**PCB 4674 Evolution** 5/23-25/2016

**Lab # 3: Phylogenies Ⅱ**

**PHYLOHENIES Ⅱ**

Reference chapters (in textbook): Ch. 4

**Lab objectives:**

* Explore NCBI database
* Examine different tree construction methods

**Please type your answers to the bolded questions. You will be submitting this on canvas by next week. It would help us out if you used a different font or color.**

**Protip:**

* Use alt+tab to quickly switch back and forth between your most recently used applications in windows (i.e. your web browser and Microsoft word) (It’s command+tab on a Mac I think)
* Ctrl+c to copy selected text (and other things)
* Ctrl+v to paste

1. **Introduction to NCBI**

The NCBI (National Center for Biotechnology Information) is an online database created to house the huge amounts of biomedical and genomic data that were amassing rapidly in the late 1980s. It was designated as a division of the National Library of Medicine (NLM), itself a division of the National Institutes of Health (NIH), and to this day is curated (updated, verified, etc) by researchers and data curators at the NLM. The mission of the NCBI is to:

“…*develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease. More specifically, the NCBI has been charged with creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules*.”

To learn more about the inception, maintenance, and mission of the NCBI, visit the website (<http://www.ncbi.nlm.nih.gov/>) and follow the links for **More about NCBI**, **Mission**, etc. The NCBI is by no means the only genetic database of its kind. Comparable databases are maintained at top-notch research facilities worldwide and include the EMBL (Europe) and DDBJ (Japan) databases (both of these databases share data with NCBI, so scientist can deposit data in any of these databases and it will appear in all three). In addition, there are many databases with specific foci that are housed and maintained by universities and private research facilities. Together, these online resources bring the international wealth of genetic data to the fingertips of anyone with internet access.

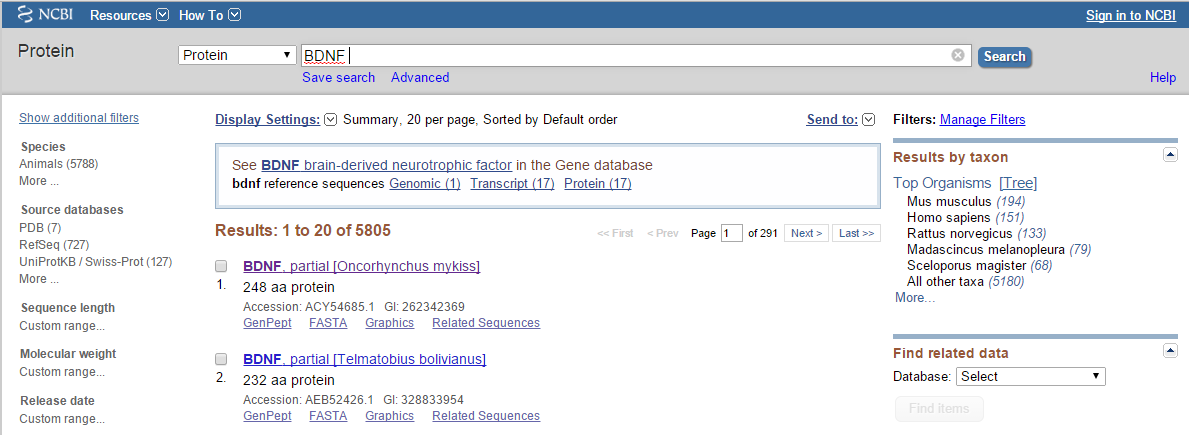
Here, we will use the human protein BDNF to familiarize you with the use of NCBI.

The NCBI homepage can be found at the web address listed above. Note that the NCBI logo appears in the upper left corner of the screen (double helix next to the letters NCBI); if you would like to return to the homepage at any time during an NCBI session, simply click this logo (which appears on all pages within the NCBI website). The homepage also provides links to a number of resources including the mission statement, various types of genetic analysis software, and PubMed – the information database in which all medically-relevant publications are cataloged. The databases that most frequently use are located on the right side of the homepage in a panel labeled: **Popular Resources**.

To find something you are interested, simply type the name or the accession code in the **Search** box at the top of the page. This search box is available on almost every page in the NCBI website which allows you to switch easily from, for example, protein searches to nucleotide searches simply by selecting the appropriate database from the dropdown menu.



Now, meet BDNF (**B**rain-**d**erived **N**eurotrophic **F**actor). This protein plays a special role in supporting growth and differentiation of neurons in the vertebrate brain and is encoded by the gene *bdnf* (please make a note of the nomenclature: proteins are signified BDNF, genes are signified bdnf). In the search box at the top of the **Protein** homepage,enter **BDNF** and click **Search**. A search for the term “BDNF” returns over 5000 results, as indicated at the top of the page. It’s important for you to keep in mind that not all sequence files are created equally; because this database is open-access, anyone can submit sequences to be cataloged here. To find only the best and most highly curated/maintained sequences, you will need to **Filter your Results** by clicking on the **RefSeq** link on the panel of the left side of the page (circled in red).



**How many sequences are available on the RefSeq tab?**

The number in parentheses after RefSeq indicates the number of sequences available using that filter. Because all of the sequences in the RefSeq database have been checked for accuracy, there will be fewer sequences available when using this filter than simply searching for **All** sequences that matched your query. Whenever possible, you should limit your searches to only those sequences available in the RefSeq database.

Expand the “Top Organisms” panel on the right and you should see a number of organisms with sequences for BDNF that were isolated from those species. **Click on *Gallus gallus***. Looking at the list on this page, you can discern several pieces of useful information without even clicking on a single record. For example, record 1 looks like this:

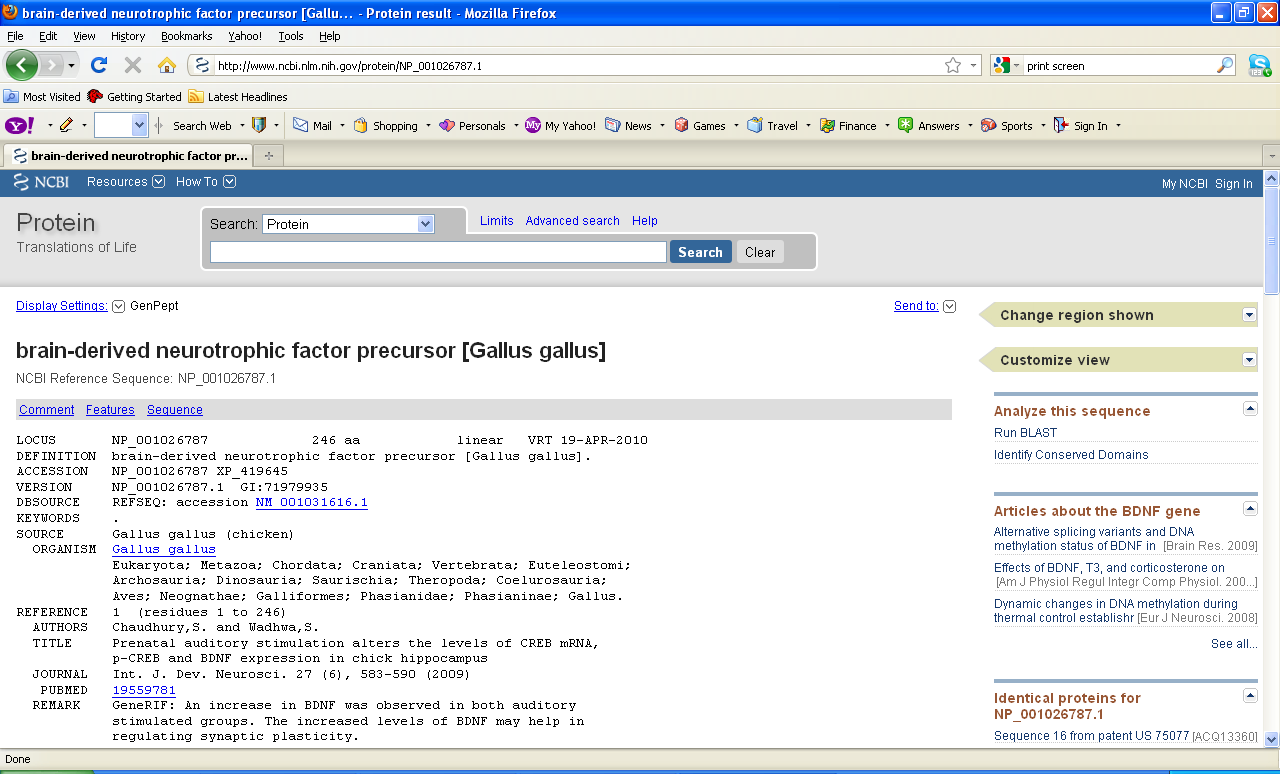
[1.] brain-derived neurotrophic factor precursor [Gallus gallus]

246 aa protein

Accession: NP\_001026787.1 GI:71979935

This record provides information about the organism from which this sequence was isolated (*Gallus gallus,* i.e. chickens), the length of the protein (246 amino acids), and the accession number (NCBI’s system of keeping track of sequence data).

Click on the name of the queried protein and you will be directed to the Protein page for that entry, a page that looks like this:



Scroll down the page, you will see several pieces of information, outlined on the left side of the screen. There is a lot of information contained within those records.

**In which section of the Protein record would you find the names of the researchers who first studied/published information on a protein of interest?**

Reference:

Authors

REFERENCE 1 (residues 1 to 246)

AUTHORS Loria MJ, White SW, Robbins SA, Salmeto AL, Hymel KA, Murthy SN,

Manda P and Sufka KJ.

REFERENCE 2 (residues 1 to 246)

AUTHORS Suzuki K, Maekawa F, Suzuki S, Nakamori T, Sugiyama H, Kanamatsu T,

Tanaka K and Ohki-Hamazaki H.

REFERENCE 3 (residues 1 to 246)

AUTHORS Taylor AR, Gifondorwa DJ, Robinson MB, Strupe JL, Prevette D,

Johnson JE, Hempstead B, Oppenheim RW and Milligan CE.

REFERENCE 4 (residues 1 to 246)

AUTHORS Yamaguchi S, Aoki N, Kobayashi D, Kitajima T, Iikubo E, Katagiri S,

Matsushima T and Homma KJ.

REFERENCE 5 (residues 1 to 246)

AUTHORS Dai P and Xiyang YB.

REFERENCE 6 (residues 1 to 246)

AUTHORS Zhou X, Nai Q, Chen M, Dittus JD, Howard MJ and Margiotta JF.

REFERENCE 7 (residues 1 to 246)

AUTHORS Chytrova G and Johnson JE.

REFERENCE 8 (residues 1 to 246)

AUTHORS Maisonpierre PC, Belluscio L, Conover JC and Yancopoulos GD.

TITLE Gene sequences of chicken BDNF and NT-3

REFERENCE 9 (residues 1 to 246)

AUTHORS Isackson PJ, Towner MD and Huntsman MM.

REFERENCE 10 (residues 1 to 246)

AUTHORS Hallbook F, Ibanez CF and Persson H.

**In which section of the Protein record would you find the taxonomic classification of the organism from which the protein was isolated?**

ORGANISM [Gallus gallus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9031)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;

Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;

Phasianidae; Phasianinae; Gallus.

**What chromosome in *G. gallus* is gene found on?**

source 1..246

/organism="Gallus gallus"

/db\_xref="taxon:[9031](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9031)"

/chromosome="5"

/map="5"

Now you have a basic idea about NCBI. As you can see, you can easily get amino acid, DNA or RNA sequences of same functional protein/gene from all sorts of different organisms within a short amount of time. This is a powerful tool for research. For instance, you hypothesize that since the BDNF is known to play a vital role in the maintenance of neurons in parts of the brain associated with higher thinking, selection should then act to maintain this function throughout the evolution of vertebrates. And one way to test this hypothesis is evaluate the sequence of the BDNF protein to look for evidence of from various vertebrate taxa. Guess where you can find all various sequences within clicks? NCBI it is!

1. **BLAST it all!**

BLAST stands for Basic Local Alignment Search Tool, and though that’s a rather dry boring sounding name, it’s an incredibly useful and powerful tool (with an awesome acronym!). In the BLAST search, the query sequence is compared to every other sequence in the appropriate databases (i.e. nucleotide, protein, etc) and a list of the most similar sequences is generated. **Try running a BLAST search now by clicking “Run BLAST” in the “Analyze this Sequence” panel on the right of the NCBI page for the BDNF protein.** This opens the blastp tab for a standard protein BLAST. You can search using accession numbers or even raw sequence data and there are a number of parameters that you can control to relax or tighten the stringency of the search. **For our purposes, let’s leave these alone and just run the search by clicking the BLAST button on the bottom left of the screen**.

Scroll down to look at the list generated of similar sequences to your query from *Gallus gallus* and pay attention in particular to two columns: “Query cover” and “Ident” (query coverage and Max Identity. Query coverage tells you how much of the query sequence the resulting sequence actually includes and Max identity tells you the highest percent similarity between the query and result sequences. **Note the percentages for these values below as well as the accession number for the sequence you think is the best match of your query sequence (EXCLUDE chicken sequences for this question).**

**Accession number:** [XP\_012957404.1](http://www.ncbi.nlm.nih.gov/protein/874480682?report=genbank&log$=prottop&blast_rank=4&RID=M6J5NW9701R)

**Query coverage: 99%**

**Maximum identity: 513**

**What gene sequence most closely matches the gene we have selected for *Gallus gallus* in this alignment?**

[PREDICTED: brain-derived neurotrophic factor isoform X1 [Anas platyrhynchos]](http://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_874480682)

**Explain what it means if you have 20% query coverage but 100% for maximum identity. How certain would you be that these two sequences are from the same gene? What additional information might you need to ascertain this?**

100% max identity means that there are the same DNA sequences have the same amount matching, the matching means the possibility is high or low. 20% query coverage means the actual DNA sequences that do match.

A useful aspect of BLAST is that it lets us search with raw sequence data even if we know nothing about it. Run a blast search for the DNA sequence given below (use blastn tab in BLAST)

GATGTGGAGGTTGCACCTCCTAAGGCTTATGAAGTTCGCATTAAGATGGTGGCTGTAGGAATCTGTCGCA

CAGATGACCACGTGGTTAGTGGCAACCTGGTGACCCCCCTTCCTGTGATTTTAGGCCATGAGGCAGCCGG

CATCGTGGAGAGTGTTGGAGAAGGGGTGACTACAGTCAAACCAGGTGATAAAGTCATCCCGCTCTTTACT

**Based on the results, what gene and species do you think this sequence comes from?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Select seq ref|XM\_004040191.1| | [PREDICTED: Gorilla gorilla gorilla alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), mRNA](http://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_426345056) | 388 | 388 | 100% | 2e-104 | 100% | [XM\_004040191.1](http://www.ncbi.nlm.nih.gov/nucleotide/426345056?report=genbank&log$=nucltop&blast_rank=2&RID=M6JPX42M01R) |

1. **Sequence Data, Phylogenies, and Epidemiology**

Here, we will use DNA sequences of HIV from a true case in France to explore different tree construction methods. A patient was hospitalized for an operation. Prior to hospitalization, she was HIV-negative and had no risk factors typically associated with HIV. Shortly after hospitalization, however, she was found to be HIV-positive. Among the hospital staff, two of the nurses are HIV-positive and so one of them may have infected the patient. Our goal here is to use phylogenetic tree to determine which nurse, if either, is responsible for infecting the patient.

Even short DNA sequences can have hundreds of base pairs – that’s hundreds of potential characters! Rather than have you build these trees by hand, we’re going to use the computer.

We’re using a program called MEGA (http://www.megasoftware.net). Google or use the link to download the program to your own laptop, and then install it.

Step 1: We need to get the DNA sequences into MEGA. From the main MEGA window, select Align > Query Databanks. An NCBI webpage should jump out.

Step 2: Make sure the pull-down menu says “**Nucleotide**”. Then, search for the following accession code: **AF125604**. This should pull up the information for the HIV strain isolated from the patient. Click the “Add to Alignment” button with the red cross at the top of the screen. After you click “Add to Alignment” a window will pop up, rename the “Sequence Label” in the bottom field to help yourself keep the sequences straight.

Step 3: Repeat the previous step with the accession numbers for the two nurses (just keep searching the accession numbers in the same window as before). In addition, you need a sample of HIV sequences from the French population at large in order to test whether the patient acquired the disease outside the hospital.

Nurse 1: AF125605 Nurse 2: AF125606

Sample 4: AF125607 Sample 5: AF125608 Sample 6: AF125609

Sample 7: AF125610 Sample 8: AF125611

Step 4: Check the alignment window in MEGA to see that everything has imported properly. You should have 8 sequences, labeled with their respective individuals.

Step 5: In the toolbar, click the “W” icon. This will align the sequences so that there are the fewest incongruities between your sequences. Don’t worry about the settings at this point, they are set appropriately.

Step 6: Click the “Data” menu at the top left and select “Phylogenetic Analysis” In the main mega window select the button with the large “T A ..” this will open another window with additional information and options.

**How many base pairs are in the sequence we’re analyzing? Does that match the length of the sequences you downloaded from NCBI? Why might there be a different number of base pairs after you align the sequences?**

We are analyzing 498 base pairs in the patient. We downloaded up to 528 sequences in total and the aligned data is the same as the original data set. There is not a difference in the aligned data sequence base pairs and the original data set, but it could be different if there are mutations such as deletions or insertions.

**How many of them are conserved (click the “C” button on the menu bar and look for the proportion of the conserved sites on the bottom of the window)?**

There are 262 out of 528 conserved.

**How many of them are variable (click the “V” button) across the samples?**

There are 230 out of 528 variable.

**Building Phylogenies:**

**(1) Neighbor-joining (a distance-based method):**

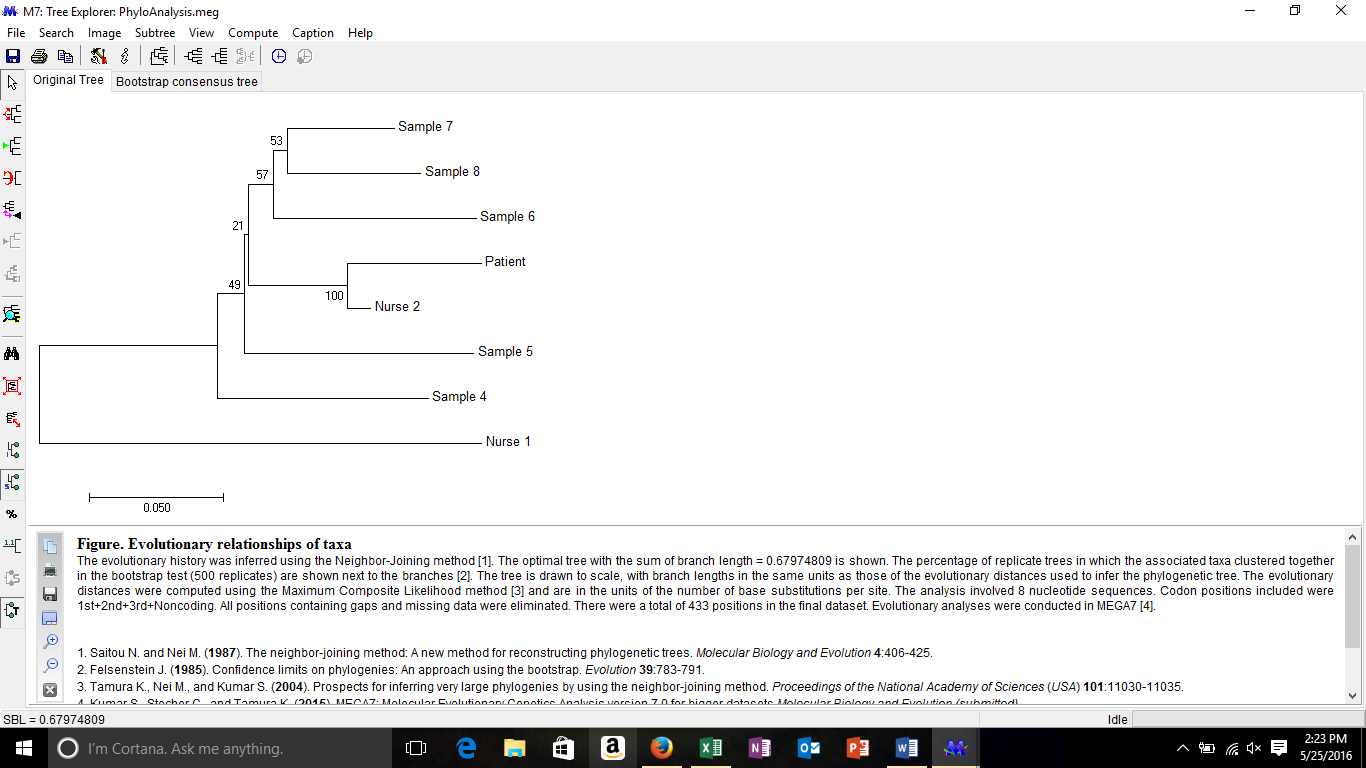
Distance methods are based on the idea that the best tree should minimize evolutionary distance between taxa. This is the fastest kind of tree to generate, as the computer only has to calculate the distances between each pair of taxa, rather than generating all of the character states for all characters over the whole tree.

The most commonly-used distance-based method is neighbor-joining (NJ). This is an algorithm for inferring a branching tree diagram from a distance matrix. It works by successively clustering pairs of taxa together. It is related to the UPGMA method; however, NJ allows for uneven branch lengths, and is trickier to do by hand. Since lineages are often evolving at slightly, or even dramatically different rates, NJ tends to give more informative trees.

Step 7: To select distance-based methods, in the main MEGA window, go to Phylogeny > Construct/Test Neighbor Joining Tree. In the “Analysis Preferences” window, select “Bootstrap Method” in the box to the right of the “Test of Phylogeny” option box. You can leave the rest of the setting at their default values. (We’ll come back to bootstrapping in a minute)

Take a minute to play around with the display tools – changing the tree type (slanting/square), rotating sub-trees around a node, etc. Convince yourself that nothing you are doing changes the relationships between the samples.

**In the “Tree Explorer” window, click on the “Image” menu and select “Copy to Clipboard” and paste your tree diagram here.**



Step 8: To prove that this tree really is based on distance methods, have MEGA calculate the distance matrix. From the main MEGA window, got to Distances > “Compute Pairwise Distances”, and then on the options screen, select “No. of Differences” from the menu to the right of the “Model/Method” option.

**At how many sites does the sequence of the patient’s HIV differ from that of the most closely related strain? How about between the patient’s strain and the next-most-closely related strain?**

The most closely related is Nurse 2 at 24 site differences. The patients strain next most closely related strain would be Sample 7 at 56 site differences.

Now, let’s see how robust our sequence data is, using a technique called “bootstrapping”. Bootstrapping is a way of generating multiple, similar data sets from our original sequence data, by randomly re-sampling our data for each replicate dataset. Each dataset is then used to create a new tree, and the different trees are compared for consistency. Branches that appear consistently in all replicate trees are given high levels of bootstrap support. 100 is the maximum, 50 is a reasonable cut-off for minimal levels of certainty. We already calculated our bootstrap values in step 7 so just go back to your tree and look at the values calculated for the nodes.

**After looking at the bootstrap values on the tree above, which groups are particularly well-supported (BS>85)? Which branches collapse to a polytomy (a node with more than 2 branches, or groups of poorly supported branches)?**

The group that is well supposed is the Patient and Nurse 2 with BS= 100. The branch group that is poorly supported is BS= 57 and BS=53 which is displaying Sample 6,7,8.

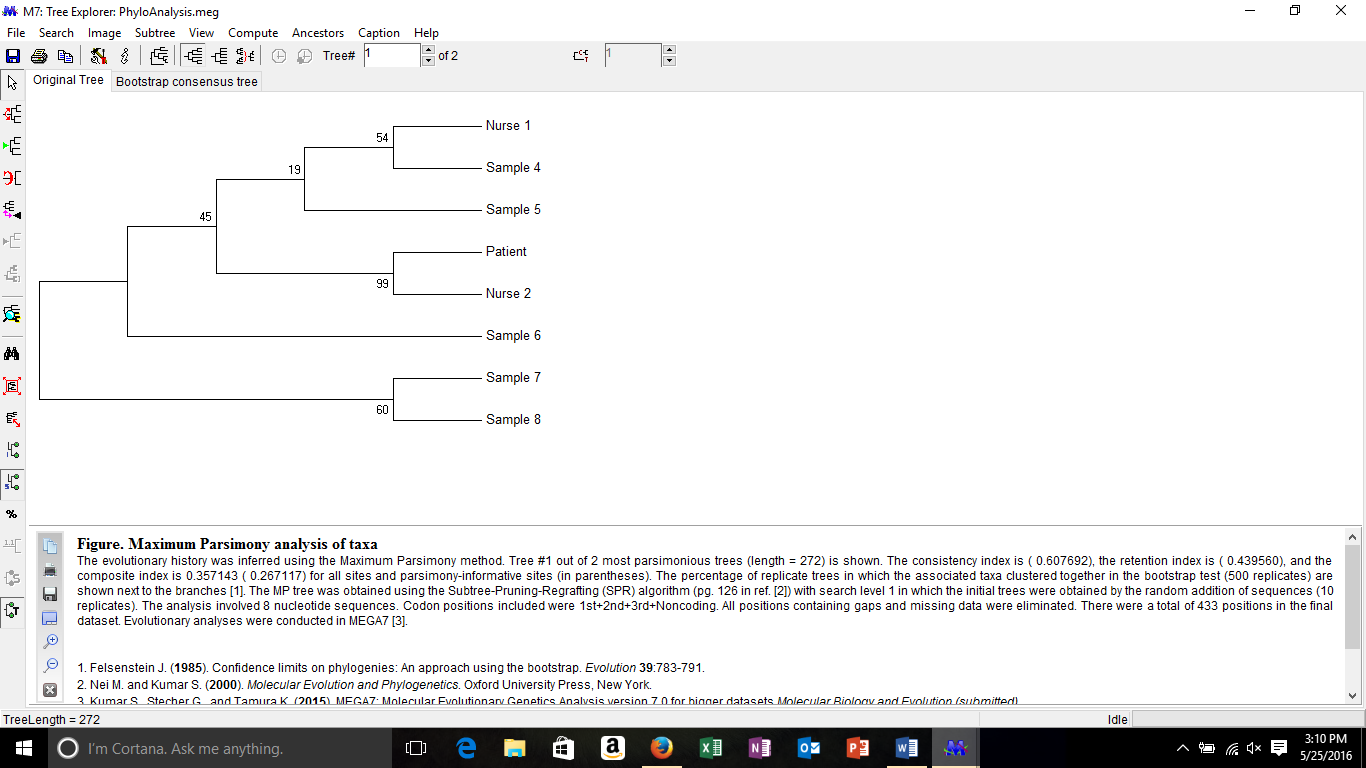
**(2) Maximum parsimony (MP)**

We have experienced this method last week. Instead of looking at distances between all of the individuals in our sample, it evaluates trees by determining how many mutations would have to occur for these individuals to evolve from a common ancestor. A “good” tree is assumed to require very few evolutionary steps, while poorer trees involve many more steps.

As the number of taxa increases, the number of potential trees increases to the point where even for a computer, it’s impractical to check every one, so a heuristic search is necessary for this dataset. Heuristic searches use various rules to determine which sets of trees are unlikely to include a good tree, and thus is able to eliminate many trees at once without having to look at all of them. It then focuses on searching through the other trees. This makes it possible to rapidly analyze even relatively large data sets.

Step 9: To select parsimony-based methods, in the main MEGA window, go to Phylogeny > Construct/Test Maximum Parsimony Tree(s). Again select “Bootstrap Method” in the box to the right of the “Test of Phylogeny” option box. This method is far more computationally intensive so it’ll take a bit longer. If your computer is having problems performing the calculations, you can lower the “No. of Bootstrap Replications”

**Again copy and paste your phylogeny here.**



MEGA won’t map characters onto the tree for us, but it will tell us something about how many characters we have. Go back to the “Sequence Data Explorer” window. The variable sites (“V”) come in two flavors, singletons (“S”) and Parsimony-Informative (“Pi”).

**What is the difference between Singleton sites and Parsimony-Informative sites? Why are PI sites useful for determining phylogentic relationships, while S sites are not?**

A site is parsimony-informative if it contains at least two types of nucleotides (or amino acids), and at least two of them occur with a minimum frequency of two.

A singleton site contains at least two types of nucleotides (or amino acids) with, at most, one occurring multiple times. *MEGA* identifies a site as a singleton site if at least three sequences contain unambiguous nucleotides or amino acids.

PI sites are useful for determine phylogenetic relationships you can compare many nucleotide relationships. With singleton sites it just shows one amino acid in one group, instead of comparing between different amino acids.

Step 10: Finally, let’s look at the bootstrap values for this tree that we calculated in step 9.

**Which groups are particularly well-supported (BS>85)? Which branches collapse to a polytomy?**

The group BS=99 between Patient and Nurse 2 are very well supported. BS= 19 can collapse in a polytomy, especially since BS is not greater than 85.

**Compare your two trees. Which relationships are the same regardless of the method of phylogenetic reconstruction? Which relationships change?**

It looks as though all the relationships stay the same. The differences are the BS numbers. They are similar but they are not exactly the same.

**Based on these trees, what can you say about the closest relative to the patient’s strain of HIV? Where (from whom) is the patient most likely to have acquired the infection (note that they told us at the beginning of this analysis that the patient did not have HIV when she was hospitalized)?**

Based on the trees I can tell that Nurse 2 sequences are closest to the Patients. They most likely acquired the infection from Nurse 2.

**Let’s say the patient is suing the hospital for damages related to their infection, and you’ve been called in to court as an expert witness. From the phylogenies we have, can you say for certain that the individual in question directly infected the patient? Is there an alternative explanation that would lead to a similar phylogeny?**

You could say that they are related both having HIV, but if they are not related then there is only one possibility for the Patient having HIV. There could be an alternative of somehow Nurse 2s blood contaminated the Patient without Nurse 2 knowledge.

**Obviously these phylogenetic methods are very useful for studying epidemiology and the movement of diseases. Where might you apply these methods today to better understand and combat public health issues?**

Yes you can understand how disease are spread from person to person and where they originate from.