Practicum Assignment 4 Due Fri, Dec. 8th

Hand in your assignment by email to comp-bio@cs.cmu.edu or in MI646 between noon and 4:00pm on Dec.8th.

The goals of this assignment are to give you experience interpreting a gene tree in the context of a species tree and to introduce you to a software package, NOTUNG, that supports this sort of analysis. The tasks are designed to give you skills that will be helpful in future phylogenetic analyses, including your course projects.

NOTUNG juxtaposes a gene tree and a species tree to infer events in the history of the gene family. If a gene family evolves solely through substitution and small insertions and deletions, the gene family tree will agree with the species tree. Larger scale events, including gene duplication, gene loss and horizontal gene transfer, result in a gene tree that disagrees with the species tree. *Reconciliation* is the process of fitting a gene family tree to a species tree to infer (a) the correspondence between each ancestral gene and an ancestral species and (b) the events that occurred in the history of the gene family.

You will need three files which you can download from the course web site:

1. The gene tree, **Prac4-genetree.nwk**

   To simplify the analysis in Practicum 4, we have constructed a gene tree from a substantially reduced subset of the proteins you used in the first three practica. This gene tree was generated with *MrBayes* (http://mrbayes.sourceforge.net/), a program that we did not discuss in class. MrBayes gives branch support scores expressed as probabilities rather that bootstrap replicates. In cases where the sequence data does not provide strong support for a particular branching order, MrBayes will output a tree with a polytomy, instead of binary tree with one or more very weak branches. You will see that this gene tree contains a polytomy.

2. The species tree, **Prac4-speciestree.nwk.**

   To provide context, this species tree contains a leaf for each of the major taxonomic groups in prokaryotes, in addition to the species represented in **Prac4-genetree.nwk.** The leaf names in the gene and species trees are in a format that is Notung-compatible.

3. An annotation file, **A.txt**

To do this assignment, you will need NOTUNG 2.9, which runs on all operating systems and can be downloaded from http://www-2.cs.cmu.edu/~durand/Notung. Click on “Downloads” to be redirected to the distribution page and then download **Notung-2.9.zip**. (The site will ask for your email address. This is optional. If you don't want to be on the mailing list for update announcements, just type "Enter" to continue.)

Unzip this file to obtain a directory that includes the NOTUNG executable, “Notung-2.9.jar”. The distribution also includes a comprehensive manual (manual_2.7_draft.pdf) and a directory with example trees. To make things easy for yourself, we recommend that you put the gene and species tree files in the same directory as the NOTUNG executable (or the NOTUNG excecutable in the same directory with your practicum files.)

Hints for using NOTUNG:

- The functions in the "Fonts" and "Zoom" pull-down menus at the top of the NOTUNG window are useful for adjusting the size and readability of a tree. Here are some shortcuts:
  - Cntl-t will adjust the zoom so the entire tree is visible in the tree panel.
  - You can zoom in on the horizontal axis using the "Ctrl+]" key combination and zoom out using "Ctrl+[" combination.
  - Similarly, "Ctrl+}" (i.e., "Ctrl+Shift+]") and "Ctrl+{" will zoom in and out on the vertical axis.
- NOTUNG interprets a tree as a gene (resp. species) tree if it is opened with the "Open Gene Tree" (resp. "Open Species Tree") command. If you open your gene tree with the "Open Species Tree" command or vice versa, you will not be able to proceed. If that happens, close the tree and reopen it with the correct command.
- Note that the leaves of the gene tree in this assignment are of the form *SPECID-geneID,* where SPECID is a five letter code for the species from which the gene was sampled. The same five letter species identifier codes were used to label the species tree. Embedding the species name in the leaf labels of the gene tree allows NOTUNG to match leaves in the gene tree with leaves in the species tree. In general, it is not necessary to use five-letter species identifiers, but the same species naming system must be used for both gene tree labels and species tree labels.
- When opening a file, NOTUNG'S file open dialog window shows you files with the .nhx extension only. You can see all other files in the current directory by changing the file type to "All files" using the pull-down menu. NOTUNG recognizes three tree file formats: newick (.nwk), extended newick (.nhx), and its own format, which is a super-extended newick. Some programs generate trees in Newick format but use other file extensioins. NOTUNG doesn't care what file extension is used (.phy, ,txt, .nwk, .nh….). If the file contains a tree in one of the three tree formats, NOTUNG will open it.

1) Open the tree files

- Click "File → Open Gene Tree" and open **Prac4-genetree.nwk.**

  A warning message will pop up informing you that this tree has edge weights in multiple locations. The warning appears because **Prac4-genetree.nwk** has two sets of edge labels: 1) branch support scores and 2) branch lengths. This is not an issue for this analysis.

- Click "**OK**."

  The gene tree appears in the tree panel. The branch support scores on the branches. These range from 0 to 1.

- Select "**File → Import Annotations**" and open **A.txt.**

The labels on the leaves of the tree will appear in color. Red and orange tones correspond to Archaeal species, blue and purple tones correspond to Firmicute species and green tones correspond to Deltaprotebacteria and Spirochaetes.

To keep the assignment short, the file **A.txt** was prepared for you in advance using Notung's Annotations feature and saved using the "**File → Export Annotations**" command. For your projects, you may find it helpful to annotate taxa in color using the Annotations feature.

- Select "File → Open Species Tree" and open **Prac4-speciestree.nwk.**
- Select "File → Import Annotations" and open **A.txt.**

  The labels on the leaves of the tree will appear in color.

2) Reconcile the gene tree with the species tree using the Duplication-Loss (DL) model.

- Click on the **Prac4-genetree.nwk** tab to select the gene tree before proceeding.
- Select the "**Reconciliation**" tab from the lower window and click "**Reconcile/Rereconcile**."

  The **Reconciliation** dialog box will appear. In this dialog box, NOTUNG asks you to specify which species tree to use for the reconciliation and what naming convention is used in the gene tree to specify the species associated with each gene.
- Select **Prac4-speciestree.nwk** in the drop-down menu labeled "Please select a species tree to reconcile with." Select "**Prefix**" under "Specify Species Label" and click "**Reconcile**."

  The reconciled gene tree now appears in the tree panel. Duplication nodes are highlighted with a red square and the letter 'D.' All other nodes are speciation nodes. For each inferred loss, a leaf, shown in gray, is added to the tree to represent the missing gene(s).

  NOTUNG assigns a numerical value, called *the Event Score*, to each reconciled tree. The Event score is the weighted sum of the number of duplications (N_D) and the number of losses (N_L). By default, Event score = $1.5\ N_D + N_L$, although the weights can be changed by the user. The values of the event weights have no impact on reconciliation with duplications and losses, but do affect other functions including reconciliation with transfers and rearrangement.

  The number of duplications and losses associated with the current reconciliation, as well as its Event Score, are shown in the lower left of the Reconciliation task panel.

Record the number of duplications and losses and the Event Score in Table I in the Practicum 4 worksheet.

- Loss nodes can expand the size of your tree substantially. For some interpretation tasks, displaying loss nodes is very useful; for others, losses can make the tree

look cluttered. It is helpful to be able to turn them off, temporarily. Uncheck the "**Display Loss Nodes**" box in the Reconciliation task panel.

The loss nodes will disappear.

- Select "**Display Options → Display Internal Node Species Names**."

  At each internal node in the gene tree, the name of the species that corresponds to that node will appear in italics. Note that this function only works after the tree has been reconciled and a mapping has been established between internal nodes in the gene and species trees.

  Find the polytomy in the tree and inspect the species that are mapped to the polytomy and to each of its children. This is easier to see when the loss nodes are not displayed. (Note that the children of the polytomy in this tree are the roots of binary subtrees.)

Record the following information in the Practicum 4 worksheet:

- ***Enter the numbers of inferred duplications and inferred losses and the Event score in Table I.***
- ***The species associated with the polytomy.***
- ***The species associated with each of its children.***

3) Resolve the polytomy in the gene tree

- Make sure that the **Prac4-genetree.nwk** tab is still selected.
- Select the "**Resolve**" tab from the lower window and click "**Resolve Polytomies**."

  A binary resolution of the gene tree will appear. The branch that was introduced to resolve the gene tree is shown in turquoise.

  Find the "resolved node"; that is, the node that is the most recent common ancestor of the three subtrees that were the descendants of the polytomy in the original, unresolved gene tree. This resolved node is easy to find, because it is the root of the turquoise branch. Inspect the species that are mapped to this node and to each of its children. (Again, this is easier to see when the loss nodes are not displayed.)

  The resolved node will have two children. One of the children corresponds to a node that was also a child of the polytomy; that is, it is the root of one of the tree subtrees descending from the polytomy. The other child is the root of a subtree that was formed by merging the other two subtrees descending from the polytomy.

Record the following information in the Practicum 4 worksheet:

- ***Enter the numbers of inferred duplications and inferred losses and the Event score in Table I.***
- ***The species associated with the resolved node.***
- ***The species associated with each of its children.***

Based on this information, and information that you recorded in step 2, which node is the root of the subtree resulting from the merger?

4) Rearrange the reconciled tree

Poorly supported areas in the gene tree can be corrected with "duplication-loss parsimony." The rationale is that if the sequence data does not contain enough information to infer the branching order unambiguously, information about gene duplications and losses can be brought to bear on the phylogeny reconstruction process. Under the assumption that gene duplication and loss are rare events, the tree that implies the fewest duplications and losses in the history of the family is the best hypothesis.

If an incongruent branch is weakly supported, NOTUNG will rearrange that branch to reduce the Event score. Here, "weakly supported" means that it has an edge weight below some threshold. The default is 90% of the highest edge weight in the tree. This threshold can be changed using the pop-up menu in the lower left hand corner of the NOTUNG GUI.

- Select **Prac4-genetree.nwk**, if it is not still selected. Click the "**Rearrange**" tab.
- Click the "**Highlight weak edges**" checkbox.

  Several edges in the reconciled tree are highlighted in yellow. These are edges with weights below the Edge Weight Threshold and are considered "weak." Weak edges may be rearranged to reduce the number of duplications and losses in the tree. Edges with weights above the threshold will not be rearranged.

  A threshold of 0.9 is quite liberal; it allows NOTUNG to change any branch with branch support that is less than 0.9. We will initially try rearranging the gene tree with a more conservative threshold.
- Click the "**Edit Values**" button in the bottom-right corner of the program.
- In the dialog box, change the Edge Weight Threshold to 0.75.
- Click "**Apply Changes**."

  The yellow highlighting on branches with support scores between 0.75 and 0.9 will disappear.

- Click "**Perform Rearrangement**." The rearranged tree will appear in the tree panel; the inferred events and the new score will appear in at the bottom of the task panel.

Enter the numbers of inferred duplications and inferred losses and the Event score in Table I.

- Click the "**Edit Values**" button in the bottom-right corner of the program.
- In the dialog box, change the Edge Weight Threshold back to 0.9.
- Click "**Apply Changes**."

  A pop-up window with a diagnostic message will appear. In this case, you can ignore this diagnostic. Click "**OK**"

  Branches with scores greater than 0.75 and less than 0.9 will turn yellow.

- Click "**Perform Rearrangement**." The rearranged tree will appear in the tree panel; the new score will appear in at the bottom of the task panel.

Enter the numbers of inferred duplications and inferred losses and the Event score in Table I.

5) Investigate horizontal transfers

The default event model in NOTUNG is a duplication-loss (DL) model. In prior steps, we inferred the most parsimonious history of *duplications and losses* needed to explain the disagreement between the gene tree and the species tree. However, *horizontal gene transfer*, the transmission of genetic material from an organism in one species to the genome of an organism in another species, is a common phenomenon in prokaryotes. Since our data set consists of prokaryotic sequences, it is possible that horizontal transfer played a significant role in the history of this family.

In this step, you will rereconcile the gene tree using a duplication-loss-transfer (DTL) model. NOTUNG infers the most parsimonious history of these events by minimizing the weighted sum of the number of duplications (N_D), the number of transfers (N_T) and the number of losses (N_L). By default, Event score under the DTL model is

$$1.5\ N_D + 3.0\ N_T + 1.0\ N_L.$$

The weights can be changed by the user and we will do so for this analysis. With the default event costs, a transfer will be inferred if it can replace three or more losses. (If it replaces three losses, then there will be two equally good histories, one with three losses and one with a transfer. If it replaces four or more losses, then the history with the single transfer will be preferred.) We will be slightly more conservative and set the transfer cost to 5.0.

- Click on the **Prac4-genetree.nwk** tab to select the gene tree before proceeding.
- Select the "**Reconciliation**" tab. If not already turned on, select "Display Options → Display Internal Node Species Names."
- Select "**Infer Transfers**" to enable the DTL model.

- Click the "**Edit Values**" button in the bottom-right corner of the program. In the dialog box, change the Transfer Cost to 5.0. Click "**Apply Changes**."

  Two diagnostic pop-up windows will appear. Again, you can ignore these. Click "**OK**" twice.

- Click "**Reconcile/Rereconcile**."

  As before, the **Reconciliation** dialog box will appear. In this dialog box, NOTUNG asks you to specify which species tree to use for the reconciliation and what naming convention is used in the gene tree to specify the species associated with each gene. Verify that the settings are correct, change them if needed, and click "**Reconcile**"

  The reconciled gene tree now appears in the tree panel. Duplications and losses are displayed as before. For each inferred transfer event, the gene tree edge associated with that transfer will appear in yellow, with a yellow triangle in the middle of the edge.

- Click on the yellow arrow of any transfer.

  The donor species and recipient species of all transfers will be displayed

- *Multiple optimal solutions:* Under the DL model, the most parsimonious history of duplications and losses is unique. In contrast, under the DTL model, more than one series of duplications, transfers and losses may result in the same pattern of gene tree incongruence. After reconciliation with the DTL-model is performed, the number of optimal solutions (histories) found is displayed at the bottom of the task panel.

Record the number of optimal solutions in your worksheet.

- When there is more than one optimal event history, a green circle is displayed on the root of each the subtree that can be explained by more than one event history. It is easy to browse through all feasible histories by clicking the corresponding green circle repeatedly.

  Make sure that Display Loss Nodes is selected. Find the green circle that is closest to the leaves (i.e., the rightmost green circle). Click on it. The label on transfer to the right of the circle will change.

  Clicking this green circle will toggle between the optimal histories for the subtree to the right of the circle. In this case, there are two histories: In one history a loss occurs before the transfer to the right of the green circle. (There is a second transfer in this subtree that does not change.) In the other history, the loss occurs after the transfer.

Toggle back and forth between the two histories and record the two labels on the transfer associated with the alternate histories in Table II

- Now we will focus on alternate event histories near the root of the tree. There are five such histories Unselect Display Loss Nodes so you can focus on changes in transfer events near the root.

Find the green circle that is closest to the root (i.e., the leftmost green circle). Click on it repeatedly to cycle through all of the optimal solutions. (Note you will see a change near the root on every second click. The intermediate clicks toggle back and forth between the two histories associated with the green circle near the leaves.)

For each optimal event history (i.e., each solution), enter the following information in the Practicum 4 worksheet. (*Tables III and IV have one line for each optimal solution. You can enter the optimal solutions in any order, but use the same order in both tables.)*

- ***Table III***
  - ***The number of duplications, transfers and losses associated with this solution.***
  - ***The event score.***
  - ***The species associated with the root of the gene tree.***
- ***Table IV***
  - ***Is there a transfer at the root of the tree in this event history (Y/N)?***
  - ***The donor and recipient species of the transfer at the root of the gene tree. If there is no transfer at root, leave blank.***
  - ***The donor and recipient species of the transfer at a child of the root of the gene tree. If there is no transfer at root, leave blank.***

Tables II – IV provide information the optimal solutions recovered by this reconciliation. How many rows are there in Table II? How many rose are there in Tables III and IV? What is the relationship between the number of rows in the tables and the total number of optimal solutions reported at the bottom of the task panel?

6) Interpreting the history of the PylB family in light of your analysis

- Consider the number of events inferred in Steps 2- 4 of your analysis, which you recorded in Table I.
  - ***What impact did rearrangement have on the number of duplications? On the number of losses? Was the change in the number of events similar for both duplications and losses, or was the impact greater on one of them? How do you explain what you observe?***
  - ***What impact did switching to a model with duplications and transfers have on the number of duplications? On the number of losses? Was the change in the number of events similar for both duplications and losses, or was the impact on one of them greater? How do you explain what you observe?***

- Given duplications and losses only, there is only one most parsimonious reconciliation of a gene tree and a species tree. However, when the history of the gene family includes transfers, there may be more than one most parsimonious history of duplications, losses and transfers that explains the disagreement between a gene tree and a species tree. In fact, you recovered several optimal solutions for the trees in this practicum.

  Equally parsimonious histories can differ in several ways. Two histories may have the same number of events of each type, but those events may occur in different taxa. Since both histories have the same numbers of events of each type, they must have the same event score. It is also possible to have two histories with the same event cost, but different numbers of events of each type.

Considering Table II only:

- **When you click on the green circle closer to the leaves to inspect a different optimal event history,**
  - **does the number of transfers in the subtree rooted at that green circle change?**
  - **does the number of duplications in the subtree rooted at that green circle change?**
  - **does the number of losses in the subtree rooted at that green circle change?**
  - **do the events occur in different locations in the *gene* tree in the two histories?**
  - **do the events occur in different locations in the *species* tree in the two histories?**

Considering Tables III and IV only:

  - ***How many optimal histories listed in those tables had the same number of events, but in different locations in the tree?***
  - ***Did you observe any histories where the number of events of each type changed?***
  - ***If so, which events types increased and which events types decreased? Explain what you observe in terms of the reconciliation costs?***

- When there is more than one most parsimonious history, it can be difficult to interpret the results. In this case, it is often helpful to identify common and distinct features of the optimal histories:
  - ***How many histories had a transfer at the root of the tree?***
  - ***How many histories involved more than one transfer between bacteria and archaea?***

- ***How many histories involved more than one transfer within the Archaea?***
- ***How many histories imply that the pyrrolysine originated in Archaea? Of those, do they agree on the taxonomic group within the Archaea?***
- ***How many histories imply that the pyrrolysine originated in Bacteria? Of those, do they agree on the taxonomic group within the Bacteria?***
- ***Which history do you find most convincing? Why?***

- Pyrrolysine is found in three distantly related groups of bacteria belonging to different Bacterial phyla. It is primarily observed in the Firmicutes and the Detalproteobacteria, but has also been found in one Spirochaete species.
  - ***Does the taxonomic group that is most closely related to the Spirochaete belong to the Firmicutes, the Deltaproteobacteria or the Archaea?***
  - ***The history of events in the clade comprising the Spirochaete and its sister taxon is the same in all of the most parsimonious histories reconciliations. What are those events? Does this event history seem plausible to you or did you expect a different event history for the Spirochaete? Why or why not?***