CHEM 436 / CHEM 630: Molecular Modelling of Proteins
TUTORIAL #1a: Protein sequence alignment

INTRODUCTION

Tutorial #1 is divided into three parts:

- Tutorial #1a: BLAST protein sequence alignment (this document)
- Tutorial #1b: Comparison and validation of alignments
- Tutorial #1c: Protein visualization

It will be done mostly using the BLAST web server as the National Center for Biotechnology Information (http://blast.ncbi.nlm.nih.gov), but will also require you to explore the Protein Data Bank (http://www.rcsb.org/pdb) and the primary scientific literature. Please note that, while the BLAST server and the PDB are publicly available, some of the journal websites will be accessible only through the university network.

The results and discussion for all three parts should be combined into a single lab report. Note that points marked with the symbol ✦ should be specifically addressed in the lab report.

REQUIRED PRE-LAB READING

Alexander Pertsemlidis and John W. Fondon III, Having a BLAST with bioinformatics (and avoiding BLASTphemy), *Genome Biology* 2001, **2** (10), 2002.1–2002.10.
http://dx.doi.org/10.1186/gb-2001-2-10-reviews2002

PRE-LAB REPORT

Create a PDF document containing the following:

- A reference chart showing, on a single page, the chemical structures of the 20 standard amino acids, grouped according to their chemical properties. Indicate their three-letter codes and one-letter codes. Provide the reference of any material obtained from the internet.
- On a separate page, show the 20 × 20 version of the BLOSUM62 matrix (describing only substitutions of one specific standard amino acid by another) and identify the matrix elements corresponding to the most likely and least likely substitutions for amino acid “Asp”.

READING

On databases:
Chapter 1 of Tramontano (“The Data: Storage and Retrieval”): All sections.
Chapter 3 of Zvelebil & Baum (“Dealing with Databases”): All sections.

On sequence alignment:
Chapter 3 of Tramontano (“Protein Evolution”): Sections 3.1 to 3.6.
Chapter 4 of Tramontano (“Similarity Searches in Databases”): All sections.
Chapter 4 of Zvelebil & Baum (“Producing and Analyzing Sequence Alignments”): All sections.

ABOUT THE COMPUTER WORKSTATIONS

To log in

Press Ctrl+Alt+Delete and log in using your Concordia Netname and Password. (Please log out where you are done.)

PROCEDURE

STEP 1: Get your query sequence and identify it

Go to http://faculty.concordia.ca/glamoure/teaching.html and get the protein sequence corresponding to the number assigned to you.

Perform a BLAST sequence alignment on this sequence. Use the blastp algorithm on the “Non-redundant protein sequences (nr)” database.

- ✦ What protein does the best alignment correspond to? Provide the sequence accession code, including the name of the database (GenBank, PDB, etc.) Does your sequence correspond to the full protein? Are there any missing residues?
- ✦ Which entry in the Conserved Domain Database (CDD) and which Position-Specific Scoring Matrix (PSSM) does it correspond to? (The CDD entry is a serial number that looks like “CDD: 123456” that you will find in the “FEATURES” section of the primary sequence database entry, following the link at the beginning of a sequence description. You will find the PSSM information by following the CDD link, in the "Statistics" section.)
- ✦ Do any of the aligned sequences have structural information? (If you don’t see any, redo the BLAST search with a larger maximum number of target sequences. The default value of 100 is not always enough.)

STEP 2: Bibliographic research

Using either PubMed or Web of Science, do a quick bibliographic research to find more information about the biochemistry of the protein (or protein domain) matching your sequence. (Note that in your lab report, the bibliographic research will have to be more thorough. See below.)

- ✦ What organism is it extracted from?
- ✦ What is the (likely) biochemical/biological function of the protein?
- ✦ Based on similar proteins discussed in the biochemical literature, and on the sequence alignment of these protein sequences with yours, which residues from your sequence are likely to be functionally important?

Useful links:
PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
Web of Science: http://apps.webofknowledge.com
Google Scholar: http://scholar.google.com
(Please note that most of the electronic journals are accessible only within the university network, so you may want to email the PDF files of the articles to yourself for later reference.)

Note that similar protein sequences do not necessarily use the same residue numbering scheme throughout the literature. The easiest way to find how your sequence fits with that of homologous proteins from the literature is to perform a blastp alignment specifically on them. The procedure is the following:

- In the blastp web form (where you enter the parameters for your search), check the box called "Align two or more sequences".
- This will make a new section appear in the form, called "Enter Subject Sequence". You can enter more than one FASTA sequence in the text box, as long as they are separated with a line starting with ">". You can also provide GenBank or UniProt accession numbers instead of FASTA sequences.
- If you click on the "BLAST" button you will then see your sequence aligned to as many subject sequences as you entered in the second text box.

STEP 3: Find conserved residues

Go to the webpage of the conserved domain you identified in STEP 1 and reformat the multiple sequence alignment so that only identical residues are highlighted in red.

✦ Where are the functionally important residues you have identified (from your bibliographic research) located on the multiple sequence alignment?

✦ How well-conserved are these positions in the multiple sequence alignment?

STEP 4: Find a structural template

Redo the alignment using only the "Protein Data Bank proteins (pdb)" database and list all sequences that have high identities to the query sequence. (Make sure you record the PDB ID and the chain of each structure.)

Note that the highest-scoring sequence may not necessarily be the best structural template, so you should report all sequences that produce "acceptable" alignments.

An ideal structural template is a protein that (1) has a low *E*-value and a high sequence identity (especially around the functionally important residues), (2) has no unresolved loops (especially if those loops align well to the sequence to be modelled), (3) has a high resolution, and (4) is biologically related to the query sequence.

✦ Which sequence (PDB ID and chain) would you retain as a structural template and why?

✦ What is the level of sequence identity between the query sequence and the structural template? Are there any gaps in the alignment? Where are these gaps located in the secondary and tertiary structure of the protein?

STEP 5: Test the robustness of your alignment

Redo the analysis of STEP 4 using slightly different scoring matrices and gap penalties.

✦ Are the structural templates making your "short list" always the same?

✦ Are the functionally important residues always properly aligned?

INSTRUCTIONS FOR THE LAB REPORT

GENERAL

As always, a good report is clear, concise, and rigorous, and presents all the information needed for the reader to assess the validity and significance of the results.

Whenever you mention a sequence alignment, show the aligned sequences and all the statistical details. (The easiest is just to "copy-paste" the output from individual alignments.) Always use a fixed-width font when displaying aligned sequences, so that corresponding residues are lined-up.

Provide the DOI URL of every reference you cite. Use the URL format "http://dx.doi.org/xxx/xxx" (example: http://dx.doi.org/10.1186/gb-2001-2-10-reviews2002). Make sure the links are working on the PDF file you are submitting!

STEP 1 and STEP 2

Present the biological context and significance of the protein (or protein domain) best matching your sequence. In which species, organ, or organelle is this protein found? What is its main function? What is the interest of studying it? Which of its residues are functionally important? Is it known to undergo any functionally important conformational changes? Does it have any known interaction partners? This is the protein you will be working on the entire term, so the more information you can find about it, the better. Relate as much of this information as possible to your own sequence.

Refer to at least one textbook (or review article) and two scientific articles.

STEP 3

Include the multiple sequence alignment (MSA) from the CDD entry to your report (with the color coding) and indicate how your query sequence falls within it. The best is to "copy-paste" the text of the MSA into your report and manually add your sequence to it.

STEP 4 and STEP 5

Summarize your results in a table containing the bit scores, *E*-values, identity percentages, and gap statistics for all alignments.