Bioinformatics: The Retrieval and Analysis of DNA and Protein Sequences

Tongjia Yin and Janice Lovett

State University of New York College of Arts and Science
1 College Circle
Geneseo, New York 14454

The exercise is designed for introductory students with only basic backgrounds in computer skills and molecular genetics. The objective of the laboratory exercise is 1) to demonstrate and reinforce major concepts and principles of DNA and protein sequences, 2) to teach the skills of using Web browsers and the availability of biological resources on the Internet, 3) to use a simple DNA analysis program, and 4) to recognize differences between prokaryotic and eukaryotic gene organization that will affect the use of such resources and software.

The concepts of the Internet, World Wide Web, and Web browsers are introduced by having students use a Web browser, NetScape, to search for information in the WWW and to access one of the many sequence databases, GenBank at NIH. The students are initially walked through a carefully scripted step-by-step procedure which includes images of the computer screen as they see it. During the first part of the exercise, students access a particular sequence and learn the format and related information of the sequence entries. Sequences are then downloaded or copied to a local machine for analysis using the Strider software.

For DNA sequences the analysis includes: searching for particular sequences, such as a start codon, and finding any ORFs (open reading frames), codon usage, and the anti-parallel, reverse and complementary sequences. The DNA sequences can be converted to RNA sequences and protein sequences to demonstrate their relationship. Using the software the students generate full restriction endonuclease (RE) maps, assess the number of cleavage sites from all or selected sets of a RE library, identify RE with unique or no sites and calculate the probability of sites in a random sequence.

The analysis of the corresponding proteins includes obtaining the molecular weight of the protein, the most frequent amino acid, the total number of occurrences of an amino acid and its percentage, comparing the most frequent codon and amino acid and explaining any discrepancies. From the primary sequence of the protein, the basic, acidic and hydrophobic regions are identified and their biological significance is discussed.

This format was developed to integrate the student's exposure to the Internet and analysis software with their knowledge of molecular biology at the introductory level. The use of the software allows several nuances of genes and prokaryotic vs. eukaryotic gene organization to be revealed while using the functions of the software. For example, the search for an ORF emphasizes the start and stop codons but also the importance of the reading frame, or when choosing a sequence for analysis the importance of introns in eukaryotic sequences is considered.

The laboratory write-up itself and the DNA analysis software are available from J. Lovett (lovett@uno.cc.geneseo.edu) in Macintosh compatible versions by e-mail or by sending a disk to Dr. Janice Lovett, Biology Department, SUNY College, 1 College Circle, Geneseo, NY 14454.