

All rights reserved. This document may be reproduced for use at one's own not-for-profit institution, with the following statement included in the reproduction: Pytel BA, Hedrick J. Bioinformatics II: Introduction to Molecular Phylogenies. Center for Bioengineering and Computational Science, Union College, Schenectady, NY (<http://bioengineering.union.edu>). Funding provided by the Howard Hughes Medical Institute.

Bioinformatics II: Introduction to Molecular Phylogenies

Barbara Ann Pytel, Ph.D
Biology Department
Union College
Schenectady, NY 12308
518-388-6746
pytelb@union.edu

James Hedrick
Dept. of Electrical & Computer Engineering
Union College
Schenectady, NY 12308
518-388-8027
hedrickj@union.edu

Summary of Activity

Students studying Molecular and Cell Biology will learn the basics of constructing phylogenetic trees based on a data set composed of DNA sequences. They will be given an accession number from a sequence already present in **Genbank**, then use BLAST! to search government databases for similar sequences. They will then use ClustalX to align sequences from four closely related taxa and one putative “ancestral” taxon. The results of their alignment will be applied to the TreeViewX program and a hypothesis of relationships (phylogenetic tree) will be generated.

Integration of Disciplines

With the advent of high-throughput molecular methodology such as the polymerase chain reaction (PCR) and automated protein and DNA sequencing, the field of systematics has undergone a transformation. Once based on strictly morphological, behavioral, or developmental characteristics, phylogenies are now being based on molecular information in the form of protein or DNA sequences.

Many laboratories around the world are collecting molecular sequence data on organisms representing every taxonomic group. Most of these data are stored at NCBI (National Center for Biotechnology Information) or EMBL (European Molecular Biological Laboratory). Complex computer software is necessary to explore these data. This has led to the birth of the burgeoning field of **bioinformatics**, the goal of which is to uncover biological information hidden in the mass of protein and DNA databases. Integration of biology and computer science is necessary because the student needs to understand the biology behind the software’s assumptions **AND** the step-wise functioning of the particular algorithms.

Learning Objectives

Students will be introduced to the successive steps involved in constructing phylogenetic trees based on protein or nucleotide sequences.

A brief description of the BLAST! program explains the step-wise approach of a heuristic algorithm that compares their **query sequence** to the millions of sequences in **GenBank**, and does so in less than 30 seconds. This algorithm is similar to that used by the search engine **Google**.

Students will then use **ClustalX** to align the homologous sequences and construct a phylogenetic tree using **TreeViewX**. Students will probably have learned about protein and DNA “family trees” in high-school classes, but to really grasp the import of these hypotheses they must understand the methodology.

Target Level

Students should have had some experience with bioinformatics concepts in introductory biology classes or from the lab module Bioinformatics I: Introduction to Bioinformatics

(http://www.union.edu/academic_depts/bioengineering/docs/bioinfinst1.pdf). Most of the students we have worked with are majors in the natural sciences and many plan to attend medical or dental school.

Tools and Materials

Computers with access to the Internet are used to access resources on the web:

- **National Center for Biotechnology Information (NCBI)**
Comparison of DNA sequences to sequences in databases: **BLAST**
<http://www.ncbi.nlm.gov/blast>
- **Plate-Forme Bio-Informatique de Strasbourg**
For the latest version of ClustalX for aligning homologous sequences:
<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/>
- **TreeView X Home**
For the version of TreeView X to run on Mac OS 10.3
<http://darwin.zoology.gla.ac.uk/rpage/treeviewx/>

Programs for Clustal X and TreeView X can be loaded onto the computers before lab and placed as icons on the desktops.

Instructors might find it useful to read *Phylogenetic Trees Made Easy, A How-To Manual 2nd edition*, 2004, by Barry G. Hall.

Theory and Background

The first part of this exercise is to reacquaint students with the wealth of information that is available on government databases. A massive amount of sequence data (nucleotide and protein) from many organisms has been collected over the past few decades and it is freely available to anyone with an Internet connection. NCBI websites are beautifully designed, but very complex. Most of the pages contain much information and it is often difficult to know where to begin. It is a daunting task for a student to attempt understanding these pages without some guidance. Interpreting output is equally daunting and knowing the next step and how to take it requires some instruction.

Students have viewed phylogenetic trees in high-school classes. They have accepted the hypotheses associated with phylogenetic trees and the relationships that they infer without understanding many of the assumptions inherent in the process. This exercise gives the students an appreciation of the science of systematics from a molecular standpoint.

This exercise can be performed using any organisms for which an accession number is available.

Safety Precautions

None. The entire lab involves computers.

Miscellaneous Advice to Instructors

This exercise can be performed with any group of organisms.

The student handout can be found at

http://www.union.edu/academic_depts/bioengineering/docs/phylogstudnew.doc