Module 8: Horizontal Gene Transfer

Objective

The objectives of this module are:

1. To determine if there is evidence that the gene under investigation has arisen by horizontal transfer rather than vertical transfer by constructing phylogenetic trees, looking at gene neighborhoods and evaluating GC content of the gene under investigation compared to the genome as a whole.

Materials

To perform this activity you will need:

- Access to the internet on a computer equipped with the most recent version of Firefox, Chrome or Safari.
- To have completed the sign up for GENI-ACT described in the Signing Up for GENI-ACT section of the manual.

Background

Horizontal (or Lateral) Gene Transfer is the transfer of a gene from one organism into another organism that is not its offspring. In bacteria and archaea, a number of mechanisms for horizontal gene transfer have been experimentally confirmed. Horizontal gene transfer is a very important factor in microbial evolution. Organisms that receive a gene from horizontal gene transfer may develop a new phenotype. A horizontally transferred gene may cause a new phenotype immediately, or over time the horizontally transferred gene may accumulate many mutations that result in a new function of that gene product and a new phenotype in the organism. New phenotypes that result from horizontal gene transfer significantly increase the speed and efficiency of evolution. In many microbes, a large portion of the genome consists of genes obtained in the microbes' histories through horizontal gene transfer.

A concise description of horizontal gene transfer can be found at : http://en.wikipedia.org/wiki/Horizontal_gene_transfer.

To determine whether a gene in an organism is the result of horizontal gene transfer, phylogenetic trees and speciation relationships are evaluated.

There are two types of phylogenetic trees: rooted and unrooted. Rooted phylogenetic trees show relationships with a timeline. This looks similar to a person's family tree. Often the length of the lines in a

rooted phylogenetic tree is related to time. By contrast, an unrooted phylogenetic tree shows a snapshot of relatedness without a timeline. An unrooted phylogenetic tree clusters similarities together but does not show an ancestral lineage. An unrooted person's family tree would show clusters of blood relatives, but would not show who was ancestor or progeny in those clusters.

The T-Coffee website that you have already used in the Sequence Based Similarity Module allows for the construction of basic cladograms and phylogenetic trees (defined below) Phylogeny.fr is a website that brings a number of distinct tools that are very useful for phylogenetic analysis into one location. Though the site is often slow, it gives more informative outputs for phylogenetic analyses.

Procedures

1. Copy the FASTA formatted amino acid for the protein under investigation from your Basic Information module notebook.
2. Navigate to http://blast.ncbi.nlm.nih.gov/Blast.cgi and perform a BLAST analysis as you did in the Sequence Based Similarity Module. **Using the nr database will typically give you a larger number of significant BLAST hits for phylogenetic analysis.**
3. When results are obtained, pick approximately 15-20 sequences for analysis, but do not simply choose the 15-20 or so best hits. Pick 4-5 hits with the highest BLAST scores, but then scroll down the list of significant BLAST hits and pick some that are toward the middle of the list and some closer to the bottom of the list. Try to pick organisms from different Genera if possible get the most variability. The idea here is not so look only at the best hits, but to see whether those proteins from organisms that most closely group with your gene are from close relatives of *Kytococcus*.

Sequences producing significant alignments:

Select: All None Selected:17

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	1031	1031	99%	0.0	99%	WP_012801520.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]	610	610	99%	0.0	63%	WP_022920049.1
chromosomal replication initiator protein DnaA [Serinicoccus profundi]	589	589	98%	0.0	60%	WP_010147278.1
chromosomal replication initiator protein DnaA [Serinicoccus marinus]	565	565	97%	0.0	59%	WP_022923463.1
chromosomal replication initiator protein DnaA [Knoellia aerolata]	560	560	96%	0.0	60%	WP_035938084.1
chromosomal replication initiator protein DnaA [Janibacter sp. HTCC2649]	552	552	96%	0.0	58%	WP_009776970.1
chromosomal replication initiator protein DnaA [Knoellia sinensis]	548	548	96%	0.0	57%	WP_035917766.1
chromosomal replication initiator protein DnaA [Tetrasphaera japonica T1-X7]	546	546	96%	0.0	59%	CCH80159.1
chromosomal replication initiator protein DnaA [Knoellia subterranea]	546	546	96%	0.0	55%	WP_035905061.1
chromosomal replication initiator protein DnaA [Knoellia flava]	544	544	96%	0.0	55%	WP_035946395.1
chromosomal replication initiator protein DnaA [Arsenicicoccus bolidensis]	539	539	96%	0.0	56%	WP_029212190.1
chromosomal replication initiator protein DnaA [Cellulomonas cellasea]	536	536	97%	0.0	58%	WP_034626490.1
chromosomal replication initiator protein DnaA,DNA-binding transcriptional dual regulator [Tetrasphaera elon	537	537	98%	0.0	55%	CCH68940.1
chromosomal replication initiator protein DnaA [Mobilicoccus sp. SIT2]	536	536	97%	0.0	59%	WP_040155375.1
chromosomal replication initiator protein DnaA [[[Cellvibrio] gilvus]	536	536	98%	0.0	58%	WP_013882065.1
chromosomal replication initiator protein DnaA [Tetrasphaera elongata]	536	536	97%	0.0	55%	WP_040753674.1
chromosomal replication initiator protein DnaA [Promicromonosporaceae bacterium W15]	532	532	96%	0.0	57%	WP_036955733.1
chromosomal replication initiator protein DnaA [Mobilicoccus pelagius NBRC 104925]	531	531	97%	0.0	59%	GAB48836.1
chromosomal replication initiator protein DnaA [Mobilicoccus pelagius]	530	530	97%	0.0	59%	WP_040633896.1
chromosomal replication initiator protein DnaA [Demetria terragena]	529	529	97%	0.0	59%	WP_040385338.1

Figure 8.1. BLAST results for Ksed_00010. The nr database was used and Ksed_00010 was found as the first hit (arrow). Seven of the 15 additional sequences used for analysis are shown as checked.

4. Download the selected hits in FASTA format as you did when you used T-Coffee in Module 2 and paste it into your notebook, a text editor or Word document (Figure 8.2).

```
>gi|502479361|ref|WP_012801520.1| chromosomal replication initiator protein DnaA [Kytococcus sedentarius]
>gi|256687299|gb|ACV05101.1| chromosomal replication initiator protein DnaA [Kytococcus sedentarius DSM
20547]
MSQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIR
LAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAVDGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERD
YSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTN
QFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP
GGDSGQITPTMILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQL
LGEDRRVYDEVSELTSIIRKKAARGR
>gi|551300082|ref|WP_022920049.1| chromosomal replication initiator protein DnaA [Ornithinimicrobium
pekingense]
MTSQSPAESAEVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLETTLRQPIVDALAGELGHDVR
LAITVDEDLRRQVEDEGDPAPGPAVTEQVPSDPDRTPYRSNGAGPGEPRSDGHRTPSGAVQTASAEDARLNPKYTFDTFV
SGSSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRDDKAGAF
QRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETR
IAILKKKAAQEGMQLPDEVLELIGSKISTNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMA
EVADYFQISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLIGERRALYDQITEL
TGIIRKASAR
>gi|551303511|ref|WP_022923463.1| chromosomal replication initiator protein DnaA [Serinicoccus marinus]
MSQPAPTSEDVWARVVDELETGGIGARERAFLQLTQMVGLLDTTVLLAVPYSHTKEMLETSLRRPIEDGLSRELNREIRV
AITVDDALRQRVEDEADDSEDDSLTRESLTRPASGPSSSSAPDPGEPDIPAVSRSAATSGIPRPATPAGPAVTGAADEAR
LNPKYTFDTFVSGPSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARRLYPGVRVRYVNSEEFTNDFI
NSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLL
TDVQPPDLETRIAILRKKAAQEGMQLPDEVLEHIASRITTNIRELEGALIRVTAFASLSSQRADADLAAHVLKDIVPGSD
TAQITVATIIREVSEFFQITVDELCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQAFGGRDHTTVMHAERKIRAQIGE
RRALYDQIAELTGSIRRASQR
>gi|737975618|ref|WP_035938084.1| chromosomal replication initiator protein DnaA [Knoellia aerolata] >gi|
700180054|gb|KGN40755.1| chromosomal replication initiation protein [Knoellia aerolata DSM 18566]
MDQIWRTTLDALDSDGIPVQQRAFLSLARLVGLLDETALIAVPNDFTKDIVETRLRDRVTETLSSQLGHTVRLAVTVDSS
LGDVPVLDPPADAPSGSTTTEPRPAAGTEGDGRHAERRAELDGIALVEDDDDGDSSRTGRSVAHTRSPGALRPRPGVTVP
EQVELTRLNPKYTFDTFVIGASNRFANAAALAVAETPAKAYNPLFIYGESGLGKTHLLHAIGHYARNLFPHVKVRYVNSE
EFTNDFINSIRDDKAANFQRRYRDVDVLLIDDIQFLQGKVQTQEEFFHTFNTLHNANKQVVITSDLPPKLLSGFEERMRS
RFEWGLMTDVQPPDLETRIAILRKKAAQEKLSVPDDVLEFIASRISTNIRELEGALIRVTAFASLNRQPVDISLAEIVLK
DLIPHDSSSQITSATIMAQTAAYFGLTLEDLQGQSRSRVLVTARQIAMYLCRELTDLSLPKIGQQFGGRDHTTVMHADKK
IRQLMAERRAIYNQVTELTNRIKQQSR
>gi|497462772|ref|WP_009776970.1| chromosomal replication initiator protein DnaA [Janibacter sp. HTCC2649]
>gi|84382082|gb|EAP97964.1| chromosomal replication initiator protein [Janibacter sp. HTCC2649]
MDQIWRTTLDALDSDGIPVQQRAFLSLAKLVGLLDETALIAVPNDFTKDIVETRLRDRVTETLSSQLGHDVRLAVTVDHS
LADVPVTIPADTTTVDGAGADQVPRTATTIGLEPGPADADGRRAKRRAELDGIALVEDDEGEDDSRNNGAIGRTRSPGAL
```

Figure 8.2. A portion of the downloaded BLAST hits for Ksed_00010. Note the long FASTA descriptor lines present for each sequence (an example of which is shown by the arrow)

5. Since T-Coffee will truncate the FASTA descriptor line to be no more than 15 characters in length, we must edit the FASTA header to have the genus and species of the homologs right after the > symbol in the FASTA header in order to clearly see the organism name in the results of the phylogenetic analysis. To achieve this simply delete all text between the last occurrence of the genus and species name of the organism from which the sequence is from and the first > symbol in the FASTA header (Figure 8.3).

```
>Kytococcus sedentarius DSM 20547]
MSQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIR
LAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAVDGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERD
YSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTN
QFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP
GGDSGQITPTMILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQL
LGEDRRVYDEVSELTSIIRKKAARGR
>Ornithinimicrobium pekingense]
MTSQSPAESAEVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLETTLRQPIVDALAGELGHDVR
LAITVDEDLRRQVEDEGDPAPGPAVTEQVPSDPDRTPYRSNGAGPGEPRSDGHRTPSGAVQTASAEDARLNPKYTFDTFV
SGSSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRDDKAGAF
QRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETR
IAILKKKAAQEGMQLPDEVLELIGSKISTNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMA
EVADYFQISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLIGERRALYDQITEL
TGIIRKASAR
>Serinicoccus marinus]
MSQPAPTSEDVWARVVDELETGGIGARERAFLQLTQMVGLLDTTVLLAVPYSHTKEMLETSLRRPIEDGLSRELNREIRV
AITVDDALRQRVEDEADDSEDDSLTRESLTRPASGPSSSSAPDPGEPDIPAVSRSAATSGIPRPATPAGPAVTGAADEAR
LNPKYTFDTFVSGPSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARRLYPGVRVRYVNSEEFTNDFI
NSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLL
TDVQPPDLETRIAILRKKAAQEGMQLPDEVLEHIASRITTNIRELEGALIRVTAFASLSSQRADADLAAHVLKDIVPGSD
TAQITVATIIREVSEFFQITVDELCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQAFGGRDHTTVMHAERKIRAQIGE
RRALYDQIAELTGSIRRASQR
>Knoellia aerolata DSM 18566]
MDQIWRTTLDALDSDGIPVQQRAFLSLARLVGLLDETALIAVPNDFTKDIVETRLRDRVTETLSSQLGHTVRLAVTVDSS
LGDVPVLDPPADAPSGSTTTEPRPAAGTEGDGRHAERRAELDGIALVEDDDDGDSSRTGRSVAHTRSPGALRPRPGVTVP
EQVELTRLNPKYTFDTFVIGASNRFANAAALAVAETPAKAYNPLFIYGESGLGKTHLLHAIGHYARNLFPHVKVRYVNSE
EFTNDFINSIRDDKAANFQRRYRDVDVLLIDDIQFLQGKVQTQEEFFHTFNTLHNANKQVVITSDLPPKLLSGFEERMRS
RFEWGLMTDVQPPDLETRIAILRKKAAQEKLSVPDDVLEFIASRISTNIRELEGALIRVTAFASLNRQPVDISLAEIVLK
DLIPHDSSSQITSATIMAQTAAYFGLTLEDLQGQSRSRVLVTARQIAMYLCRELTDLSLPKIGQQFGGRDHTTVMHADKK
IRQLMAERRAIYNQVTELTNRIKQQSR
>Janibacter sp. HTCC2649]
MDQIWRTTLDALDSDGIPVQQRAFLSLAKLVGLLDETALIAVPNDFTKDIVETRLRDRVTETLSSQLGHDVRLAVTVDHS
LADVPVTIPADTTTVDGAGADQVPRTATTIGLEPGPADADGRRAKRRAELDGIALVEDDEGEDDSRNNGAIGRTRSPGAL
RPRPGATVPEQVELTRLNPKYTFDTFVIGASNRFANAAALAVAETPAKAYNPLFIYGESGLGKTHLLHAIGHYARNLYPH
VKVRYVNSEEFTNDFINSIRDDKAANFQRRYRDVDVLLIDDIQFLQGKVQTQEEFFHTFNTLHNANKQVVITSDLPPKLL
SGFEERMRSRFEWGLMTDVQPPDLETRIAILRKKAAQEKLSVPDDVLEFIASRISTNIRELEGALIRVTAFASLNRQPVD
ISLAEIVLKDLIPHDSANQITSATIMAQTAAYFGLTLEDLQGQSRSRVLVTARQIAMYLCRELTDLSLPKIGQQFGGRDH
TTVMHADKKIRQLMAERRAIYNQVTELTNRIKQQSR
```

Figure 8.3. The same sequences as in Figure 8.2 with the FASTA headers edited to only include the genus and species names of the organisms from which the hits were obtained.

6. Navigate to http://www.ebi.ac.uk/Tools/msa/tcoffee/ , paste the edited sequences into the search box and hit submit.

7. Figure 8.4 shows a portion of the T-Coffee results. Note the genus names of the hits at the left. **You should copy and paste the entire alignment, including the Clustal W header, into your notebook or a**

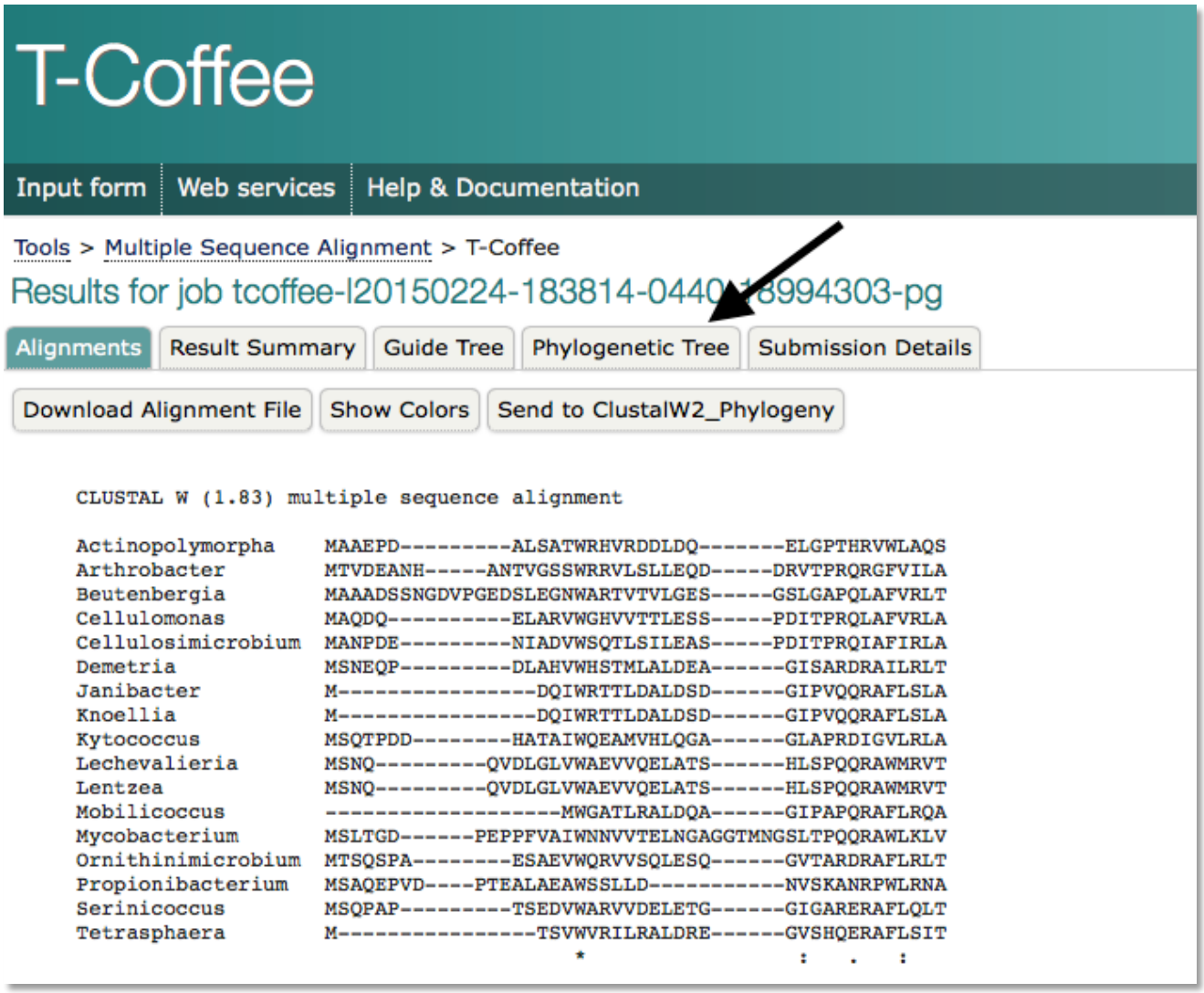

Figure 8.4. The T-Coffee results page with genus names clearly identified for each protein in the alignment. The arrow points to the Phylogenetic Tree menu, which will generate both a cladogram and phylogram based on the alignment as described below.

Word document and save it. You will use the alignment again later in this module.

8. Click on the Phylogenetic Tree tab and scroll to the bottom of the resulting page. You will see a cladogram indicated by default (Figure 8.5A), but if the "real" branch length button is clicked the image will appear as in Figure 8.5B).

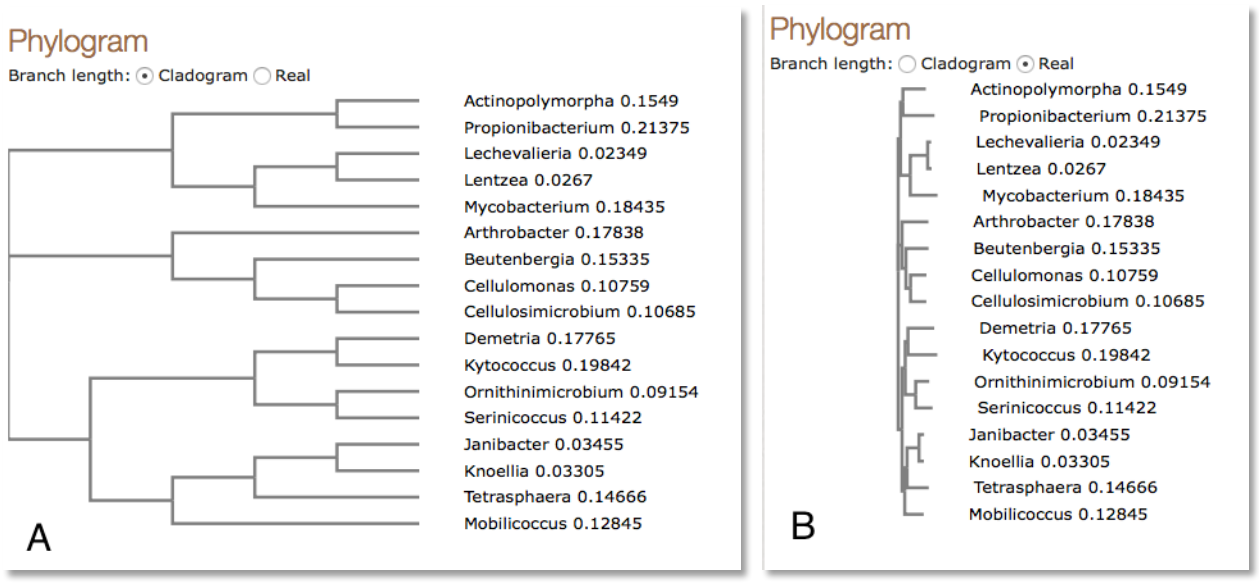

Figure 8.5. The results of phylogenetic tree analysis in T-Coffee. A. Results expressed as a cladogram. B. Results expressed as a phylogenetic tree. See the text for a description of the meaning of the two different outputs.

9. Interpretation of Cladograms and Phylograms: (http://evolution.berkeley.edu/evolibrary/article/phylogenetics_02)

9.1. A cladogram (Figure 8.5A) is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length. In other words the distances on the x axis are meaningless. Organisms in Figure 8.5A are in clades. For example *Kytococcus* and *Demetria* are in one clade in this figure. The branch point closest to them represents a hypothetical ancestor that gave rise to each. *Ornithinimicrobium* and *Serinicoccus* are in another clade with a common ancestor predicted. The next branch point to the left forms a higher order clade that includes all 4 of these species and predicts a common ancestor further back in evolutionary time that gave rise to all 4 of these organsisms. The length of the lines between the branch points is not an estimate of evolutionary time. The arrangement of genera from top to bottom (the y-axis) also has no bearing on relationships.

9.2. A phylogram (Figure 8.5B) is a branching diagram (tree) that is assumed to be an estimate of a phylogeny. The branch lengths are proportional to the amount of inferred evolutionary change. Thus, while at first glance the cladogram in Figure 8.5A seems to suggest large differences in evolutionary time among the genera, we can see in Figure 8.5B that this is not the case.

9.3. Once the phylogram is produced in T-Coffee, it can be evaluated to make predictions on whether the gene under investigation might have been transferred to *Kytococcus* by horizontal gene transfer. **However, more informative true phylogenetic trees can be prepared at another site. If your instructor tells you to stop at this point, you will interpret your phylogram as described below, but you are encouraged to generate the more informative phylogenetic tree diagrams described beginning in item below prior to making your conclusions.**

10. Possible interpretations of your T-Coffee phylogram:

10.1. The Ksed protein is clustered with proteins from closely related bacteria – *no evidence for horizontal gene transfer.*

10.2. The Ksed protein is clustered with proteins from bacteria that are NOT closely related – *horizontal gene transfer is possible.*

10.3. Ksed protein is clustered with proteins from related and unrelated bacteria – *can't interpret the results.*

11. Visit the NCBI Taxonomy webpage at http://www.ncbi.nlm.nih.gov/Taxonomy/ and determine whether the organisms clustered with *Kytococcus* are closely related by comparing the lineages of each to that of *Kytococcus.*

11.1. Write the name *Kytococcus sedentarius* in the search window (Figure 8.6) and type go.

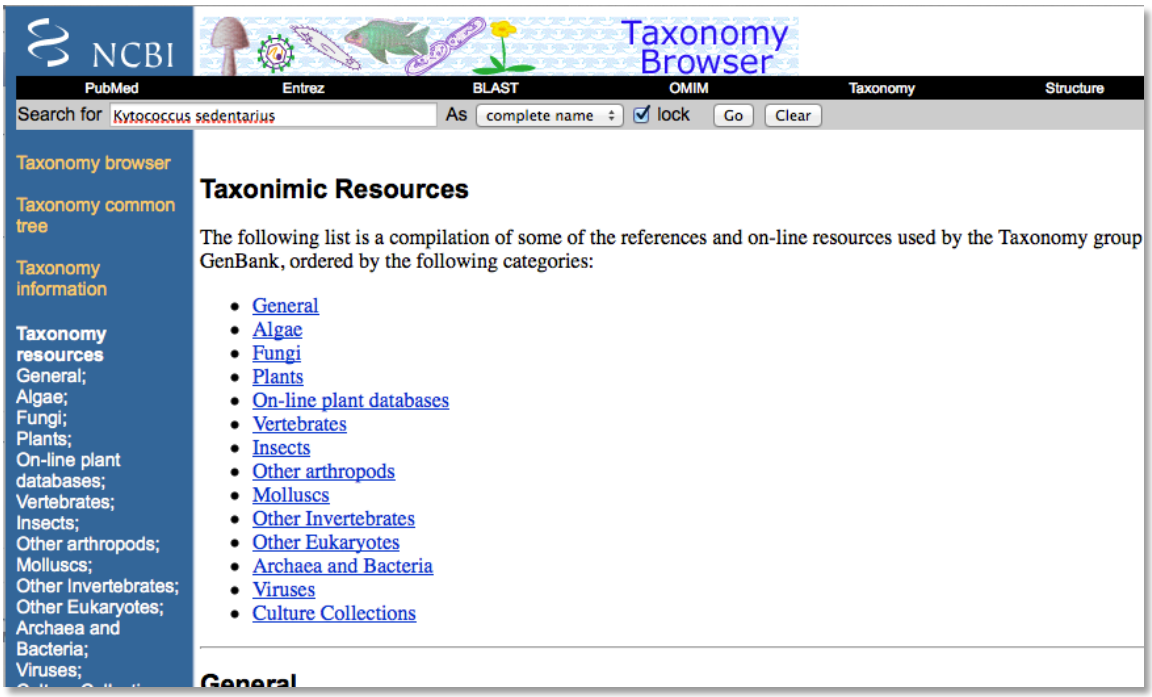

Figure 8.6. The NCBI Taxonomy Browser entry page.

11.2. The results page is shown in Figure 8.7, in which the levels of classification of *Kytococcus sedentarius* indicated. Hovering over the hyperlinked names will tell you the level of classification of each.

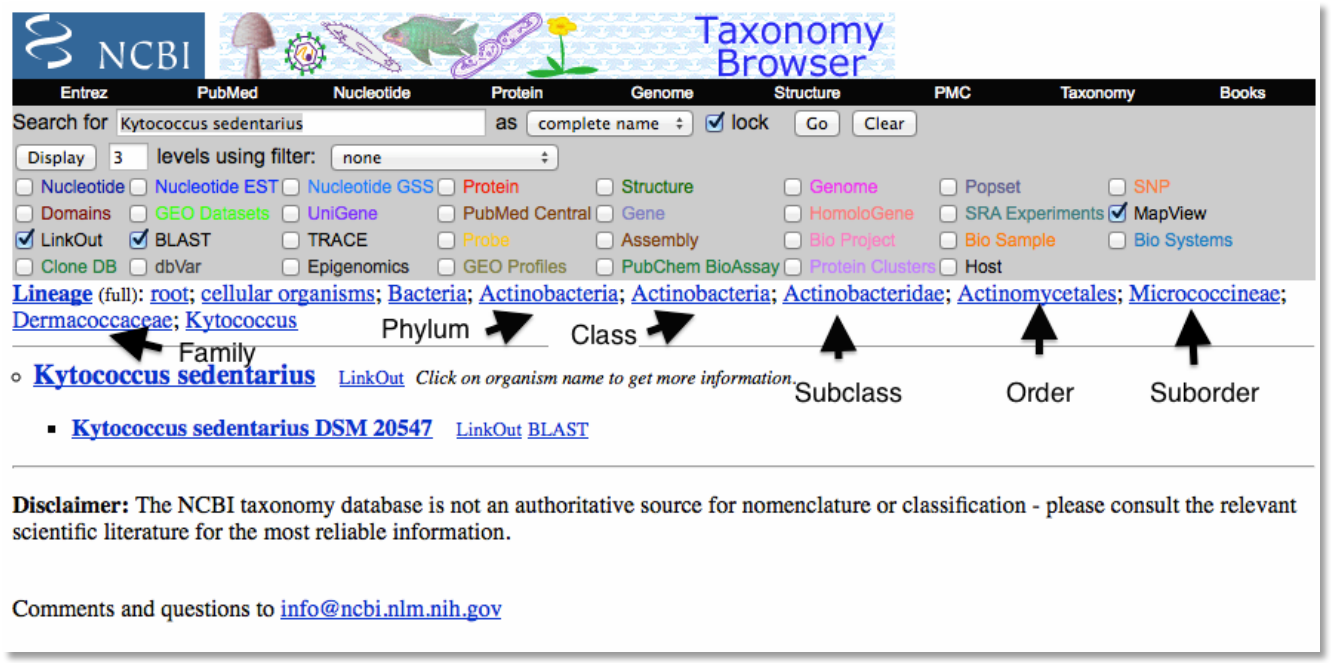

Figure 8.7. The NCBI Taxonomy Browser results for *Kytococcus sedentarius*. The classification of *Kytococcus* down the Family level is indicated. The genus *Kytococcus* and species *sedentarius* would come next.

11.3. The same process should be repeated for the organisms closely clustered with your query protein in *Kytococcus.* Figure 8.8 shows the classification of the genus *Demetria*, which is in the same clad as *Kytococcus* in the phylograms in figure 8.5 above. You will note that the genus *Demetria* has identical classification to *Kytococcus* down to the Family level. If this process was repeated for *Ornithinimicrobium* and *Serinicoccus*, which were in grouped with *Demetria* and *Kytococcus* in the next higher level clad, it would be found that *Ornithinimicrobium* and *Serinicoccus* were classified in the same manner as *Kytococcus* down to the suborder level of classification, and that each of these belonged to the same Family.

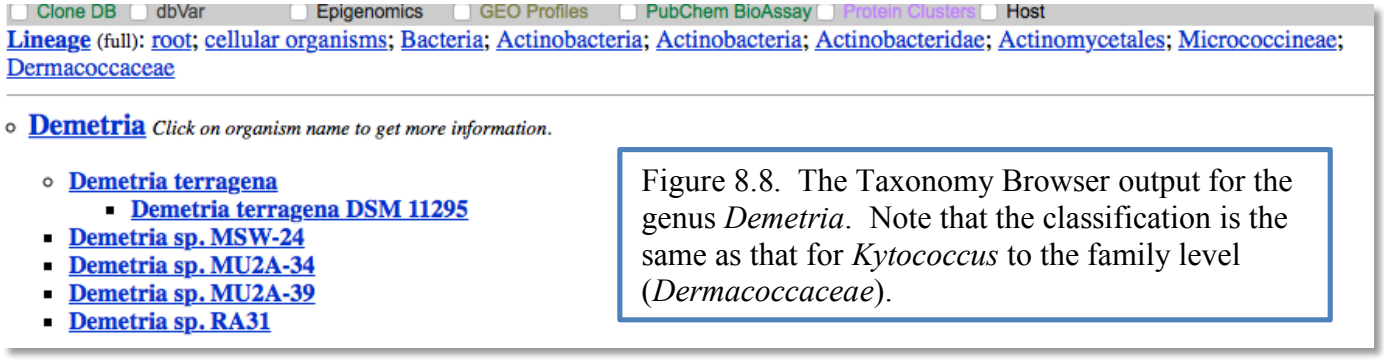

Figure 8.8. The Taxonomy Browser output for the genus *Demetria*. Note that the classification is the same as that for *Kytococcus* to the family level (*Dermacoccaceae*).

11.4. The interpretation of the tree will vary depending on how many organisms closely taxonomically related your organism have been sequenced. **Hits in the same or Family or Order will might be considered very close relatives for organisms who have not had genomes of close relatives sequenced, and those in the Suborder, or even Class or Subclass should most often be considered close relatives in the phylogenetic tree analysis.** This is, in fact, the case for the *Kytococcus* example given above. If your organism's gene clusters with genes from organisms with identical lineage down to these levels of classification, you would **NOT** have evidence of horizontal gene transfer. Use the information you find from the Taxonomy browser to decide whether or not you have found information to support the possibility of horizontal gene transfer as described in step 10 above.

Phylogeny.fr

The phylogeny.fr site will allow you to produce more advanced phylogenetic trees than available on T-Coffee. The following information is provided for those who wish to construct such trees, but your instructor may have you opt out of using this site. If instructed to skip this section you can jump to the chromosome neighborhood viewer section below.

1. Navigate to http://phylogeny.lirmm.fr/phylo_cgi/index.cgi . Select the Online Programs pull down menu and choose the PhyML option listed under Phlogeny on the top section (Figure 8.9)

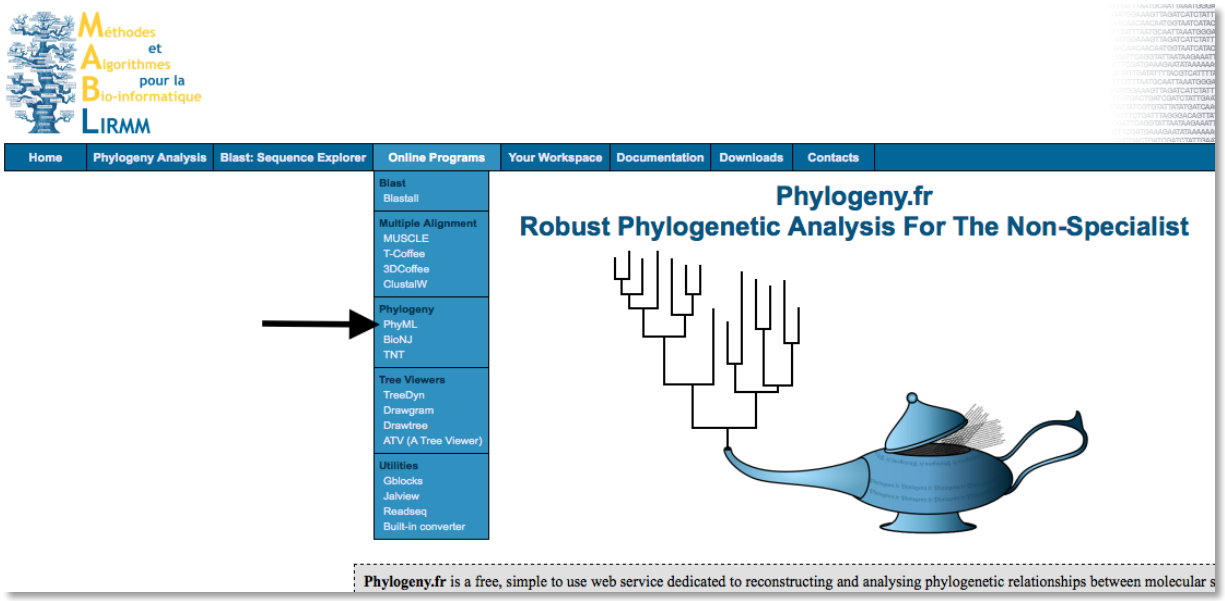

Figure 8.9. The Phylogeny.fr entry page. Click on the PhyML option (arrow) to begin.

2. Paste the T-Coffee alignment that you generated earlier in this module into the box, including the Clustal W header as shown (Figure 8.10) click submit. **NOTE: this server is often very slow.**

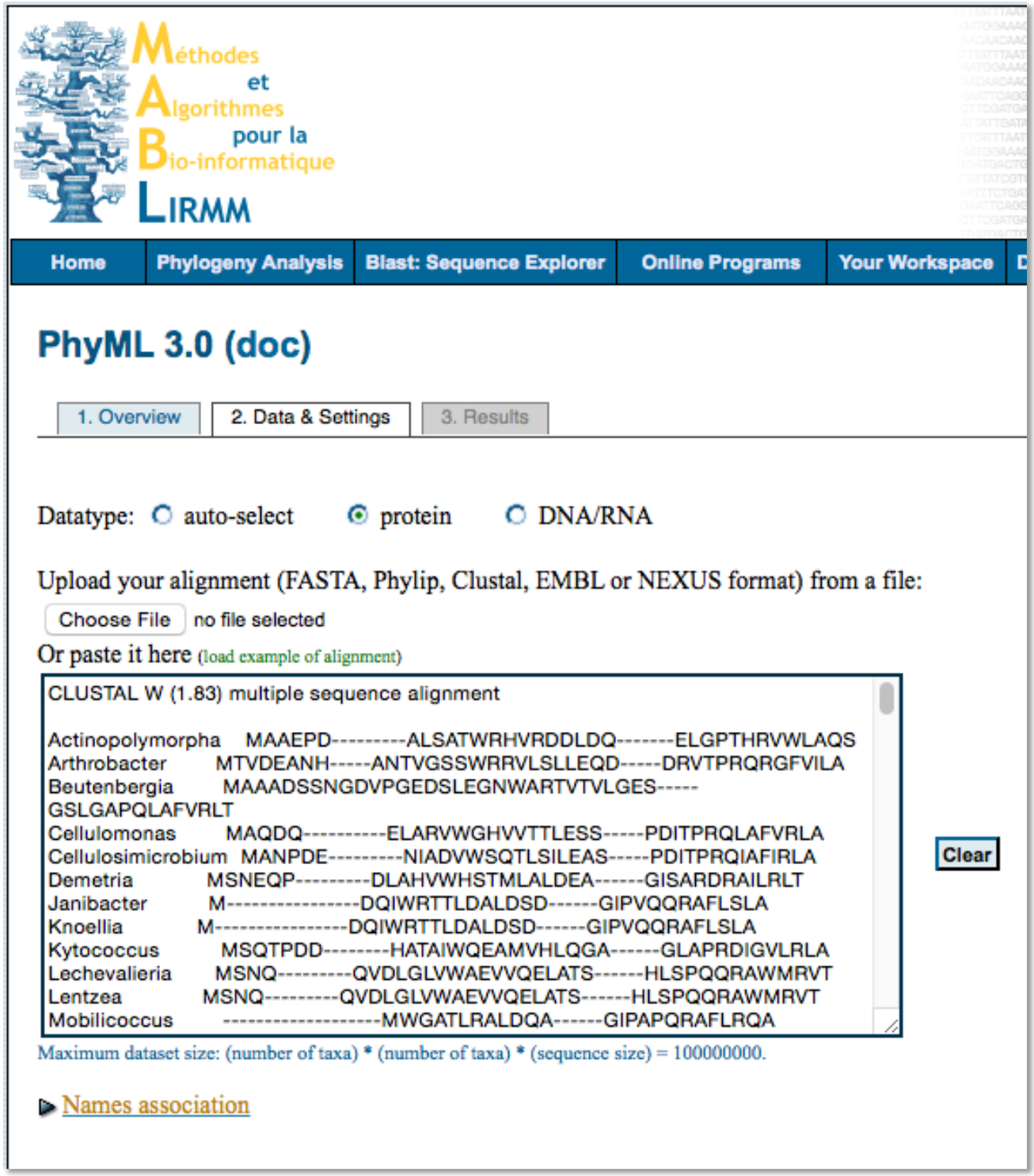

Figure 8.10. The PhyML start page with the T-Coffee alignment generated from Ksed_00010 pasted into the alignment box. Select the protein option and submit to being the analysis.

3. It make take a few minutes to be completed, but a rudimentary unrooted phylogenetic tree will be generated based on the alignment (Figure 8.11). We are not interested in the tree itself at this point, but rather the Tree in Newick format file indicated by the arrow.

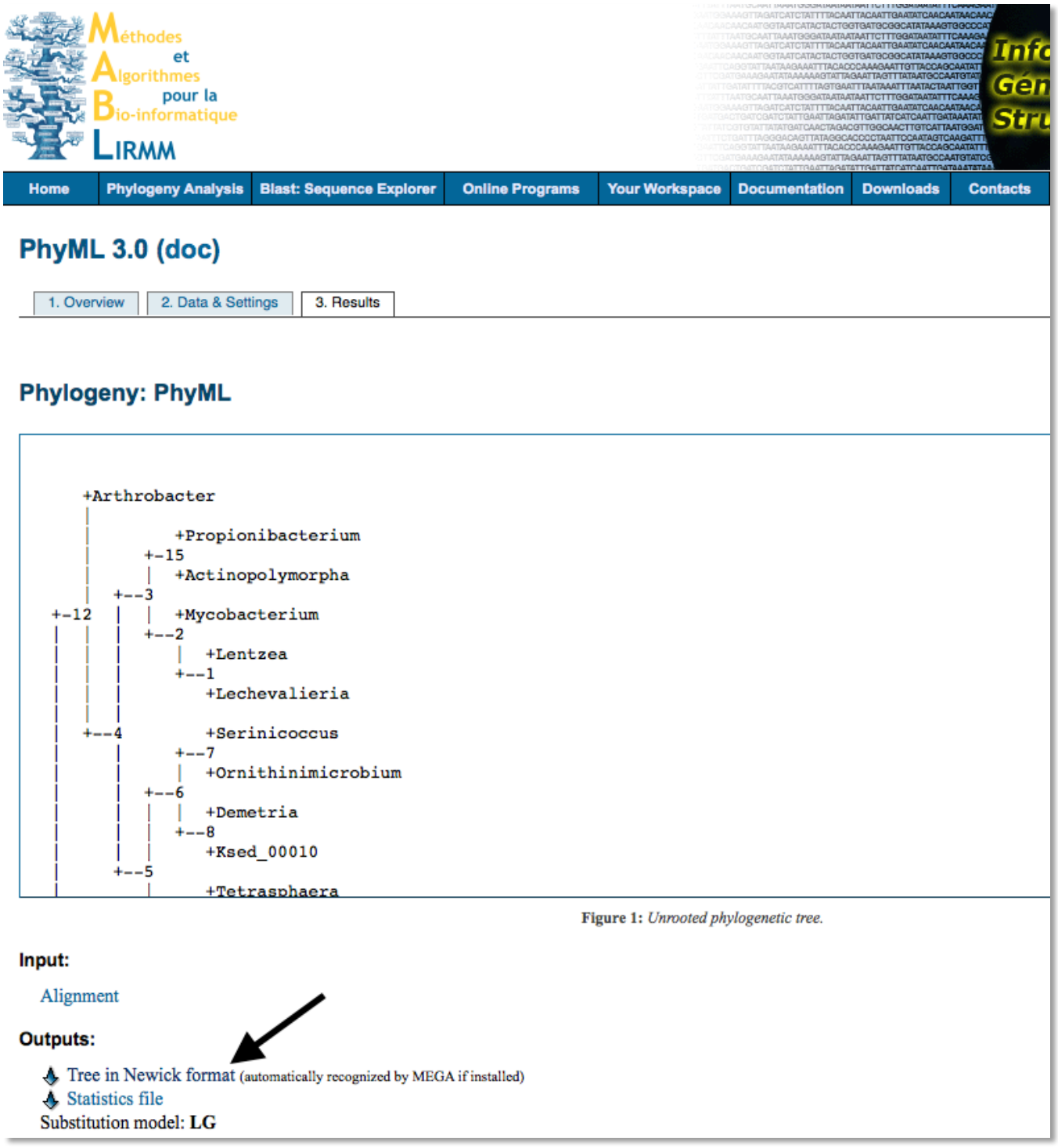

Figure 8.11. The output from the PhyML tool. A crude unrooted phylogenetic tree file is shown at that top, as well as a hyperlink to the results in Newick format as shown by the arrow.

4. Clicking on the Tree in Newick link will open another window as shown in Figure 8.12. This file will be used in higher resolution phylogenetic tree drawing programs available through phlogeny.fr.

```
((((((Cellulosimicrobium:0.2936459041,Cellulomonas:0.2412694841)
0.9220000000:0.0921152442,Beutenbergia:0.5017551797)0.2180000000:0.0690189538,
Arthrobacter:0.7785780943)0.9660000000:0.1249665617,(((Tetrasphaera:0.4917173111,
(Knoellia:0.0424342705,Janibacter:0.0698143629)0.9780000000:0.1361079432)
0.9570000000:0.1394947530,Mobilicoccus:0.3691031781)0.6750000000:0.0486927429,
((Serinicoccus:0.3191061226,Ornithinimicrobium:0.1529821893)
0.9980000000:0.3759242413,(Kytococcus:0.7652702145,Demetria:0.6245865885)
0.6160000000:0.0996593747)0.9990000000:0.3335979306)0.8980000000:0.1034392703)
0.9730000000:0.1702258519,(Propionibacterium:0.9744162000,
Actinopolymorpha:0.2990113739)0.9620000000:0.2016282530)
1.0000000000:0.6204767276,Mycobacterium:0.7884713727)0.9570000000:0.2232154702,
Lentzea:0.0390734776,Lechevalieria:0.0414063111);
```

Figure 8.12. The PhyML results in Newick format. This file should be copied and pasted into your notebook or a Word document and saved, as it will be pasted into two different tree drawing tools.

5. From the Online Programs menu at the top of the page at phylogeny.fr, select the TreeDyn tool and paste the Newick formatted file into the text box as shown in Figure 8.13. Then click "submit".

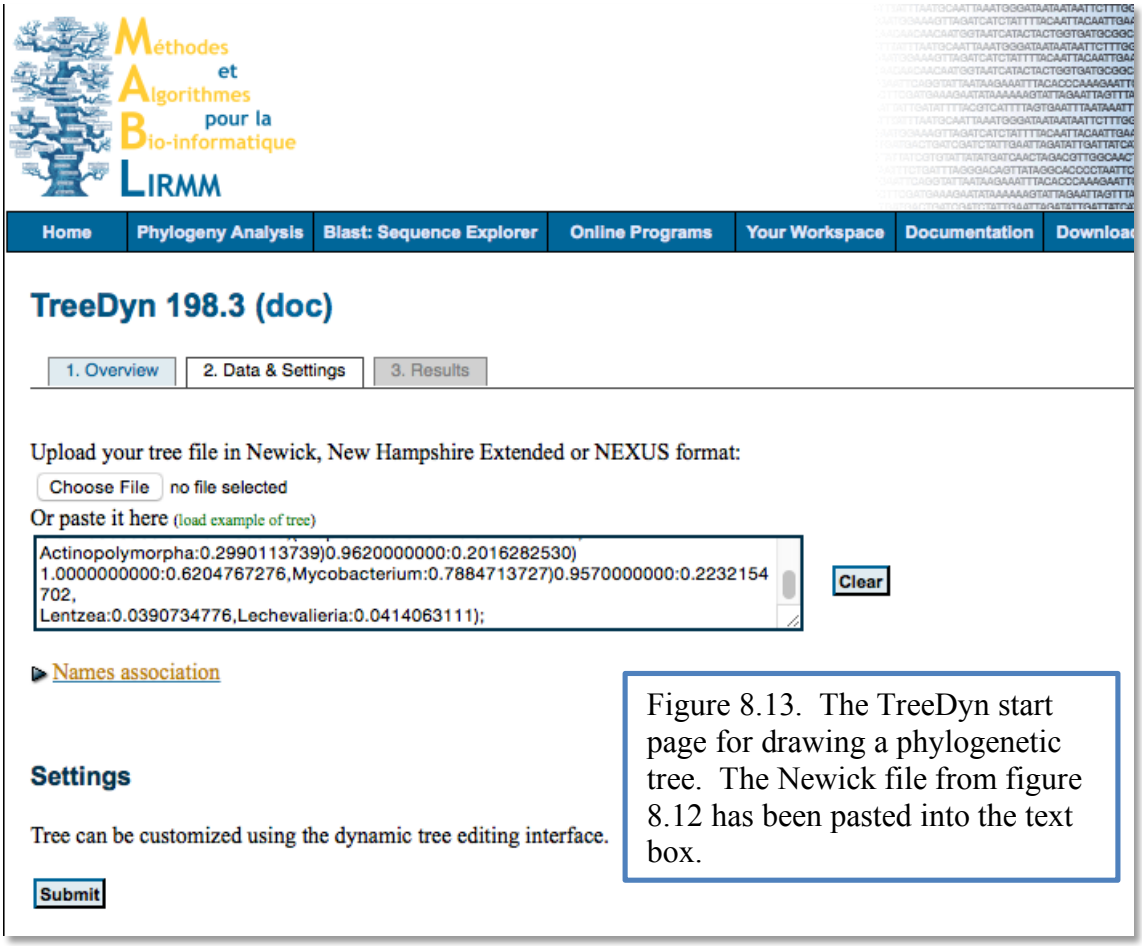

Figure 8.13. The TreeDyn start page for drawing a phylogenetic tree. The Newick file from figure 8.12 has been pasted into the text box.

6. An output will be generated as shown in Figure 8.14. This may take several minutes, so once the program is running you can open another window and proceed to the next section of the manual below to save time.

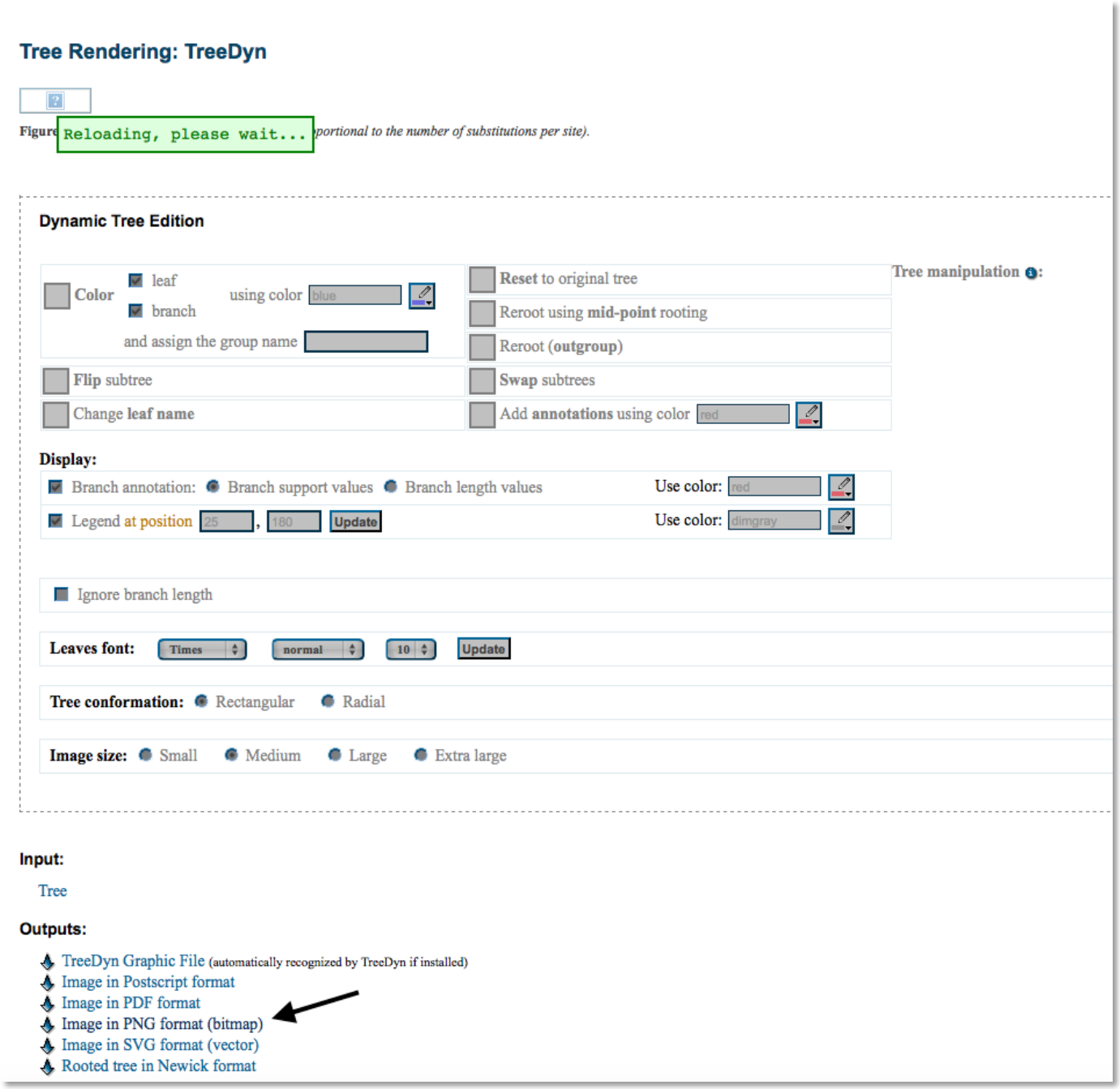

Figure 8.14. The TreeDyn results page. The "reloading, please wait" message in the green box may persist for some time. Once you see this page appear, click on the Image in PNG format (bitmap) link shown by the arrow. This will show the TreeDyn phylogenetic tree.

7. After clicking on the Image in PNG format (bitmap) in Figure 8.14 has been clicked, an unrooted phylogenetic tree similar to the one shown in Figure 8.15 will appear. A concise review of the interpretation of figure 8.14 can be found at: http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny. A summary of the main points, as taken from the hyperlink, is presented here. As before, the vertical dimension has no meaning, but the horizontal dimension is meant to represent the extent of genetic difference. *Demetria* and *Kytococcus* are still in one clade, but now we can interpret that *Kytococcus* has slightly more genetic change from the hypothetical precursor to each due to the fact that the line is longer from the branch point to *Kytococcus* than it is from the branch point to *Demetria*. The red numbers indicate a type of statistical measurement of the significance of the grouping to the right of the number. The closer to 1 that the number is means that there is strong evidence that the sequences to the right of the number cluster together to the exclusion of any other. The line at the bottom is serves as scale of the level of genetic change represented per unit length of lines in the phylogenetic tree.

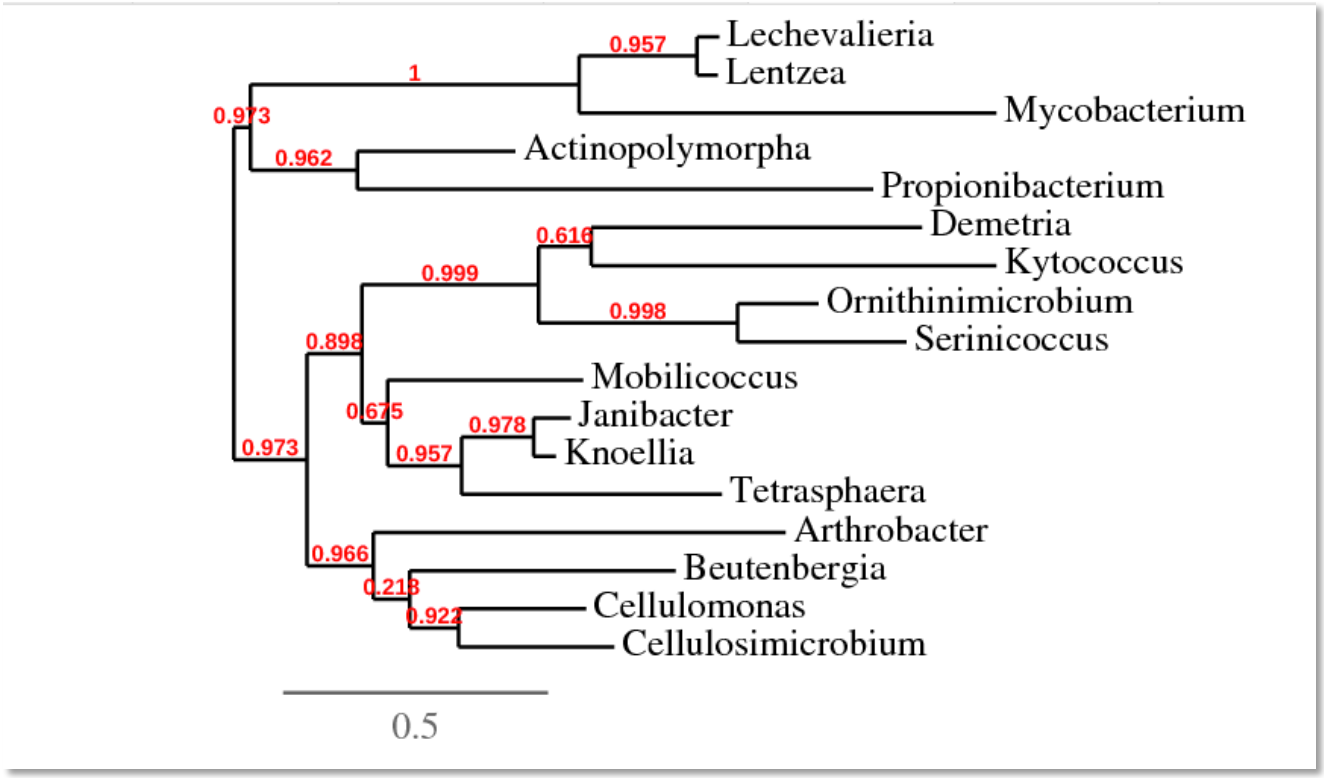

Figure 8.15. The TreeDyn phylogenetic tree. See the text for details.

8. You save the image to your storage device and then upload it to your GENI-ACT notebook

9. Another version of a phylogenetic tree, a so called radial tree, can be produced by another program on phylogeny.fr. To access the program, select Drawtree from the Online Programs menu at phylogeny.fr and paste the Newick formatted tree file into the text box as shown in figure 8.16. Hit submit to begin drawing the tree. As with the TreeDyn program, it may take several minutes for your tree to appear.

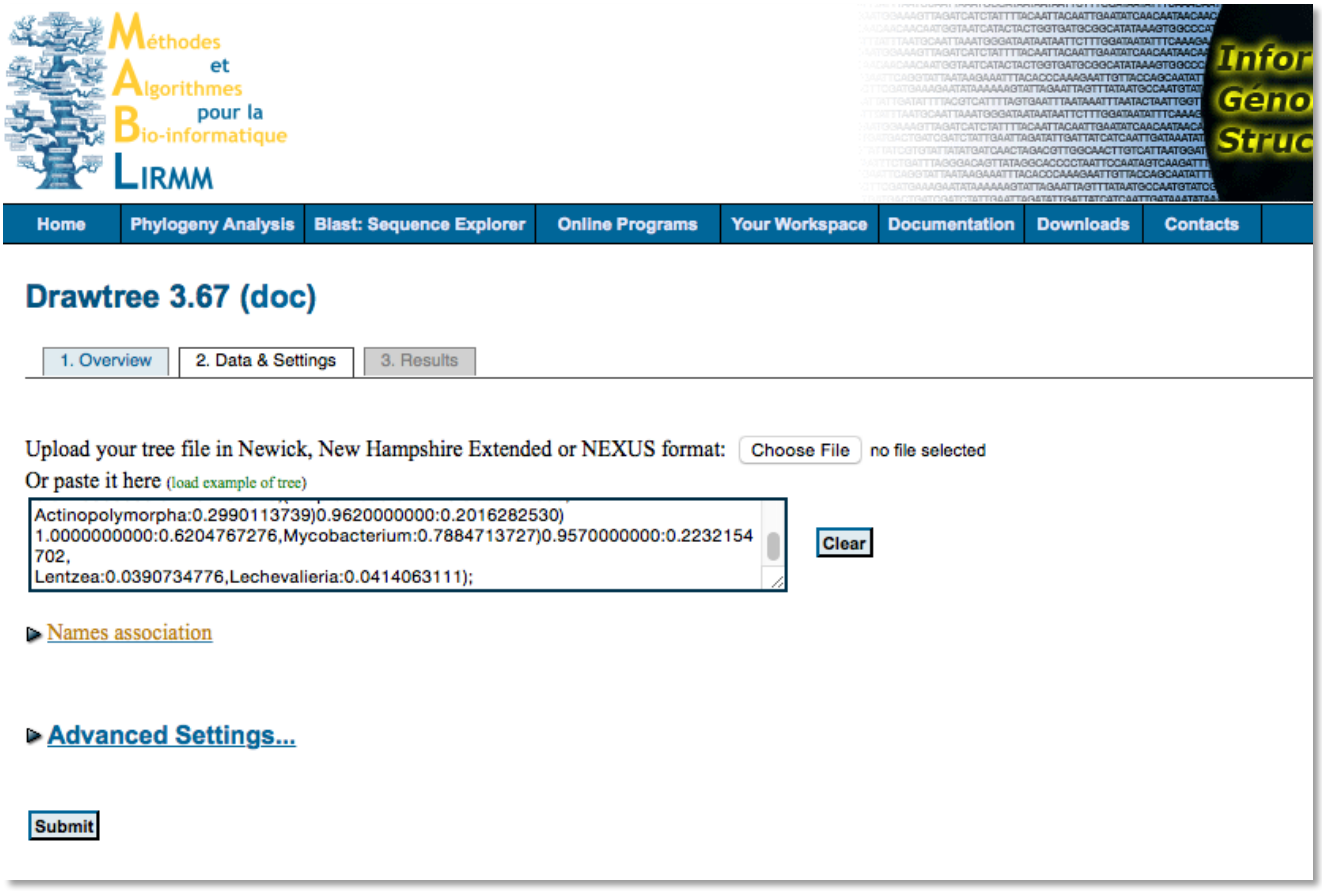

Figure 8.16. The Drawtree start page. The Newick formatted tree file generated earlier from PhyML is pasted into the text box. Hit submit to begin drawing the radial tree.

10. Figure 8.17 shows the Drawtree results for the Ksed_00010 alignment used throughout this section. The grouping of *Kytococcus* and *Demetria* can be seen as in the previous trees, and the line connecting *Kytococcus* to the tree is longer than that for *Demetria*, just as was shown in the TreeDyn tree in figure 8.15. The length of the lines again reflects the amount of genetic change between the organisms in the tree. One shortcoming of this type of tree at times is the overlap of names of organisms, as can be seen at the bottom of figure 8.17.

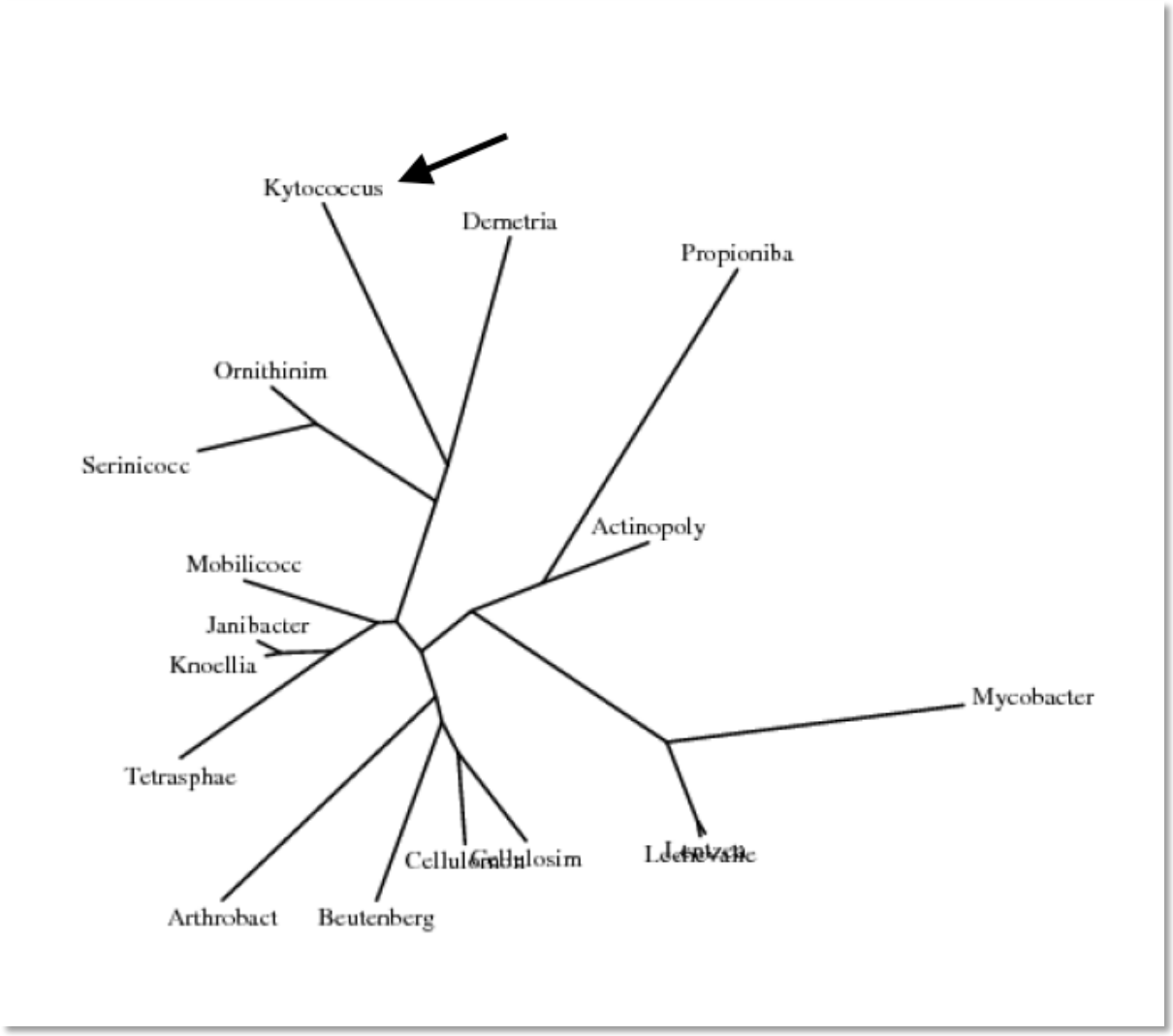

Figure 8.17. The Drawtree radial tree output. The arrow points to the position of *Kytococcus* in the tree. This version of the tree clusters the hits in a similar way as in Figure 8.15, but the length of the branches reflect the degree of relatedness of the tree proteins.

11. Save Drawtree image and upload it as a PNG file in your lab notebook.

12. Review the trees you have constructed and come to a conclusion about whether you feel they show evidence of horizontal gene transfer, as described earlier in this module.

Gene Context

The next piece of information you will consider in order to decide if there is any evidence of horizontal gene transfer is to look not only at your gene, but at the genes found in the immediate neighborhood of your gene on the *Kytococcus sedentarius* genome. Genes tend to be inherited in blocks and if you find that your gene exists in a neighborhood in *Kytococcus* that looks similar in related species, it would support the conclusion that the gene was inherited by vertical transfer. If, on the other hand, you see your gene in a neighborhood more similar to that of unrelated species you would have evidence to support the possibility that the gene was acquired by horizontal transfer.

1. Navigate to IMG/EDU at http://img.jgi.doe.gov/cgi-bin/edu/main.cgi. Click on the Find Genes tab at the top of the page as you did in the Alternative Open Reading Frame Module and select the Gene Search option from the pull down menu.
2. In the Gene Search window paste the locus tag for your gene in the keyword box, select Locus Tag (list) from the filters pull down menu and click Go.
3. On the resulting IMG gene detail page for your gene, scroll down to the Evidence for Function Prediction section.
4. Click the hyperlink for Show neighborhood regions with this gene's bidirectional best hits (Figure 8.18) to determine if IMG has already recognized similar proteins in the same neighborhood in other organisms.

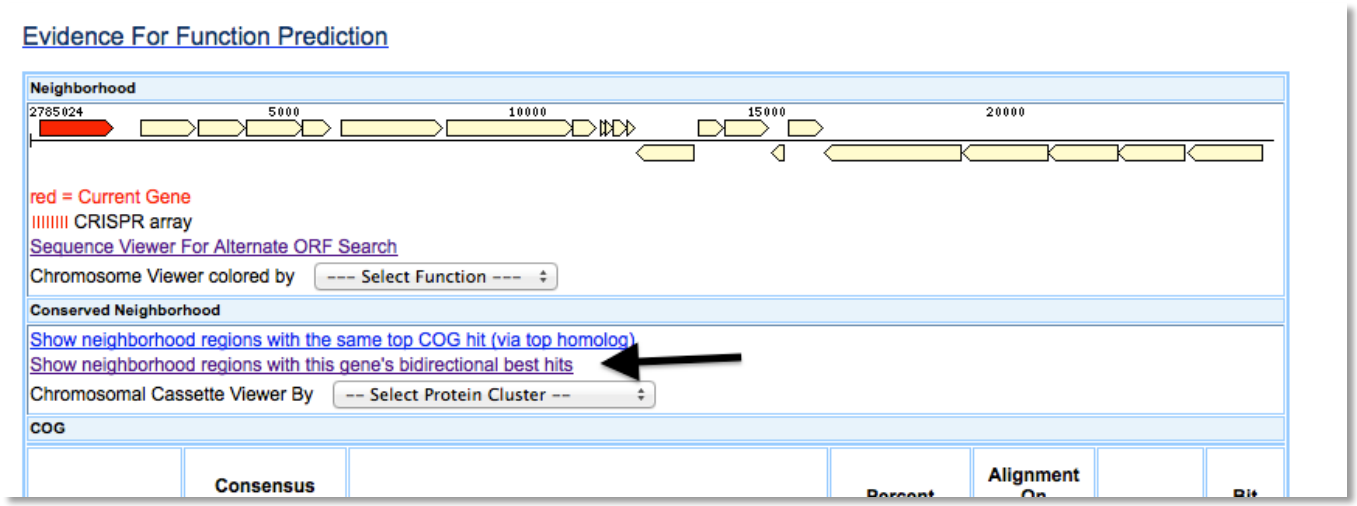

Figure 8.18. The Evidence for Function Prediction section of the Gene Details page on IMG/EDU. The link to show neighborhood regions with genes bidirectional best hits is indicated by the arrow.

5. In the example in Figure 8.19, the arrows show that the gene neighborhood in which Ksed_00010 is found has hits that are very similar in the organisms shown. The query gene (Ksed_00010) will be shown in red and neighbors that are similar will have the same color in each organism. Note also that the genus and species names shown are ones that we have previously determined to be closely related to *Kytococcus* when constructing the phylogenetic trees earlier in this module.

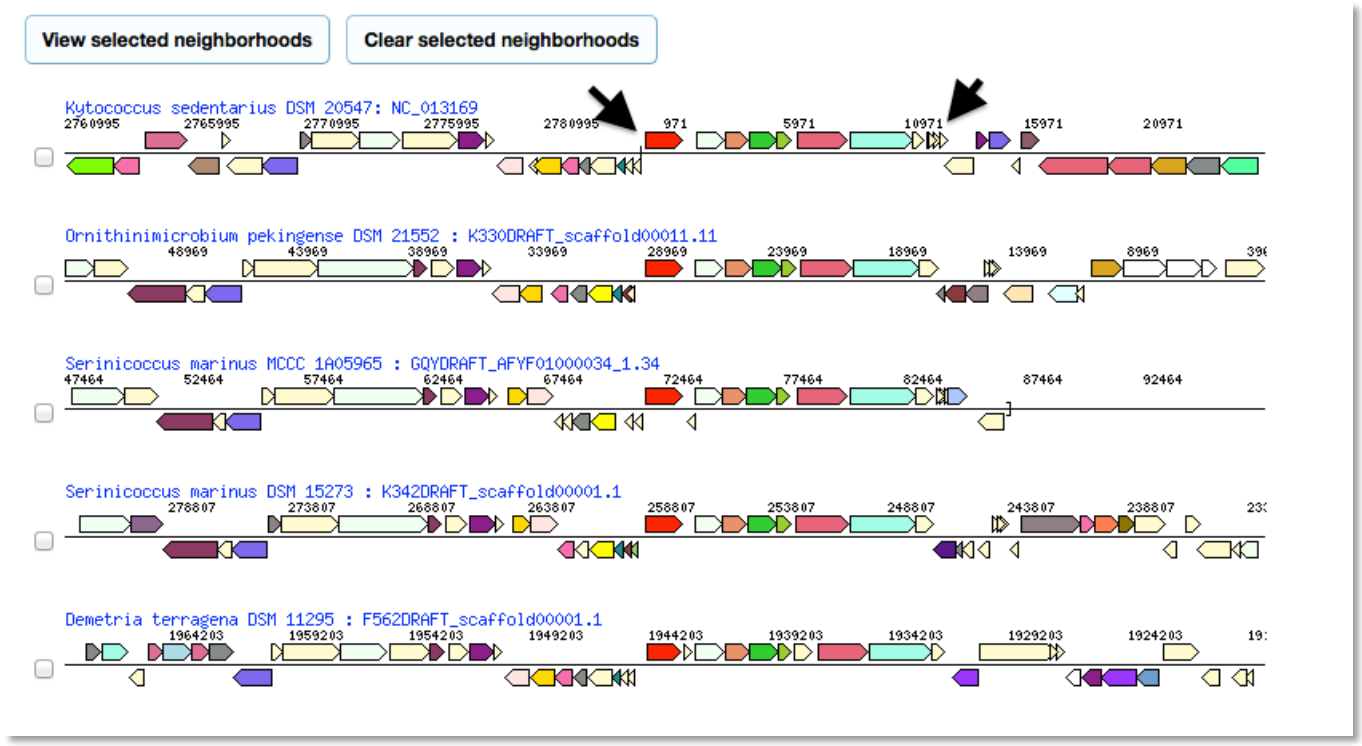

Figure 8.19. The results of the show neighborhood regions with gene's bidirectional best hits. Arrows indicate a series of genes in different organisms with the same organization. The query gene (Ksed_00010) is shown in red.

6. Possible interpretations of the gene context investigation:

   6.1. The gene neighborhood from your organism looks similar to the same neighborhood closely related bacteria – *no evidence for horizontal gene transfer.*

   6.2. The gene neighborhood from your organism looks similar to bacteria that are NOT closely related – *horizontal gene transfer is possible.*

   6.3. The gene neighborhood from your organism looks similar to neighborhoods from related and unrelated bacteria – *can't interpret the results.*

7. Snip (PC) or Capture (Mac) the image of the neighborhood regions with the your gene's best bidirectional hits and upload it to the lab notebook, along with comments about your interpretation of the findings.

Chromosome Viewer Heat Map

The third, and final, piece of information you will consider is how the GC (guanidine and cytosine) content of your gene compares with the average GC content of the genome. If the GC content of the gene you are annotating has a GC content significantly different from the average GC content of the genome, it suggest that there is a possibility that the gene was transferred to your organism from an organism that has a different average GC content.

1. Go to the gene detail page and scroll down to the Evidence for Function Prediction section. Click on the hyperlink "Chromosome Viewer colored by" and select GC. This will map the chromosome region with a colored coded GC % according to the legend at the bottom of the page (not shown). The query gene will have red bar under it (Ksed_00010 in Figure 8.20), and the average GC% of the genome (72%) will be indicated at the top of the page (Genome average GC content in Figure 8.20). Hovering your cursor over the query gene (not clicking on it), will result in a popup box displaying the GC content of the gene. You can also find the GC content of the gene on gene details page of IMG/EDU.

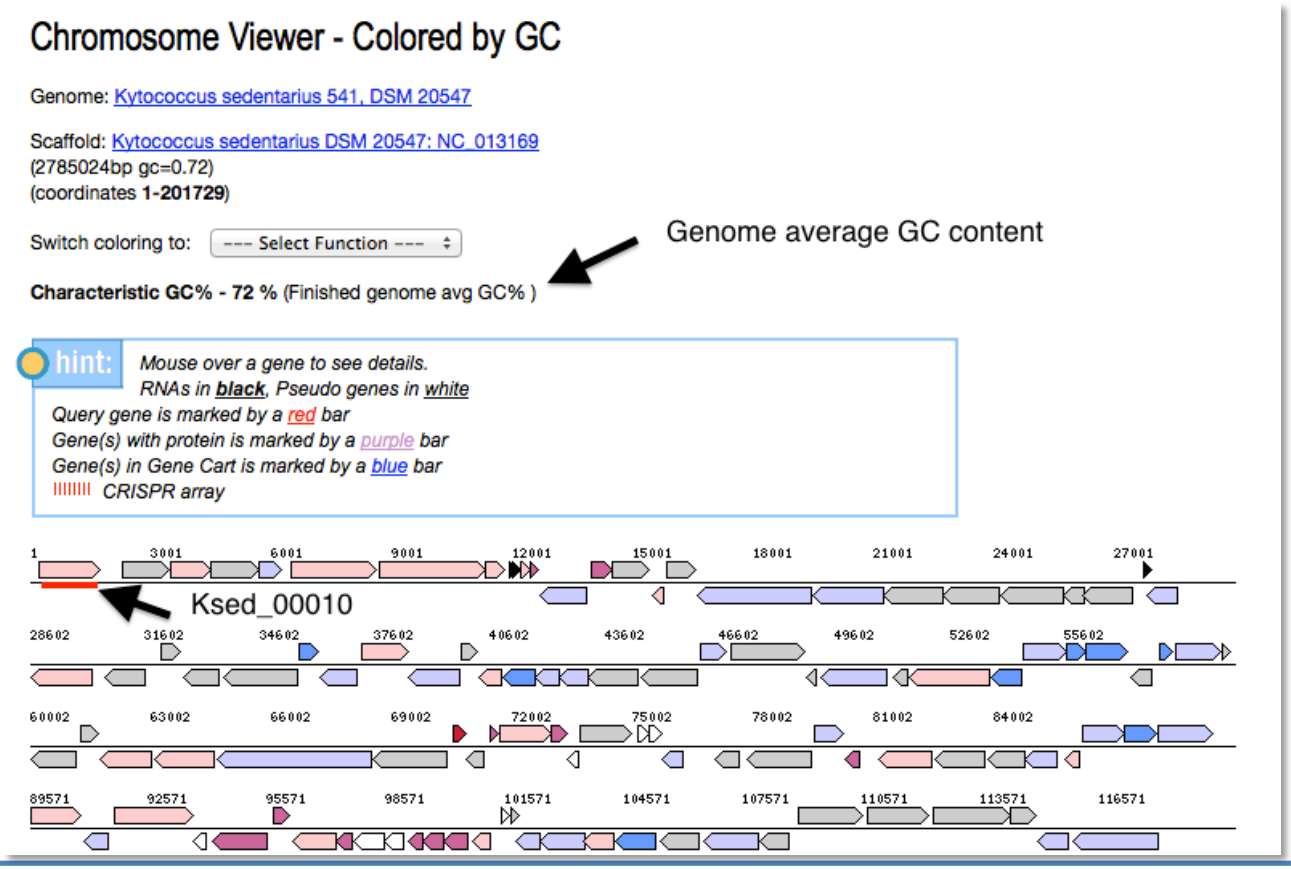

Figure 8.20 . The Chromosome viewer colored by GC content results page. Hovering the cursor over the query gene (Ksed_00010 in the example above) will result in a box popping up with the GC content of the gene, which can then be compared to the average GC% of the genome shown a the top of the figure.

2. Look to see if the genes in the neighborhood share similar GC percentages. If the query gene has a greatly different GC percentage than the genome as whole it could indicate horizontal transfer from the genome of an organism with a GC% different from that of your organism.
3. Check the individual gene for GC percentage. Record the information in the lab notebook.

A hypothesis should be formed about the likelihood of horizontal gene transfer having taken place for your query gene in your organism. Summarize the evidence (phylogenetic tree, gene neighborhood, GC content data) to support horizontal gene transfer if you hypothesize that may have occurred. State that there is no evidence for horizontal gene transfer if you feel the results of this module do not support horizontal gene transfer.