizoaceae, and some, but not all, members of the genera *Passiflora* (Passifloraceae) and *Medicago* (Fabaceae) (Downie et al. 1996).

Mapping of the nuclear genome is in its infancy, and is only rarely applied outside the model systems for molecular biology and major crop plants. Techniques for mapping are developing rapidly, however, and comparative studies using the nuclear genome have been done in the grasses and in the Solanaceae. In addition, nuclear mapping studies are just beginning to address questions of speciation. In a study of *Mimulus*, Bradshaw et al. (1995) found that the shift from bee pollination to bird pollination involved eight genes, which they were able to localize to linkage groups (see Chapter 4). In a similar study of *Helianthus*, Rieseberg and his colleagues (1995, 1996) found that two species (*H. annuus* and *H. petiolaris*) differed by at least ten genomic rearrangements (three inversions and at least seven translocations), which affected recombination and possibilities for introgression. The genome of their hybrid derivative, *H. anomalus*, was rearranged relative to that of both parents, so that the species was partially reproductively isolated from both. Rieseberg's group then created new hybrids of *H. annuus* and *H. petiolaris*, and found that the chromosomal rearrangements in the experimental hybrids were similar to those in the naturally occurring hybrid species, *H. anomalus*. They concluded that certain combinations of genes and gene rearrangements were selectively favored in the hybrid.

It is not clear whether this laborious approach will ever become simple enough to apply to the multiple species usually covered by a systematic study. It may, however, become extremely useful for systematists interested in the mechanics of the speciation process.

SEQUENCE DATA

By far the majority of molecular systematic studies have used DNA (or RNA) sequences. Initially this was done indirectly by using restriction enzymes to sample the underlying sequence, generally as part of the same studies that generated gene or genome mapping data. Because restriction enzymes cut DNA at a particular sequence, any time a particular restriction site is found, the sequence can be inferred at that position. Conversely, if the site is not found, a mutation in one of the four or six bases in the restriction site can be inferred. This relatively simple inference has been turned into a powerful tool for systematics. It has been used most notably in studies of the chloroplast genome. Sequences (restriction sites) are scored as present or absent, and these scores are then used as characters in phylogenetic analysis. Methods for this sort of study are the same as those used for genome mapping, now most commonly done using Southern blots. Yet another method, devised once PCR was widely available, is to amplify a particular piece of DNA and then cut it with restriction enzymes.

The advantage of using any restriction site approach is that it potentially covers a large stretch of DNA, and thus is thought to be less sensitive to local vagaries of selection or differences in mutation rate. This is also a disadvantage, of course; generally one does not know exactly where the restriction sites are (e.g., inside a gene

or outside it, in the third position of a codon or not, etc.). Therefore it is not possible to be absolutely certain that a restriction site gain or loss is in exactly the same place across several taxa. Using standard methods, estimates of the size of restriction fragments are accurate only to 50 or 100 bp, which means that two sites very close to each other could easily be confused. There is an additional problem: A restriction site is a four- or six-base sequence of DNA, and it can be lost by mutations in any one of its four or six bases. This means that different mutations will all look the same, and cannot be distinguished. With a complete DNA sequence, some of these ambiguities are no longer a problem. Nonetheless, sequence data have their own set of complexities. These are described briefly in the next section.

Analysis of Molecular Data

There is a huge literature on the use of DNA sequences in phylogeny reconstruction. Here we will discuss some of these uses and some examples that have affected our current view of phylogenetic relationships. The major issues to be addressed are mutation rate, alignment, analytic technique, and the relationship between the history of genes and the history of organisms (gene trees vs. species trees).

Genes accumulate mutations at different rates. This is in part because the gene products (RNA or protein) differ in how many changes they can tolerate and still function. Histones, for example, generally cease to work if many of their amino acids are replaced with different ones, whereas the internal transcribed spacer of ribosomal RNA (ITS) can still fold properly even if many of the nucleotides are changed. Thus genes for histones do not accumulate mutations rapidly, whereas genes for the ITS do, reflecting the different functional constraints on their gene products. This simple observation has implications for the use of particular genes in phylogenetic reconstruction. If a gene is changing slowly, then a lot of data will have to be generated to find mutations from which a phylogeny can be constructed. At a very low mutation rate, the level of variation will approach the irreducible level of sequencing error (often estimated at about 3 in 10,000 bp for a double-stranded sequence), and inferences will become unreliable. Conversely, if a gene is changing too fast, parallelisms and reversals will accumulate to the point that all phylogenetic information is lost—the history of the sequence will be obliterated. The latter problem is particularly acute when working with noncoding sequences or remotely related taxa.

Many of the methods used to analyze molecular data, and the limitations that apply to them, are similar to those for morphological data. There are some methods, however, that were developed specifically for use with molecular data (e.g., neighbor joining, maximum likelihood), and some problems that, while present in all data sets, become more acute with molecular data.

ALIGNMENT OF SEQUENCES

Once sequences are generated, they must be aligned. This is a critical step in that it determines which bases will be compared. It is the stage at which the scientist makes the initial assessment of similarity of nucleotide sites. Alignment is by far the greatest difficulty in using sequence data, and one for which there is no good analytic solution at the moment.

There are many computer programs that will produce alignments, although in practice most systematists rely heavily on alignment "by eye." For many molecules in current use (e.g., *rbcL*), alignment is not a serious problem. For other molecules, such as genes encoding RNAs, alignment can be guided by models of secondary structure (the way the molecule folds). In this case, the secondary structure is used as a template, and the sequence is mapped on it. This ensures that the proposed alignments maintain the structure of the molecule. (Methods for inferring secondary structure, however, have their own limitations.) In protein-coding sequences, alignments generally need to maintain the reading frame, so that any postulated insertions or deletions occur in sets of three, corresponding to the gain or loss of an amino acid. Addition or subtraction of a single base pair will change the entire structure of the protein encoded by the sequence. For example, the sequence ATGTCTCCTGAA codes for the four amino acids Met-Ser-Pro-Glu (the first four amino acids of the large subunit of RuBisCO). If a single base is deleted from the second codon, for example, ATGCTCCTGAA, the protein changes and now is Met-Leu-Leu followed by Asn or Lys.

HOMOPLASY AND LONG BRANCHES

A reversal or convergence at a particular nucleotide is undetectable except via a phylogenetic analysis; more detailed study of the character won't help. Multiple mutations (sometimes called "multiple hits") at a site are another aspect of this problem—such multiple changes may well be invisible and can be corrected for only by assuming particular models of evolutionary change. Discussion of this problem is beyond the scope of this text; it is described in Swofford et al. 1996.

With high rates of mutation, or long evolutionary times between speciation events, parallelisms and reversals will increase purely because of random changes. For example, an adenine at a particular position in a gene may have changed to guanine and then back to adenine. Thus, the actual amount of evolutionary change will be underrepresented by the observed differences. If a high rate of mutation appears in only a few taxa, it creates what has become known as "long branch attraction." If one or more of the species studied have accumulated many mutations since diverging from other taxa, they will appear on long branches in phylogenetic trees (where length equals number of mutations). Because there are only four nucleotides, some mutations will make the sequence of the divergent taxon look more

similar to the sequence(s) of other taxa than it really is, purely by chance. Random mutations tend to make those taxa (species) look alike, and they may appear in the analysis to be closely related even if they are not; the long branches "attract" each other. This can be a particular problem if there is more than one long branch. This could occur in principle with morphological data, but it is more likely with molecular data because the potential number of characters is so large and the available character states (A, C, G, T) are so few.

METHODS OF PHYLOGENY RECONSTRUCTION

Some of the methods of phylogeny reconstruction that have been described in Chapter 2 are particularly appropriate for use with DNA sequence data. These are generally methods that rely on statistical models of how DNA has changed over time. As noted there, for data with little homoplasy, virtually all methods will produce the same phylogenetic tree. In some cases, however, the choice of method will affect the result, and this is particularly true in the case of unequal rates of evolution described above. Because of the problem of multiple mutations at the same site, the observed number of mutations, and thus the apparent divergence between sequences, can easily be less than the actual number of changes. In this case, a correction factor can be applied to estimate the actual evolutionary distance. Which correction factor to use depends on estimates of the probability of particular types of mutations. (A full discussion of this sort of analysis can be found in Swofford et al. 1996.)

GENE TREES VERSUS SPECIES TREES

If a species has a single history, then we expect that all parts of the plant should reflect that history. This implies that any phylogeny based on any gene will reflect the history of the organisms bearing the genes, but we now know that this is not necessarily true. Nuclear genes may or may not track the history of the nucleus, and chloroplasts and mitochondria may or may not have a history different from that of the nucleus. There are three main reasons for this.

1. Mutation is a random process; therefore the phylogeny reconstructed for a particular gene may differ from other genes by chance alone.
2. Hybridization or introgression may transfer some DNA into a different lineage. This is particularly true for organelles, which are not linked to particular nuclear genomes.
3. Polymorphisms in an ancestral species can be lost in descendant species. This can, by chance, happen so that the history of the genes is actually different from the history of the organisms (Figure 5.3).

We are rapidly approaching a time when there will be multiple gene trees for a group of organisms and none of those gene trees will necessarily be exactly the same as the species tree (Box 5A). An early hope was that the relationships among species indicated by DNA were more likely to be correct than those based on morphology; this now seems naive. There are now many examples of plants that have the "wrong" chloroplast, presumably because of introgression.

Molecular Characters

Good systematic work requires detailed knowledge of characters, their underlying biology, and the nature of variation. For morphological characters, this leads naturally into studies of developmental morphology. For molecular characters, it directs our attention to molecular biology and the structure and function of particular molecules. Each molecule has its own role in the cell, and its sequence is constrained according to that role. Each molecule, like each set of morphological characters, has its own natural history, reflecting historical contingency, developmental constraints, past and current adaptations (to both intra- and extracellular factors), and stochastic changes, whether fixed or transient. This means that future molecular systematists will need to become as familiar with the structure and function of the molecules they study as they are with the plants themselves. (At the same time, of course, they must be careful not to overlook the plants for the molecules!) Molecular genetics and biochemistry are becoming increasingly important as tools for understanding evolution.

In the following discussion, we will describe some of the major molecules used in systematic studies and what they do in the cell. The literature on biochemistry and molecular biology, however, should be explored for each molecule used.

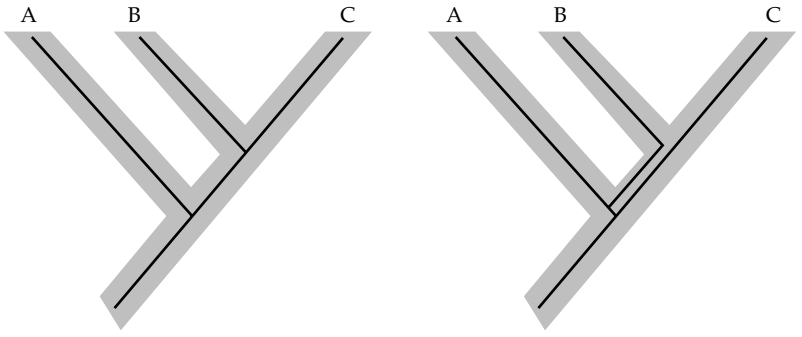

Figure 5.3 A comparison of gene trees and species trees. The gene tree is shown by a dark line, the species tree by gray shading. In the left-hand tree, the genes have the same history as the species that bear them. In the right-hand tree, a polymorphism appears in the lineage leading to species B and C. One of the two gene copies is more closely related to A. Sampling of this gene will lead to incorrect inferences about the species tree. (After Avise 1994.)

BOX 5A *Gene Trees and Species Trees*

The grass family has been the subject of many molecular systematic studies. Consider the subfamily Pooideae of the family Poaceae (Figure 5.4). This group was identified as monophyletic by cladistic studies of morphology, but there were several genera or small tribes, including the genus *Brachyelytrum* and the Stipeae, that were sometimes placed with the pooids and sometimes placed in other subfamilies. We now have five molecular phylogenies of the pooid clade, and these all show the Stipeae as an early-diverging lineage. The morphological characters of the

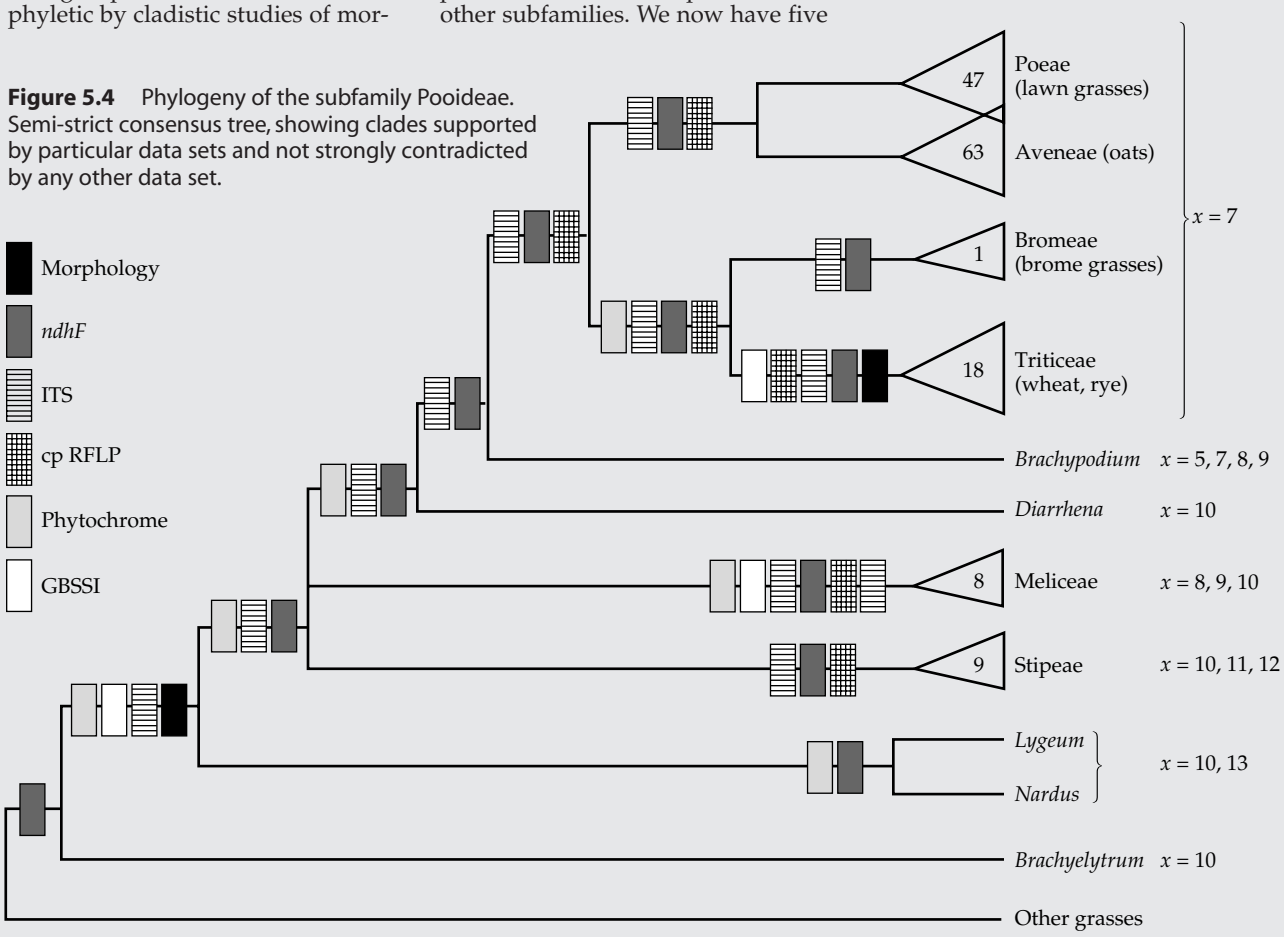

Figure 5.4 Phylogeny of the subfamily Pooideae. Semi-strict consensus tree, showing clades supported by particular data sets and not strongly contradicted by any other data set.

For historical reasons, the most data are available on chloroplast DNA and on nuclear genes for ribosomal RNA. These pieces of DNA are abundant in the cell, making their detection on Southern blots easy. Ribosomal RNA was initially sequenced directly, without reference to the genes; this was possible because its high copy number made it easy to extract and sequence. The strengths and limitations of each of these molecules, as well as those of other molecules, are discussed below. The choice of molecule for phylogenetic analysis is a difficult one and will continue to be a topic of much discussion for the foreseeable future.

CHLOROPLAST DNA STRUCTURE

Many phylogenetic studies of plants have used chloroplast DNA (see reviews by Olmstead and Palmer 1994; Sytsma and Hahn 1997). Analysis of restriction site variation was the method of choice in the 1980s and remains a good way to assess relationships among species that have diverged recently. For taxa that are more distantly related, restriction site variation often becomes hard to interpret—once there have been too many mutations or rearrangements, it is impossible to infer which mutations occurred in which order.

Because restriction site variation is generally lower within than between species, systematists have been somewhat slower to realize the potential of restriction site studies for assessing population histories. In one extensively studied example, Soltis et al. (1991) found that individuals of *Tellima grandiflora* (Saxifragaceae) have two distinct chloroplast genomes, with distinct geographic ranges. The two differ by at least 18 restriction sites. The "northern" genome type occurs in plants from northern Oregon to Alaska, and the "southern"

Stipeae are thus a mixture of synapomorphies linking them with the pooids and symplesiomorphies, which they share with many other grasses. Two of these studies were based on chloroplast DNA, using restriction site polymorphisms (cp RFLP; Davis and Soreng 1993) and sequences of *ndhF* (Catalán et al. 1997). We would expect these to give the same phylogeny because the chloroplast does not recombine and thus has a single history. The other three studies were based on nuclear genes—those for the ITS (Hsiao et al. 1994), phytochrome B (Mathews and Sharrock 1996), and granule-bound starch synthase I (Mason-Gamer et al. 1998). These support the same placement of the Stipeae. The fact that all data from both nuclear and chloroplast genomes suggest the same relationships indicates that the gene trees are probably good estimates of the organismal phylogeny. These data are also congruent with information on chromosome number.

A different result appears when we investigate relationships within the tribe Triticeae (Figure 5.5). For this group we have five molecular phylogenies, all dealing with the diploid genera. The two chloroplast phylogenies, based on restriction site polymorphisms (Mason-Gamer and Kellogg 1996) and *rpoA* sequences (Petersen and Seberg 1997), suggest the same groupings, as expected. However, the three nuclear gene trees (based on three different chromosomes) are significantly different (Kellogg et al. 1996). The explanation for this is not clear, but may involve a history of limited gene flow among the genera. The important point is that not all genes have identical histories. This means that one gene tree needs to be compared with a second one, preferably from a different genome, if we are to begin to infer organismal histories.

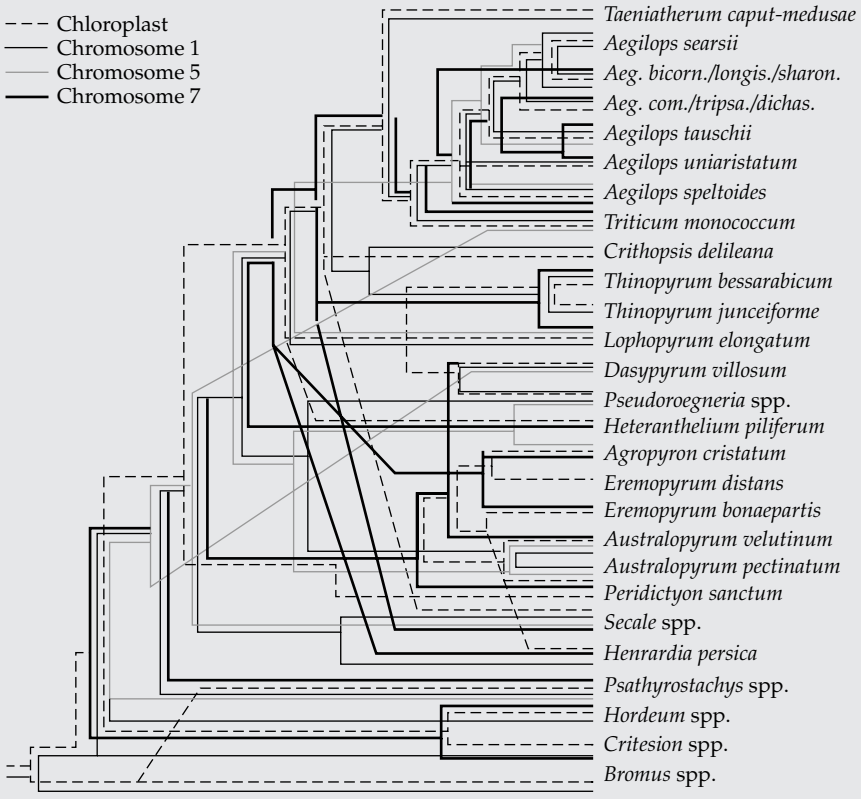

Figure 5.5 Phylogeny of the tribe Triticeae. Data for the two chloroplast phylogenies are from *rpoA* sequences and restriction site polymorphisms. Chromosome 1 and 5 histories are based on sequences of independent sets of 5S DNA spacers. Chromosome 7 history is based on sequences of granule-bound starch synthase I.

type mostly in plants from northern Oregon south into California; there are also a few "southern" plants on Prince of Wales Island in the Alaskan panhandle and on the Olympic Peninsula in Washington. Because the "southern" chloroplast genome is most closely related to that of the genus *Mitella*, it is likely that there was some ancestral introgression from *Mitella* into *T. grandiflora*.

The situation in *Tellima* is not an isolated example. In *Coreopsis grandiflora* (Asteraceae), the chloroplast genome is highly polymorphic and can be divided into two types that differ by at least 19 restriction sites (Mason-Gamer et al. 1995). Both types can be found co-occurring in populations and within different varieties of the species. Other species of the genus *Coreopsis* have one chloroplast type or the other, but it is not known how many are polymorphic like *C. grandiflora*. The chloroplast polymorphism suggests either that *C. grandiflora* is the result of hybridization between two morphologically distinct species, or that both chloroplast types are shared by many members of the genus. The latter explanation suggests that polymorphism can be retained for many years, even through multiple speciation events.

Many other examples of chloroplast DNA variation are cited by Soltis et al. (1992). Such studies illustrate how dynamic plant populations are, and how much gene exchange can occur over evolutionary time.

rbcL Many plant systematists have been involved in a community-wide effort to generate a large database of sequences of the chloroplast gene *rbcL*. This gene encodes the large subunit of the photosynthetic enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), which is the major carbon acceptor in all photosynthetic eukaryotes and cyanobacteria. The

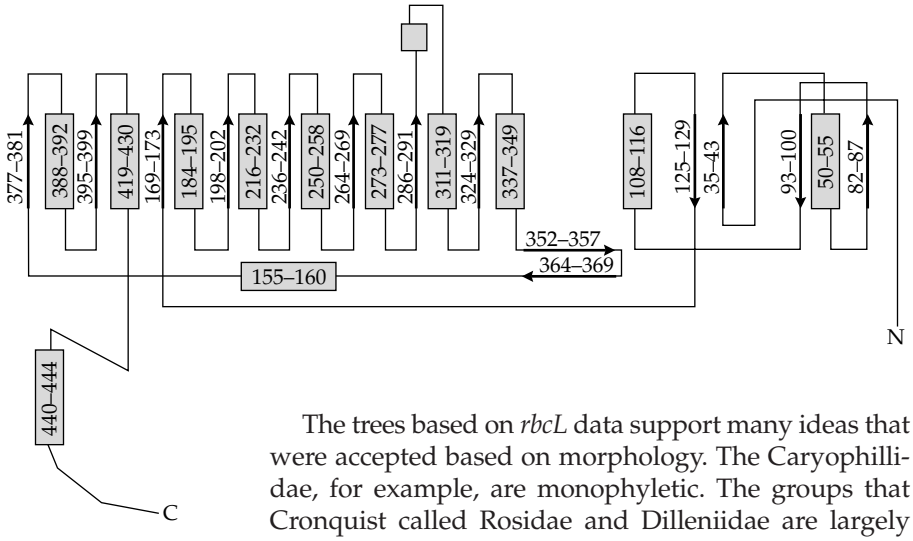

Figure 5.6 Secondary structure of *rbcL*. The numbers refer to numbered amino acid residues. Rectangles represent alpha helices; arrows represent beta sheets. (From Kellogg and Juliano 1997.)

secondary structure of the gene is known (Figure 5.6), and amino acids can be assigned to particular structural components of the gene. This gene was chosen because it is almost universal among plants (excepting only the parasites), it is fairly long (1428 bp), it presents no problems of alignment, and as part of the chloroplast, it is present in many copies in the cell. The enthusiasm for sequencing the *rbcL* gene was aided by the generosity of Gerard Zurawski (University of Georgia), who designed a set of near-universal PCR primers that he distributed freely to anyone who wanted them. The availability of these primers has encouraged many plant systematists to generate *rbcL* sequences and has resulted in well over 2000 sequences, primarily of seed plants. The power of this broadly collaborative approach should not be underestimated.

The gene trees generated from these *rbcL* sequences have had an enormous influence on our view of relationships among angiosperm families, and they are referred to throughout this book. In particular, several studies presented in a single issue of the *Annals of the Missouri Botanical Garden* in 1993 have generated many hypotheses of relationship, which are now available for testing with other molecular and morphological data. These studies are remarkable for their tremendous heuristic value.

Like any heuristic study, however, these studies have limitations. Much attention has been focused on the work of Chase and his co-authors (1993), who attempted to generate a phylogeny for all seed plants using 499 *rbcL* sequences. The published trees turned out not to be the shortest available for that data set, a few of the sequences proved to be pseudogenes, and whole families were represented by single sequences, among other problems. Reanalyses of the 499-taxon data set have found many equally parsimonious trees, many of which are quite different from the ones presented in the original study (Rice et al. 1997).The results, in other words, should be interpreted with caution—a point that the authors of the paper recognized. Nonetheless, *rbcL* data, and the Chase et al. tree in particular, are widely cited and are taken as the starting point for many current research projects. We will therefore discuss the results briefly as an entrée into the current literature.

The trees based on *rbcL* data support many ideas that were accepted based on morphology. The Caryophillidae, for example, are monophyletic. The groups that Cronquist called Rosidae and Dilleniidae are largely intermingled, as had been suspected by anyone who had tried to teach according to Cronquist's system. Several well-known family pairs (e.g., Asclepiadaceae/Apocynaceae, Araliaceae/Apiaceae, Brassicaceae/Capparaceae, here united to preserve monophyly) are supported by *rbcL* data, although the exact relationships between members of the pairs differ. Families long suspected of being polyphyletic (e.g., Saxifragaceae, Caprifoliaceae) appear polyphyletic based on *rbcL* data. The eudicot (tricolpate) clade is supported, albeit weakly, by *rbcL* data.

In other cases, *rbcL* data have helped to resolve relationships that were previously ambiguous. The Ericaceae, for example, were included in Engler and Prantl's Sympetalae, but current thought had placed them well outside it. Trees based on *rbcL* data support the placement of an Ericalean clade in a larger clade with the Asteridae, reuniting much (but not all) of the Englerian Sympetalae. These data also support a monophyletic Rosaceae, within which three of the four subfamilies are monophyletic; some genera will need to be realigned, however (Morgan et al. 1994).

Finally, there are a few cases in which *rbcL* data suggest something quite surprising. We have mentioned the placement of Vochysiaceae and Limnanthaceae in the Myrtalean and glucosinolate clades, respectively. Another striking example is the finding that the nine families with nitrogen-fixing members fall into a single large clade, along with a restricted set of families that are not nitrogen-fixing (Soltis et al. 1995). Because previously the N-fixing families had appeared to be completely unrelated, this finding suggests that these families may have more in common than previously believed.

OTHER CHLOROPLAST GENES

One of the limitations of *rbcL* as a phylogenetic marker is its slow rate of change. It is fundamentally a conservative molecule that is highly constrained at the amino acid level. It is therefore not particularly useful for inferring relationships within or between closely related genera. Instead, other chloroplast genes have been used for

such purposes, notably those for subunit F of NADP dehydrogenase (*ndhF*, in the small single-copy region), those for the α and β″ subunits of RNA polymerase II (*rpoA* and *rpoC2*, in the large single-copy region), and a maturase gene in the intron separating the coding region of *trnK* (*matK*, previously known as ORF, for *o*pen *r*eading *f*rame, K). Because these are all part of the same nonrecombining genome as *rbcL*, they all track the same (generally maternal) history.

AtpB, the gene encoding the β subunit ot ATP synthase, is used increasingly to address the same problems as *rbcL*. It appears to evolve at about the same rate and thus provides additional phylogenetically informative characters.

NUCLEAR GENES

Ribosomal RNA Historically, the only nuclear genes with a high enough copy number for easy study were the ribosomal genes. These are arranged in tandem arrays of several hundred to several thousand copies. The general arrangement of these genes is shown in Figure 5.7. The small subunit (18S) and large subunit (26S) genes are separated by a smaller (5.8S) gene, and the whole set of genes is transcribed as a single unit. There are short transcribed spacers (ITS) between the three genes. Each set of three genes is separated from the following set by a large spacer, variously referred to as the intergenic spacer (IGS), extragenic spacer (EGS), or nontranscribed spacer (NTS). The last is something of a misnomer, in that some of the sequence immediately upstream of the 18S gene and downstream of the 26S is actually transcribed; these regions are sometimes called the external transcribed spacers (ETS). The middle portion of the spacer is not transcribed, and is made up of variable numbers of short repeated sequences (ca. 100–300 bp each). These are thought to play a role in gene regulation.

A completely separate rRNA array encodes only 5S RNA (not to be confused with 5.8S rRNA), a molecule with an unknown function in the ribosome. The 5S rRNA genes are in tandem arrays of several thousand copies, and are separated by nontranscribed spacers.

Such highly repetitive sequences undergo homogenization processes known as **concerted evolution**. If a mutation occurs in one copy of a sequence, it is generally corrected to match the other copies, but sometimes the nonmutated copies are "corrected" to match the mutated one, so that nucleotide changes propagate throughout the array. This means that the many copies of the sequence are generally more similar to one another than they are to copies in other species. Within-species variation does occur, however, because concerted evolution is slower than the mutation rate. This means that some highly repetitive sequences can be used to assess variation within and among populations of the same species.

Ribosomal genes were studied initially by restriction site mapping of the intergenic region. This region, particularly the short repeated sequences, is quite variable even within populations. Its major value has therefore been in studies of closely related plants, where it has been useful in determining population structure and in assessing patterns of hybridization (see Box 5B). King (1993), for example, studied North American and European species of dandelions. All known North American dandelions are asexual, but European ones are both sexual and asexual. All rDNA and chloroplast DNA genotypes found in asexual plants were also found in sexual plants, but in different combinations. This finding is evidence that asexual plants are produced frequently by hybridization between sexual ones.

Methods for direct sequencing of RNA (without recourse to the DNA that encodes it) were developed earlier than rapid DNA sequencing methods, and well before PCR. This led to some early hope that the 5S rRNA gene sequences would be the key to evolutionary history. Researchers soon recognized, however, that the 5S rRNA genes were too short (at 120 bp) and too conservative to illuminate relationships. The nontranscribed spacer region has been used with some success in closely related groups, however.

Sequences of the 18S and 26S genes are more promising. These genes are large (about 1800 and 3300 bp respectively). They have some regions that are highly conserved, which helps in alignment, and others that are quite variable, which helps to distinguish phylogenetic groups. A large cooperative effort, analogous to the *rbcL* study, is now under way to generate a large database of 18S sequences (Soltis et al. 1997). The results of this study are helping to test some of the hypotheses that emerged from the *rbcL* study. In broad outline, phylogenies of the two genes have found similar groups among the angio-

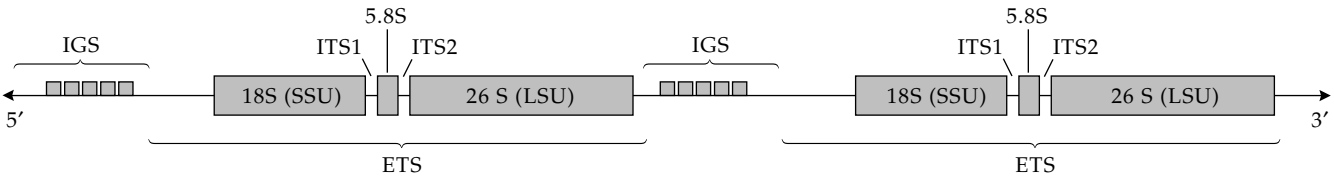

Figure 5.7 Structure of the ribosomal array. Coding regions of the 18S small subunit (SSU), 5.8S, and 26S large subunit (LSU) are shown as heavy open rectangles; the transcription unit is bracketed. Spacers are indicated by black lines, and the short repeats in the intergenic spacer (IGS) are indicated by small boxes. ETS = external transcribed spacer, ITS = internal transcribed spacer.

BOX 5B *Molecular Data Reveal Ancient Hybridization*

Wendel and his colleagues (1995) studied the evolution of the genus *Gossypium*, which includes all the species that produce cotton. They used isozymes, nuclear ITS sequences, and chloroplast restriction site analysis to study the history of both diploid and tetraploid species. Most of their data indicate that the New World diploids (with a genome designated D; Figure 5.8) are monophyletic, as are the Old World diploids (genome groups A, B, and F). The surprise came in analyzing the New World tetraploids, including *Gossypium hirsutum*, the source of most of the world's commercial cotton. These were formed by allopolyploidization of A and D genomes. Wendel and his colleagues found that *G. hirsutum* has a chloroplast derived from one of the African species, and that it must have acquired it only about 1–2 million years ago, well after the formation of the Atlantic Ocean. One other New World species has elements of an Old World ITS sequence as well. We do not know how such long-distance gene flow might have occurred.

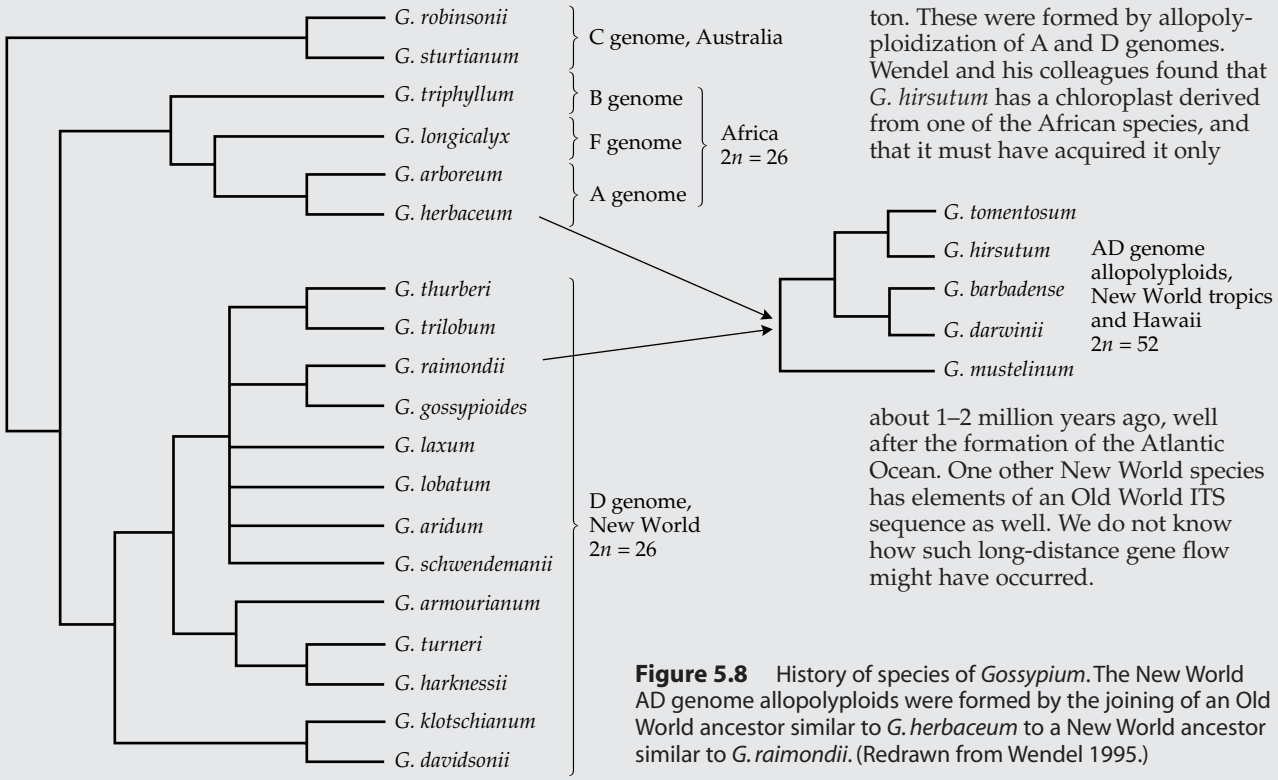

Figure 5.8 History of species of *Gossypium*. The New World AD genome allopolyploids were formed by the joining of an Old World ancestor similar to *G. herbaceum* to a New World ancestor similar to *G. raimondii*. (Redrawn from Wendel 1995.)

sperms. Many exciting insights are forthcoming as details of these gene trees continue to be worked out.

Recently some researchers have used the ITS region to determine relationships among species. In general, the ITS region has supported relationships inferred from the chloroplast or from morphology. In other cases, however, it has proved to be polymorphic within species, or even within individual plants, suggesting that concerted evolution has not completely homogenized the repeats. Although this polymorphism may be a problem in some cases, it also provides tools for understanding gene flow and population-level variation.

Low copy number genes Work on genes with low copy numbers is in its infancy in plant molecular systematics, and it is safe to say that no nuclear gene has yet been used across enough groups to provide a really clear comparison with chloroplast or ribosomal data. The difficulties of working with low copy number genes are appreciable, but by no means insurmountable. In order for a gene to be a useful phylogenetic indicator, it must not be easily confused with any other gene. This requirement can be a problem because many nuclear genes are duplicated or exist as a part of a small set of genes (a gene family). If there are multiple genes in the gene family, they must be distinguishable and not undergo any sort of concerted evolution. Population genetic theory suggests that allelic variation should not be misleading above the species level because alleles in one species should be more closely related to each other than they are to alleles in other species. Data on the genes for alcohol dehydrogenase in *Arabidopsis* and granule-bound starch synthase I (*waxy*) in grasses and Rosaceae support this expectation, but it still needs to be tested more broadly.

The gene for phosphoglucoisomerase (PGI) has been studied in some detail in the genus *Clarkia*, by Gottlieb and his colleagues (1996). There are actually two PGI genes in angiosperms, one used in the cytosol and one in the plastid. The two can easily be distinguished. Isozyme studies indicated that some species of *Clarkia* had two copies of the cytosolic gene, rather than just one. Studies

of DNA sequences showed that both copies were present in *all* species of the genus, but that in some species the extra copy had accumulated so many mutations that it could no longer be translated properly and thus would not appear on an isozyme gel. Such nonfunctional genes are known as **pseudogenes**. The sequences themselves, however, provided robust and identical gene trees, which clarified relationships among the species.

In general, nuclear genes and gene families need to be extensively characterized before they can be used to infer relationships. Other nuclear genes that are good candidates for phylogeny reconstruction are those for the phytochromes, the small heat-shock proteins, and glutamine synthetase. Before such genes can be used reliably for phylogenetic studies, considerable data must be acquired to determine the taxonomic level at which the genes vary, the copy number of the genes, and whether the copies tend to correct each other. The latter process, concerted evolution, can lead to confusion about which gene copies are most closely related to each other. Over the next several years, low copy number nuclear genes will give us much new information on plant phylogeny.

High copy number noncoding nuclear sequences

Unlike chloroplast, ribosomal, and nuclear protein-coding genes, high copy number noncoding sequences appear to evolve rapidly, and are thus useful for addressing questions at the species level or below. These high copy number sequences are generally short sequences that are repeated many times, often at many locations in the genome. So-called **minisatellites** or **variable number tandem repeats** (**VNTR**) are made up of repeated sequences that are generally tens of base pairs in length. There are also similar regions, known as **microsatellites**, in which the repeats are much shorter, consisting of only two or three nucleotides. Such repeated sequences are unstable and prone to errors in replication usually deriving from **replication slippage** (although unequal crossing-over is also a possibility in some cases). Replication slippage occurs as the DNA is being copied. The strands separate for replication, but reanneal out of register, leading to a loop in the DNA. Mismatch repair mechanisms then either remove the loop (leading to a loss of the repeat unit) or insert extra bases on the opposite strand (leading to a duplication). Because of this instability, individual organisms often vary in the number of repeats at a particular satellite locus. This variation can be used to determine a DNA "fingerprint" unique to a particular plant or closely related group of plants. Studies of population structure generally depend on accurate assessment of relationships among individual plants, and these markers are useful for such assessment.

Another method often used in studies at the population level is the **random amplified polymorphic DNA** (**RAPD**) method. In this technique, short (10 bp) PCR primers are designed with arbitrary sequences. These short random sequences will generally match one or more sequences somewhere in the genome of the plant, and the primers will bind to and amplify a fragment of DNA. By doing many such PCRs with random primers, one can generally find fragments that distinguish individual plants or populations. This allows for a rapid assessment of how many genotypes are present in a population and a rough estimate of how different they are. The technique is limited, however, because the identity of the fragments is not known. In other words, a fragment of 150 bp in one plant may not actually represent the same part of the genome as a 150 bp fragment in another plant. Verifying the identity of the fragments requires Southern blotting or restriction site analysis, at which point the technique may become as laborious as restriction site studies or sequencing. Other techniques, such as AFLP (amplified fragment length polymorphisms), have been developed to circumvent the problems of RAPD, but a full discussion of these is beyond the scope of this book.

Summary

Molecular techniques provide powerful tools for the study of evolution and phylogeny. Most data on relationships at the species level and above have so far come from the chloroplast genome and the highly repeated sequences of ribosomal RNA. Future data are likely to come also from low copy number nuclear genes. New tools are continually being developed for the study of variation within and among conspecific populations, including methods of genome mapping. As these tools come into more widespread use, they will provide new insights into the processes of population-level differentiation.

No matter how powerful the molecular data, however, morphological data will remain critical for phylogenetic studies. The major questions in plant systematics are still morphological. Questions about the origin of species, the mechanisms of diversification, and the best way to classify that diversity all require understanding of morphology as well as phylogeny. We can now envision a time when robust phylogenies will have been constructed for all groups of plants, and the question of systematics will shift from "What is the phylogeny of my group?" to "How did the morphological diversity arise?"

Literature Cited

Avise, J. C. 1994. *Molecular markers, natural history and evolution*. Chapman & Hall, New York.

Bradshaw, H. D., S. M. Wilbert, K. B. Otto and D. W. Schemske. 1995. Genetic mapping of floral traits associated with reproductive isolation in monkeyflowers (*Mimulus*). *Nature* 376: 762–765.

Catalán, P., E. A. Kellogg and R. G. Olmstead. 1997. Phylogeny of Poaceae subfamily Pooideae based on chloroplast *ndhF* gene sequences. *Mol. Phylog. Evol.* 8: 150–166.

Chase, M. W. and 41 others. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann. Missouri Bot. Gard.* 80: 528–580.

Crawford, D. J. 1990. *Plant molecular systematics: Macromolecular approaches*. John Wiley & Sons, New York.

Davis, J. I. and R. J. Soreng. 1993. Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. *Am. J. Bot.* 80: 1444–1454.

Downie, S. R., E. Llanas and D. S. Katz-Downie. 1996. Multiple independent losses of the *rpo*C1 intron in angiosperm chloroplast DNAs. *Syst. Bot.* 21: 135–151.

Gottlieb, L. D. and V. S. Ford. 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred from the nucleotide sequences of *Pgi*C. *Syst. Bot.* 21: 45–62.

Hillis, D. M., C. Moritz and B. K. Mable. 1996. *Molecular systematics*, 2nd ed. Sinauer Associates, Sunderland, MA.

Hsiao, C., N. J. Chatterton, K. H. Asay and K. B. Jensen. 1994. Molecular phylogeny of the Pooideae (Poaceae) based on nuclear rDNA (ITS) sequences. *Theor. Appl. Genet.* 90: 389–398.

Jansen, R. K. and J. D. Palmer. 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Nat. Acad. Sci., USA* 84: 5818–5822.

Kellogg, E. A., R. Appels and R. J. Mason-Gamer. 1996. When genes tell different stories: The diploid genera of Triticeae (Gramineae). *Syst. Bot.* 21: 321–347.

Kellogg, E. A. and N. D. Juliano. 1997. The structure and function of RuBisCO and their implications for systematic studies. *Am. J. Bot.* 84: 413–428.

King, L. M. 1993. Origins of genotypic variation in North American dandelions inferred from ribosomal DNA and chloroplast DNA restriction enzyme analysis. *Evolution* 47: 136–151.

Mason-Gamer, R. J., K. E. Holsinger and R. K. Jansen. 1995. Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae). *Mol. Biol. Evol.* 12: 371–381.

Mason-Gamer, R. J. and E. A. Kellogg. 1996. Chloroplast DNA analysis of the monogenomic Triticeae: Phylogenetic implications and genome-specific markers. In *Methods of genome analysis in plants: Their merits and pitfalls*, P. Jauhaur (ed.), 301–325. CRC Press, Boca Raton, FL.

Mason-Gamer, R. J., C. F. Weil and E. A. Kellogg. 1998. Granule-bound starch synthase: Structure, function, and phylogenetic utility. *Mol. Biol. Evol.*, 15 1658–1673.

Mathews, S. and R. A. Sharrock. 1996. The phytochrome gene family in grasses (Poaceae): A phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. *Mol. Biol. Evol.* 13: 1141–1150.

Miyamoto, M. M. and J. Cracraft. 1991. *Phylogenetic analysis of DNA sequences*. Oxford University Press, Oxford.

Morgan, D. R., D. E. Soltis and K. R. Robertson. 1994. Systematic and evolutionary implications of *rbcL* sequence variation in Rosaceae. *Am. J. Bot.* 81: 890–903.

Olmstead, R. G. and J. D. Palmer. 1994. Chloroplast DNA systematics: A review of methods and data analysis. *Am. J. Bot.* 81: 1205–1224.

Petersen, G. and O. Seberg. 1997. Phylogenetic analysis of the Triticeae (Poaceae) based on *rpoA* sequence data. *Mol. Phylog. Evol.* 7: 217–230.

Rice, K. A., M. J. Donoghue and R. G. Olmstead. 1997. Analyzing large data sets: *rbcL* 500 revisited. *Syst. Biol.* 46: 554–563.

Rieseberg, L. H., C. VanFossen and A. Desrochers. 1995. Genomic reorganization accompanies hybrid speciation in wild sunflowers. *Nature* 375: 313–316.

Rieseberg, L. H., B. Sinervo, C. R. Linder, M. Ungerer and D. M. Arias. 1996. Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. *Science* 272: 741–745.

Rodermel, S. 1993. Genetic map of *Zea mays* plastid chromosome. *Maize Newsletter* 67: 167–168

Soltis, D. E., M. Mayer, P. S. Soltis and M. Edgerton. 1991. Chloroplast DNA variation in *Tellima grandiflora* (Saxifragaceae). *Am. J. Bot.* 78: 1379–1390.

Soltis, D. E., P. S. Soltis and B. G. Milligan. 1992. Intraspecific chloroplast DNA variation: Systematic and phylogenetic implications. In *Molecular systematics of plants*, P. S. Soltis, D. E. Soltis and J. J. Doyle (eds.), 117–150. Chapman and Hall, New York.

Soltis, D. E., P. S. Soltis, D. R. Morgan, S. M. Swensen, B. C. Mullin, J. M. Dowd and P. G. Martin. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Nat. Acad. Sci., USA* 92: 2647–2651.

Soltis, D. E. and 15 others. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Ann. Missouri Bot. Gard.* 84: 1–49.

Soltis, D. E., P. S. Soltis and J. J. Doyle (eds.). 1998. *Molecular systematics of plants. II. DNA sequencing*. Kluwer, Boston.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. 1996. Phylogenetic inference. In *Molecular systematics*, 2nd ed., D. M. Hillis, C. Moritz and B. K. Mable (eds.), 407–514. Sinauer Associates, Sunderland, MA.

Sytsma, K. J. and W. J. Hahn. 1997. Molecular systematics: 1994–1995. *Prog. Bot.* 58: 470–499.

Wendel, J. F., A. Schnabel and T. Seelanan. 1995. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. *Mol. Phylog. Evol.* 4: 298–313.