Beyond Simple Homology Searches: Multiple Sequence Alignments and Phylogenetic Trees

Rebecca A. Zufall[1]

[1]University of Houston, Houston, Texas

ABSTRACT

Phylogenetic trees represent hypotheses about evolutionary relationships between organisms or nucleotide or amino acid sequences. Because the best BLAST hit often does not represent the most closely related sequence, phylogenetic analyses are an essential extension of inquiry into any new protein or gene. In this unit, the reader will first learn how to create a multiple sequence alignment using ClustalX. He or she will then learn how to use that alignment to build a neighbor-joining phylogeny using the program Geneious. Finally, the user will learn how to interpret the phylogeny in light of the research questions. *Curr. Protoc. Essential Lab. Tech.* 1:11.3.1-11.3.17. © 2009 by John Wiley & Sons, Inc.

Keywords: Phylogeny • alignment • neighbor-joining • homology • ClustalX • Geneious

OVERVIEW AND PRINCIPLES

Phylogenetic trees, or phylogenies, are graphical representations of hypotheses about the relationships between organisms or sequences and their evolutionary history. Biologists have long used phylogenetic reconstruction to address a variety of issues in evolutionary biology. Recently, the utility of phylogenetics in other fields of biology has also become apparent, and its use has become widespread. For example, phylogenies are used to infer gene function from sequenced genomes (Eisen, 1998) or identify unclassified organisms from metagenomic data (McHardy and Rigoutsos, 2007). Unfortunately, this tool is not always applied or interpreted appropriately by those not trained in evolutionary biology (or even by those who are). The purpose of this unit is to give an introduction to some of the basic methods of sequence alignment and phylogenetic analysis; this will hopefully allow you to appropriately infer and interpret phylogenetic trees to address questions relevant to your research.

Phylogenies are commonly used to ask questions of the type “What are the relationships among a group of organisms?” or “How are a set of genes related to each other?”. All phylogenetic analysis relies on the fact that organisms, and hence their DNA and protein sequences, share a common origin. The goal of phylogenetics then is to reconstruct the evolutionary history that has occurred since that common origin.

For example, Figure 11.3.1 shows the phylogeny of four species. The tips, labeled with species names, represent extant taxa. The circled node labeled “A” represents the most recent common ancestor of all four species. The branching pattern, or topology, of the phylogeny indicates that flies diverged from the lineage that led to the other three species before any of those species diverged from each other. Because humans and chimps shared a common ancestor (node C) more recently than any other species on the tree, we can say that humans and chimps are more closely related to each other than they are to any of the other species on the tree. Likewise, mice are equally related to humans and chimps

Current Protocols Essential Laboratory Techniques 11.3.1-11.3.17, June 2009
Published online June 2009 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/9780470089941.et1103s01

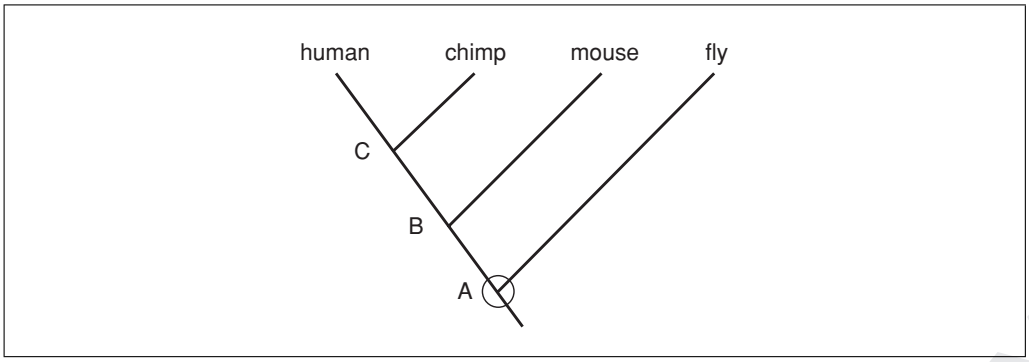

Figure 11.3.1 This phylogeny shows the evolutionary relationships between four extant species. The nodes labeled A, B, and C represent the most recent common ancestor: A, common ancestor of all four species; B, common ancestor of human, chimp, and mouse; C, common ancestor of human and chimp.

because mice shared a most recent common ancestor with each of those species at the time represented by node B. For a nice set of exercises to get you familiar with "tree thinking," see Baum et al. (2005).

Perhaps the most important concept to understand before you begin to align sequences or build a phylogeny is that of homology. "Homologous" sequences (or other traits) are those that share a common ancestry. Since the goal of phylogenetic reconstruction is to understand ancestral relationships, you can see why it is important to use only sequences that share a common origin in your analysis. Unfortunately, determining whether or not two sequences are homologous is not always straightforward. For example, two proteins may be very similar due to similar functions, but are not derived from a common origin. Likewise, two sequences may share very little similarity because they have rapidly acquired new functions, but if they are derived from a common origin, they are nonetheless homologous.

As you can see from the above example, similarity is not the same as homology. Similarity indicates how much two sequences look like each other, or how many residues they share in common. Similarity can be expressed in degrees, e.g., "sequences A and B are highly similar." Homology indicates that such similarity is due to shared ancestry. Sequences are either homologous or not; you cannot express degrees of homology. However, in practice, we often make hypotheses about homology based on similarity (which likely leads to the confusion between these terms); the more similar two sequences are, the more confident we are that they are homologous. However, this is not always the case; there are many circumstances where sequences can be similar, but not homologous. For example, short sequences may be similar just by chance, or sequences may be similar due to selection for a similar function, such as binding a particular substrate. Thus, the key distinction to make is whether sequence similarity is due to shared ancestry (implying homology) or not.

We can further consider two different types of homologous relationships: orthologous and paralogous. Orthologs are homologs that diverged following speciation events (e.g., human A and chimp A in Fig. 11.3.2). Orthologs often retain the same function in different species. For example, α-tubulin in humans and α-tubulin in rats are orthologous. Paralogs are homologs that have diverged following gene-duplication events (e.g., human A and human B in Fig. 11.3.2). For example, in humans α-tubulin and β-tubulin are paralogs. Multiple gene duplication events result in a family of paralogs, or gene family; members of a gene family generally have related but not identical functions. There are many well known examples of gene families, such as globins, G-proteins, and a variety of kinases.

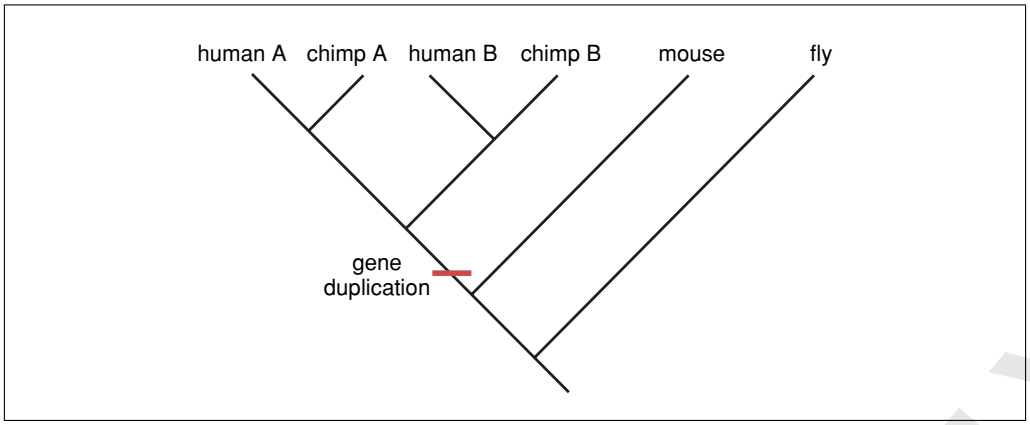

Figure 11.3.2 This phylogeny shows the relationship between various orthologs and paralogs of a gene. Prior to the divergence of humans and chimps, this gene underwent a gene-duplication event (indicated by the horizontal bar). That gene duplication event resulted in two paralogs: A and B. Speciation between humans and chimps resulted in the orthologs "human A" and "chimp A," and the orthologs "human B" and "chimp B."

As discussed above, phylogenies are generally used to answer two types of questions, one about relationships among organisms and the other about relationships between genes. These two types of questions result in two different kinds of phylogenies: a species tree represents relationships between organisms and a gene tree represents relationships between genes. Sometimes species trees are identical to gene trees, i.e., the phylogenetic history of a single gene can adequately represent the history of the species that carry those genes. However, there are many circumstances under which gene trees are not congruent with species trees. For example, if paralogs, rather than orthologs, are used to infer a phylogeny, the resulting gene tree will likely not be a good representation of species relationships. Likewise, if a gene has been transmitted laterally rather than vertically, i.e., to an individual that is not the offspring of the gene donor, and often not of the same species as the gene donor (lateral or horizontal gene transfer), then the resulting phylogeny will not represent the true species relationships.

The concept of homology can be further extended from the level of the gene to sites within a gene. In particular, the goal of multiple sequence alignment is to determine positional homology, i.e., which sites within a gene share a common ancestry, across all of the sequences being analyzed. When you perform a BLAST search (*UNIT 11.1*), the search algorithm may be implicitly making predictions about positional homology between two sequences, but when comparing many sequences for phylogenetic analysis, it is necessary to determine positional homology across all of the sequences; this is what multiple sequence alignment algorithms do. Multiple sequence alignment thus allows us to ask questions about how particular sites within a gene are evolving across all of the species under study.

Once you have built a phylogeny, it is informative to ask how confident are you that the relationships depicted in the tree are accurate. A common way to assess the reliability of a phylogeny is to use a resampling method called bootstrapping. In general, bootstrapping is a statistical tool that relies on random resampling of data to make parameter estimates. Bootstrapping in phylogenetics involves resampling the sequence data to build a new tree. This procedure is replicated many times to get many trees, each built with a different subsample of sequence data. These bootstrap replicates are then compared with each other to determine whether different subsamples of the data produce phylogenies with similar relationships. In particular, each branching relationship, or node, is compared between trees and assigned a score (bootstrap value) that indicates the percentage of replicate trees in which that node was found. The more of these replicate trees that

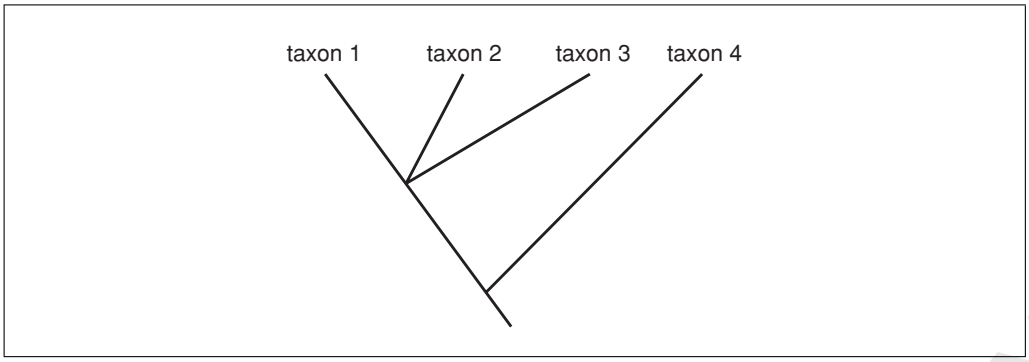

Figure 11.3.3 Phylogeny showing the relationships between taxa 1, 2, and 3 as a polytomy, representing unresolved relationships between these taxa.

give the same branching relationship, the more confident we are that this relationship is accurate. In other words, the higher the bootstrap value of a node in the tree, the more confident we are that that node reflects the true evolutionary history. By convention, if bootstrap support is less than 50%, indicating very weak support for that node given the data, that relationship is not used in the final tree. Instead, the relationships between the taxa for which there is weak support are shown as an unresolved polytomy, or node with more than two descendent lineages (Fig. 11.3.3), indicating that we cannot confidently determine the dichotomous branching relationships between these taxa.

STRATEGIC PLANNING

Preparing Your Data

Find the sequences for your analysis

Using the tools you learned in the previous units (*UNITS 11.1 & 11.2*), retrieve a set of homologous sequences, either nucleotide or amino acid, using either a BLAST search or text search on your gene of interest. Determine which sequences you will use in your analysis based on what question(s) you want to address and the similarity, as measured by E-value, among available sequences. In general, you can hypothesize that two sequences are homologous if a BLAST E-value is less than 10^{-5} (Hall, 2004), but this is by no means a hard and fast rule.

Many factors will determine which is the best set of sequences that you could use in your analysis. The most important consideration (once you understand that you can only use homologous sequences) is the question that you are trying to address. For example, if you are trying to elucidate the function of a protein that you have sequenced, then it is essential that you include related protein sequences that have their functions identified.

The set of sequences must also include the appropriate amount of diversity. For example, if you do a BLAST search of a sequence and the first 10 hits are 99.9% identical, there will not be enough differences among these sequences to determine their evolutionary history. On the other hand, you don't want sequences that are so different that it becomes impossible to determine which sites are homologous (e.g., see E-value cutoff described above). Ideally, your dataset can contain sequences with a range of pairwise identity from about 30% for amino acids or 50% for nucleotides to 100%.

Determining how many sequences to include is also important. In general, the more sequences you include, the more robust your phylogeny will be (Zwickl and Hillis, 2002). However, for every additional sequence you include, you increase the amount of time required for all stages of the analysis. The number of sequences to include, however, will ultimately depend on the question that you are addressing.

Format and download the sequences

To align and further analyze your sequences, you will need to export your sequences in a file type that is recognized by the alignment software you will use. The simplest format for this is called FASTA. You can recognize a FASTA-formatted file of sequences because ">" is the first character at the beginning of each sequence (See *UNIT 11.1*).

If you retrieved your sequences using a BLAST search, you will first need to check the box next to the sequences you want to use in your analysis, then click the "Get Selected Sequences" button. When these sequences are displayed, you will see a list of sequences similar to that from a text search, as in Figure 11.3.4. Select the checkbox next to those sequences that you want to download. Change the "Display" pull-down menu to "FASTA" and the "Send To" menu to "File." A file containing the sequences you selected, in FASTA format, will be downloaded to your computer. The location of the download will depend on the settings you have selected for your browser. You will probably want to rename the sequence file and move it to a useful folder.

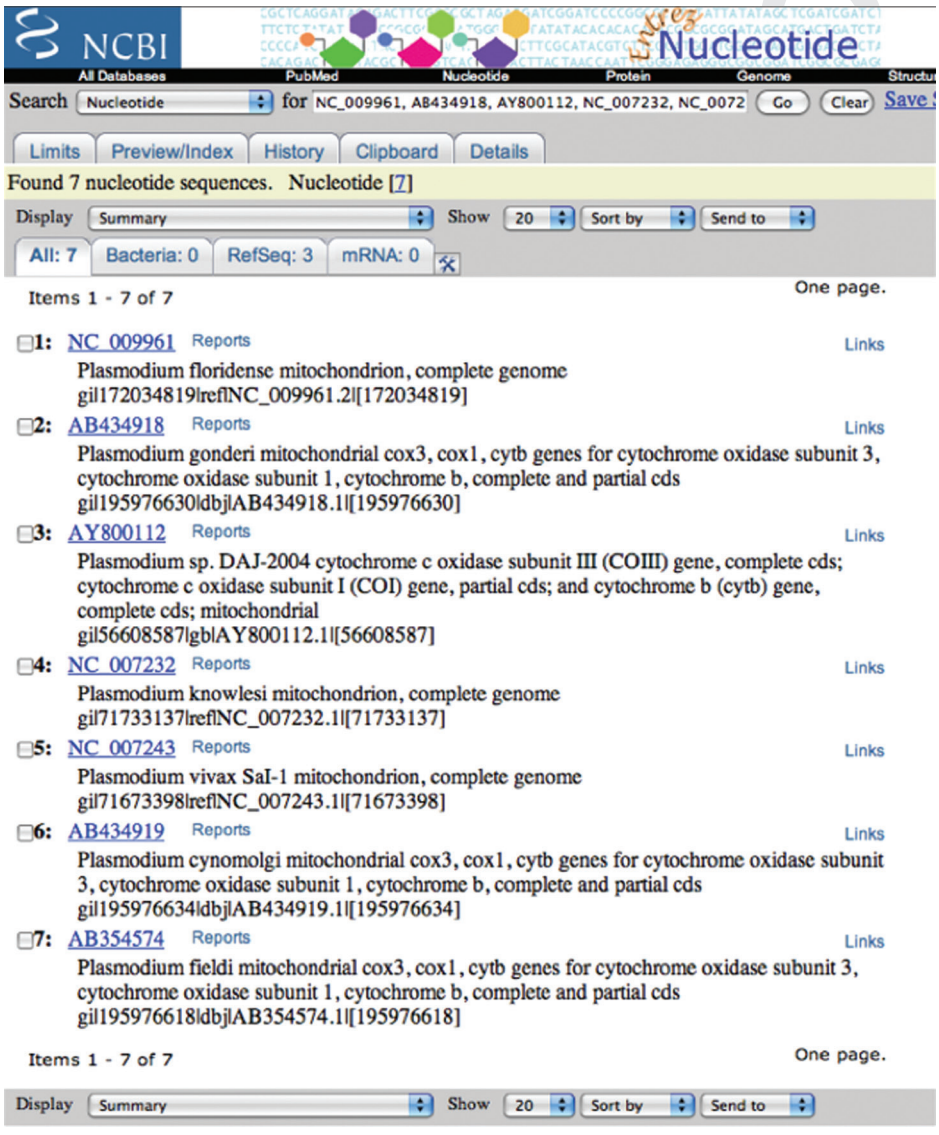

Figure 11.3.4 Results of a GenBank query for cytochrome oxidase I sequences from select *Plasmodium* species. This search returns several full-length mitochondrial sequences. The genes of interest can be extracted from these sequences as described in the text.

If your search returns entries that contain more genes than the one that you are interested in, e.g., complete mitochondrial genomes (as in the example described in Understanding Results; also see Fig. 11.3.4), you will need to select just the portion of the entry you are interested in. Click on the accession number of an entry, or change "Display" to "GenBank." Next, find the gene that you are interested in and click on "gene" or "protein" next to the entry. This will display the entry for just that gene. You can then download these sequences in FASTA format as described above.

The downloaded files are text files that can be opened in any word processing program so that you can examine and manipulate the contents of the file. You can rename the sequences if desired to make them easier to recognize. If you have your own sequence that you want to add to the file, you can do that now. Just be sure to start each new sequence with ">" followed by the sequence name and to save the file as "text only."

Multiple Alignment

The purpose of sequence alignment is to arrange your sequences so that each column of the alignment represents a homologous site in the gene or protein. This allows you to determine where in the sequence an evolutionary change happened, either a point mutation or an insertion or deletion. Multiple sequence alignment algorithms make predictions of positional homology by adding gaps to one or more sequences until the sequences match the "best." This "best" matching is determined based on user-defined penalties assigned to different types of evolutionary events. This process is similar to pairwise alignment (*UNIT 11.2*); however, with multiple alignment, positional homology is inferred across all of the sequences in your analysis (rather than just two sequences at a time).

There are many programs available for sequence alignment. ClustalX (Larkin et al. 2007) is probably the most commonly used. ClustalX is available for download for Macintosh, Windows, and Unix from *http://www.clustal.org*. Documentation and additional help is available at this site. In addition, if you prefer to run a Web-based version of Clustal, you can find links to online servers from this site. To run ClustalX on your computer, you must download the appropriate version for your operating system and decompress it.

Once you have downloaded the appropriate sequences for your analysis, creating a multiple sequence alignment is quick and easy. However, making sure that the alignment makes sense, and making any adjustments, can take some time.

Building the Tree

Geneious (Drummond et al., 2008)

There are many software packages that you can use for phylogenetic reconstruction. MEGA is often used for building neighbor joining trees, but an excellent description of how to use MEGA for this purpose is given in Hall (2007), so there is no point in repeating that here. Instead, we will use the Geneious package. Geneious is available for free download from *http://www.geneious.com/*. A version with more features (Geneious Pro) requires purchase of a license; those additional features will not be required to follow the discussion in this unit. Download and install the appropriate version of Geneious for your computer. The Geneious interface is fairly intuitive, and you will find a nice tutorial in the Help panel on the right-hand side of the Geneious window (Fig. 11.3.5). You will notice that Geneious is also capable of performing multiple sequence alignment; however, unless you purchase the Pro version, the Clustal algorithm is not available from within Geneious.

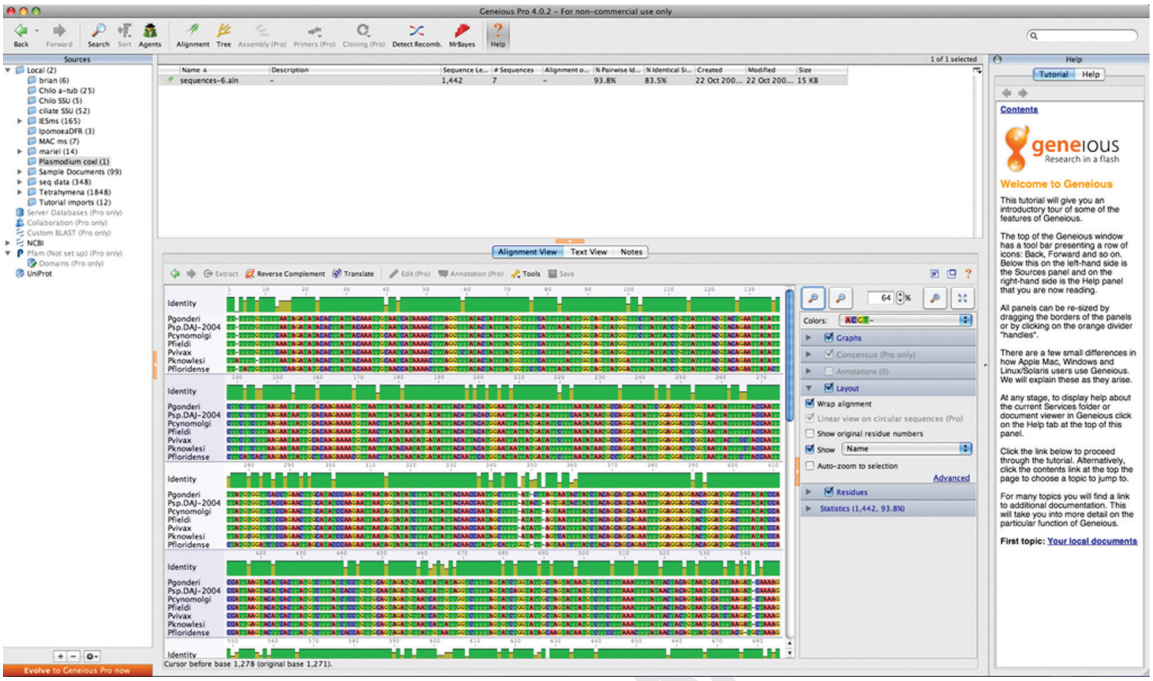

Figure 11.3.5 The Geneious interface. Across the top is the toolbar. The left panel shows folders of the local documents and links to NCBI database searches. The right panel contains a tutorial and Help files. The top panel shows the contents of the selected documents folder. The bottom panel is the sequence/tree viewer. To the right in the bottom panel are a series of options that allow you to change the way you view the sequences. This screenshot shows the alignment of *Plasmodium* sequences that were imported from ClustalX.

Tree-building method

There are several different methods for reconstructing phylogenies: distance methods, parsimony, maximum likelihood, and Bayesian are the most popular. Distance methods build trees by sequentially connecting taxa that have the fewest differences between them. Parsimony methods find the tree that requires the fewest number of evolutionary changes to explain the differences between the taxa. Maximum likelihood methods evaluate trees based on the probability that the tree would give rise to the observed data, given a particular model of evolution. Bayesian methods use a simulation technique to produce a posterior probability distribution of trees given the data and any prior information on relationships among taxa. For more on these methods, see Hall (2007).

Only distance methods are available in the basic version of Geneious. Distance algorithms compare each pair of sequences in an alignment and determine the number of changes between them. These pairwise distances are then used to construct a tree by sequentially joining pairs of taxa that have the smallest distances between them. This method works because, on average, the longer two taxa have been diverging, the greater the number of differences will be between them. For more on distance methodology see Hall (2007) or Felsenstein (2004). We will use the neighbor-joining distance method.

Genetic Distance Model

Neighbor-joining methods require you to specify a model of sequence evolution. These models differ in the way that they parameterize various aspects of sequence evolution, such as the rate of different types of mutations. For nucleotide alignments, Geneious gives you three options for this model: Jukes-Cantor, HKY, and Tamura-Nei. For protein alignments, the only option in Geneious is Jukes-Cantor. Under Jukes-Cantor (Jukes

and Cantor, 1969), all mutations occur at equal rates and equal base (or amino acid) frequencies are assumed. The HKY model (Hasegawa et al., 1985) allows for different rates of transitions and transversions and does not assume equal base frequencies. The Tamura-Nei model (Tamura and Nei, 1993) adds one parameter in addition to HKY that allows different rates for the two different types of transitions. It is best to use the simplest model that adequately explains your data. Ideally, you would determine which model this is using a program such as Modeltest (Posada and Crandall, 1998); however, this cannot be done in the basic version of Geneious. In the examples described here, we will use HKY, which we hypothesize fits the data better than Jukes-Cantor, but does not include more parameters than are necessary to adequately explain the data.

Outgroup

If you know that one of your taxa or sequences is more distantly related to all of the other taxa than those taxa are to each other (e.g., fly in Fig. 11.3.1), then designate that taxon or sequence as the outgroup. This allows you to root your phylogeny, and thus determine the order of branching events. Outgroups are not necessary; if you do not have a taxon that you can use as an outgroup, you will be able to determine the relationships between species in your phylogeny, just not the branching order.

BASIC PROTOCOL 1

CREATING A MULTIPLE SEQUENCE ALIGNMENT

When you open ClustalX, you will see an empty window waiting for you to import sequences. Under the File menu, select "Load Sequences." A dialog box will allow you to select the file that contains the sequences that you have prepared for analysis. When you select an appropriate file and click "Open," you will see your sequences in the ClustalX window. If your sequences are in multiple files, you can import additional sequences using the option under the File menu, "Append Sequences."

Align sequences in ClustalX

When all of your sequences are loaded (or appended), you are ready to align your sequences.

1. Go to the "Alignment" menu and select "Do Complete Alignment."

   You will be asked to specify output files for the guide tree and the alignment. The default is to save these files to the same folder and name as your input file, with different extensions (.dnd for the guide tree and .aln for the alignment).

2. After you have specified output files, click "OK."

   When the alignment is complete, you will see the alignment file displayed in the ClustalX window (Fig. 11.3.6). Nucleotides or amino acids are colored so that you can easily see where aligned sequences have the same residue at a site and where they differ. You can also see this in the gray graph below the sequences; the higher the peaks, the more sequences that are identical at that site.

   The goal of sequence alignment is to arrange your sequences such that homologous nucleotide or amino acid positions are aligned, i.e., nucleotide or amino acids that are derived from a common ancestor are at the same position in the alignment. In practice, this means introducing gaps into some of the sequences to maximize the similarity between sequences at each site. It is suggested to start with the default alignment parameters, but see "Adjust the alignment," below, for ways to change the default parameters that will let you introduce more or fewer gaps. Multiple sequence alignment is fraught with many complications that this treatment will ignore; for more on this topic, see Thompson et al. (1999) and Landan and Graur (2008).

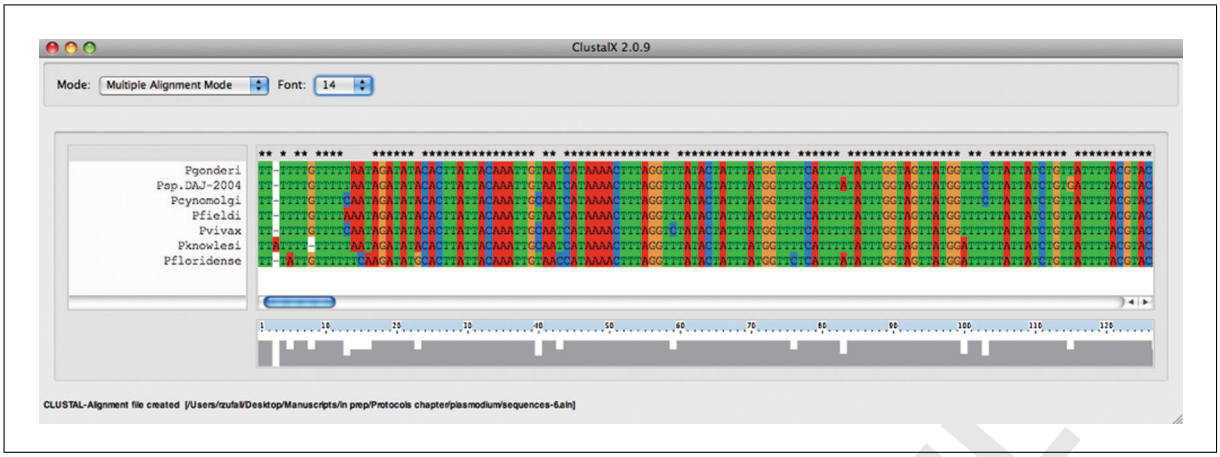

Figure 11.3.6 Alignment of sequences of the cytochrome oxidase I gene from seven species of the malaria parasite *Plasmodium*. Sequences were loaded into ClustalX from FASTA-formatted files, and aligned using the default parameters. Sequence positions in each column are hypothesized to have positional homology. Shown is the first 130 bp of the alignment.

Adjust the alignment

3. Examine your alignment to see if it makes sense (see Troubleshooting).

   ClustalX will align sequences even if they are not homologous; thus, it is your responsibility to make sure that they are. You cannot make adjustments to the alignment in ClustalX, but editing can be done in your original FASTA-formatted text file. There are also several freely available sequence editors that you could use (e.g., BioEdit for Windows (http://www.mbio.ncsu.edu/BioEdit/page2.html) or eBioX for the Macintosh (http://www.ebioinformatics.org/index.html). Geneious (see Basic Protocol 2) has a nice sequence editor, but only in the Pro version, which requires purchase of a license.

Interpret the alignment

4. Pause and consider what you can learn from the alignment itself.

   If you are interested in protein function, for example, look for regions of the gene that are highly conserved, i.e., have identical residues, across all or most of the sequences; these may represent sites that are involved in substrate binding or catalysis. Other protein functions, such as antigen activity, might be inferred from residues that are highly variable across sequences. While such interpretations are not conclusive, they can lead to hypotheses about protein function that can be further tested experimentally.

MAKING A PHYLOGENETIC TREE

BASIC PROTOCOL 2

There are several types of methods that can be used for phylogenetic reconstruction: distance, parsimony, maximum likelihood, and Bayesian. Each method has its own benefits and drawbacks. Here we will only cover distance methods, in particular the distance method called neighbor-joining. Neighbor-joining algorithms are very quick. The more taxa you include, the longer the analysis will take, but even with many taxa, getting a tree will still be fast. It will, however, take a little longer to run a bootstrap analysis to determine the support for your phylogeny.

Import your ClustalX alignment into Geneious

1. Make a new folder in Geneious by selecting “New Folder” under the File menu. At the prompt, give this folder a meaningful name.
2. Highlight the newly created folder in the “Sources” pane on the left side of the window by clicking on it.
3. Under the File menu, select “Import,” then “From File. . . ”. From the resulting list of possible file formats that Geneious can import, select “Clustal (*.*aln*).”

4. Navigate to where you saved your alignment file, highlight it, and click "Open."

   When your alignment opens (Fig. 11.3.5), in the top panel you will see the name of your alignment file with some additional information. If the alignment is highlighted in the top panel, you will see your aligned sequences in the bottom panel (if your alignment is not highlighted, click on it). On the right of the bottom panel are several options that will change the way you can view the sequences.

5. *Optional:* Change the zoom using the magnifying glass buttons. Under the "layout" menu, wrap the text so you can see more of the sequences on the screen at a time.

   Above the aligned sequences, you will see a graph, similar to the graph in ClustalX, that shows you how many of the sequences are identical at each site.

Build a neighbor-joining tree in Geneious

You are now ready to build a tree.

6. With your alignment selected in the top panel, click the "Tree" button on the tool bar.

   The Geneious tree builder gives you a few options that you can consider (Fig. 11.3.7); also see Strategic Planning.

7. Once you have selected all of your options, click "OK."

   When your tree is finished, you will see a new item in the top panel called "Tree of <alignment>."

8. Select the "Tree of <alignment>" item.

   You will see your tree in the panel below (Fig. 11.3.8). To the right of the tree, you will see a new set of options that will allow you to change how you visualize your tree.

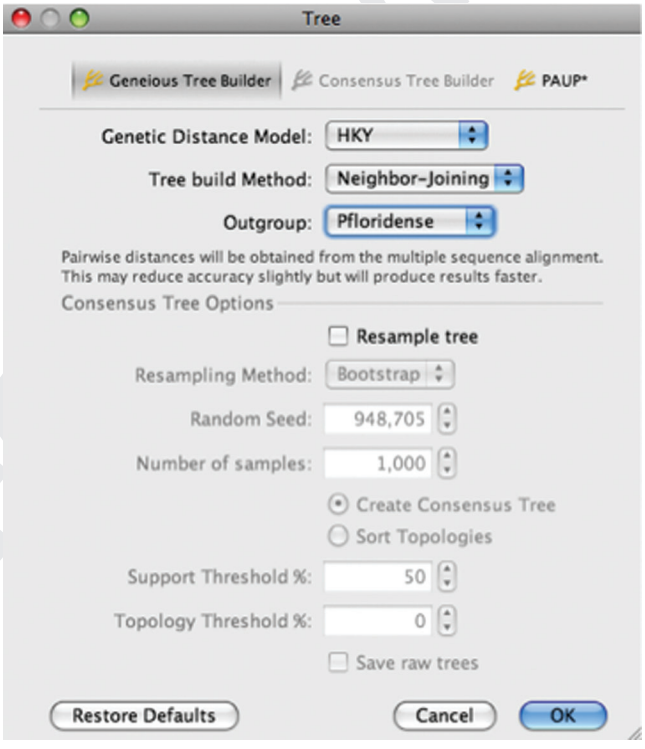

Figure 11.3.7 Tree building options in Geneious. To build a phylogeny as described in the text, select HKY as the distance model, neighbor-joining as the tree building method, and an outgroup (if you have one) from the list of sequences in the alignment.

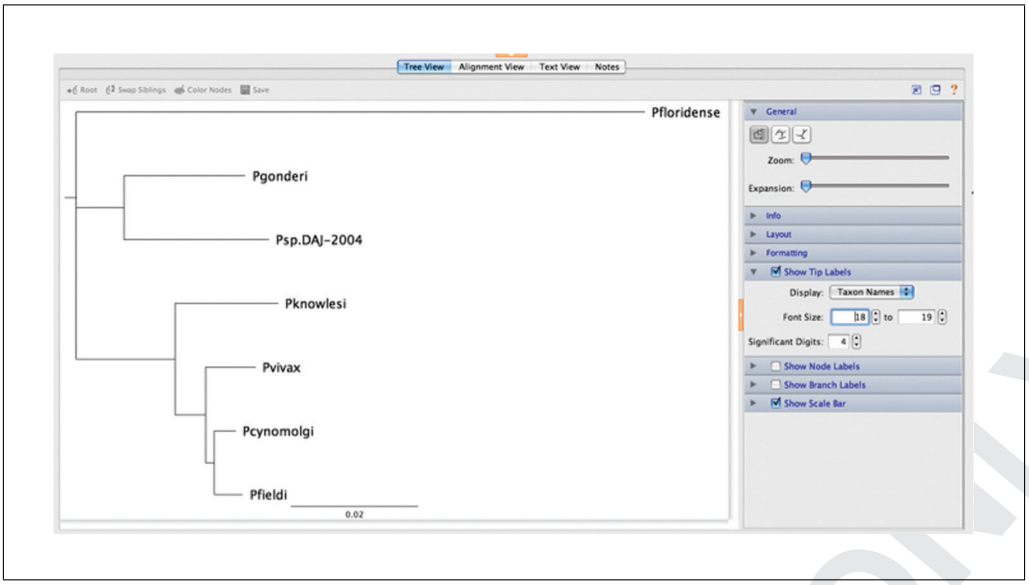

Figure 11.3.8 Neighbor-joining phylogeny of seven species of *Plasmodium* based on the cytochrome oxidase I gene shown in the Tree viewer panel of Geneious. You can return to the alignment used to build this phylogeny by clicking on "Alignment View" at the top of the panel.

Determine support for the tree with bootstrapping

9. With either the alignment or tree selected, click the "Tree" button. Under "Consensus Tree Options," check the box labeled "Resample tree" (Fig. 11.3.9).
10. Make sure that the resampling method is set to "Bootstrap." There is no need to change the Random Seed.
11. Set the number of samples to 1000.

    If your dataset is very large, this will take a long time and you can set the number of samples to 100.

12. Select the "Create Consensus Tree" radio button.
13. Set the Support Threshold to 50%.

    This value can be set lower if you want to see a fully resolved phylogeny including nodes with less than 50% bootstrap support, or higher if you want to only resolve nodes with higher bootstrap support (see Overview and Principles).

14. Click "OK."
15. Select the resulting "Consensus tree of <alignment>."
16. Click the arrow next to "Show Node Labels" to the right of the tree; then, change the display option to "Consensus support(%)."

    You will then see your bootstrap support on your tree (Fig. 11.3.10).

Save the tree

17. The tree will be saved automatically in Geneious.

    The tree can also be exported in a variety of formats for use in other programs.

18. *Optional:* Save the tree as an image by selecting "Save as image file. . . " under the File menu.

    This is useful, for instance, when saving the tree for use in a presentation.

    The image can be saved as .png or .jpg.

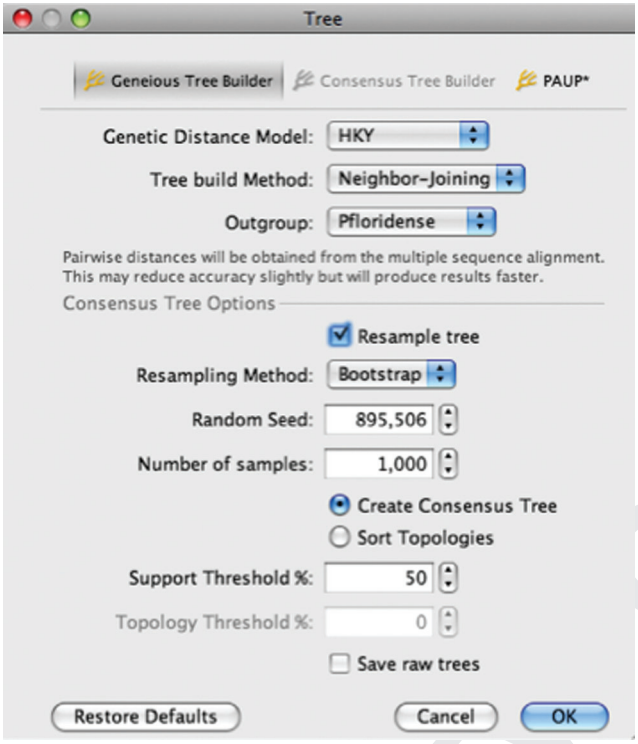

Figure 11.3.9 Tree building options in Geneious. To bootstrap a phylogeny, select the box labeled "Resample tree."

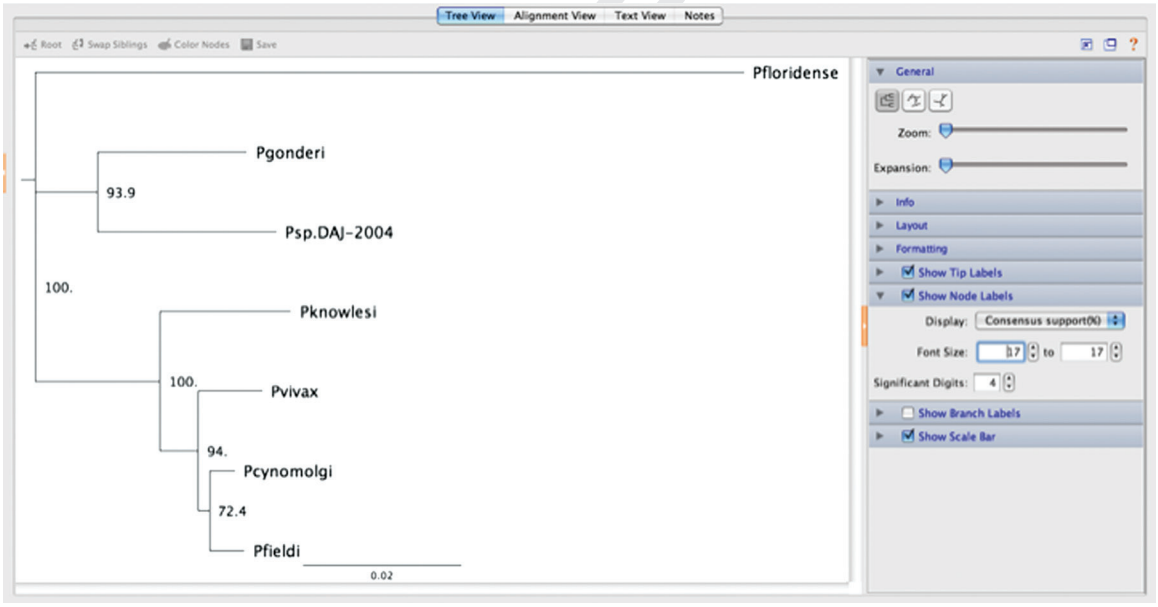

Figure 11.3.10 The bootstrapped phylogeny of *Plasmodium* species. Bootstrap support for each set of relationships is shown to the right of the node where two lineages diverge.

COMMENTARY

Understanding Results

Now that you have a phylogeny, you need to know what it means. We will work through two examples that will help you understand what your phylogeny means.

What's that bug?: Relationships between organisms

The impetus for and analysis of this example are drawn largely from Hayakawa et al. (2008).

First, we will address the question: what is the relationship between various species of malaria parasites? We chose six species of *Plasmodium* that infect primates, and a species that infects lizards as an outgroup (*P. floridense*). We retrieved the cytochrome oxidase I mitochondrial gene from each species from GenBank. You can recreate this example by downloading these sequences from Entrez (see *UNIT 11.2*); GenBank accession numbers are shown in Figure 11.3.4. Remember that it is useful to rename the sequences at this point.

Using the methods described above (Figs. 11.3.4 to 11.3.9), the bootstrapped phylogeny shown in Fig. 11.3.10 was constructed. Now we want to know what this phylogeny says about the relationships between these organisms. In particular, we might think about the identity of the unclassified species, *Plasmodium sp.* (Psp.DAJ-2004 in Figs. 11.3.6 to 11.3.10; Mu et al. 2005). According to our phylogeny, this species is more closely related to *P. gonderi* than to any of the other species in our analysis. Based on the high bootstrap support (93.9%), we are fairly confident of this grouping.

We also see that, according to this phylogeny, *P. cynomolgi* and *P. fieldi* are more closely related to each other than either are to *P. vivax*. But if we look at a study of this same group that includes more taxa and more informative sites from the mitochondria, the relationship between these three species is not as clear (Hayakawa et al., 2008). This points to the caution that we should employ when interpreting phylogenies based on limited data (in this case, limited informative sites and few taxa).

Finally, we have our tree depicted in such a way that the length of each branch is proportional to the number of evolutionary changes that have occurred on that branch. Thus, we see that while *Plasmodium sp.* is most closely related to *P. gonderi*, the length of the branches separating them, and thus probably the time since their divergence, is greater than that of the branches separating *P. cynomolgi* and *P. fieldi*. This supports the hypothesis that *Plasmodium sp.* and *P. gonderi* are separate species.

Orthologs versus Paralogs: Elucidating Patterns of Gene Duplication and Gene Function

As discussed above, homologous genes are genes that are derived from a common ancestor, and homologs can arise by different processes resulting in orthologs and paralogs (see Overview and Principles).

From Figure 11.3.2, it is clear that phylogenies are invaluable in distinguishing paralogs from orthologs and in allowing us to reconstruct the history of gene duplications. And, as you will see below, determining orthologous relationships is a key step in using phylogenetics to elucidate gene function.

Suppose you have sequenced a previously uncharacterized protein from your favorite organism: you can now use data on the function of related proteins to elucidate its function (Eisen, 1998). Using the tools you just learned, you can do a BLAST search of that protein, download the sequences of homologous proteins to a FASTA-formatted file, align the sequences, build a phylogeny, and identify orthologs of that protein.

For example (Zufall and Rausher, 2004), we sequenced a gene for dihydroflavonol reductase (*DFR*) from the cypress vine morning glory (*Ipomoea quamoclit*). *DFR* genes are part of a gene family, each member of which has different functions. Without performing any enzyme assays, we can develop a reasonable hypothesis about the function of our *I. quamoclit DFR* (called *DFR** until its relationship to other members of the gene family is known) based on its phylogenetic position with respect to DFRs from other species where the functions are known.

We first performed a BLAST search of this sequence (Fig. 11.3.11). Then, we downloaded related sequences for which there are published activity assays (Des Marais and Rausher, 2008) and imported these into ClustalX (GenBank accession numbers AB011667, which includes *I. purpurea* DFR-A, B, and C, AY463156, EU189073, EU189076, EU189080, and EU189082). A ClustalX alignment reveals that some of the

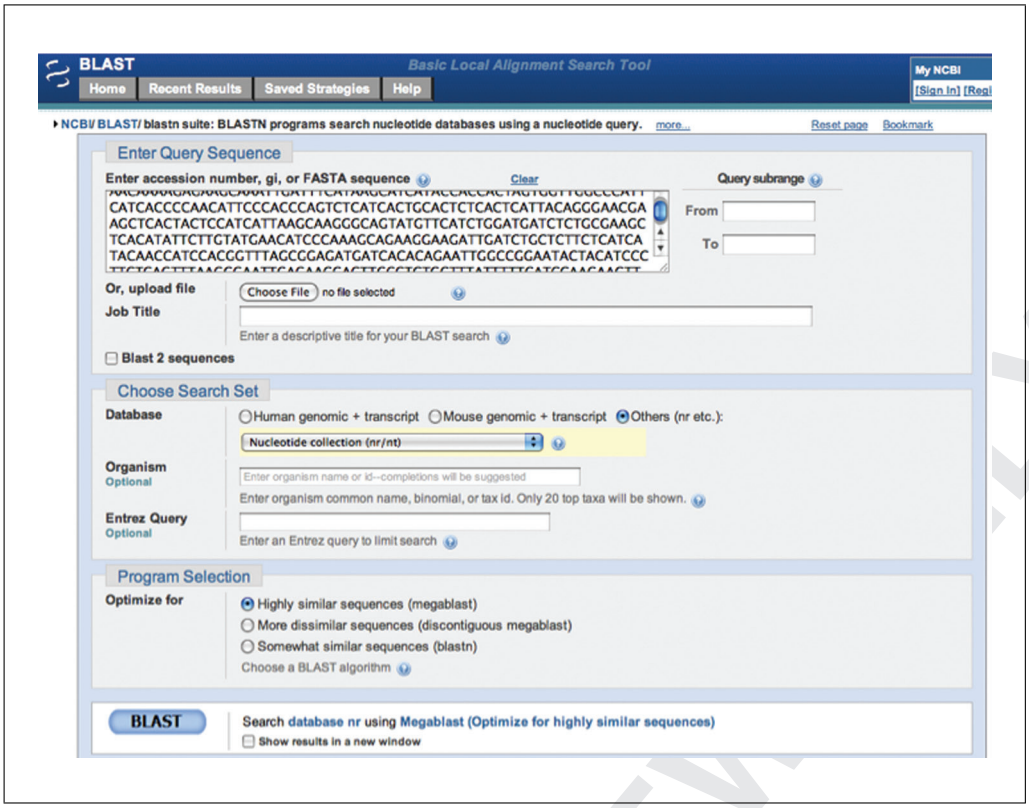

Figure 11.3.11 BLAST search of *I. quamoclit DFR**. The results of this BLAST search will be used to select sequences to use in building a phylogeny to assess the function of *DFR**.

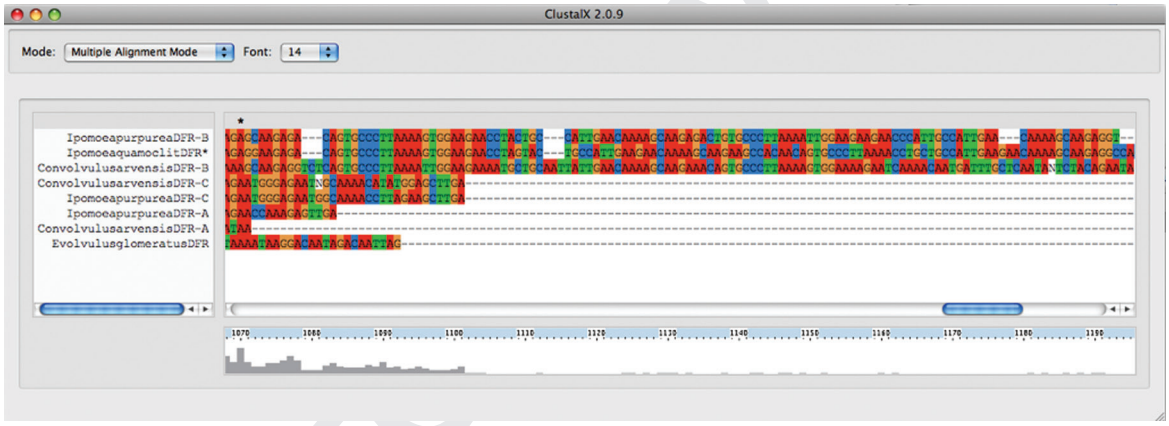

Figure 11.3.12 Alignment of *DFR* sequences. The first three sequences are much longer than the other sequences. These sequences must be trimmed, and the alignment repeated, before using them in a phylogenetic analysis.

sequences are much longer than the others (Fig. 11.3.12), so we trimmed these sequences, repeated the alignment, and checked to make sure the new alignment looked reasonable. We then imported this alignment into Geneious (Fig. 11.3.13) and created a neighbor-joining tree with bootstrap support (Fig. 11.3.14).

The resulting phylogeny (Fig. 11.3.14) shows that the *DFR* from *I. quamoclit* is most closely related to *DFR-B* from related species of plants, i.e., *DFR** from *I quamoclit* is an ortholog of DFR-B from *I. purpurea* and *C. arvensis*. Des Marais and Rausher (2008) demonstrate that *DFR-B* functions efficiently in the reduction of dihydroflavonols, whereas DFR-A and DFR-C do not. Thus, I can hypothesize that *DRF** is also active in dihydroflavonol reduction.

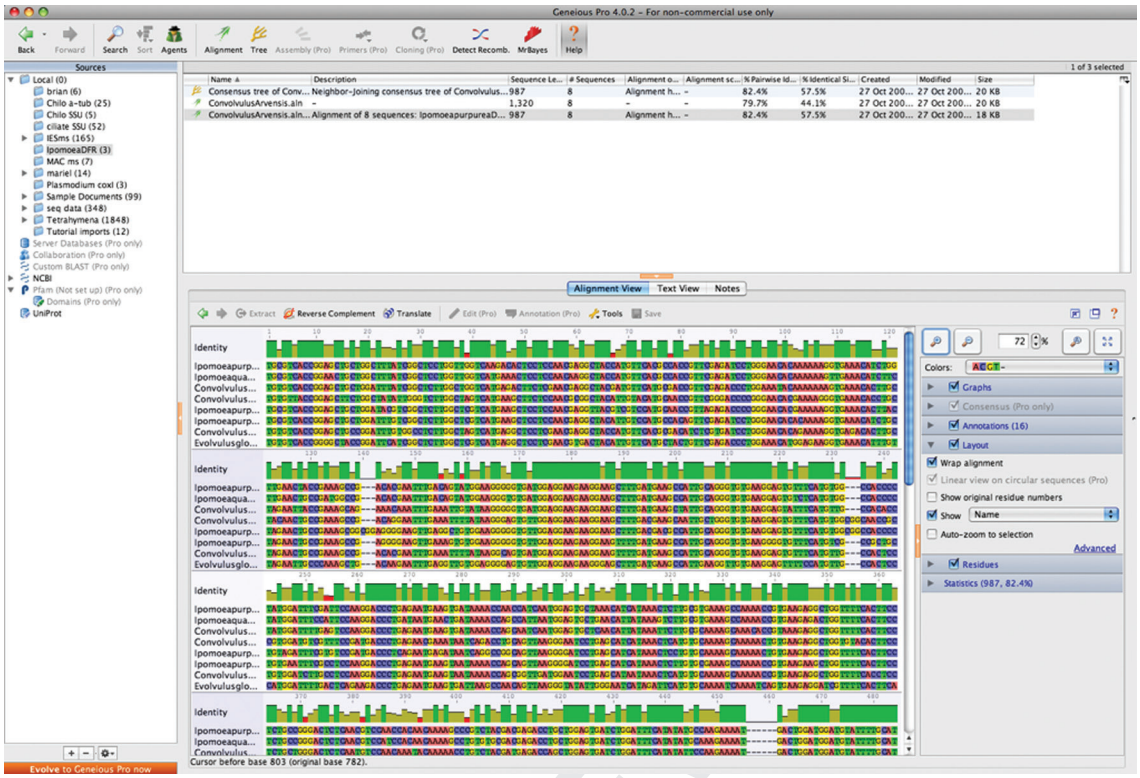

Figure 11.3.13 Alignment of *DFR* sequences imported into Geneious.

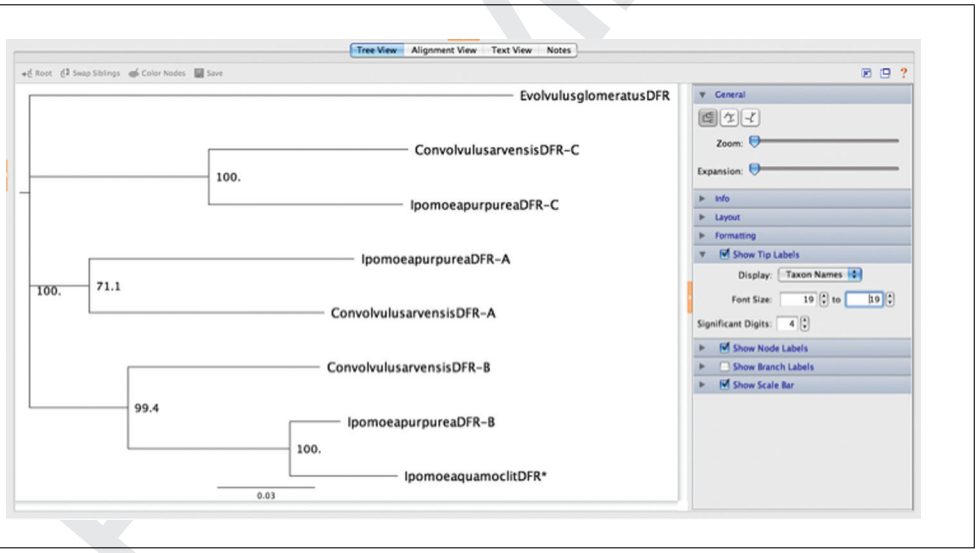

Figure 11.3.14 Bootstrapped phylogeny of *DFR* shows that *DFR** is most closely related to related species' *DFR-B*. Thus, we can hypothesize that the function of *DFR** is most like the function of DFR-B.

Troubleshooting

Sequences That Are Too Long or Too Short

Errors in alignment can occur when some sequences are longer or shorter than others (e.g., Fig. 11.3.12). If one sequence is longer than the other sequences, edit the sequence so that it contains only the homologous region of the gene. If one sequence is shorter than the others, you must either truncate all of the other sequences or remove the short sequence. If this editing involves long regions of sequence or whole sequences, you must repeat the alignment on the edited file.

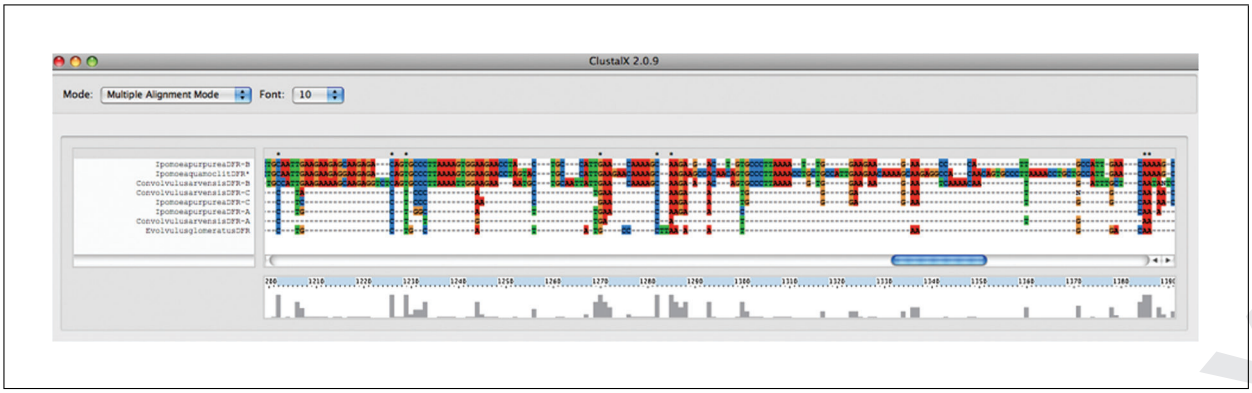

Figure 11.3.15 Sequence alignment with an unrealistic number of gaps.

Too Many or Too Few Gaps

If you were to introduce enough gaps in an alignment, you would be able to have identical residues or gaps at every site. However, this would not represent a very realistic evolutionary scenario. Figure 11.3.15 shows an alignment with many more insertions and deletions, i.e., gaps, than is realistic. A more realistic scenario would involve some insertions and deletions (represented by gaps in an alignment) and some point mutations (represented by different residues at a site). The problem is figuring out the frequencies of each event that most closely resemble what actually happened during evolution. This is not an easy problem to solve, but often it is worth redoing an alignment with different gap-opening and gap-extension penalties. This will change the relative frequencies of insertions and deletions versus point mutations. Under the "Alignment" menu, select "Alignment Parameters," then "Multiple Alignment Parameters," and try different values for "Gap Opening" and "Gap Extension." Then, repeat the alignment and compare the results. Note that each time you do a new alignment, you should make sure to change the name of the output files; the default is to save to the same name each time.

Literature Cited

Baum, D.A., Smith, S.D., and Donovan, S.S.S. 2005. Evolution: The tree-thinking challenge. *Science* 310:979-980.

Des Marais, D.L. and Rausher, M.D. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454:762-765.

Drummond, A.J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T., and Wilson, A. 2008. Geneious v4.0. http://www.geneious.com/.

Eisen, J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163-167.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Mass.

Hall, B.G. 2004. Phylogenetic Trees Made Easy: A How-To Manual, 2nd Edition. Sinauer Associates, Inc., Sunderland, Mass.

Hall, B.G. 2007. Phylogenetic Trees Made Easy: A How-to Manual, 3rd Edition. Sinauer Associates, Inc., Sunderland, Mass.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160-174.

Hayakawa, T., Culleton, R., Otani, H., Horii, T., and Tanabe, K. 2008. Big bang in the evolution of extant malaria parasites. *Mol. Biol. Evol.* 25:2233-2239.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. *In* Mammalian protein metabolism (H.N. Munro, ed.), pp. 21-132. Academic Press, New York.

Landan, G. and Graur, D. 2008. Characterization of pairwise and multiple sequence alignment errors. *Gene.* Epub ahead of print.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

McHardy, A.C. and Rigoutsos, I. 2007. What's in the mix: Phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.* 10:499-503.

Mu, J., Joy, D.A., Duan, J., Huang, Y., Carlton, J., Walker, J., Barnwell, J., Beerli, P., Charleston, M.A., Pybus, O.G., and Su, X. 2005. Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* 22:1686-1693.

Posada, D. and Crandall, K. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14:817-818.

Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512-526.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682-2690.

Zufall, R. and Rausher, M. 2004. Genetic changes associated with floral adaptation restrict future evolutionary potential. *Nature* 428:847-850.

Zwickl, D.J. and Hillis, D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51:588-598.

Key References

Hall, B.G. 2007. See above.

A "cookbook" for phylogenetic reconstruction. Recommended for beginning users interested in learning parsimony and maximum likelihood methods, in addition to distance methods. Relies largely on the program MEGA.

Felsenstein, J. 2004. See above.

A comprehensive guide to phylogenetic methodology and application. Recommended for those who want to delve deeply into the subject of phylogenetic inference.

Graur, D. and Li, W. 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland, Mass.

Recommended reading for an understanding the evolutionary biology behind the methods of phylogenetic reconstruction.