

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: Willing P. and Hedrick J. 2005. Bioinformatics I: Introduction to Bioinformatics. Center for Bioengineering and Computational Biology, Union College, Schenectady, NY (<http://bioengineering.union.edu>). Funding provided by the Howard Hughes Medical Institute.

Title: Bioinformatics I: Introduction to Bioinformatics

Authors: Paul Willing, Ph.D.
Department of Biology
Union College
Schenectady, NY 12308
518-388-6713
willingp@union.edu

James Hedrick, Ph.D.
Department of Electrical and Computer Engineering
Union College
Schenectady, NY 12308
518-388-8027
hedrickj@union.edu

Summary of Activity:

Students will use instructional computer programs on the worldwide web (www) to review the genetic processes of replication, transcription, and translation. Students will then be introduced to the emerging field of bioinformatics in three steps: 1) the types and extent of molecular information that is stored in large databases; 2) the nature of the stored information and how it is transferred; and 3) some basic tools for searching, manipulating, and analyzing sequence information. Students will also explore how bioinformatics is used to study genetic diseases and genetic components of diseases.

Integration of Disciplines:

Biology is being transformed from a purely lab- or field-based science to a field that includes the use of computers to store and analyze large amounts of information, called **bioinformatics**. Scientists around the world are continually discovering more information about DNA and protein sequences; most of these data are stored in computer databases at the NCBI (National Center for Biotechnology Information; Bethesda, MD, USA). Bioinformatics involves the application of computer and information science to the management and analysis of DNA and protein data. Sophisticated software tools are necessary to

explore these data; thus, bioinformatics is an interdisciplinary field that integrates biology and computer science.

The next generation of biologists should be introduced to bioinformatics for several reasons. First, no matter what branch of biology a student ultimately pursues, it will be necessary that s/he develop basic skills in searching databases. Second, the student may find bioinformatics so interesting that s/he chooses to seek a career in that field. Third, students who appreciate the need for collaboration with experts in computational skills will ultimately become better scientists and better collaborators.

People with a wide variety of interests and training – biologists, physicians, students, teachers, and members of the general public – are interested in accessing databases containing information about DNA and protein sequences. The software programs used to access the data must therefore be flexible and user-friendly, and they often present the data in graphical or three-dimensional form to give the user a visual image of the structures. These requirements present challenges to the software designers.

The introduction to bioinformatics must be a joint effort between biologists and computer scientists. The role of the biologist is to teach the student about the processes of replication, transcription, and translation and the relationships among DNA, RNA, and proteins. The initial role of the computer scientist(s) is to explain about the storage of data and the efficient and rapid transfer of these data. In later courses, the computer scientist(s) will help students understand the software programs and the algorithms for these programs.

Learning Objectives:

Students will develop the following skills:

- A better understanding of the Central Dogma of molecular biology
- Understanding of the basic features of the genetic code: A, T, C, G; triplet codons; start and stop codons; reading frames; open reading frames
- Understanding the nature and extent of data that are now available in databases
- How to search the databases for specific information
- How to use electronic resources to study genetic diseases, evolutionary relationships, gene functions, DNA, forensic science, and others

Students will also learn that the information is only as good as the programs used to access and manipulate the data and that these programs

require collaboration between biologists and computer scientists. Finally, we hope our students will see that they can play an active part in contributing to data and its analysis.

Target Level:

This module has been developed for use by two different groups of students: introductory biology students and non-science majors. Introductory biology students are usually college freshmen. Most of these students have learned some of the concepts of molecular biology and some have been exposed to the idea of bioinformatics. More than half of these students are biology, chemistry, or other science majors, and many intend to apply to medical or graduate school.

Non-science majors usually have about the same level of experience as our introductory biology students, but are more typically sophomores, juniors, or seniors. They are generally less interested in biology and science in general than are the introductory biology students.

Tools and Materials:

The introduction to molecular biology, the Central Dogma, and the introduction to bioinformatics in this lab module are entirely computer-based. Computers with high-speed internet access are used to access resources on the www:

The Dolan DNA Learning Center, Cold Spring Harbor, NY, USA

- Sequence translator:
http://www.dnalc.org/bioinformatics/2003/2003_dnalc_nucleotide_analyzer.htm
- Program to find genes, or *open reading frames*:
<http://www.dnai.org/geneboy/index.html>
- Comparison of random DNA sequences to genes:
http://www.dnalc.org/bioinformatics/2003/2003_dnalc_nucleotide_analyzer.htm

National Center for Biotechnology Information (NCBI; Bethesda MD, USA)

- Comparison of DNA or protein sequences to sequences in the databases (BLAST)
<http://www.ncbi.nlm.nih.gov/blast/>
- Human genome resources:
<http://www.ncbi.nlm.nih.gov/genome/guide/human/>
- Databases for human genetic diseases:
<http://www.ncbi.nlm.nih.gov/disease/>

Theory and Background:

The first part of the exercise is to reinforce the concepts of molecular biology by using instructional software provided by the Dolan DNA Learning Center: <http://www.dnalc.org/home.html>. This site contains instructional material that will allow students to practice using the concepts of molecular biology to which they were introduced in their introductory textbooks and lectures.

Voluminous DNA data collected from many organisms over the last few decades, including human data from the Human Genome Project, are available at the NCBI sites free of charge to anyone with an internet connection. The NCBI web pages are wonderfully designed, but very complicated. Most of their pages are packed full of information and it would be daunting for a student to attempt to understand these pages without guidance. Our students are therefore introduced to just a few of the simpler types of information on these pages. We feel that nucleotide comparisons are the easiest type of data for the beginning student to understand. For more advanced students, the protein and other databases at NCBI are also useful.

Our first DNA example is a forensic science exercise. DNA, in the form of a text file listing the DNA bases (A, T, G, C), is compared to DNA in the databases to determine the likelihood that it is human DNA. Some of the other examples involve comparisons to other primate species, which gives the students an appreciation of how DNA analysis illustrates evolutionary relationships. (Evolutionary relationships based on DNA is the focus of another lab module, entitled ?? .) In the next exercise of this lab module, students work with text files containing the sequences for DNA segments that are involved in genetic diseases and in cancer. This exercise gives the students an appreciation for the role that bioinformatics has in medicine.

ITS User 12/6/05 11:55 AM

Comment: Barb Pytels?

References:

Reading material about the Human Genome Project:

<http://www.genome.gov/10001772>

Francis Collins was head of the publicly funded portion of the Human Genome Project. Below is a link to a paper by Francis Collins, et al., that states the significance of this accomplishment, as well as the probable directions of future research.

<http://www.genome.gov/11007524>

Beekman, George. *Computer Confluence*, Upper Saddle River New Jersey: Prentice Hall, 2001.

Krane, Dan E., Raymer Michael L. *Fundamental Concepts of Bioinformatics*, San Francisco: Benjamin Cummings, 2002.

Safety Precautions:

None. The only equipment required for this lab exercise is a computer with internet access.

Miscellaneous Advice to Instructors:

The student handouts are provided in a separate document, http://www.union.edu/academic_depts/bioengineering/docs/bioinfstudent.pdf, which is a MSWord document outlining six separate exercises; the first four exercises can be completed in class, while the last two can be assigned as homework. Students should read the background material, go to the first link, do the first exercise, and type their answers into the document. They should then proceed to the next exercise, and so on. When finished, students should save the file under their name; this is their completed in-class assignment.

Extensions and Options:

After the two-to-three-hour lab module is completed, students can be given additional assignments to do on their own. For example, we gave students the entire reverse-transcribed DNA sequence from HIV as an “unknown DNA” text sequence (Exercise 5). By accessing the Dolan DNA Learning Center and NCBI sites, students were able to determine that the sequence belonged to HIV and also how many genes it contained. Another assignment could be to give the students a sequence(s) involved in a genetic disease(s). Students can be asked to determine what chromosome the DNA is from, what disease(s) it is involved with, and the nature of the mutation in that disease (Exercise 6).

If you have trouble accessing any of the documents for this module please contact christel@union.edu .