Introduction

Bioinformatics is an applied science that uses computer programs to access molecular biology databanks to make inferences about the information contained within the data archives. This lab introduces you to some of the main databanks, their applications, and programs. You will learn how to retrieve information from the databases, and analyse the information to obtain useful knowledge about a DNA sequence and its protein product.

There are two components to this lab. In the first part (Part I), you will search some of the most established public biological databanks using a query DNA sequence. These databanks provide a variety of analysis programs and are accessible without charge on the Web. They represent only a small number of the many sources of information available; many other programs, applications and databanks are available on the Web and programs designed for specific analysis can be purchased commercially.

Also in Part I, you will examine the predicted amino acid sequence from the nucleotide base sequence provided. You will use this amino acid sequence to conduct a search for similar amino acid sequences in the database.

The second component (Part II) of this lab is to utilize sequence data from your bacterial DNA PCR products to identify the bacteria from which you obtained this DNA. To carry out this activity you will conduct a BLAST search on the Internet and identify the sequence that most closely matches your sequence.

We will do Part I using computers together in class. For Part II, you will work independently.

DATABANKS

Nucleic acids

The archive of nucleic acid sequences is a worldwide database collaboration developed and maintained by a partnership of The National Center for Biotechnology Information (NCBI), a component of the United States National Library of Medicine, the European Bioinformatics Institute in the UK (EMBL), and the DNA Data Bank of Japan (DDBJ), a component of the National Institute of Genetics, Japan. These three groups exchange data daily so each databank has the same raw data. However, storage and annotation of the data in the individual databases varies slightly.

The gi|number| is the identifier number (geninfo number) for a DNA sequence in the database of nucleic acid sequences. An abbreviation of the databank the sequence was originally lodged with follows the gi number. The ENTREZ databank of the NCBI is identified by the letters |gb|, for GenBank, the European Bioinformatics Institute by |emb|, and the DNA Data Bank of Japan by |dbj|.

The URL (Uniform Resource Locator) or website of ENTREZ, the NCBI databases which you will use in this lab, is http://www.ncbi.nlm.nih.gov/sites/gquery.

Proteins

Similar to DNA, protein sequences are lodged in protein databases, which are identified by a databank abbreviation. Some of the main protein databanks and their abbreviations are: SWISS-PROT |sp|, in collaboration with EMBL |emb|, NCBI protein database |gb|, Protein Identification Resource |pir|, of the National Biomedical Research Foundation of Georgetown University in Washington DC, and the Munich Information Center for Protein Sequences (MIPS), Munich, Germany. An accession number, a unique identifier given to the sequence in order to distinguish it from other similar sequences in the database, follows the database abbreviation.

Other derived databases such as PROSITE and ENZYME DB and PDB offer information on more specific areas such as sequence motifs, enzymes and structures.

Procedures and Assignment

I. Using an Example Sequence (Full Gene) to Do BLAST Searches

You will use the WCC Natural Science Department laptop computers to carry out this assignment. We will work through these steps together, each of you working individually with a computer.

1. Open one new Microsoft Word document. Save it as Bioinfo_Assignment(your last name).doc. Be sure to save this file frequently during the exercise.
2. Using an Internet browser to navigate to the following URL:
   http://www.wcc.hawaii.edu/paces/summerfiles/DNAsequences/

3. Download the “example1.txt” sequence text file to your desktop.

4. Copy and paste the DNA sequence in this file into your Bioinfo Assignment Word document under the heading Requirement 1. Change the font of the sequence to "Courier New". Font size should be 9.

   Notice that the first column of the sequence begins with “>” followed by the gi number. The “>” indicates that the sequence is in FASTA format and that the information on this line is not part of the sequence. This line usually contains the data base identifier and annotation that provides information about the sequence. All but the gi number has been deleted from this line: >gi|142814|.

**NCBI BLASTn**

BLASTn is a search in which a nucleotide query sequence is compared with the contents of a nucleotide sequence database to find regions of sequences similar to regions of the original (query) sequence. Matching sequences are called BLAST hits. Database query: "given this DNA sequence, what sequences in the database are similar?"

1. Use your Internet Browser to go to the NCBI BLAST (Basic Local Alignment Search Tool) page at http://www.ncbi.nlm.nih.gov/ and select the BLAST link under "Popular Resources.

2. Under the “Basic BLAST” header choose the “nucleotide blast” link.

3. After the Nucleotide BLAST window opens, paste a copy of your FASTA DNA sequence into the “Enter Query Sequence” search box (be sure the blastn tab is selected).

4. Under “Choose Search Set” choose “Others” and “Nucleotide collection (nr/nt).”

5. Under “Program Selection” choose “Somewhat similar sequences (blastn).”

6. Then click the “BLAST” button.

   Your request will be placed in the BLAST queue and given a request ID that can be used later to retrieve the data. Wait until the results page appears.
7. Click on the “Citation” link which provides the specific bibliographic details for the paper in which the BLAST program is published. Copy and paste this citation into your Bioinfo Assignment Word document under the heading: References. The reference section should later be moved to the end of the assignment.

8. Click on the “Search Summary” link to reveal a report on the search itself. Refer to this summary report to answer some of the questions below.

9. Throughout this lab you will be proposed with sets of questions (italicized). Answer these questions in your Bioinfo Assignment Word document under the designated headings.

**Blastn Questions:**

a. What databases were searched for your query? (See “Description” under “Database Name”)
b. What databases were not searched?
c. What was the query length of the DNA sequence searched for?
d. How many DNA bases, in the NCBI database, were compared for the query?
e. How many sequences were examined from the database?
f. How many BLAST hits were obtained in the search?
g. What is a BLAST hit?

10. Scroll down to the sequences producing significant alignments section (under the “Descriptions” header). Using the Macintosh utility Grab, select, copy and paste the title lines of sequences with an E-value of zero into your Bioinfo Assignment Word document under the heading Requirement 2.

The score (bits) is an evaluation of the sequence alignment between the query sequence and the hit sequence. To determine whether the alignment is significant, the BLAST program applies a scoring matrix to describe the probability of a pair occurring in alignment. The score is the sum of the logarithms of the ratio of the probability that a base has occurred as the result of an evolutionary event and the probability that it has occurred randomly for each pair in the matrix. A higher positive score indicates more significant alignment. The entries are sorted in decreasing order of statistical significance.

The E-value (expectation value) is the probability that the observed degree of similarity (score) could have arisen by chance. An E-value equal to zero indicates the sequences are identical or nearly identical. For the purposes of this lab, an E-value greater than 0 but less than or equal to 0.05 will be considered a significant hit. Depending on the sequence, researchers will consider E-values up to 0.1 as significant.

11. Within the identical hits (E-value equal to zero) is your query sequence. Referring to Requirement 1 in your Word Document, identify and click on the gb number (Accession column) in the Blast Window for the query sequence. By clicking on this gb number, you will open the Genbank file for the sequence.

12. Under the heading Requirement 3, record the gi, the accession number, source organism, and PubMed reference number. Also include the name of the gene and protein product, this information can be found under features. If any of the information is not available for your unknown sequence indicate it by writing N/A.

13. Copy and paste the complete reference information including the names of the authors, title of the submission and journal into the Reference section of your Bioinfo_Assignment Word document.

Referencing guidelines differ depending on the publishing journal. You can find the accepted format for any journal by going to the journal website, and clicking on Instructions for Authors. For our purposes use the format below:


**ORF Finder: Open Reading Frame Search**

Regions of DNA that encode proteins are first transcribed into mRNA and then translated into protein. In translation of mRNA into protein, codons of three nucleotides translate into an amino acid sequentially from the start until a stop signal is reached. In most species, the open reading frame starts with an ATG, which codes for the amino acid methionine (Met) and ends with a stop codon (TAA, TAG, or TGA).

To determine the amino acid sequence of the protein directly from the DNA sequence the nucleotide that starts the translation and the stop signal must be determined. This sequence is the open reading frame (ORF). There are six possible reading frames for each gene: three reading frames in the forward direction (5’–3’) and three reading frames in the reverse direction (3’–5’). The reading frame that results in the longest protein is usually the ORF that encodes the expressed protein.

Examples of ORFs for a DNA sequence are presented below. Frame 1 results in the longest protein and therefore is the ORF.

**Examples of Reading Frames:**

There are three reading frames in the 5’–3’ direction for this sequence:

ATGCCCCAAGCTGAATAGCGTAGAGGGGTTCATCATTTGAGGACGATGTATAA

Frame 1 begins with “A”

**ATG** CCC AAG CTG AAT AGC GTA GAG GGG TTT TCA TCA TTT GAG GAC GAT GTA **TAA**

M P K L N S V E G F S F E D D V STOP

Frame 2 begins with “T”

**TGC** CCA AGC **TGA** ATA GCG TAG AGG GGT TTT CAT CAT TTG AGG ACG ATG TAT

C P S STOP I A STOP R G F H H L R T M Y

Frame 3 begins with “g”

**GCC** CAA GCT GAA **TAG** CGT AGA GGG GTT TTC ATC ATT **TGA** GGA CGA TGT ATA

A Q A E STOP R R G V F I I STOP G R C I

2. Copy and paste your DNA sequence in FASTA format into the search box. Click on OrfFind. The reading frames and possible sequences will be displayed in turquoise.

ORF Finder Questions:

a. What does the acronym ORF stand for?

b. What are you asking the ORF Finder program to do?

c. How many ORFs were predicted for the given sequence?

d. What frame produces the longest ORF?

e. How many nucleotides long is the longest ORF?

3. Click on the longest ORF. The ORF colour will change from turquoise to pink. In addition, you will observe this reading frame plus its corresponding amino acid sequence below the graphics display.

ORF Finder Questions (cont):

f. How many amino acids long is this ORF?

4. Click on the “Accept” button. The ORF changes from pink to green.

5. Choose “3 FASTA protein” view in the options pull-down menu and press “View.” Copy and paste the protein sequence in FASTA format into your Bioinfo_Assignment Word document under the Requirement 4 heading.

The output from this search will be the input for the next search.

BLASTp

Blastp results are a one pair-wise alignment between the query sequence and sequences in the database. The results are presented in a manner similar to a BLASTn query. The gi number, accession number, name of protein, score and E-value for the alignment is given. Some of the hit sequences may have the letter “S” in the right hand column indicating that a structure associated with the sequence is on file in the Protein Data Bank (pdb).

1. Return to the NCBI BLAST homepage (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and click on the “protein blast” link.

2. Paste your FASTA format protein sequence into the “Enter Query Sequence” search box.

3. For “Database” under the “Choose Search Set” header select “Non-redundant protein sequence (nr).”

4. Under the “Program Selection” header, select the “blastp (protein-protein BLAST)” algorithm.

5. Click the “BLAST” button.

6. After the blastp results window opens, scroll down to view matched amino acid sequences.
7. **Requirement 5.** Using Grab, select, copy and paste all identical (E-value equal to zero) and significant hits (0 < E-value < 0.05) into your Bioinfo_Assignment Word document.

*Blastp Questions:*

a. *How many hits did you receive on this query?*

b. *What is the gi number of the last protein considered significant in this query?*

II. **BLAST Search of Your PCR DNA Product.**

For this activity you will work on your own to identify the organism in the database that is most similar to the bacteria from which you extracted your DNA. To do this, you will be presented with the sequence data for your PCR product. Remember the primers used targeted a 500 bp region of the gene that codes for the 16S ribosomal subunit.

1. Download your DNA sequence for the DNA sequence website onto your desktop:

   http://www.wcc.hawaii.edu/paces/summerfiles/DNAsequences/

2. Open FinchTV and drag this sequence file (has “ab1” as an extension) from the desktop into the FinchTV window. Note, FinchTV is downloadable from http://www.geospiza.com/finchtv/ as "freeware."

3. Following the instructions presented in class, edit this sequence to obtain the region with high-quality sequence.

4. Export this edited sequence as a text file (FASTA format).

5. Copy and paste this edited sequence into your Bioinfo_Assignment Word document under **Requirement 6.**

6. Also copy and paste this edited sequence into the BLASTn search box (see directions in Part I above).

7. Conduct your BLAST search for sequences that most closely match yours.

8. Select the sequence that most closely resembles yours by clicking on the gi number.

9. Set the display for “Genbank.”

10. **Requirement 7.** Copy (from “LOCUS” to “FEATURE” only) and paste the information presented into your Bioinfo_Assignment Word document. If necessary, adjust the font of the pasted text as done previously.

**Laboratory Summary**

Submit your Bioinfo_Assignment Word document divided appropriately into its two parts (include headings to identify these):

- Include each of the requirements, with a brief title for each these.
- Answered questions, with headings.