Project 3: Mapping Protein Evolution

Due Date: 12:01am, May 3rd + 8 hours grace period

Project Goals:

The purpose of this project is to familiarize you with the algorithmic and experimental basis of phylogenetic analysis, and to demonstrate one of most fascinating aspects of phylogenetic applications of this method family: active site discovery. Evolution is not simply for reconstructing trees of life.

Background:

The theory of divergent evolution, that populations of nearly identical organisms can independently specialize under varying environmental conditions, is an observation with a very human perspective. Within such populations, the fittest individuals spread their genetic information from generation to generation because the unfit do not survive. This is true in part because multicellular organisms are complicated systems that cannot so easily adapt to a wide range of environments, but also because simpler organisms can even exchange genetic information in the same generation: the unfit can become the fit.

A monocellular Darwin would have seen a fundamentally different universe. Less complex organisms can exhibit considerable tolerance for environmental hardship, and populations of bacteria and viruses, for example, exhibit considerably more variation than populations of multicellular life. In fact, divergent evolution is not so accurate a model when members of a population can literally exchange DNA, as bacteria do when they exchange plasmids, or when members of a population require part of a host's genome to reproduce, and thus react differently in different hosts, as viruses do. A monocellular Darwin would have seen divergent evolution as a special case – a simplified case – of a far more complex tapestry of genetic information transfer.

It is surprising, then, that software designed with divergent evolution in mind remains so effective at certain tasks. One task in particular, the identification of functional sites in proteins based on the measurement of amino acid conservation, is surprisingly effective, even on viral and bacterial proteins.

Implementation Project: Compute a Phylogenetic Tree using UPGMA and Neighbor Joining

A phylogenetic tree is a critical tool in bioinformatics. Constructing phylogenetic trees is one of the only ways to shed light on the deep similarities and differences between the organisms we share this planet with. Among infectious agents, phylogenetic trees are also a fundamental tool for building classifications that can reveal potential vulnerabilities when developing new treatments.

More broadly, the construction of phylogenetic trees is an exercise in clustering, where similarities and variations in chaotic high dimensional data is exploited to create a comprehensible order. Clustering techniques are everywhere: internet search, computer vision, artificial intelligence, even robotics. A familiarity with clustering, as a class of methods, can open important doors in many industries.

Phylogenetic trees are binary trees that serve as hypothetical models of past evolution, built from data in the present. Present day sequences are represented by leaf nodes, while putative common ancestors are non-leaf nodes. As we will see in upcoming lectures, the topology of a parsimonious phylogenetic tree forms subtrees with high pairwise sequence identity, while separating sequences with low sequence identity.

Recommended Steps: (50% of the total project)

1) Parse the distance matrix for processing (5%)
2) Compute a UPGMA tree from the input data (40%)
3) Compute a Neighbor Joining tree from the input data (40%)
4) Output a text illustration of the updated tree (15%)
    a. The text illustration should be a topological representation of the tree's layout, and the edge lengths do not need to be represented.

The matrix of sequence identities is provided for you

Tools at your disposal: (see /proj/cse308/Project3)

1) ClustalW: we've been using this software for a while in class
2) PHYLIP: A suite of phylogenetic analysis tools, containing:
    a. neighbor: This software computes an unrooted neighbor-joining tree or a UPGMA tree, based on a matrix of sequence identity data. Thus, you can compare your results with PHYLIP's results, using any distance matrix.
3) Clustal2Matrix.py: a python script that parses the stdout from ClustalW and generates a distance matrix that is compatible with phylip.

Implementation Report: (50% of the total project)

1) Recidivism was a critical problem in the early treatment of HIV, until the development of Highly Active Anti-Retroviral Therapy (HAART), and remains a problem due to the prevalence of medicinal noncompliance. (40%)
    a. Imagine a hypothetical patient infected by one strain of genetically identical HIV capsids. The patient does not seek treatment for several years after the initial infection, and is never exposed to another strain of HIV. Following these years, if we could sequence the RNA genome in every HIV capsid in the patient, would we find fewer, more, or the same number of genetically identical HIV strains in this patient?
    b. Suppose it is the late 80s, and the patient is one of the fortunate few to receive a course of Zidovudine. What is Zidovudine? If we if we could sequence the RNA genome in every HIV capsid in the patient, how would the number of strains afflicting this patient differ from situation (a), above?
    c. Compare and contrast, approximately, a phylogenetic tree (there is no need to illustrate) of HIV genomes within this patient before and immediately after the treatment. What property do surviving virii have in common?

    d. Why does HAART apply several drugs at once? How would the phylogenetic tree of RNA genomes differ if the drugs were applied sequentially, rather than in parallel?
    e. If HAART is so effective, why is medicinal noncompliance a problem?

2) Proteins with very different sequences (e.g. 25% sequence identity) have been known to exhibit very similar "folds", or overall shape, while proteins with very different folds never exhibit very similar amino acid sequences. Abstractly, how might incorporating protein shape into the alignment of protein sequences, especially for very different sequences, yield additional information for reconstructing evolutionary history? (20%)
3) Given a phylogenetic tree (arranged vertically, so that leaves are at the bottom and the root is at the top) that relates a representative set of homologs that evolve at a similar rate, you can draw a horizontal line across the tree at any point in its height. Subtrees below a point of intersection conserve properties within the subtree that differ from those of other subtrees. If the line is drawn just below the root, fewer, though more complex, subtrees are found, while if the line is drawn just above the leaves, perhaps the only subtrees found are individual leaves. Wherever the line is drawn, a set of amino acids is conserved among the members of each subtree. (40%)
    a. What might be the functional significance of an amino acid that is conserved within the members of a subtree, but not conserved in the whole tree?
    b. The most conserved amino acids (e.g. those conserved throughout the tree, or those that vary very little between branches) have been observed to cluster spatially near active sites: regions on the molecular structure that are essential for some chemical function. Why might this be the case?

Applications Project: Identify a protein active site.

Proteins, the worker molecules of the cell, are under constant evolutionary pressure to perform their biological functions. These functions can include catalyzing essential chemical reactions, or transporting critical resources, or many other tasks that involve the specific recognition of other molecules. In almost all of these cases, certain amino acids are critical for enabling these functions, and if they are lost, so too is the function of the protein. These amino acids are said to reside at the active site of a protein.

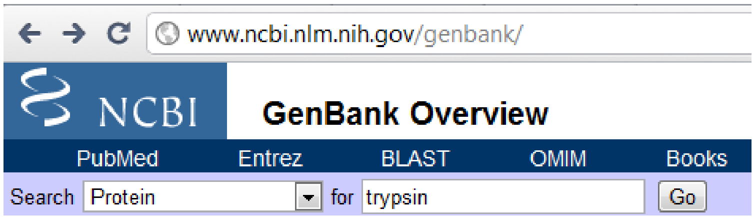
Evolutionary pressure on functionally significant amino acids at the active site is significant, and they are strongly conserved among all proteins that perform the same function. This effect can be observed in alignments of multiple protein sequences, such as those computed with ClustalW.

However, proteins selected from a nondiverse population of organisms have other similarities as well, because they have recent common ancestors. Here, functionally significant amino acids may be conserved, but they may also be indistinguishable from other amino acids that are simply conserved among the nondiverse population. For this reason, identifying functionally significant amino acids depends on truly diverse input populations.

One way to measure diversity among protein structures is by observing their phylogenetic tree. Parent nodes with short edges to their child nodes, indicate considerable similarity between the children; the opposite is true of nodes with long edges. By building a population with a large and balanced phylogenetic tree, the corresponding multiple sequence alignment will conserve the most amino acids essential for biological function, while avoiding incidentally conserved amino acids.

Tasks (50% of total project):

1) Gather a population of your designated proteins from as diverse a set of organisms as possible, using genBank:



www.ncbi.nlm.nih.gov/genbank/

Search Protein for trypsin

2) Compute a multiple sequence alignment of these proteins using ClustalW
3) Compute and visualize a phylogenetic tree describing these proteins, using protpars
4) Identify the most conserved amino acids, by studying your multiple sequence alignment
5) Using the same number of proteins from similar (rather than diverse) organisms, repeat steps 1-4.

Note: 1-4 may need to be repeated several times, while increasing the number of proteins, so that the most significant amino acids become visible. Only a few amino acids are truly conserved – less than 10 in many cases, less than 5 in most cases. If you see many amino acids conserved, then you need a more diverse set.

Also, there may not be such a large number of proteins that you can be 100% sure that each amino acids is fully conserved. One way to make sure that you have exactly the most important amino acids would be to consult the scientific literature on these proteins, to see which amino acids are functionally significant.

Tools at your disposal: (see /proj/cse308/Project3)

1) NCBI protein search:
    a. See the image above describing how to get here, noting that the search pulldown has been set to "protein". Put your protein name next to it, and hit go.
2) ClustalW: we've been using this software for a while in this class, now
3) PHYLIP: A suite of phylogenetic analysis tools, containing:
    a. protpars: This software attempts to compute a maximum parsimony tree based on a multiple sequence alignment.

Protpars computes trees with high parsimony. Use these trees to build a balanced representation of the space of available protein sequences homologous to your designated protein

4) Clustal2Matrix.py: a python script that parses the stdout from ClustalW and returns a distance matrix necessary for running PHYLIP.

Applications report questions (50% of total project)

1) Explain, in detail, how a UPGMA tree is computed, beginning by describing the input distance matrix and what it represents. While explanations do not require repetitive or trivial details, they should be detailed enough that a programmer could read your description and implement the method, without further questions. (40%)
2) Explain, in detail, how a neighbor joining tree is computed, and why it differs from a UPGMA tree, both in its basic topology and in its underlying assumptions about the rate of evolution. While explanations do not require repetitive or trivial details, they should be detailed enough that a programmer could read your description and implement the method, without further questions. (40%)
3) A UPGMA tree is not frequently used for phylogenetic analysis, but is frequently used as a guide tree for pairwise multiple sequence alignment. Describe how such a guide tree is used. Why is multidimensional dynamic programming not used instead? (20%)

Using Clustal2Matrix.py:

Clustal2Matrix.py is a simple script that takes the stdout generated by clustalw and uses it to generate a distance matrix based on sequence identities. This distance matrix is fed to the 'neighbor' program in the PHYLIP package. The command line to run this script is as follows:

```
python clustal2Matrix.py [clustalOutput] [matrixFile]
```

Where [clustalOutput] is the stdout generated by clustalW, and [matrixFile] is the distance matrix you want. You generate the clustalOutput file as follows:

```
./clustalw myFile.fasta > clustalOutput
```

Which automatically runs clustalW with the slow alignment and pipes the file (via ">" to the file called clustalOutput, which you can obviously name anything you want)

Values in the matrix are not exactly sequence identities but actually 1-(seqId/100), so that the distance at zero is equivalent to identical sequences (e.g. 100%) and as distance increases, sequence identity decreases, to zero.

Important note: for those doing the applications project that will require building a considerable multiple sequence alignment, note that the names that 'neighbor' accepts can have at most 10 characters (sorry, PHYLIP is very old software). Hence, my script will truncate your sequence names down to 10 characters. The names are passed on from the fasta header lines into the ClustalW output, into the matrix file generated by this script.

Groups Rosters:

	Roster	Proteins for Applications Project
1	Bicher, Romero	Trypsins
2	Chu, Gu	Chymotrypsins
3	Wu, Feinstein	P53
4	Shardt, Guo	Hemoglobin
5	Wallace, Dodd	BRCA1

Extra Credit:

Identify an unbiased set of sequences that represents the family of proteins assigned for the Applications Project in your group. This is typically achieved by identifying the largest possible set where no pair of sequences exhibit greater than 90% sequence identity. Compute a neighbor joining tree of this set. How does this tree compare or differ to a neighbor joining tree of a 50% sequence identity representative set? How does the set of available sequences affect the topology of your tree? Answer these questions in a separate section of your report.