Using Sequence to Determine Phylogenetic Trees

Now we will explore using sequences to search databases for similar sequences and to create a phylogenetic tree. The first thing we need to do is establish a good point of reference for determining inheritance. One of the most primitive cellular processes is that of translation of messenger RNA to protein product using ribosomes. Bacterial ribosomes, specifically the small 16S RNA component, have been used for years as a phylogenetic reference source. This is because this ribosomal RNA sequence contains highly conserved regions that have changed very little over millions of years.

Now what we need is a good organism to study. For this we will choose one of the completed genomes currently displayed on the TIGR web site. This can be found on the NCBI web site:  http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome

On this site:

1. Navigate to the list of bacteria genomes:
2. Scroll down to “Thermotoga maritima MSB8”

This genome was sequenced in 2001 and contains almost 2 million base pairs. Click on the link labeled: “NC_000853”.

Shortcut: Go to the NCBI web site. In the search box enter the accession id: “NC_000853”. Change the pull down menu next to this textbox to “Genome”. Click the “Search” button. This will bring up a result page that contains one hit. This is the link to the genome page for the Thermotoga maritima.

This page contains summary information of the genome. There is a lot of information, but for our purposes we will discuss only a few details. Under the “Genome Information” column there are links to two databases. The first is the RefSeq database. This is an archive generated from Genbank data to provide reference sequences of genes and genomes. It therefore represents a set of non-redundant sequences and can serve as a reference database for functional and diversity studies.

In the genomic dataset for Thermotoga maritima there are 1928 known genes of which 1858 are protein coding and 49 are structural RNA’s. Click on the link for these 49 RNA sequences. The resulting page shows a list of various tRNA’s and rRNA’s. The first Ribosomal RNA is 1559 nucleotides long. This is the length of the 16S unit of the Ribosome. The other rRNA’s do not come close to 1500. Under the “links” column on this page there is a yellow icon (diamond) that links to a text file of the sequence of nucleotides. The file format is called FASTA and is a universal file format used in bioinformatics for describing DNA, protein sequences.
The FASTA format always looks the same and is nothing more than a comment line in the first line of the file followed by the sequence (either peptide or nucleotides).

>`gi|15642775:188968-190526 Thermotoga maritima MSB8, complete genome
TATATGGAGGGTTTGATCCTGGCTCAGGGTGAACGCTGGCGGCGTGCCTAA etc… ..`

Sequence comparisons (BLAST)

**Shortcut:** Go to the NCBI web site: [http://www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov). In the search text box enter the ID “15642775” and in the pull-down menu select the “Nucleotide” option. Click the “Go” button.

Over time sequences change, or mutate, and it is this change that allows us to identify relationships between species. For example comparing this ribosomal gene to another distant relative will show slight changes. A sequence that doesn’t show changes over time is called a “highly conserved” sequence. Let’s see this in action:

Copy this sequence to your computer clipboard. (“Copy and paste” is a process that is heavily used in bioinformatics, so getting to know how to do this quickly is important).

1. Select the sequence of the FASTA file. In other words select everything after the comment line (the first line with the ‘>’ character).
3. Click on the “BLAST” link at the top of the page.

In the Nucleotide box click on the “blastn” link.

1. Paste the sequence into the text window labeled “Search”.
2. In the “database” pull-down menu select the “refseq_genomic” option
3. Click the “Blast” button.

**NOTE:** The search will take about a minute to three minutes to return a result. If you do not see anything after one minute click on the “Format” button.

We are searching for similar sequences in a database called refseq_genomic. This database was generated from the Genbank database and contains representative gene sequences that the community uses as a reference point. What this means is that we will be able to see diversity better than if we did a sequence similarity search on all available sequences on Genbank.
The blast results will have graphical, and interactive figure that provides a qualitative view of the similarity of each sequence. Since this is a search on refseq_genomic we should get results that show a different genomic sequence on each line.

Qualitatively, we can see the red lines are “conserved”, or stay the same across organisms. The thin lines show regions of variability. If we scroll down past this figure we can get a list of available sequences where we can view the actual alignments. If we scroll down and view the top BLAST hits we can see that the first one is “Thermotoga maritima MSB8”. This is what we expected since this is our query sequence. It’s an exact match to itself. As we scroll down the list we see the match becomes less and less. There are several ways of looking at this information but we will just look at the expectant value (e-value). Simply put this is the chance that the queried sequence will randomly get this hit. The e-value is a good metric because it takes into account the size of the database.

We can make the following assumptions about the blast e-value:

1. A score of $1 \times 10^{-5}$ or less is generally a highly unique sequence and not due to error.
2. A value greater than this means the hit might be strong but not entirely reliable and more research is necessary.

Blast searches can be run on DNA, RNA and protein sequences. NCBI offers several flavors of the program to search their databases. For example, blastp uses protein sequences to query a protein database, PSI-BLAST is used to find distant relatives of a protein, blastx compares the six-frame translated protein products with a protein database, tblastn compares a protein query against the six frame translations of a nucleotide database and megablast will run multiple input sequences against a database.

Now let’s draw a phylogenetic tree of the organisms in the blast results.

Under the image description of the results click the link that says, “Distance tree of results”.
The lines show relative distances of each organism based on the sequence comparisons. Our query sequence for Thermotoga is at the bottom. If we click on any of the nodes in the tree we can see a comparison of the sequence with our Thermotoga sequence:

```
ATGGAGCGTTTGATGCTACGGTGAAACGCTGCGCGTGCTCTAACACATGCAAGTC
```

The letters in the middle show mismatched regions.

Thermotoga maritime was originally isolated from a geothermal heated marine sediment in Italy and optimally grows at around 80°C (175°F). Much more work is needed to get a completely reliable tree, but for this purpose we can get a reasonable understanding of what could be an evolutionary route.

There are a couple of notable things in this tree. We can see deep in the lineage there are chloroplasts of plants and even a genome from a mitochondrion. Though this is not a rigorous test it does fall in line with the theory that chloroplasts and mitochondria evolved from bacteria.