

## Lab 3A: Bioinformatics, part I

Computer databases of genetic information are available to researchers and students on the internet. These data require sophisticated search and pattern recognition algorithms to present and interpret them. The data and analysis software are our paths to understanding the huge and increasing amount of sequence and other data generated by genome sequencing and other large scale biology efforts.

### Objectives

- 1) Practice using the resources at NCBI and other public databases.
- 2) Understand database accessions and the tools to compare them, and,
- 3) Obtain useful answers to questions using database query tools.

### Lab Questions

How can basic tools for database search and comparison be used to explore questions relating to genetics, molecular biology, evolution, and biochemistry?

### Introduction

NCBI, the National Center for Biotechnology Information was established in 1988 as a repository for the growing amount of sequence and other data about genes and proteins. Although NCBI is a multifaceted organization, most scientists interact with NCBI largely through remote access to databases. Figure 1 shows a graph of the content of GenBank, one of the NCBI databases, over time. In this figure, each sequence represents one GenBank entry or **accession**. Together, the NCBI databases are one of the cornerstones of modern genetics and molecular biology. The textbook tutorials

you have completed over the last several weeks have given you an introduction to the resources available at NCBI.

This week, our interaction with the NCBI databases will be based on practicing information retrieval, understanding the data we access, and comparing sequence data with each other. Remember that evolution (descent with modification) occurs at the genetic level. Sequences that have diverged from a common ancestor retain evidence of that ancestor in their sequence, with the degree of divergence reflecting (in a complex way) the time that has passed since the most recent common ancestor. In cases of duplication and divergence within a genome, the ancestral gene itself can be considered the ancestor, and the modern derivatives its progeny. These ideas will be important as you investigate families of genes and proteins.

Growth of GenBank

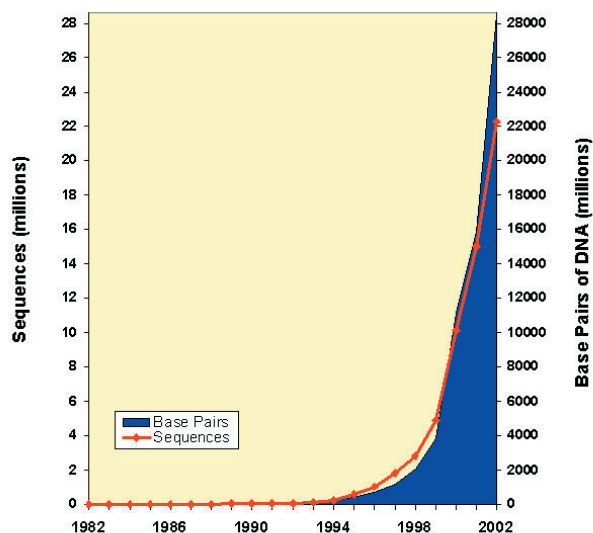


Figure 1, Growth of GenBank sequence data.

You may complete this exercise and Part II with a lab partner if you would like. No one should work in a group of more than two people. Turn in assignments from part I at the end of the class period. Then, each student should write his or her own lab report for the assignment in part II. Your lab report is due Wednesday 24 March in lecture.

## Procedure

Go to the NCBI web site,

<http://www.ncbi.nih.gov/>

and use the gene and nucleotide databases to answer the following questions. Hand in your paper at the end of class. This exercise may take longer than the allotted class time, especially if you are unfamiliar with the NCBI tools.

1. Find the sequence of the region of the *Drosophila* genome where the *reaper* gene, genetic control of programmed cell death, is located. Without printing out the entire sequence, answer the following questions on a paper to hand in or on the screen:

a) what does “complement (47242..47439)” mean in this context?

b) highlight the reaper coding sequence and point out the start codon.

Get your instructor’s signature when you are finished.

---

2. Find a 4233 base pair genomic accession containing the Human CDK4 gene. Print out the sequence, and neatly annotate it to show all exons, splice site consensus sequences, and the start and stop codons. Then find a 45998 base pair accession containing this gene and its neighboring sequences. On your paper to hand in, make a to-scale map of this region showing all relevant information about these genes. Think carefully before you begin this map so you do not have to start over.

3. Find a 2323 base pair genomic accession containing the *E. coli ilvIH* operon. On your paper to hand in, indicate the accession number, the authors of the sequence, and the enzyme(s) encoded by *ilvI* and *ilvH*. Use the definition of an operon to explain the probable role of the sequence from 1765 - 1787 in this sequence.

4. Find a 5175 base pair accession containing the *Saccharomyces cerevisiae* PPS1 gene (previously known as ORF YBR276c). Use the *Saccharomyces* Genome Database link in the accession to find out more about this gene and protein. What is the phenotype of a strain that overexpresses this protein?

5. What is the role of the *Saccharomyces cerevisiae* STE7 gene and protein? What is the evidence linking STE7p to the regulation of pseudohyphal growth? What is the accession number for the closest human homolog of STE7. What is the p value for this match? What does this p value mean?

6. Create a Word document saved to the desktop of your group workspace named “B331L3 xxxx” where xxxx are the initials of two people in your lab group. Paste the sequences for yeast STE7, and the closest homologs from human, mouse, *Drosophila*, and *C. elegans*. Save the file. Go to the Baylor ClustalW site:

<http://searchlauncher.bcm.tmc.edu/multi-align/Options/clustalw.html>

Format your data in the Word file as requested by the ReadSeq Format link (hint - look at the example in FastA format) and submit it for an alignment. Print this data and use it to answer the following questions: Which two sequences are most closely related? Which is most distant from the others? What is the meaning of these relationships? Which parts of the protein sequence are likely to be critical for catalytic function?

7. There is no number 7!