

Sequence Alignment and Analysis, Database Searching

Practical Exercises, Monday 9th September

Introduction

The number of analyses and searches that one can do with a biological sequence is practically countless. In the practical exercises today, you will attempt just a few.

Most of the programs you will use will be run from the web interface of the Australian National Genomic Information Service (ANGIS.) The web site for our student accounts on this service is at <http://mell.angis.org.au/>.

Exercise 1: Pairwise Sequence Alignment

The comparison of two sequences is a fundamental operation that tells us so much about how those two sequences may be related functionally and genetically. In this exercise you will align a number of sequences, both nucleic acid and protein; learn the distinction between local and global alignment; and learn the effect of changing the more important parameters associated with the alignment techniques.

Preliminary Work: Retrieve sequences

Use BrowseCode to retrieve the following protein sequences from the PIR database, **HBMQC**, **HBHU**, and **LGB1_PEA**. Use BrowseCode to retrieve the following nucleic acid sequences from Genbank, **MMHOX7R** and **MMHOX8**.

Sequence comparison: Dot plot

A dot plot was one of the first computer-based methods applied to sequence analysis, by the Australians Gibbs and McIntyre in 1970. It can be a very useful tool for identifying inexact repeat sequences or similarities that only cover part of the two sequences under consideration. It is a useful analysis to perform before using any sequence alignment program.

To create a dotplot run the WAG program **Compare**, followed by **DotPlot**.

Compare the sequences **MMHOX7R** and **MMHOX8**.

Compare the protein sequences **HBMQC**, **HBHU** and **LGB1_PEA**, two at a time.

What is the appearance of the dot plot?

How does the length of the longest diagonal compare with the length of any structural motif of the gene product?

Can you see any evidence of sequence duplication or reversal?

Sequence comparison: Pairwise alignment

Sequence alignment programs come in many forms. They can, however, be placed into two main categories. Global alignment methods attempt to align the complete length of one sequence with the complete length of the other. Local alignment tools attempt to find the longest stretches of highest similarity between the two sequences. Global alignment

programs should not be used when the two sequences to be compared differ significantly in length. Think about why this may be so!

The WAG program for global alignment is named **Gap**, and for local alignment is named **BestFit**.

Global pairwise alignment: Gap

Align the two nucleic acid sequences using the Gap program. Deselect the option to abbreviate large gaps. Perform the following five alignment runs. *You will find it useful to change the name of the output file before proceeding with each run.*

Alignment 1. Accept all of the default gap penalties.

Alignment 2. Alignment 1 plus Select the option to penalize end gaps.

Alignment 3. Alignment 1 plus set the gap opening and gap lengthening penalties to 8 and 0.6 respectively.

Alignment 4. Alignment 1 plus set the gap opening and gap lengthening penalties to 3 and 0.2 respectively.

Alignment 5. Alignment 1 plus set the gap opening and gap lengthening penalties to 1 and 0.15 respectively.

Create a table containing the quality scores, percent identities, lengths and number of gaps for each of these five alignments.

How has Alignment 2 changed the alignment at either end?

How has Alignment 2 changed the alignment of the protein-encoding region of each sequence?

Which of the five alignments do you prefer? Why?

In what way does changing the gap penalties change the appearance of the alignment?

How does the alignment compare to the appearance of the respective dotplot?

Perform global alignments on each pair of the globin protein sequences, using the default alignment parameters. For one of the comparisons involving the pea globin request the statistics on 100 randomisations of your sequences.

How does each alignment compare to the respective dotplot?

What is "quality"?

Local pairwise alignment: Bestfit

Align the two nucleic acid sequences using the BestFit program and its default parameters.

What is the quality score and percent identity of this alignment?

Which regions of each sequence are included in the alignment?

What proportion of the coding region does the alignment include?

What does this tell you about the functional similarity of these two sequences?

Exercise 2: Sequence Similarity Database Searching

Transthyretin is a thyroid hormone-binding protein that transports thyroxine from the bloodstream to the brain. The molecule has a mainly beta-sheet fold, appearing like a sandwich with 7 strands in 2 sheets with a Greek-key motif in the directions of the chains. According to the SCOP hierarchy there are many families of all-beta-sheet folds. Transthyretin belongs to the prealbumin group along with beta-amylase, VHL, and some glycotransferases and dioxygenases.

The most common suite of programs for sequence database searching is the Blast suite written at NCBI, available there, through the Angis interface and almost everywhere else. The purpose of this exercise is to learn about some of the parameter settings on the Blast interface and how they affect database searching. To complete this exercise you will perform four searches within the Angis system.

The majority of sequences in the databases are nucleic acid sequences and consequently the majority of searches use nucleic acid sequences as the query. You will no doubt notice that most of the different Blast programs perform their searches by translating either the query or the database entries into protein sequences. A search conducted in this way is much better able to find distantly related sequences, or put another way, to explore more deeply into evolutionary history. However, searching the nucleic acid databases in this way introduces a number of problems associated with ESTs, so today the searches will be restricted to BlastP.

Retrieve the chicken transthyretin protein sequence S17827, and the human transthyretin sequence TTHY_HUMAN using BrowseCode. Select Blast2 from the Development menu. Select BlastP from the Blast program menu. The next step requires the selection of a database to search. Choose the one named NR Proteins. A menu of BlastP program options will now be presented.

Search 1:

Select the file that contains the chicken transthyretin sequence. This will probably be named something like S17827.pep. When the file is highlighted a name will appear for a file to take the output of the search. Append a number to this name. Type a different number for each of your searches. Angis runs on a unix computer and will overwrite files if the same name is used. Modify the output format options to select **500** scores and **500** alignments.

Search 2:

Use the same sequence file and output format options as Search 1, but make sure to specify a different output file name. In the Optional Search Parameters section set the Filter query sequence option to **False**.

Search 3:

Use the same sequence file and output format options as Search 1, but make sure to specify a different output file name. In the Optional Search Parameters section change the value of the E parameter to 40.0.

Search 4:

Use the human sequence file, specify the output file and use the output format options of Search 1. In the Optional Search Parameters section set the Alignment view option to **Flat Master-Slave, show identities**.

To answer these questions you will have to view the output of each search from the WebFM interface.

Did you see 500 scores listed in the output? If so, do you think there may be some matches not reported? If not, why not?

TTHY_HUMAN and AAA36784 are both human transthyretin sequences. How do you explain the difference between their scores?

What is the cause of the XXXs in the output of search 1?
What is the sequence buried by the XXXs? What is special about this sequence block?
Did the search without filtering report any additional hits compared to with filtering? If so which molecule has been hit by this search?

What was the result of changing the value of the expectation parameter?
For this search where are any new results added to the list?

Look at the output for the search with TTHY_HUMAN. List any three amino acid differences to the human sequence that are shared by bacteria. Give the human and bacterial sequences and the position number within the alignment.

Exercise 3: Sequence Analysis

There are many different analyses that can be performed on any sequence. The exercises here provided examples of just a few.

Translation

Translation of nucleic acid sequence to give a resultant amino acid sequence requires the correct reading frame, translation table, beginning and stopping point. The matter is further complicated if the nucleic acid sequence is genomic and contains introns and exons. Most programs that perform translation expect the user to enter the first and last base to be translated. Often this is not known at the beginning. The GCG MAP program, a restriction enzyme mapping program, can be used to sidestep this problem by running the analysis without selecting any enzymes. Another way to start is to identify all of the open reading frames. The latter approach allows the identification of frame shifts

Use the sequence for MMHOX8. Run the WAG FRAMES program. Run the WAG MAP program, select all six reading frames and select the NO ENZYMES options (or you get a very large output file.)

What is the correct reading frame in order to translate the protein product of MMHOX8?
Which bases should be included in the translation?
Which reading frame contains the **next** longest ORF, and approximately how many bases long is it?

Hydropathy Profiles

Protein secondary structure prediction and hydrophobicity profiling were amongst the first methods of sequence analysis to be developed. They take advantage of the relationship between sequence and structure.

Retrieve the membrane-bound kinase receptor protein sequence EPA3_HUMAN. Run the WAG program PepPlot. This program allows the user to plot many properties of the sequence. Try selecting any of the interesting ones. Make sure that a hydrophobicity/hydrophilicity plot is amongst them.

What does a hydrophobicity plot plot?

On the basis of the plot where is the transmembrane region of this protein?

Describe the shape of the plot immediately surrounding this region.

Motifs and Patterns

One purpose of sequence analysis is to identify the function of a molecule, or to identify the nature of the domains present in a large multi-domain protein. Many, many, many sites exist on the Web that permit one to perform a search for domain signatures or patterns or motifs within a sequence. Some look for similarity to whole protein domains, some look for signature patterns perhaps only a few (not necessarily consecutive) amino acids long. Each has its own algorithm and database of patterns. In many cases the information you obtain at different sites may be complimentary. This is one area where it is beneficial to try more than one analysis and to combine the results from all of the sites visited before making a final conclusion. Unfortunately today we will only have time to use one method, Prosite. It is the most widely available of all the methods.

Run the WAG MOTIFS program on EPA3_HUMAN. Ensure that your search excludes those patterns that are frequently found in many proteins.

List the motifs that have been found by this program.

Which motif is probably a false-positive, and why?

What is the function of this molecule?

Exercise 4: Information Retrieval

Exercise 4.1: Sequence database entries

The purpose of this exercise is to investigate the format of entries in different sequence databases.

Use SRS to retrieve databases entries from Genbank (ID: HSBCDIFFI), EMBL (HSBCDIFFI), Swissprot (IL5_HUMAN), PIR (a28477), PRF (Z1401202A) and PatchX (G790823).

Observe the different formats for presenting the data.

Identify the accession number, locus name, gi number, citations, database cross references, comments and sequence features.

What is a gi number?

Which databases have the most information?

Which databases have the most links to other databases?

Follow some of the links to other databases!

Exercise 4.2: Web Resources

For this question you will be using a number of resources on the World Wide Web. In most circumstances you will be expected to follow links from one page to another. Where this is not so you will be given the relevant URL. The purpose of this question is to provide you

with a guided tour of a number of the most valuable web-based genomic databases. Answer the questions as you go along and keep an eye out for answers to the miscellaneous questions listed at the end.

Go to <http://www.gdb.org>, and perform a simple search for the name IL5.

- What organism is the concern of this database?
- What is the name of this gene?
- What is the symbol of this gene?
- What is the chromosomal location of this gene?

Follow the link under Homology Links to MGD

- What is the URL of the home page of MGD?
- What organism is the concern of this database?
- What is the name of the murine homolog of IL5?
- What is the chromosomal location of this murine homolog?

Go to the main menu for this database. Select Mammalian Homology and Comparative Maps, then select an Oxford Grid using mouse and human as the two organisms. Select the cell in the grid that includes the IL5 homology. (Do you remember on which chromosome it can be found for each organism?)

List two neighboring genes from this syntenic region.

Go back to the GDB page for IL5 and follow the link to Phenotype Links.

- What database have you entered now?
- Is there a link between IL5 and asthma, discuss?

Select the citation for Pereira et al. and retrieve it from Medline.

- From what city did these researchers publish their work?
- Give the citation of one related article?

Go back to the GDB page for IL5 and follow the link to Protein Structure.

- A crystal structure exists for IL5, what is its PDB accession number?
- According to SCOP, what is the structural fold of this protein? (Give the full hierarchy please.)
- Name two other proteins with a similar fold to IL5.

You may want to visit the GeneCard entry for further information to complete the following questions.

- What alternative names (or aliases) does this molecule have?
- What is the function of this molecule?
- How many amino acids in the IL5 protein?
- What is its sub-cellular location?
- Give the accession numbers for IL5b entries in the following databases.

Swissprot	NCBI nucleic acid reference sequence
OMIM	NCBI protein reference sequence
GeneCards	

Exercise 4.3: SRS – The sequence retrieval system

The purpose of this exercise is to use QueryManager in SRS to build more complicated Boolean searches of a sequence database.

In Genbank find nucleic acid patent sequences for the interleukin 6 receptor molecule.
[Hint: Find all the Patent sequences, find the interleukin 6 receptor sequence, combine these two searches.]

The purpose of this exercise is to learn how to explore the links between different databases using SRS.

Find all the insulin sequences in Swissprot.. Select the link option from QueryManager and find links to this set of entries. Look for links in nucleic acid databases and PDB, Prosite, or OMIM?
[Hint: Insulin entries in Swissprot have locus names starting with the string INS.]

Exercise 4.4: Sequence entry retrieval

The purpose of this exercise is to learn about the differences between a number of sequence database entry retrieval programs. ANGIS provides BrowseCode, QueryDB and Lookup. NCBI contains a number of interconnected services.

Choose any database entry name from above or elsewhere. Try to retrieve the entry using each service. Use NCBI Entrez to retrieve database entries. Find related Medline, nucleic acid and protein sequence.

How do these services compare?

Which is the most flexible for finding entries with unknown names?