Building Phylogenetic Trees from Molecular Data with MEGA

Barry G. Hall*
Bellingham Research Institute, Bellingham, Washington
***Corresponding author:** E-mail: barryghall@gmail.com.
Associate editor: Joel Dudley

Abstract

Phylogenetic analysis is sometimes regarded as being an intimidating, complex process that requires expertise and years of experience. In fact, it is a fairly straightforward process that can be learned quickly and applied effectively. This Protocol describes the several steps required to produce a phylogenetic tree from molecular data for novices. In the example illustrated here, the program MEGA is used to implement all those steps, thereby eliminating the need to learn several programs, and to deal with multiple file formats from one step to another (Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 28:2731–2739). The first step, identification of a set of homologous sequences and downloading those sequences, is implemented by MEGA's own browser built on top of the Google Chrome toolkit. For the second step, alignment of those sequences, MEGA offers two different algorithms: ClustalW and MUSCLE. For the third step, construction of a phylogenetic tree from the aligned sequences, MEGA offers many different methods. Here we illustrate the maximum likelihood method, beginning with MEGA's Models feature, which permits selecting the most suitable substitution model. Finally, MEGA provides a powerful and flexible interface for the final step, actually drawing the tree for publication. Here a step-by-step protocol is presented in sufficient detail to allow a novice to start with a sequence of interest and to build a publication-quality tree illustrating the evolution of an appropriate set of homologs of that sequence. MEGA is available for use on PCs and Macs from www.megasoftware.net.

Protocol

A phylogenetic tree is an estimate of the relationships among taxa (or sequences) and their hypothetical common ancestors (Nei and Kumar 2000; Felsenstein 2004; Hall 2011). Today most phylogenetic trees are built from molecular data: DNA or protein sequences. Originally, the purpose of most molecular phylogenetic trees was to estimate the relationships among the species represented by those sequences, but today the purposes have expanded to include understanding the relationships among the sequences themselves without regard to the host species, inferring the functions of genes that have not been studied experimentally (Hall et al. 2009), and elucidating mechanisms that lead to microbial outbreaks (Hall and Barlow 2006) among many others. Building a phylogenetic tree requires four distinct steps: (Step 1) identify and acquire a set of homologous DNA or protein sequences, (Step 2) align those sequences, (Step 3) estimate a tree from the aligned sequences, and (Step 4) present that tree in such a way as to clearly convey the relevant information to others.

Typically you would use your favorite web browser to identify and download the homologous sequences from a national database such as GenBank, then one of several alignment programs to align the sequences, followed by one of many possible phylogenetic programs to estimate the tree, and finally, a program to draw the tree for exploration and publication. Each program would have its own interface and its own required file format, forcing you to interconvert files as you moved information from one program to another. It is no wonder that phylogenetic analysis is sometimes considered intimidating!

MEGA5 (Tamura et al. 2011) is an integrated program that carries out all four steps in a single environment, with a single user interface eliminating the need for interconverting file formats. At the same time, MEGA5 is sufficiently flexible to permit using other programs for particular steps if that is desired. MEGA5 is, thus, particularly well suited for those who are less familiar with estimating phylogenetic trees.

Step 1: Acquiring the Sequences

Ironically, the first step is the most intellectually demanding, but it often receives the least attention. If not done well, the tree will be invalid or impossible to interpret or both. If done wisely, the remaining steps are easy, essentially mechanical, operations that will result in a robust meaningful tree.

Often, the investigator is interested in a particular gene or protein that has been the subject of investigation and wishes to determine the relationship of that gene or protein to its homologs. The word "homologs" is key here. The most basic assumption of phylogenetic analysis is that all the sequences on a tree are homologous, that is, descended from a common ancestor. Alignment programs will align sequences, homologous or not. All tree-building programs will make a tree from that alignment. However, if the sequences are not actually descended from a common ancestor, the tree will be

meaningless and may quite well be misleading. The most reliable way to identify sequences that are homologous to the sequence of interest is to do a Basic Local Alignment Search Tool (BLAST) search (Altschul et al. 1997) using the sequence of interest as a query.

Step 1.1

When you start MEGA5, it opens the main MEGA5 window. From the **Align** menu choose **Do Blast Search**. MEGA5 opens its own browser window to show a nucleotide BLAST page from National Center for Biotechnology Information (NCBI). There is a set of five tabs near the top of that page (blastn, blastp, blastx, tblastn, and tblastx). By default the **blastn** (Standard Nucleotide BLAST) tab is selected. If your sequence is that of a protein click the **blastp** tab to show the Standard Protein BLAST page.

Note that NCBI frequently changes the appearance of the BLAST page, so it may differ in some details from that described here.

There is a large text box (**Enter accession number** . . .) where you enter the sequence of interest. You can paste the query sequence directly into that box. However, if your query sequence is already itself in one of the databases, you can paste its accession number or gi number. If your DNA sequence is part of a genome sequence, you can enter the genome's accession number then, in the boxes to the right (**Query subrange**) enter the range of bases that constitute your sequence. (You really do not want to use a several megabase sequence as your query!)

The middle section of the page allows you to choose the databases that will be searched and to constrain that search if you so desire. The default is **Nucleotide collection (nr/nt)**, but the drop-down text box with triangle allows you to choose among a large number of alternatives, for example, Human Genomic or NCBI genomes.

The optional **Organisms** text box allows you to limit your search to a particular organism or to exclude a particular organism. For instance, if your sequence is from humans you might want to exclude Humans from the search, so that you do not pick up a lot of human variants when you are really interested in homologs in other species. To include more organisms click the little + sign next to the options box.

The Exclude option allows you to exclude, for instance, environmental samples.

Step 1.2: Which BLAST Algorithm to Use?

The bottom section of the page allows you to choose the particular variant of BLAST that best suits your purposes. For nucleotides, the choices are megablast for highly similar sequences, discontiguous megablast for more dissimilar sequences, or blastn for somewhat similar sequences. The default is blastn, but if you are only interested in identifying closely related homologs tick megablast. This is the first choice that really demands some thought. The sequences that will be on your tree are very much determined by the choice you make at this point.

At the very bottom of the page click the BLAST button to start the search; do not tick the "show results in a new window" box. A results window will appear, possibly with a graphic illustrating domains that have been identified, typically with a statement similar to "this page will be automatically updated in 5 seconds." Eventually, the final results window will appear. The top panel summarizes the properties of the query sequences and a description of the database that was searched. Below that is a graphic that illustrates the alignment for the top 100 "hits" (sequences identified by the search). Scroll down below that to see the list of sequences producing significant alignment scores. For each sequence, there is an Accession number (a clickable link), a description, a Max Score (also a clickable link), a total score, a Query coverage, and *E* value and a Max ident. You use that information to decide which of those sequences to add to your alignment and thus to include on your tree.

The description helps decide whether you are interested in that particular sequence. There may be several sequences from the same species; do you want all of those or perhaps only one representative of a species—or even of a genus? If you are possibly interested in that sequence look at Query coverage. Are you interested in a homolog that only aligns with 69% of the query? If not, ignore that sequence and move on. Are you interested in a sequence that is 100% identical to your query? If you are only interested in more distantly related homologs, you may not be. If you want the most inclusive tree possible, you may be. *You* must decide; there is no algorithm that can tell you what to include.

If you decide that you are interested in a hit sequence, click the "**Max score**" link to take you down to the series of alignments. What you see depends on whether your query was a DNA sequence or a protein sequence.

Step 1.3: DNA Sequences

The alignment of the query to the hit begins with a link to sequence file via its gi and accession numbers. If that link is to a genome sequence, or even to a large file that includes sequences of several genes, you will not want to include the entire sequence in your alignment. There are two ways to deal with the issue. 1) Look at the alignment itself and note the range of nucleotides in the subject. Be sure to notice whether the query aligns with the subject sequence itself (**Strand = plus/plus**) or with its complement (**Strand = plus/minus**). Click the link to bring up the sequence file. At the top right click the triangle in the gray **Change region shown** box, then enter the first and last nucleotides of the range, then click the **Update View** button. In the gray **Customize view** region, below, tick the **Show sequence** box, and if Strand = plus/minus also tick the **Show reverse complement** box, then click the Update View button. Finally, click the **Add to Alignment** button (a red cross) near the top of the window. (2) If your query is a coding sequence or is some other notable feature you may see `Features in this part of subject sequence`: just below the sequence description with a link to the feature. Click that feature link to bring up the sequence file already showing the region of interest. Check to be sure whether the sequence shown is the reverse complement of the query, and if it is tick the **Show reverse complement** box in the **Customize view** region, update the view, then click the

Add to Alignment button (a red cross) near the top of the window.

Step 1.31. When you click the **Add to Alignment** button, MEGA5's **Alignment Explorer** window opens and the sequence is added to that window. After adding a sequence to the Alignment Explorer use the back arrow in the BLAST window to return to the list of homologous sequences and add another sequence of interest.

Step 1.4: Protein Sequences

The main difference from nucleotide searches is that you may see accession number links to several protein sequence files. These all have the same amino acid sequence, although their underlying coding sequences may differ. Click any one of the links to bring up the protein sequence file, then click the **Add to Alignment** button.

Things that May Go Wrong

1) You may find that all the hits that are returned from your search are from very closely related organisms; that is, if your query was an *Escherichia coli* protein, all the hits may be from *E. coli*, *Salmonella*, and closely related species. If the hits all show a high maximum identity and you are pretty sure the sequence occurs in more distantly related sequences you have probably come up against the default maximum of 100 target sequences. Repeat the search, but before you click the BLAST button to start the search notice that immediately below that button is a cryptic line "**+ Algorithm Parameters**." Click the plus sign to reveal another section of the BLAST setup page. Set the **Max Target Sequences** to a larger value and repeat the search. You may also want to exclude some closely related species in the **Choose Search Set** section above. Enter a taxon, for example, *E. coli*, in the box and tick the **Exclude** box. If you want to exclude more than one species click the plus sign to the right of **Exclude** to add another field. You can exclude up to 20 species.
2) When you try to return to the list of hits you may get a page that says "**How Embarrassing!** Error: —400 Cache Miss." Click the circular arrow next to the **Add to Alignment** button. You will be sent to the main BLAST page but do not despair. At the top right of that page is a **Your Recent Results** section. The top link in the list is your most recent search. Just click that link to get back to your results.

When you have added all the sequences that you want to, just close the MEGA5 browser window.

In the Alignment Editor window save the alignment by choosing **Save Session** from the **Data** menu. I like to use a name such as Myfile_unaligned just to remind myself that the sequences have not been aligned. The file will have the extension .mas.

Step 1.5: Alternatives to MEGA5 for Identifying and Acquiring Sequences

Step 1.51. You can access NCBI BLAST through any web browser that NCBI supports at http://blast.ncbi.nlm.nih.gov/Blast.cgi. In the **Basic BLAST** section click the **nucleotide blast** or **protein blast** link to get to the page identical to that described earlier. Everything is the same as when using MEGA5's browser except that you cannot click a convenient button to add the sequences to the Alignment Editor.

Step 1.52. Open a new file in a text editor. You can use MEGA5's built in text editor by choosing **Edit a Text File** from the File menu. That editor has several functions for editing molecular sequences, including reverse complementing and converting to several common formats including Fasta. Alternatively, use Notepad for Windows or TextWrangler for Mac (http://www.barebones.com/products/textwrangler/). Save the file with a meaningful name with the extension.fasta, for example, myfile.fasta. Do **not** use Microsoft Word, Word Pad, TextEdit (Mac), or another word processor!

Step 1.53. When you have identified the sequence that you want to add and clicked the link to take you the page for that sequence file, adjust the Region Shown and Customize View if necessary. Notice the **Display Settings** link near the top left of the page. The default setting is GenBank (full). Change that to **Fasta (text)**, select everything, copy it then paste into the text editor file. As you add sequences to the file, it is convenient, but not necessary, to leave blank lines between the sequences.

Identifying and acquiring sequences is discussed in more detail in Chapter 3 of *Phylogenetic Trees Made Easy*, 4th edition (PTME4) (Hall 2011).

The next section explains how to import those sequences into MEGA5's alignment editor.

Step 2: Aligning the Sequences

If the Alignment Explorer window is not already open, in MEGA5's main window choose **Open a File/Session** from the **File menu.** Choose the MEGA5 alignment file (.mas) or the sequence file (.fasta) that you saved in Step 1. In the resulting dialog choose **Align**.

The Alignment Explorer shows a name for each sequence at the left, followed by the sequence, with colored residues. Typically the name is very long. That name is what will eventually appear on the tree, and long names are generally undesirable. This is the time to edit those names, in fact it is the only practical time to edit the names, so do not miss the opportunity. Simply double click each name and change it to something more suitable.

If your sequence is DNA you will see two tabs: **DNA Sequences** and **Translated Protein Sequences**. The DNA sequences tab is chosen by default. Click the Translated Protein Sequences tab to see the corresponding protein sequence.

Step 2.1

Now is the time to align the sequences. Two alignment methods are provided: ClustalW (Thompson et al. 1994) and MUSCLE (Edgar 2004a, 2004b). Either can be used, but in general MUSCLE is preferable. In the tool bar, near the top of the window, Clustal alignment is symbolized by the **W** button, and MUSCLE by an arm with clenched fist to "show a muscle." Click one of those buttons or choose **Clustal** or **Muscle** from the **Alignment** menu. If your sequence is DNA you will see two choices: **Align DNA** and

Align Codons. If your sequence is a DNA coding sequence it is *very important* to choose **Align Codons**. That will ensure that the sequences are aligned by codons, a much more realistic approach than direct alignment of the DNA sequences because that avoids introducing gaps into positions that would result in frame shifts in the real sequences.

Step 2.2

Choosing an alignment method opens a settings window for that method. For MUSCLE, I recommend that you accept the default settings. For ClustalW, the default settings are fine for DNA, but for proteins, I recommend changing the Multiple Alignment Gap Opening penalty to 3 and the Multiple Alignment Gap Extension penalty to 1.8.

Step 2.3

Click the **OK** button to start the alignment process. Depending on the number of sequences involved and the method you chose, alignment may take anywhere from a few seconds to a few hours. When the alignment is complete **Save** the session. I like to save the aligned sequences under a different name, thus if my original file was Myfile_unaligned.mas, I would save the aligned sequence as just Myfile.mas.

Step 2.4

MEGA5 cannot use the .mas file directly to estimate a phylogenetic tree, so you must also choose **Export Alignment** from the **Data** menu and export the file in MEGA5 format where it will get a .meg extension. You will be asked to input a title for the data. You can leave the title blank if you wish, but it is helpful to add some sort of title that is meaningful to you. If it is an alignment of DNA sequences you will also be asked whether they are coding sequences.

Alignment is discussed in more detail in Chapter 4 of PTME4 (Hall 2011).

Step 2.5: An Alternative to Aligning with MEGA5

Once the alignment is complete, you will see that gaps have been introduced into the sequences. Those gaps represent historical insertions or deletions, and their purpose is to bring homologous sites into alignment in the same column. It should be appreciated that just as a phylogenetic tree is an "estimate" of relationships among sequences, an alignment is just an estimate of the positions of historical insertions and deletions. The quality of the alignment can affect the quality of a phylogenetic tree, but MEGA5 offers no way to judge the quality of the alignment. The web-based program **Guidance** (http://guidance.tau.ac.il/) provides five different methods of alignment, but more importantly, it evaluates the quality of the alignment and identifies regions and sequences that contribute to reducing the quality of the alignment. Discussion of **Guidance** (Penn et al. 2010) is beyond the scope of this article, but the topic is covered in detail in Chapter 12 of PTME4 (Hall 2011).

Guidance requires that the unaligned sequences are provided in a file in Fasta format. See Hall (2011) for a detailed description of the Fasta format. If you downloaded the sequences through your favorite web browser and saved them as a .fasta file that file can be used as the input for **Guidance**. If you used MEGA5 to download the sequences into the Alignment Explorer you can export the unaligned sequences in FASTA format by choosing **Export Alignment** from the **Data** menu, then choosing **FASTA** format. If you forgot to keep the unaligned sequences you can select all the sequences (Control-A), then choose **Delete Gaps** from the **Edit** menu before you export the sequences in FASTA format.

Step 3: Estimate the Tree

There are several widely used methods for estimating phylogenetic trees (Neighbor Joining, UPGMA Maximum Parsimony, Bayesian Inference, and Maximum Likelihood [ML]), but this article will deal with only one: ML.

Step 3.1

In MEGA5's main window choose **Open a File/Session** from the **File** menu and open the .meg file that you saved in Step 2.

Step 3.2

ML uses a variety of substitution models to correct for multiple changes at the same site during the evolutionary history of the sequences. The number of models and their variants can be absolutely bewildering, but MEGA5 provides a feature that chooses the best model for you. From the **Models** menu choose **Find Best DNA/Protein Models (ML)**. . . . A preferences dialog will appear, but you are safe enough accepting the default setting. Click the **Compute** button to start the run. Models can take quite awhile to consider all the available models, but a progress bar shows how things are coming along.

When complete a window appears that lists the models in order of preference. Note the preferred model, then estimate the tree using that model. For the examples below, the WAG + G + I model was the best.

Step 3.3

From the **Phylogeny** menu choose **Construct/Test Maximum Likelihood Tree**. . . . A preferences dialog similar to that in figure 1 will appear.

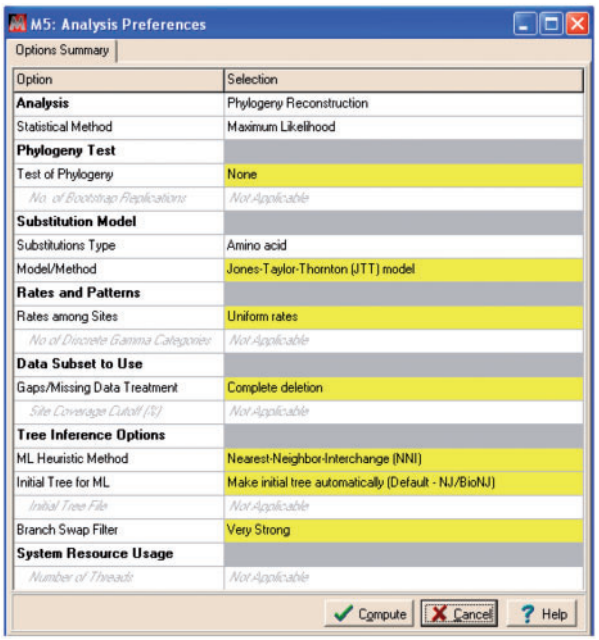

Fig. 1. ML analysis preferences.

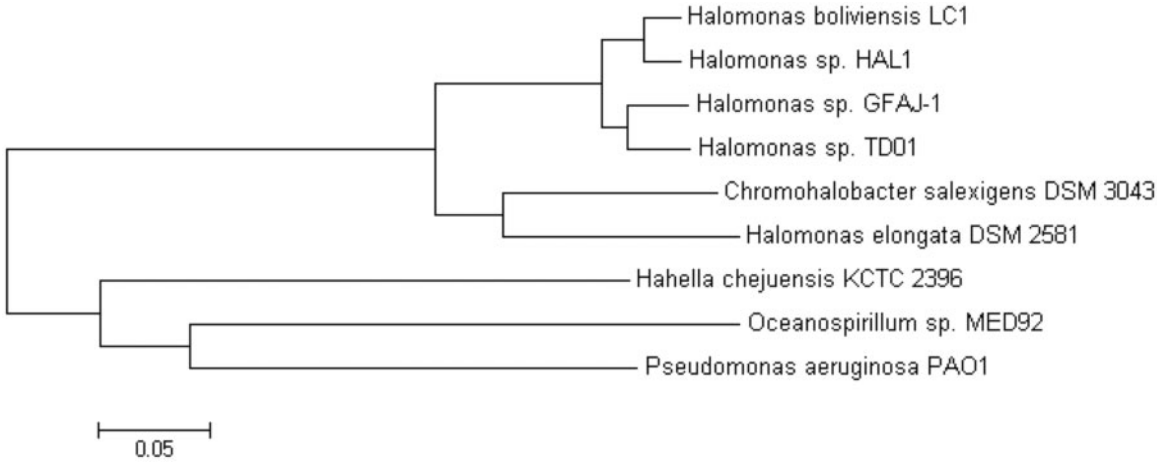

FIG. 2. An ML tree.

The yellow areas are parameters that you can modify. You need only be concerned with three parameters: 1) **Model/Method**, 2) **Rates among Sites**, and 3) **Gap/Missing Data Treatment**. Click at the right end of a yellow area to reveal a drop-down menu.

- For **Model/Method**, the WAG model would be selected.
- For **Rates among Sites**, the Gamma distributed with Invariant Sites (G + I) option would be selected. Together with the Models/Method selection above, this matches the best model found by the Models function.

Gaps/Missing Data Treatment determines how gaps are handled. **Complete deletion** means that MEGA5 ignores all columns in which there is a gap in any sequence. Unless there are very few gaps, that option can lose a lot of information because it removes a lot of sites from consideration. I prefer the **Partial Deletion** option in which sites with missing data are removed only as the need arises because that option retains more information.

When the preferences are set, click the compute button to compute the tree. Eventually, a tree explorer window will open that displays the tree (fig. 2).

Step 3.4

It is important to save the tree, so that it can be modified later if necessary. Save the tree from the File menu. The file will have the extension .mts.

You can also Export the tree for input into other tree drawing programs (see Step 4). From the **File** menu choose **Export Current Tree (Newick)**.

Step 3.5: Estimating the Reliability of the Tree

The tree that you estimated is almost certainly not a true representation of the historical relationships among the taxa and their ancestors. Instead, it is an estimate of those relationships. As with any estimate, it is desirable to know the reliability of that estimate. The most common way to estimate the reliability of a phylogenetic tree is by the **bootstrap** method. A detailed description of the bootstrap method is beyond the scope of this protocol, but the method is discussed in some detail on page 82–88 of Hall (2011).

To perform the bootstrap test return to the Analysis Preferences dialog shown in figure 1. Under **Phylogeny Test**, set Test of Phylogeny to "Bootstrap Method," then set No. of Bootstrap Replicates to an integer between 100 and 2,000. The higher the number, the longer it will take to perform the test. Click compute. A window with a progress bar shows how the analysis is proceeding. When the analysis is complete, a tree will appear with numbers on every node. Those numbers, bootstrap percentages, indicate the reliability of the cluster descending from that node; the higher the number, the more reliable is the estimate of the taxa that descend from that node. In general, we do not take seriously nodes with <70% reliability. The bootstrap test does not estimate the overall reliability of the tree; instead it estimates the reliability of each node. That is actually advantageous because it tells you which parts of the tree you should trust and which parts you should not take seriously.

Step 3.6: Alternatives to MEGA5 for Estimating the Tree

PhyML (http://www.atgc-montpellier.fr/phyml/binaries.php) (Guindon et al. 2010) is another program that estimates ML trees, and it can also be used over the web http://www.atgc-montpellier.fr/phyml/. SeaView (http://pbil.univ-lyon1.fr/software/seaview.html) (Gouy et al. 2010) is another multipurpose program that aligns sequences, estimates trees by several methods, and draws trees.

Step 4: Present the Tree

A drawing of a phylogenetic tree conveys a lot of information, both explicit and implicit. Some of that implicit information can be misleading, so it is up to the investigator to ensure that the information conveyed, both explicit and implicit, is correct.

A phylogenetic tree consists of external nodes (the tips) that represent the actual sequences that exist today, internal nodes that represent hypothetical ancestors, and branches that connect nodes to each other. The lengths of the branches represent the amount of change that is estimated to have occurred between a pair of nodes. That is the explicit information conveyed by a tree drawing. The tree in figure 2 is

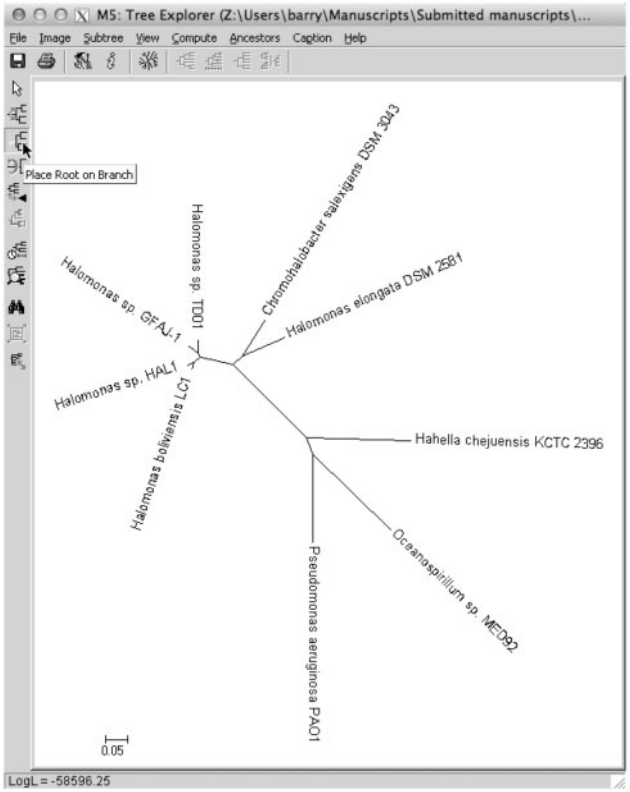

Fig. 3. ML in radiation (unrooted) format.

in the "rectangular phylogram" format in which internal nodes are represented by vertical lines.

In figure 2, the leftmost node appears to represent the root, the common ancestor from which all sequences are descended. That implicit information is incorrect and misleading. In fact, the ML method, in common with the Neighbor Joining, Parsimony, and Bayesian Inference methods, is incapable of determining the root of a tree; all those methods estimate **unrooted** trees.

Step 4.1

The **Radiation** or **Unrooted** format shown in figure 3 is a better way to draw an unrooted tree because it does not allow the viewer to assume a root that is unknown. To display the tree in Radiation format, in the Tree Explorer window choose **Tree/Branch Style** from the **View** menu, then select **Radiation** from the submenu. Because the Radiation format is unfamiliar to many readers, the default Rectangular Phylogram format is often published, despite the fact that it misleadingly implies a rooted tree. A rooted tree provides direction to the evolutionary process, with the order of descent from the root toward the tips. Assuming that directionality can easily lead to incorrect assumptions about the evolutionary history of those sequences. To avoid the unjustified implication of directionality, it is important to specify in the figure legend or in the text that the tree is unrooted.

Step 4.2

Often we do want to present a rooted tree to draw conclusions that depend upon the order of descent. To do that, we need additional information about the sequences, information that is external to the sequences themselves, that is, an **outgroup**. An outgroup is a sequence that is more distantly related to the remaining (ingroup) sequences than they are to each other. We cannot infer an outgroup from the tree itself, so we turn to other information. For the sequences in figure 2 we know that *Pseudomonas aeruginosa* belongs to the order *Pseudomonadales*, whereas the remaining organisms belong to the order *Oceanospirillales*, both of the class Gammaproteobacteria,. Thus, *P. aeruginosa* is a legitimate outgroup for the remaining sequences.

Step 4.21. We can root the tree on *P. aeruginosa* by using the rooting tool that is found in the sidebar of the Tree Explorer widow (fig. 3). In either the Rectangular Phylogram view or the Radiation view, while the rooting tool is selected, click on the branch leading to *P. aeruginosa* to root the tree on that sequence as shown in figure 4.

The rooted tree in figure 4 now correctly implies the direction of evolution of those sequences. When the tree is published, it would be important to specify that the tree was rooted on *P. aeruginosa*.

Step 4.3

MEGA5 provides a variety of tools for manipulating the appearance of the tree. I have already mentioned the Rectangular Phylogram and Radiation formats. Although those formats appear to be very different, they are drawings of exactly the same tree. In both cases, branches are drawn, so that the lengths of the lines are proportional to the branch lengths. Those formats make it obvious that there has been much more change between *Hahella chejuensis* KCTC 2396 and *Oceanospirillum* sp. MED292*a* than there has been between *Halomonas boliviensis* LC1 and *Halomonas* sp. HAL1.

Step 4.4

The cladogram, or **Topology Only** format, is another important format. Choose **Topology only** from the **View** menu to see the tree drawn, so that the lengths of the branch lines are unrelated to branch lengths. Why would one ever want to eliminate that information from the drawing? In some trees, there are some nodes that are separated by very short branches, whereas others are separated by very long branches. When the branches are too short, it may be impossible to see the branching order or topology. The Topology Only format makes it possible to see the branching order of the entire tree.

Step 4.4.1. But what about those branch lengths? Do we really want to lose that information? No, we do not, so we can simply label the branches with their branch lengths. To do that choose **Show/Hide** from the **View** menu and select **Branch Lengths** from the submenu.

Step 4.5: Publishing the Tree

Although you can Print the tree for your own purposes, to publish it you must save it in a graphics file format that is acceptable to the journal. The portable document format (PDF) is almost universally acceptable. Choose **Save as PDF File** from the **Image menu**.

You may want to manipulate the drawing in ways that MEGA5 does not provide: boldfacing some sequence names to draw attention to them, adding an arrow, etc. Such

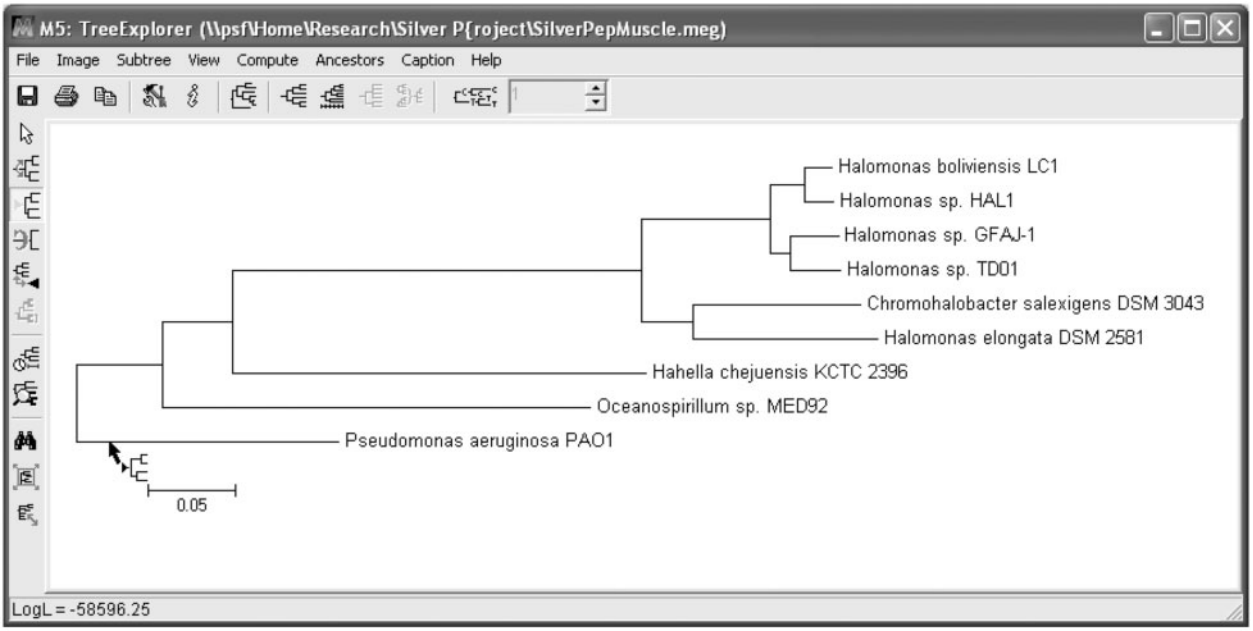

Fig. 4. ML tree rooted on *Pseudomonas aeruginosa*.

manipulations are done with a graphics drawing program. Most drawing programs will accept files in PDF format, but in case they do not, MEGA5 also allows you to save the image in PNG and Enhanced Meta File formats.

Step 4.6: Alternatives to Drawing Trees within MEGA5

The tree drawing program FigTree (http://tree.bio.ed.ac.uk/software/figtree/) is a full-featured program that offers many capabilities of the MEGA5 system and many other capabilities. It is available for Windows and Mac operating systems as a Java executable that will run on any OS including Linux. To import a tree into FigTree, export it as a Newick file as described in Step 3.

Using MEGA5 on Macintosh Computers

The **Save** and **Open** dialogs are Windows-like and may be unfamiliar to Mac users. A document detailing navigation in MEGA 5 for Mac can be downloaded from http://bellinghamresearchinstitute.com/NavigatingMEGA5/index.html.

Some Macintosh users have reported problems running MEGA5 for Mac on their machines; they need not, however, do without MEGA5. There are several "virtual machine" programs such as Parallels (http://www.parallels.com/products/desktop/) and VMFusion (http://www.vmware.com/products/fusion/overview.html) that will allow a Macintosh to run the Windows operating system. They both necessitate buying a copy of Windows and installing it in the virtual machine, but once that is done and MEGA5 for Windows is installed on that virtual machine, MEGA5 is as convenient and easy to run as it would be on a dedicated Windows computer. In addition, the user then has access to the entire world of Windows programs, some of which are actually as good as Macintosh programs.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25: 3389–3402.

Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27:221–224.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Hall BG. 2011. Phylogenetic trees made easy: a how-to manual. 4th ed. Sunderland (MA): Sinauer Associates.

Hall BG, Barlow M. 2006. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. *Ann Epidemiol.* 16: 157–169.

Hall BG, Pikis A, Thompson J. 2009. Evolution and biochemistry of family 4 glycosidases: implications for assigning enzyme function in sequence annotations. *Mol Biol Evol.* 26:2487–2497.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* 38(Web Server issue):W23–W28.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.