This document was designed and written to be accessed from the World Wide Web. There are interactive exercises, references and links that require that the reader access the page from the Web.
This document has been provided to allow users to print a reference copy. Each web page has simply been added after the previous one. No editing has been done to the text. Linked screenshots have been put below the refering paragraph. Appendix features (except the list of links) can be found at the end of this document.

If you find this guide useful, tell me!

Good Luck!

Andrew Louka

In silico biology is a thorough, expanding and complex science. This guide provides an interactive working introduction, for scientists with no working knowledge of molecular sequence analysis.

You will learn the essentials of molecular sequence analysis by performing your own searches of provided "unknown" sequences. Each database has it's

1. Moleculaa Databases with [Site Map]

Let's get an overview of how to search the molecular databases to identify a *query sequence*, before you do some searching of your own. Don't worry if this seems complicated at the moment, it will make more sense as you go through the interactive exercises.

Unfortunately, searching for nucleotide or protein sequences is not quite as

of N's (NNNNNN), the IUB code for any DNA base. Refer to the literature for more information. (References).

It is a good idea to mask almost all sequences, if you have the option to do so. Poly-A tails, for example, can give rise to artificially high scores and therefore misleading results. This is due to the large numbers of such sequences distributed throughout the genome, and therefore throughout the database.

Summary

We have learnt that to identify our unknown sequence, we can perform a sequence alignment search to see if our sequence is listed in the public databases. First, we need to choose an appropriate search program. We need to decide which database is most appropriate to search, and which matrix will give a rapid, accurate result. If the program allows, we also need to decide whether or not to allow for gaps in the sequence, and if sequences should be filtered.

Making the right choices will help to ensure that you have the best possible chance of getting a relevant and accurate answer, quickly.

An Overview of Public Molecular Databases

apply your knowledge to almost all molecular databases on the Internet. When using other databases, take care to check how regularly updated the

locate your sequence, assuming it is listed in the database being searched! BLAST and FASTA are arguably the most commonly used sequence alignment

greatly enhance your understanding.

If you get stuck, have a look at a screenshot (50.6 Kb) of the EXPASY BLAST page.

Scenario: You have been given a human gene to sequence and identify, but no clues as to what it is. The provider wants an unbiased opinion. To identify the sequence, you should:

1. Copy the human query sequence, given below:

```
AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAA
TATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTGTCTTATTTTAGAATACTGTGAT
CTATTTCTTTAGTATTAATTTTTCCTTCTGTTTTCCTCATCTAGGGAACCCCAAGA
GCATCCAATAGAAGCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATA
AAGAAGTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGA
ATATTTTTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAAT
TTCCTTTTTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGC
AAGATGTGAATTATTGACACTGAACTTAATAACTGTGTACTGTTCGGAAGGGGTTC
CTCAAATTTTTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTTAAGTTCTTA
TGAGGAGGGGAGGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCC
CCATCTAGACTAGCAATCTCTTCATTATTCTCTGCTATATATAAAACGGTGCTGTG
AGGGAGGGGAAAAGCATTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTA
ATAAGCAATCAAGGTTATAATTTTTTTTTAAAATAGAAATTTTGTAAGAAGGCAATA
TTAACCTAATCACCATGTAAGCACTCTGGATGATGGATTCCACAAAACTTGGTTTT
ATGGTTACTTCTTCTCTTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTG
GAGGGAGGGAAGGGTTTCTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTG
TAACTTGCTTAAATTTCTTATGTGACATTAACAAATAAAAAAGCTCTTTTAATATTAGATAA
```

2. Go to the EXPASY (EMBnet) BLAST Server WWW page
   If for any reason, you cannot access the EXPASY BLAST server, you can use any other BLAST server. I will refer specifically to the EXPASY server options and page layout.

3. Select the program: **BLASTN**
   This is the BLAST program that will compare a nucleotide query sequence against a nucleotide database

Further down the report, you come to a list of one line descriptions of database sequences that produced a significant alignment. The first one in the list (represented as the top long red line in the graphic) reads as follows:

emb|L37747|HSLAM11 [Homo sapiens]Homo sapiens lamin B1 gene, ex... 416 e-114

The

. The E value is givabein scientific notation. In the example, e-114 readein full as 1 times 10 to the power of minus 114. Orein other very close to zeroeindeed! This value is the number of times you xpect to see such a match (or better) merely by chance. The closer ue is to zero, the less likely the evabt is, To reiterate, the the E value, the more significant the match is.

that the sequence we tested is human, so there are only two tives from the list: the first and second. The others are not only tches with E valueseindicating that these matches are probably the of chance, but most of them are also from other organisms.

of warning! Lamin B has a special (uncommon) nucleotide sequence. u search for your own sequences, you are likely to get more than one sequence reported. The most statistically significant match (lowest) is not necessarealpmal sequees, you aookicatrchninook oe, rch est unco own cl(u

The Filter Option

BLAST version 2.0 enables the application of a filter. The filter masks regions of the query sequence that have low compositional complexity (e.g. Alu sequences), as determined by computer programs (SEG or XNU); (References). Masking is achieved by replacing the sequence with a string of N's (NNNNNN). N is the _____

should use PAM120 for generalised similarity searches. Take care! You cannot compare the alignment scores (see later) from one matrix directly against the alignment scores from another matrix!

You can choose an alternative scoring matrix for BLASTP, BLASTX, TBLASTN or TBLASTX. You can choose between PAM30, PAM70, BLOSUM80, BLOSUM45, or the default BLOSUM62. You cannot choose a matrix for BLASTN searches (instead, specify M and N, discussed below).

The EXPECT Option

You may, for example wish to set an expected score threshold (EXPECT) for the search, set to 10 by default. This means that ten matches are expected to be found by chance. If the statistical significance of a match is greater than the expected score threshold, it is not reported. Only if the statistical significance is less than this level, will the match be reported. In other words, a lower EXPECT threshold applies a more stringent search. This leads to fewer chance alignments being reported. You can enter fractional values if you wish; values are often suggested in a menu.

The Score Value Options

At the top of this page, we learnt that a [query:database] nucleotide pair was rewarded with a score depending on whether the nucleotides at that position were identical or not. The score awarded can be set by the user.

M Parameter
: The score awarded when a pair of aligned residues match. Must be a positive integer.

N Parameter
: The score awarded when a pair of aligned residues do not match. Must be a negative integer.

The ratio of M:N determines the degree of divergence (evolution) that is accepted. The default value for M is 5 and for N is 4. The ratio of 1.25 equates to around 47 nucleic acid point accepted mutations (PAMs) per 100 residues. PAMs are used as a predictor of the degrees of evolution from an ancestor (in molecular terms). If you adjust the M and N values to give a higher ratio, more nucleic acid PAMs will be accepted by the algorithm, resulting in a more divergent search.

Fetching Sequences

Fetching Sequences

instructions are all there.

You should use the EXPASY BLAST Server for these searches. Write down your answers to the proposed questions, and check them against the answers given along the way.

Exercise 1

Copy the sequence given below, and use it to run a nucleotide BLAST search. Use the default criteria except where you *must* change the options. Hint: You may need to select a program and database! Can you identify the sequence?

GTCCGGCCTGGGCGACAGAGCAAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Exercise 2

The sequence is taken from a GenBank entry whose accession number is S56967. Using the BLAST server, locate this sequence.

Seeing the actual sequence, you shouldn't be surprised that you were unable to identify it in your first search... it's an

GAATTCTAATCTCCCTCTCAACCC

Nothing found? That seems rather odd, because we know that there is a corresponding sequence in the database --we found it in the previous search!

3. Paste the sequence into the specified window and run the search using the default parameters.
4. Be patient! If your query takes time to process, please don't close the window and give up --unless you really must. The server will still be processing your query.
5. Examine the output. Could you identify the sequence if you didn't know what it was?

Even with the short sequence provided, there were many alignments reported. Did you notice how the reported alignments were presented differently to the BLAST results? If not, you might like to go back and look again. Notice that BLAST only reports the aligned bases, whereas FASTA also presents the c 1'vt ofs the alignmen i.e.s theflanksing sequencsT o ei the svidd.

the extension phase. In FASTA, the word is not scored, but must be an exact

sequence, and use the result to search the database.

FASTA: Reading the Output

The FASTA output is essentially very similar to the BLAST output. You are first presented with a list of the reported sequences: the most significant alignments. Below the list, the actual alignments are presented in context of the database sequence. The number of bases that match exactly are reported as a percentage of identity.

In the list of all reported sequences at the top of the page, the last but one value is the score, and the last number given on each line is the *expect* value (scoring E Value). The maximum (threshold) E value is 2.0 by default. As with BLAST, the smaller the expect value, the lower the probability is that the reported alignment is a chance fiot onsh%ttringit, ao thr way,s the

Finally, let's try the fastx program, which translates the DNA sequence 'on-the-fly' into the three reading frames of the given strand, and compares the amino acid residue sequence with the entries in a protein database. We will use the same sequence as elsewhere on this page. Follow the instructions given below:

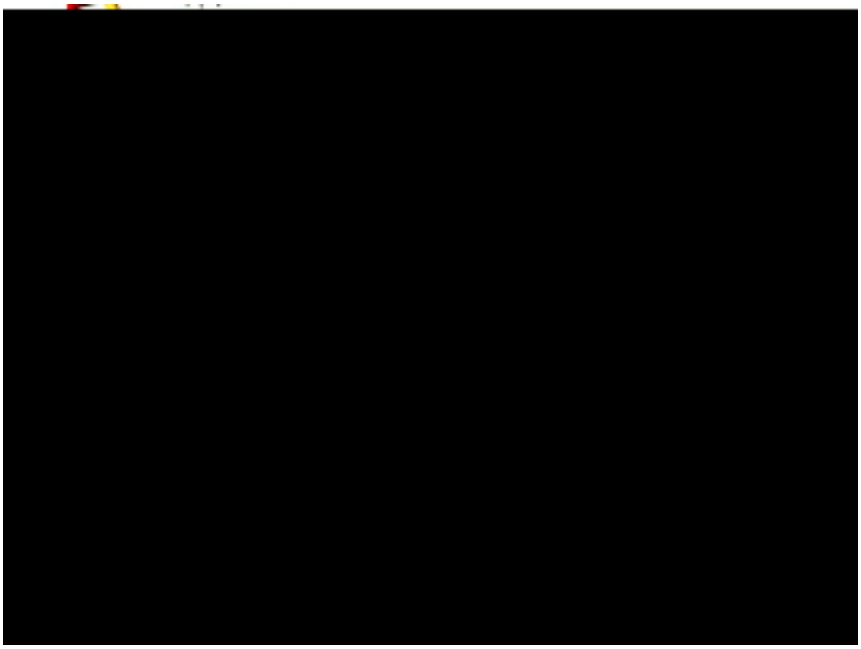
1. Copy the DNA sequence:

position (there are 20 possible amino acids at each position in a protein). The genetic code is redundant —there are several variations of most DNA codon triplets that code for an amino acid. Although a the protein product may be identical to your query sequence, you may not get an identical match with the DNA. Also, protein sequence similarity is more conserved through time than is DNA sequence similarity.

The search for protein orthologues is becoming increasingly important in molecular biology. Now that the complete *Saccharomyces cerevisiae* (yeast; a unicellular eukaryote) and *Caenorhabditis elegans* (nematode; a multicellular eukaryote) genomes have been sequenced, work is well underway to identify orthologous groups. If a novel human protein can be matched with orthologous proteins from yeast or The Worm, the investigator is likely to save a lot of time (and money!), having identified a likely function for the protein. [A good example is: Chervitz SA, *et al* (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*. **282**:2022-2028.]

The Main Protein Sequence Databases

There are two major worm and yeast:U028 -12 Td(may be ecialise past: ortd1 Tc 0



and G). We can also see an en dash (-) in the lower sequence. This indicates that a second A has been inserted into the query sequence. This is different to a mismatch because if this position is skipped (a - is inserted), the following nucleotides align perfectly.

Notice also the position of the alignment. The query sequence is reported from base 1 and aligns with the database sequence from base 71799.

Simple, isn't it? You should give it a try...

Alignment Using GCG Software

Warning! This page is only of interest to readers who have access to the Wisconsin GCG Package (referred to as GCG) on a UNIX server. If you don't have access, feel free to read on, but you should realise that you will not be able to use the licenced software.

GCG is a software package that can be run on a UNIX computer. You will usually run GCG on your remote server through a telnet session on your computer. If you don't understand this, contact your local user support or system administrator for help.

To complete the following exercises, you will need to be running GCG. If you are a registered user at the HGMP-Resource Centre, you may wish to telnet to menu.hgmp.mrc.ac.uk now. Use the menus and load the latest version of GCG. If you have an account elsewhere, you should open a telnet session now.

You should begin by loading the GCG package. At the UNIX prompt
(

When you have finished entering your description, hit ö**D** (i.e. on your keyboard, hold down the **Control** key as you hit **D**). The input cursor will move to the middle of the screen where you can paste or enter the sequence. Try it! It is normal practice to enter the nucleotides in capital letters (A, C...), and to use small letters (a, c...) to mark areas of special interest. Use N to designate a base whose identity is unclear.

When you have finished, hit ö**D** again to move the cursor to the command line at the bottom of the screen. Type **ex** and hit Enter/Return to snishand exit seqed. To exit without sniing, type **q**. For more commands, type **help**.

The snisd file is now ready to be fsd into GCG programs. To edit the file later, you can load seqed with the name of the file you just made:
$> **seqed** *filename*

To snishtime, you may wish to download an example sequence in GCG format. You can do so by sniing the contents of the following link to disk: M13mp18.seq. This file is a text file UNIX line breaks. Don't open and snish it in a non-UNIX compliant editor, or it will become unreadable by GCG!

The GCG Alignment Software

The *bestfit* program will determine the optimal alignment of two GCG formath

and paste the given sequences one at a time, into the alignment program window. The instructions given below are simply guidelines. You should feel free to explore, although there aren't many options to choose between! You might like to try aligning polypeptide sequences.

Imagine that you have made mutant clones of a DNA sequence. How do you know which clones have been successfully mutated, and what the mutations are? You could sequence the clones, and align (compare) them with the original sequence! Have a go! The instructions are given below:

1. Go to the ALIGN server at GeneStream, France.

Nucleotide Sequence Translation

The genetic code is known, and can be used to translate coding nucleotide sequences. For example, CCC codes for a proline amino acid, and GTA codes

Open Reading Frame Search

It is possibe to feed a computer program a protein sequence, and ask it to identify all of the open reading frames in that sequence. ORF Finder is an

GACTGTGGCTGCTGGCGTTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCAGAGACCTCACTCC
CGGGGAGCCAGCATGTCCACTGCGGTCCTGGAAAACCCAGGCTTGGGCAGGAAACTCTCTGACTTTGGAC
AGGAAACAAGCTATATTGAAGACAACTGCAATCAAAATGGTGCCATATCACTGATCTTCTCACTCAAAGA
AGAAGTTGGTGCATTGGCCAAAGTATTGCGCTTATTTGAGGAGAATGATGTAAACCTGACCCACATTGAA
TCTAGACCTTCTCGTTTAAAGAAAGATGAGTATGAATTTTTCACCCATTTGGATAAACGTAGCCTGCCTG
CTCTGACAAACATCATCAAGATCTTGAGGCATGACATTGGTGCCACTGTCCATGAGCTTTCACGAGATAA

Protein Motif Searching

Not only can you search databases for homologous sequences, but you can also search for protein motifs that have been conserved through evolution. If you have identified a new protein and you don't know anything about it, you would find it useful to perform a motif search to see if any other sequences have a homologous motif. The results might give you a clue as to it's function. Alternatively, you might wish to perform a motif search of your mutated sequence to see if an important motif such as a transporter signal sequence has been corrupted.

Try running a search using the PRINTS server at EBI, UK. Paste in the complete sequence and hit the button "Run PPSRCH". A screenshot (20 K) of the output is available. The ABC transporter family of carier proteins (recognised in this sequence) includes the multidrug resistance ATPase in mammalian cells, chloroquine-resistance ATPase in *Plasmodiues a9ciparum* and many others. In the CFTR protein, the ABC transporter is responsible for

encountered in bioinformatics. This glossary forms part of an online Guide to Molecular Sequence Analysis.

Some of these explanations are rather simplistic, in favour of brevity. Please refer to molecular biology text books for more comprehensive details.

Alu

most frequently occuring at that position, in the real sequences.

OMIM
: Online Mendelian Inheritance in Man. Database of genetic diseases wuth references to molecular medicine, cell biology, biochemistry and clinical details of the diseases.

ORF
: Open Reading Frame. A series of codons (base triplets) which can be translated into a protein. There are six potential reading frames of an unidentifed sequence; TBLASTN (see BLAST) transalates a nucleotide sequence in all six reading frames, into a protein, then attempts to

A non-redundant (See Redundancy) protein sequence database. Thoroughly annotated and cross referenced. A subdivision is TrEMBL.

TrEMBL
: A protein sequence database of Translated EMBL nucleotide sequences.

UniGene
: Database of unique human genes, at NCBI. Entries are selected by near identical presence in GenBank and dbEST databases. The clusters of sequences produced are considered to represent a single gene.

Upstream
: Toward the 5' end of a nucleotide sequence.

FASTA Format Explained

The genomic sequence below is in FASTA format, which is often required when searching molecular databases. **Take care!** *The first line* must begin with
= description! The description can be anything that you choose
ɔu obtce beo.on c neau0 below i62scription can be anything that you choose

>gi|576838|gb|L37747|HUMLAM11 Homo sapiens lamin B1 gene, exon 11, complete cds

search programs. *Nucleic Acids Research* **25**:3389-3402.

2.