Esercitazione 31-05-2011 laboratorio Algoritmi e Linguaggi per Bioinformatica: Matlab – Bioinformatics Toolbox

Reference: Bioinformatics toolbox 3 user guide:
http://www.mathworks.com/help/pdf_doc/bioinfo/bioinfo_ug.pdf
Solutions to the following exercises can be found at page 2.2.

Exercise 1: Sequence Tool (solution at page 2.42 of the user guide)

The Sequence Tool window integrates many of the sequence functions in the toolbox. Instead of entering commands in the MATLAB Command Window, you can select and enter options.

a. Importing a Sequence

The first step when analyzing a nucleotide or amino acid sequence is to import sequence information into the MATLAB environment. The Sequence Tool can connect to Web databases such as NCBI and EMBL and read information into the MATLAB environment.

The following procedure illustrates how to retrieve sequence information from the NCBI database on the Web. This example uses the GenBank accession number NM_000520, which is the human gene HEXA that is associated with Tay-Sachs disease.

1. In the MATLAB Command Window, type *seqtool*. The Sequence Tool window opens.

2. To retrieve a sequence from the NCBI database, select **File > Download Sequence from > NCBI**.

3. In the Enter Sequence box, type an accession number for an NCBI database entry, for example, NM_000520. Click the Nucleotide option button, and then click OK.

b. Viewing Nucleotide Sequence Information

After you import a sequence into the Sequence Tool window, you can read information stored with the sequence, or you can view graphic representations for ORFs and CDSs.

1. In the left pane tree, click *Comments*. The right pane displays general information about the sequence.

2. Now click Features. The right pane displays NCBI feature information, including index numbers for a gene and any CDS sequences.

3. Click ORF to show the search results for ORFs in the six reading frames.

4. Click Annotated CDS to show the protein coding part of a nucleotide sequence.

c. Searching for Words

The following procedure illustrates how to search for characteristic words and sequence patterns. You will search for sequence patterns like the TATAA box and patterns for specific restriction enzymes.

1. Select Sequence > Find Word.
2. In the Find Word dialog box, type a sequence word or pattern, for example, atg, and then click Find.

d. Exploring Open Reading Frames

The following procedure illustrates how to identify the protein coding part of a nucleotide sequence and copy it into a new view. Identifying coding sections of a nucleotide sequence is a common bioinformatics task. After locating the coding part of a sequence, you can copy it to a new view, translate it to an amino acid sequence, and continue with your analysis.

1. In the left pane, click ORF.

2. Click the longest ORF on reading frame 2.

3. Right-click the selected ORF and then select Export to Workspace. In the Export to MATLAB Workspace dialog box, type a variable name, for example NM_000520_ORF_2, then click Export.

4. Select File > Import from Workspace. Type the name of a variable with an exported ORF, for example, NM_000520_ORF_2, and then click Import.

5. In the left pane, click Full Translation. Select Display > Amino Acid Residue Display > One Letter Code.

e. Viewing Amino Acid Sequence Statistics

The following procedure illustrates how to view an amino acid sequence for an ORF located in a nucleotide sequence. You can import your own amino acid sequence, or you can get a protein sequence from the GenBank database. This example uses the GenBank accession number NP_000511.1, which is the alpha subunit for a human enzyme associated with Tay-Sachs disease.

1. Select File > Download Sequence from > NCBI.

2. In the Enter Sequence box, type an accession number for an NCBI database entry, for example, NP_000511.1. Click the Protein option button, and then click OK.

3. Select Display > Amino Acid Color Scheme, and then select Charge, Function, Hydrophobicity, Structure, or Taylor. For example, select Function. The display colors change to highlight charge information about the amino acid residues. The following table shows color legends for the amino acid color schemes.