Bios 533 Bioinformatics

By:

Susan Cates

Bios 533 Bioinformatics

By:
Susan Cates

Online:
< http://cnx.org/content/col10152/1.16/ >

CONNEXIONS

Rice University, Houston, Texas

Collection structure revised: September 24, 2008
PDF generated: February 4, 2011
For copyright and attribution information for the modules contained in this collection, see p. 59.

Table of Contents

1 Prerequisite Material: Database Lab

1.1 NCBI: National Center for Biotechnology Information 1
1.2 Entrez 2
1.3 PDB 4
1.4 Tour of Bioinformatics Sites 7

2 Basic Sequence Alignment Lab

2.1 Introduction to Sequence Alignment 9
2.2 PipMaker 13

3 Multiple Sequence Alignment Lab

3.1 Multiple Sequence Alignment 17

4 Substitution & Scoring Matrices

4.1 Scoring Matrices 23

5 Phylogenetics

5.1 Phylogenetic Trees 27

6 Profiles and PSSMs

6.1 PSI-BLAST 31

7 Independent Study Project

7.1 Bioinformatics Project 35

8 Supplemental Material: Advanced Topics

8.1 RNA Secondary Structure Prediction 37
8.2 Microarray Experiments 40
8.3 Expasy Proteomics Tools 45
8.4 Protein Folding and Secondary Structure Prediction 51

Bibliography 55
Index 58
Attributions 59

Chapter 1

Prerequisite Material: Database Lab

1.1 NCBI: National Center for Biotechnology Information[1]

The National Center for Biotechnology Information (NCBI) provides a comprehensive website for biologists that includes biology-related databases, and tools for viewing and analyzing the data inherent in the databases. A division of the National Library of Medicine at the National Institutes of Health, NCBI is the agency responsible for creating automated systems for storing and analyzing the rapidly growing profusion of genetic and molecular data. One of the most difficult challenges faced in the field of bioinformatics is how to store, in an easily accessible manner, the overwhelming abundance of new information, including the sequences of entire genomes, the ongoing discoveries of new genes and gene products, and the determinations of their functions and structures. NCBI was established as the government's response to the need for more and better information processing methods to deal with this challenge.

View the NCBI home page[2] . A relatively good overview of the tools and databases that can be accessed through NCBI is provided in the list along the left border of the home page. Clicking on the link entitled "About NCBI" produces a second menu containing the topics "A Science Primer", and "Databases and Tools", among others. Click on "A Science Primer" to access general definitions and introductory information regarding the branches of science included in bioinformatics. Many bioinformatics terms are defined in this section in a clear-cut and basic manner, making this Primer an excellent first resource. From the table at the top of the web page, click on "Databases and Tools" to yield a listing of accessible information. This web page containing the databases and tools menu is a good choice for those who are inclined toward bookmarking.

The first item under the "Databases and Tools" menu is "Literature Databases". PubMed is the most heavily used of the literature databases and can be used to access MEDLINE biological and medical scientific journal citations dating back to articles written in the mid-1960's. The second item under the "Databases and Tools"menu is "Entrez Databases". *Entrez*[11] (1) is a search and retrieval system developed by NCBI that is capable of accessing integrated information by searching many of the NCBI databases with just one query (instead of searching only one database per query, then having to repeat the query to find information on the same topic from another NCBI database). The NCBI databases that are included in the search when you launch an Entrez query are shown when you click on this link. Look down the list and read the description for the "Protein sequence database". Notice this database has many sources such as Swiss-Prot, PIR, PRF, PDB, and GenBank, which are all individual biological databases in their own right, containing different types of data about proteins. The Entrez accessible "Nucleotide sequence database" contains annotated collections of publicly available nucleotide sequences, many of which are cDNA and mRNA sequences. The evolution of bioinformatics data mining methods has been largely driven by the prodigious amount of sequence information collected by scientists in recent years. New sequences of unknown

[1] This content is available online at <http://cnx.org/content/m11789/1.3/>.

[2] http://www.ncbi.nlm.nih.gov/

function can be compared with sequences of well-characterized genes and proteins. Similarities can be identified between the new, unknown sequences and the well-characterized sequences, and used to postulate theories regarding function or structure.

Click on "Databases and Tools" from the table at the top of the web page. Selecting the "Tools for Data Mining" topic will show a list of data retrieval tools, including Entrez, mentioned above, and BLAST, the *Basic Local Alignment Search Tool*[5] (2). Blast is the predominant sequence alignment tool for performing rapid searches of nucleotide and protein sequence databases and detecting local, as well as global, sequence alignments between the query sequence and the database sequences.

This is a brief glimpse at some of the more widely used tools and databases presented by NCBI. As a final exercise, take a moment to select the "Outreach and Education" link from the table. There are two links on this page that may prove helpful at times, "Courses and Tutorials" and "Glossaries". There are tutorials in the use of Blast, Entrez and Pubmed, among others. Return to these as needed while learning the use of these tools. The Glossaries are particularly useful because bioinformatics has a lot of field-specific lingo and acronyms that can be relatively confusing to decipher.

1.2 Entrez[3]

Entrez[12] (1) is a search and retrieval tool developed by NCBI that is capable of searching multiple NCBI databases with just one query. Entrez returns search results that can include a combination of many types of data on the query, such as nucleotide sequences, protein sequences, macromolecular structures, and related articles in the literature. Prior to the creation of Entrez, an individual might have to place one query to a nucleotide database to find a nucleotide sequence, submit another query to a structural database to find the published structure of the gene product, and submit a final query to a literature database to find citations for journal articles on the query topic. NCBI recognized the time and effort that could be saved by a tool that could cross-link these databases and integrate all information related to a given query subject into one report. View the Entrez Database page[4] . This module contains a few problem questions, for use in a computer lab setting. The lab instructor may require that you supply answers to these questions as an indication that you have completed the module.

The Entrez Nucleotides database includes sequences from GenBank, RefSeq, and PDB. GenBank is the National Institutes of Health (NIH) genetic sequence database. GenBank, the DNA DataBank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL) comprise the International Nucleotide Sequence Database Collaboration. These three organizations exchange data on a daily basis. The number of bases in the Entrez Nucleotides database currently grows at an exponential rate. Click on the Nucleotide link listed under the heading "Nucleotide Databases".

Exercise 1.1

What is the number of bases stored in the Entrez nucleotide database, as of the last report?

Use the back arrow of the browser to return to the Entrez Database web page. Locate the MMDB (Molecular Modeling DataBase), one of NCBI's structure databases and click on the link to read about it. MMDB is a subset of three-dimensional structures obtained from the Protein Data Bank (PDB), excluding theoretical models. While the protein databases contain protein sequences, the structural database contains coordinate files (PDB files) of biological molecules with solved (known) structures. Click on the arrow next to the search box at the top of the web page and view the list of databases for selection. The literature database is accessed through PubMed, which encompasses the National Library of Medicine's journals database, MEDLINE, as well as providing some additional online services. MEDLINE is a collection of medical and life science journal citations that includes articles dating back to the mid-1960's. Entrez allows access to information such as nucleotide and protein sequences organized by species in the NCBI taxonomy database, also found on the selection list. The connectivity of the databases available on the selection list are indicated by the

[3] This content is available online at <http://cnx.org/content/m10996/2.7/>.

[4] http://www.ncbi.nlm.nih.gov/Database/index.html

diagram on the Entrez Database web page[5] . Click on the diagram to access a Flash model of Entrez database connectivity. As long as the browser has a Flash plug-in, placing the mouse over one of the nodes representing a database will highlight its connectivity. Try this on the node labeled "Protein". Actually clicking on the node will forward the user to the database home page.

Use the back arrow of the browser to return to the Entrez Database web page. There is a menu bar at the top of many NCBI web pages that contains links to the most commonly used tools and databases, such as PubMed, Entrez, and BLAST. Click on the "Entrez"link at the top of the page. The Entrez cross-database search page should be visible in your browser, now. Here, one can enter a query and click "GO" to search against all databases, or click on a database link for the search page that is specific to that database. Perform a search using the query string *Mycobacterium tuberculosis*, and click "GO".

Exercise 1.2
How many PubMed literature citations and abstracts contain the character string *Mycobacterium tuberculosis*?

Exercise 1.3
How many nucleotide sequences are returned?

Exercise 1.4
How many protein sequences are returned?

Exercise 1.5
How many 3-D macromolecular structure entries are returned?

Click on one or two of the databases that returned items in response to this query. Take a quick look at the information returned as a match. This is an overwhelming amount of information that has been returned in response to this query. It is difficult to do anything with this much information. For this reason, a good search strategy is required to limit the search as cleverly as possible in an attempt to obtain mostly records of interest, with very little excess information, without restricting the search so much that it is likely to miss important records.

There are many different ways to limit a search query. To illustrate one approach available in Entrez, from the cross-database search page, click on the Nucleotide Database link. Notice the menu just under the query box, and click on the link entitled "Limits". Under "Limited to:", select "organism". On the pull-down menus, change the limits from "molecule" to "Genomic DNA/RNA", change "segmented sequences" to "show only master of set", and change "only from" to "GenBank". This limits the search from returning records from any type of molecule, including protein, ESTs, etc., to only records of submitted Genomic DNA or RNA sequences. It furthermore limits the sequences returned to only master sequences of any sets, and it only searches the GenBank database for records. Using *Mycobacterium tuberculosis* as the query string again, perform the search with these limits.

Exercise 1.6
Now, how many nucleotide sequences are returned?

Exercise 1.7
How does this compare to the number of nucleotide sequences returned in the cross-database search?

Hopefully, this has illustrated that a general cross-database search is best used when there is very little information available related to the query, and so it is desirable to find all pieces of related data. However, when lots of data is available related to the query, it it desirable to limit your items returned. Using the "Limits" function in Entrez is not always the best way to limit a query, though. Perhaps the area of interest happens to be genes that help confer drug resistance to *Mycobacterium tuberculosis*. Deselect the previously set limits by clicking on the check mark to the left so that it disappears. Now, search "nucleotide" using the query string "*Mycobacterium tuberculosis* drug resistance".

Exercise 1.8
How many items (sequence records) are returned?

[5] http://www.ncbi.nlm.nih.gov/Database/index.html

Look at the list of results. The numbers at the head of each result are called access codes. Click on the access code of one of these records. The left column of the record contains terms that are referred to as "identifiers". The identifiers in any database are defined terms that indicate the record section and the type of data included in that section. Scroll down to the section entitled "Features". Two common identifiers found in this section are "gene" and "CDS" listings. The CDS tag identifies "coding DNA sequences", meaning these sequences have been determined (most often by bioinformatics and not experimental methods) to encode proteins, and are thus distinguished from the noncoding regions that make up a substantial amount of the DNA in the human genome. A good primer on the basic characteristics of DNA, including the differences between coding versus noncoding sequences, can be found on the *Dolan*[1] DNA Learning Center web page[6] (2). Scroll through the results, and notice that there are links embedded in this record. These links connect this record to other databases, as illustrated in the connectivity diagram discussed earlier in this module. So, even though this search was performed over the nucleotide database, the result may contain a link that takes us to a record in the protein database. Find a record that contains a "gene" link in the Features section of the record, and click on this link. In the new record, there should be a sequence of capital letters at the bottom of the CDS section.

Exercise 1.9

What does this sequence represent?

There is an additional sequence in lower case letters at the bottom of this record.

Exercise 1.10

What type of sequence is represented by the lower case letters?

If these questions regarding sequences have been difficult to answer, please review the genetic code[7] , as this is prerequisite information for this course.

Try your own search. Scroll back to the top of the web page and this time next to the Search command, choose PubMed from the menu. Pick any life sciences topic that interests you for your query. Attempt a first query with a general topic, such as protein kinase or diabetes.

Exercise 1.11

What type of results does PubMed return from a query?

Note how many items in total (not just on the first page) were returned. Make your query topic related to your original choice, but more specific. For example, change 'protein kinase' to 'protein kinase C'.

Exercise 1.12

How much did this reduce the number of items returned?

This module is intended as an introduction to performing searches of the NCBI databases using Entrez. If you are unfamiliar with Entrez, please feel free to return to this module as a resource for getting started on NCBI searches.

1.3 PDB[8]

The Protein Data Bank (PDB) is a public domain repository containing experimentally determined structures of three-dimensional biological macromolecules. The majority of these structures have been determined by x-ray crystallography, but structures determined using nuclear magnetic resonance (NMR) methods are on the rise. A very few theoretical models are also included in the PDB. The PDB was originally established at *Brookhaven National Laboratory*[23] (1) in October, 1971, with 7 structures. It is currently managed by *Rutgers,*[30] (2) The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the Center for Advanced Research in Biotechnology/UMBI/NIST, and it stores over 29,000 structures. The European Bioinformatics Institute Macromolecular Structure Database group

[6] http://www.bioservers.org/bioinformatics/dna_characteristics.htm

[7] http://www.bioservers.org/bioinformatics/Worksheets/genetic_code.htm

[8] This content is available online at <http://cnx.org/content/m10997/2.4/>.

(UK) and the Protein Research Institute at Osaka University, Japan are international contributors to the contents of the PDB.

The name Protein Data Bank is historical in origin, because the present-day PDB includes many DNA and RNA structures as well. The most important information contained in any given PDB file is a set of 3-dimensional vectors representing the atomic coordinates for each of the individual atoms that comprise the biological molecule(s) included in the structure. These coordinates can be fed into various graphics programs that allow the scientist to view the a 3-dimensional model of molecule. An example of one of these models is the molecule of the month that can be viewed by clicking on the link in the left-hand blue border of the PDB home page[9] .

Exercise 1.13

What is the featured molecule of the month today?

The PDB search engine asks the user to "Enter a PDB ID or keyword". The PDB ID, which is also referred to in journal articles as the PDB accession code, is a 4 character alphanumberic ID code assigned to the structure coordinate file when it is deposited into the PDB by the experimentalist who solved the structure. For instance, the PDB ID for a 2.2 angstrom crystal structure of the protein calmodulin is 3CLN. Perform a search using 3CLN as the query. The result is a summary page that lists the method of structure determination, as well as the authors of the structure. It is not uncommon for the authors of the primary citation to differ somewhat from the authors of the structure, as the first refers to the writer(s) of the article where the new structure first appeared and the second refers to the experimentalist(s) who determined the structure of the deposited molecule(s). The compound field identifies the common name of the protein or nucleic acid molecule in the structure, and the source identifies the genus and species of the organism from which this molecule is derived, which in this case is a rat. At the bottom of the summary, is a table entitled HET groups. HET (heteroatom) refers to any atom that is not part of the biological molecule(s) in the structure. These are often ligands, which are molecules that commonly bind the particular protein or nucleic acid in the structure. In this case, there is calcium in the structure, which may not be surprising, given that the classification listed for this protein in the summary is "Calcium Binding Protein". The formula column of this table gives the chemical formula of the ligand.

Exercise 1.14

How many calcium ions are in this structure?

Exercise 1.15

What is the date this entry was deposited to the PDB?

Scroll back up to the top of the summary page. In the blue border on the left of the page click the link entitled "View Structure". Most of the interactive 3D displays require downloading and installing software. For those who do not wish to do this, view the high (500 X 500) resolution ribbons image under "Still Images". This is a ribbons "backbone" model of a protein, that includes the calcium ions, but does not show the sidechains of the individual amino acids. Compare the ribbons model with the cylinders model, wherein all the helices are shown as cylinders. These images are both made with the same atomic coordinates from the 3CLN file, but the coordinates can be modeled in many different ways, depending on what properties the image renderer wants to illustrate.

Now, select the "Download/Display File" option listed in the blue border on the left. For those files that need to be downloaded for making images, usually the text file format is preferred because many of the graphics programs will not accept the html format file. In the table shown under the "Display the Structure File" option, choose the PDB text file format, complete with coordinates. Now, take a look at the contents of the files.

Exercise 1.16

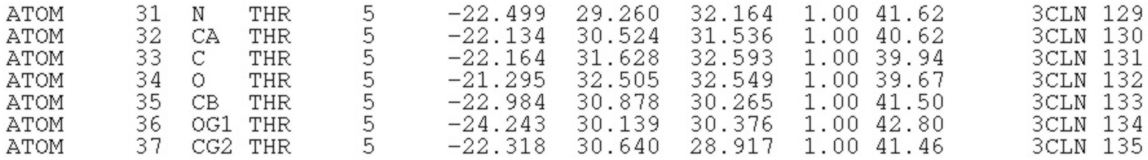
What was the experimental method for determining the structure of calmodulin in the entry 3CLN?

The first column of the PDB file contains an identifier for the type of information contained on that line. For example, JRNL, is the identifier for all lines containing the literary citation(s) for the journal(s) where

[9] http://www.rcsb.org/pdb/

the structure was published. The first section of the PDB file is the Title Section, which begins with a line containing the identifier HEADER and continues until the end of the lines containing the identifier REMARK. This section includes information about the experimental method used to obtain the atomic coordinates present in the file. The 3CLN structure is an older entry into the PDB, and does not contain a great deal of information about experimental methods and statistics, however newer entries to the PDB are required to have more information. Scroll down until you see the ATOM identifier in the first column. This is the beginning of the atomic coordinates listing and this section of the file should look something like the figure illustrated below.

PDB COORDINATE FILE FORMAT



```
ATOM      31  N   THR     5     -22.499  29.260  32.164  1.00 41.62        3CLN 129
ATOM      32  CA  THR     5     -22.134  30.524  31.536  1.00 40.62        3CLN 130
ATOM      33  C   THR     5     -22.164  31.628  32.593  1.00 39.94        3CLN 131
ATOM      34  O   THR     5     -21.295  32.505  32.549  1.00 39.67        3CLN 132
ATOM      35  CB  THR     5     -22.984  30.878  30.265  1.00 41.50        3CLN 133
ATOM      36  OG1 THR     5     -24.243  30.139  30.376  1.00 42.80        3CLN 134
ATOM      37  CG2 THR     5     -22.318  30.640  28.917  1.00 41.46        3CLN 135
```

Figure 1.1: Each separate column in a given section of a PDB file is designated as a different "field". The format of the coordinate section of the PDB file is illustrated above.

The 5th column of this file lists the residue number. Notice that in the 3CLN file the first residue number is 5. This is mentioned in the REMARKS section at the beginning of the file, as it should be.

Exercise 1.17

What explanation is given for the missing amino acid residues?

The PDB offers a syntax directory to help interpret the lines and columns in each section of a PDB file. Scroll back to the top of this page and click on the PDB icon in the upper left hand corner to return to the PDB home page. Locate the link entitled "FILE FORMATS". Under PDB File Format, select the most current version of the File Format Contents Guide. Scroll through the Table of Contents listed on the left side of the page. Beginning with the Title Section, links are provided for every possible type of line identifier that can be found within a PDB file. Under the Coordinate Section, click on the link for the line identifier ATOM to call up the Record Format listing for the atomic coordinates section. Each possible field in an item line is listed according to the character column numbers assigned to the field. Not every available field will always be used in a file. For instance, in the Record Format listing, the residue sequence number is listed in the 7th possible field, but the 3CLN file only uses 5 of the first 7 fields and so the residue number is the 5th column. This is because the fields altLoc and chainID are not required in the 3CLN file.

It is common to have more than one structure present in the PDB for a medically or scientifically important protein or nucleic acid, particularly if the structures represent genetically engineered mutants of the same biological molecule, similar molecules from different organisms, or the same molecule bound to different ligands. Return to the PDB home page (the PDB home page icon is always in the upper left hand corner). Enter the accession code 1CFC as a query to the PDB.

Exercise 1.18

What biological molecule is represented by this entry?

Exercise 1.19

View the summary page for this structure. What was the experimental method used to determine this structure?

Exercise 1.20

Are any ligands bound to this molecule?

Notice that 25 coordinate sets, representing slightly different molecular conformations, are present. This is characteristic of solution studies, where the molecule is dynamic. Use the "View Structure" link to view the ribbons image of a superposition of the 25 models. The parts of the molecule that overlap well between the models are areas that maintain a relatively rigid structure even in solution, while the parts of the molecule that do not overlap well are dynamic in solution. Use the "Download/Display File" link to display the PDB text file, header only, no coordinates.

Exercise 1.21

Why, in this case, might it be undesirable to display the file complete with coordinates? (If uncertain, choose the "complete with coordinates" option and look through the coordinate file to find out.)

Before closing this file, view the experimental methods discussed in the Title Section to see how they differ from the 3CLN entry.

Return to the PDB home page and search by keyword calmodulin.

Exercise 1.22

How many structures are found in response to the query?

Exercise 1.23

Are all of the matches returned structures of calmodulin?

Exercise 1.24

Explain the answer to problem 11 in terms of the way the search engine responds to the keyword "calmodulin" that was used as a query.

If the query results at the top of the page state that there are structures being processed or "on hold", these listings can be accessed by clicking on the link entitled "matching your query". The Processing/Hold database contains structures that are soon-to-be released. The NIH has a strict release policy, requiring that any structures derived through experiments supported by NIH grants be deposited at the time of submission of a research article to a journal, and although a hold may be placed on these coordinates during the submission process, the rules state that the coordinates are to be released upon publication.

1.4 Tour of Bioinformatics Sites[10]

This is a tour of various bioinformatics web sites and search launchers. It is a good idea to be familiar with many of the bioinformatics centers, because they tend to emphasize the methodology that they have had a hand in developing, and therefore each institution offers somewhat different tools. Most of the databases share information, particularly the Japanese, European and North American databases, so the databases essentially mirror each other. However, it is worth exploring different bioinformatics web sites to take advantages of the different tools for data analysis that they have to offer.

The European Bioinformatics Institute (EBI) is a part of the European Molecular Biology Laboratory (EMBL). The European Molecular Biology Laboratory is a non-profit, academic entity supported by sixteen countries including nearly all of Western Europe and Israel, with facilities in Heidelberg (Germany), Hamburg (Germany), Grenoble (France), Hinxton (the U.K.), and Monterotondo (Italy). The EBI is the bioinformatics arm of the EMBL, and it functions in Europe to maintain and create databases and bioinformatics tools in the same way NCBI does in the United States. EBI maintains the EBI Toolbox[11] , a selection of bioinformatics software that can be accessed on the internet. The Toolbox web page has a menu on the left of the page that lists different categories of tools such as Homology and Similarity, Protein Functional Analysis, Structural Analysis, Sequence Analysis and Miscellaneous Tools. Look through these categories to get an idea of the number of tools available at this web site alone.

[10] This content is available online at <http://cnx.org/content/m10998/2.2/>.

[11] http://www.ebi.ac.uk/Tools/

A very large selection of tools are available on ExPASy[12] , (Expert Protein Analysis System), the proteomics server of the Swiss Institute of Bioinformatics (SIB). The ExPASy tools and databases are numerous, and relatively easy to locate and use. ExPASy is also unusually good about documentation, including in their databases "SeqAnalRef", one of the few good listings of bioinformatics bibliographic references. Take a moment to inspect this web site and the array of tools it provides.

The Baylor College of Medicine Search Launcher[13] contains a nice variety of bioinformatics tools. This site benefits from a close association with the Human Genome Center at Baylor, and offers new tools that have been developed in conjunction with this collaborative effort. An example of one of the tools developed at Baylor is BEAUTY (BLAST enhanced alignment utility), an enhanced version of NCBI's BLAST tool that aids in assigning functions to matched sequences. Tour the BCM search launcher and read about the BEAUTY search.

These are only a few of the many bioinformatics sites available via the internet. Perform a Google[14] search with "bioinformatics tools" as your query.

Exercise 1.25

How many results were returned from this search?

Exercise 1.26

List at least two sites other than an NCBI, PDB, EBI, ExPASy or BCM associated site that contain a collection of bioinformatics tools, and list the sponsoring organizations. (It is acceptable to list pages that contain links to the sites above, as long as they are not sponsored by the organizations listed above.)

As is evident by the number of results returned in this google search, bioinformatics is somewhat unique in that it is one of the rare new branches of science spawned almost entirely on the internet.

[12] http://us.expasy.org/tools/

[13] http://searchlauncher.bcm.tmc.edu/

[14] http://www.google.com/

Chapter 2

Basic Sequence Alignment Lab

2.1 Introduction to Sequence Alignment[1]

Proteins that have a significant biological relationship to one another often share only isolated regions of sequence similarity. For identifying relationships of this nature, the ability to find local regions of optimal similarity is advantageous over global alignments that optimize the overall alignment of two entire sequences. *BLAST*[6], or Basic Local Alignment Search Tool (1), is an alignment tool that uses a measure of local similarity to score sequence alignments in such a way as to identify regions of good local alignment.

The basic BLAST algorithm can be implemented in DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. There are 5 BLAST options on the NCBI web site. BLASTP compares any amino acid query sequence against a protein sequence database. Similarly, BLASTN compares a nucleotide sequence against a nucleotide sequence database. BLASTX takes a nucleotide query sequence and translates it in all reading frames for comparison against a protein sequence database. TBLASTN compares a protein sequence against a nucleotide sequence database, translating the nucleotide database sequences in all possible reading frames. TBLASTX compares translations of a nucleotide query sequence against translations of a nucleotide sequence database. Sequences must be input in one of three formats; FASTA sequence format, NCBI Accession numbers, or GIs (GenBank Identifiers). The FASTA format is the only format where you can input a sequence, instead of a number or code identifying a gene that has already been deposited in GenBank, so this is the format that must be used for nonpublished sequences. Take a moment to view the FASTA format description[2] .

Consider the following nucleotide sequence, which is in FASTA format, but lacks the single-line description that identifies it source:

```
TCAAGCAGATCACTGTCCTTCTGCCATGGCCCTGT
GGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCC
CTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAA
CCAACACCTGTGCGGCTCACACCTGGTGGAAGCTC
TCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTAC
ACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCA
GGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTG
CAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCC
CTGCAGAAGCGTGGCATTGTGGAACAATGCTGTAC
CAGCATCTGCTCCCTCTACCAGCTGGAGAACTACT
GCAACTAGACGCAGCCCGCATGCAGNCCCCCACCC
```

[1] This content is available online at <http://cnx.org/content/m11026/2.13/>.

[2] http://www.ncbi.nlm.nih.gov/BLAST/fasta.html

```
GCCGNCTTCTGCACCGAGAGAGATGGAATTAAACC
CTTGAACCCAGCANANAAAAAAAAGAAATAAAA
```

Access the NCBI BLAST home page [3] , and click on the link for nucleotide BLAST (BLASTN), the standard BLAST for searching a nucleotide database with a nucleotide query sequence. Highlight and copy the FASTA sequence above, and paste it into the first box on the BLAST query page. BLAST allows the option of querying over only a subset of the query sequence, identified by numerical order in the "query subrange" boxes next to the query. If it were desirable to query with only the subsequence beginning with the 50th nucleotide of the example sequence and ending with the 250th nucleotide, then the number 50 would be entered into the "From" box and 250 would be entered into the "To" box. Here in this example the entire sequence should be used as the query, so these boxes should be left blank. The database can be selected under "Choose Search Set". There are many available database choices you can view from the pull down menu. The "nr/nt" database is the largest nucleotide database available through NCBI BLAST; select the "nr/nt" database for this exercise. It includes all GenBank, RefSeq Nucleotides, EMBL (European nucleotide database), DDBJ (Japanese nucleotide database) and PDB (Protein Data Bank) sequences, but no EST, STS, GSS, or phase 0, 1 or 2 htgs (unfinished high throughput genomic) sequences. The NCBI nr database originally got its name from the phrase "nonredundant" nucleotide database, but there is no longer any claim to nonredundancy in the sequence set. Choosing the largest database is not always best; in fact, often the goal is to find a match from a specific organism, for example, a lab that works with flies (*"Drosophila melanogaster"*) as the model organism might be only interested in their organism's sequences. In that case, there is a box entitled Organism, where a user interested can type in *"Drosophila melanogaster"* .

Under program selection, the user can choose to optimize blastn for the type of sequence search that best relates to their problem, for highly similar sequences (megablast), more dissimilar sequences (discontiguous megablast)or somewhat similar sequences (blastn). At the bottom of the web page, there is a link that allows the user to view the default values for formatting (return the descriptions for the top 100 scoring alignments) and for the algorithm parameters. If you selected megablast, optimize for highly similar sequences, the word size used in the search will automatically default to 28 nucleotides. If you select blastn to optimize for somewhat similar sequences the word size will be 11 nucleotides, the historically common default value for blastn. For the purposes of this example, select to optimize for somewhat similar sequences (blastn) and click on the BLAST button to submit your query to the BLAST queue. BLAST should state that the query was successfully submitted, and provide a request ID that accesses the results. The page updates itself automatically, you do not have to refresh it.

The results begin with a graphical overview that uses the query sequence as the rule, shown separately at the top of the graph. In this example, the rule (our query sequence) is 453 nucleotides in length. The score of each alignment is divided into one of 5 ranges by the five different colors in the color key provided, and the alignments are placed in the order of best scores first. The scores are normalized, and normalized scores are given in units of bits, to allow comparisons between alignments. More information on scores can and should be perused in the Altschul tutorial[4] on the NCBI website. Clicking on any of the first 50 lines in the graph takes the user to the associated alignment, located in the last section of this results page.

Exercise 2.1

Look at the graph of best results at the top of the page. How long is the alignment length for the best match in the graph?

The second section of the results page, immediately following the graph, contains the top 100 alignment scores and their sequence descriptions. Here, the "hits", or matched sequences returned as a list of results, contain links to their GenBank entries, their GenBank definitions under the column entitled "Accession", the normalized scores of the alignments, and the E-values. The E-values quantify the probability that the listed alignment might occur randomly. The scores are linked to the actual alignment further down the page. Clicking on a score will take you directly to the alignment that obtained that score. In some cases, there

[3] http://www.ncbi.nlm.nih.gov/BLAST/

[4] http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

are icons containing capital letters in the links column, which are links to related entries in one of the NCBI databases, such as UniGene unique gene cluster information[5] , Entrez Gene [6] and GEO gene expression and hybridization array data [7] . Look at the first 5 "hits" in the results list.

Exercise 2.2
What organism is the most common source of the sequences in the first 5 hits?

Exercise 2.3
What protein is most commonly identified in the description column of the alignments as being associated with or related to these sequences?

The third section of the results page contains the actual alignments, listed in the same order as the graphed results and the scores.

Exercise 2.4
In the first alignment, what is the ratio of identical bases to total bases in the aligned subsequence?

Exercise 2.5
What is the listed percent identity?

Exercise 2.6
What is the score?

Exercise 2.7
What is the E- (Expect) value?

Look for entry with the accession code BC005255.1 at the beginning of the alignment and click on the accession link. Note that this "hit" is a cDNA, or a DNA sequence that has been copied from an mRNA molecule by the enzyme reverse transcriptase. This is indicated in the description, but whenever the source of the sequence is unclear, click the link for the GI number and view the GenBank entry for more information. Look under the "COMMENT" section for additional details about this sequence.

Exercise 2.8
What group was responsible for the DNA sequencing for this entry?

Return to the BLAST alignment results page. Scroll down to view the actual sequence alignments. The alignments are represented with vertical lines illustrating identical matches, blanks indicate no match. In these nucleotide alignments, there is no measurement for similarity between non-identical nucleotides. In protein alignments, however, chemical or structural similarity is usually identified between the amino acids.

Return to the BLAST home page[8] , and select the blastx "translated nucleotide sequence query". Cut and paste the same nucleotide sequence used in the previous example into the "Search" box. The default translation tool is BLASTX, which translates the query sequence into a protein sequence and searches a protein database. To search a translated nucleotide database with a protein query, the TBLASTN algorithm should be selected. To search a translated nucleotide database with a translated nucleotide sequence, the choice would be TBLASTX. TBLASTX is the most computationally intensive of these three searches, therefore common practice is to use BLASTX first to search protein sequence databases, and then if there are no hits with BLASTX, consider TBLASTX to translate nucleotide sequence database entries into protein sequences. For this example, use the translation algorithm BLASTX. Use the nr database and leave the default values for all other BLAST options, and hit BLAST. View the results page.

Exercise 2.9
What protein is listed in the vast majority of the returned matches?

Look at the first alignment, which should have 100% identity between the aligned subsequences. Note that a particular alignment may have more than one sequence associated with it. Therefore, you must look at the actual alignments, not the list of scores, to answer the following questions.

[5] http://www.ncbi.nlm.nih.gov/UniGene/

[6] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

[7] http://www.ncbi.nlm.nih.gov/geo/

[8] http://www.ncbi.nlm.nih.gov/BLAST/

Exercise 2.10

Why is the score lower for the BLASTX result than for the BLASTN result, even though the percent identity between aligned subsequences is 100%?

Exercise 2.11

How many different species are listed as sources where the aligned subsequence has an identical amino acid to total amino acid ratio of 86/86, and 100% identity?

This protein sequence is highly conserved across several species, a trait that is often a sign of physiological importance. Scroll through the protein alignments. Notice that alignment between amino acids is illustrated differently than alignment between nucleotides. An identical match is shown by listing the one letter amino acid code in the middle row, between the two aligned sequences. A mismatch is indicated by a blank, but similarity is indicated by the "+" sign, meaning the amino acids are not identical, but they have some of the same chemical or structural properties. Gaps are indicated by hyphens in the sequence that contains the gaps. Gaps are penalized in an alignment, and cause the normalized score to be lower.

BLAT[33], stands for Blast-like alignment tool (2), and is used for sequence comparisons against an entire genome. Review the difference between BLAT and BLAST at the BLAT FAQ site[9] . Execute a BLAT search, using the sequence above, from the UCSC Genome Bioinformatics Site[10] . Click on the link in this web page entitled "BLAT". Search the human genome, leaving the settings at the default values. The search results will first appear in summary form. Identify the sequence our query most closely matches by the highest score.

Exercise 2.12

What human chromosome (number) contains the sequence our query most closely aligns?

Exercise 2.13

What percent identity exists between our sequence and the aligned subsequence(s)?

Click on the link entitled "details". First, the query sequence is presented, with matching bases capitalized in blue font. Bases in cyan mark the beginning and end of aligned subsequences, indicating a gap in either the reference or the query sequence. Second, the genomic sequence from the selected chromosome is shown, and blue, capitalized bases illustrate the matching region(s). Notice that the query cDNA sequence is missing a large region of DNA that is present in the chromosome. The pairwise alignment is shown below the genomic sequence. Return to the original results summary and click on the link entitled "browser". The missing region is illustrated in graphical form here, where the chromosome band is shown in grey, extending across the graph and the query sequence, labeled at the left as "YourSeq", is in black, below the STS Markers. In fact, the examples in this module use an EST, or an expressed-sequence tag, for the query sequence. ESTs are mRNA-derived representations of the genes expressed in a given tissue and/or at a given developmental stage.

Exercise 2.14

Knowing that our sequence is an EST, how could this explain the region of DNA that appears in the genomic sequence, but not in our query sequence?

If the answer to this question is not intuitive, read the EST section[11] of the NCBI primer. Scroll down to the section entitled "Genes and Gene Prediction Tracks" and change the display options under Ensembl to "pack" and the setting under GenScan to "pack". Next, scroll to the bottom of the page and hit the refresh button. (For a description of the different display options for annotation tracks, read the User Guide[12] . The GenScan predicted genes for this area of the chromosome are shown in brown, while the Ensembl gene predictions are in maroon.

Exercise 2.15

How does our sequence line up with the predicted genes?

[9] http://genome.ucsc.edu/FAQ/FAQblat

[10] http://genome.ucsc.edu/

[11] http://www.ncbi.nlm.nih.gov/About/primer/est.html

[12] http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#IndivTracks

Exercise 2.16

Now, design an independent search in order to answer the following question. What is the GI number and definition for the EST that aligns most closely with our query sequence, according to BLAST?

This has been an introduction to BLAST sequence alignment, designed to give an idea of some of the different ways to optimize searches and problems that require the use of sequence alignment tools. Subsequent modules in this course will build on this foundation, exploring advanced techniques and additional alignment tools.

2.2 PipMaker[13]

A useful graphical tool for viewing sequence alignments, particulary within the genome context, is the percent identity plot, called a pip for short. *PipMaker*[21] (1) is a web server that will perform a sequence alignment on two long DNA segments and produce a pip in pdf format and email the results to the user. Exons of genes and repetitive elements can be labeled along the horizontal axis of the reference DNA sequence, and some coloring options are available for the advanced pip user to clarify and enhance their display. A dot plot is included in the results, illustrating the locations of the query sequence segments in comparison to the reference sequence.

View the PipMaker Home Page[14] . Read the introductory information on this page, then click on the link "PipMaker Instructions" and read the section entitled "Overview". PipMaker requires that the user submit two sequences, and gives the option of submitting additional information. The reference sequence is submitted first, in FASTA format. This is the sequence that will be depicted along the horizontal axis. The query sequence must be provided second. A mask file containing repeat sequences can be submitted; the user generates this file using the RepeatMasker tool, also provided on this web site. The repeats in the mask file are indicated in the plot by various kinds of triangles. Additionally, gene and exon positions can be input by the user, for labeling the locations of exons within their respective genes in the plot. CpG (cytosine-phosphate-guanine) islands[15] in the first sequence are independently determined by PipMaker and are shown as low boxes at the top of the graphical display.

Take a look at the example plot shown on the "Instructions" page. The vertical axis of the plot is the percent identity, ranging from 50% to 100%. Identities below 50% will not be plotted. The current version of PipMaker compares the first sequence with both the second sequence and its reverse complement, so matching regions need not occur in the same orientations and relative positions in the two sequences. Read the section entitled "Input to PipMaker" and become acquainted with the format of the different input options. Return to this instruction page later for help interpreting the output that PipMaker produces, if necessary.

Use your browser's back button to return to the PipMaker Home Page. Click on the PipMaker link, located just below the phrase "The application itself can be found here:". This example will use a section of the sequence from human chromosome 11 as the reference sequence, and an EST query sequence. In the box labeled "First sequence", paste in the following reference sequence:

> Sequence from homo sapiens chromosome 11

```
GAGGACGTGGCTGGGCTCGTGAAGCATGTGGGGGTGAGCCCAGGGGCCCC
AAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGCCTGTCTCCCAG
ATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTG
GCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCA
ACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGG
AACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTG
CAGGGTGAGCCAACTGCCCATTGCTGCCCCTGGCCGCCCCCAGCCACCCC
```

[13] This content is available online at <http://cnx.org/content/m11027/2.5/>.

[14] http://bio.cse.psu.edu/pipmaker/

[15] http://www.bioinfo.de/isb/2003/03/0021/main.html

```
CTGCTCCTGGCGCTCCCACCCAGCATGGGCAGAAGGGGGCAGGAGGCTGC
CACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAGTTCTCTTGGTC
ACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTG
AGGACGGTGTTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACG
CTCCTCCCTCCACTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCC
ACCCTCATTTGATGACCGCAGATTCAAGTGTTTTGTTAAGTAAAGTCCTG
GGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGGGCGAACA
CCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAGTGGGCCAGAC
CCCTGTCGCCAGGCCTCACGGCAGCTCCATAGTCAGGAGATGGGGAAGAT
GCTGGGGACAGGCCCTGGGGAGAAGTACTGGGATCACCTGTTCAGGCTCC
CACTGTGACGCTGCCCCGGGGCGGGGGAAGGAGGTGGGACATGTGGGCGT
TGGGGCCTGTAGGTCCACACCCAGTGTGGGTGACCCTCCCTCTAACCTGG
GTCCAGCCCGGCTGGAGATGGGTGGGAGTGCGACCTAGGGCTGGCGGGCA
GGCGGGCACTGTGTCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCG
CCGCTGTTCCGGAACCTGCTCTGCGCGGCACGTCCTGGCAGTGGGGCAGG
TGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTG
GAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCAT
CTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAG
GCAGCCCCACACCCGCCGCCTCCTGCACCGAGAGAGATGGAATAAAGCCC
TTGAACCAGCCCTGCTGTGCCGTCTGTGTGTCTTGGGGGCCCTGGGCCAA
GCCCCACTTCCCGGCACTGTTGTGAGCCCCTCCCAGCTCTCTCCACGCTCTCTGGGT
```

In the box labeled "Second sequence", paste in the following query sequence:

```
TCAAGCAGATCACTGTCCTTCTGCCATGGCCCTGT
GGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCC
CTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAA
CCAACACCTGTGCGGCTCACACCTGGTGGAAGCTC
TCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTAC
ACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCA
GGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTG
CAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCC
CTGCAGAAGCGTGGCATTGTGGAACAATGCTGTAC
CAGCATCTGCTCCCTCTACCAGCTGGAGAACTACT
GCAACTAGACGCAGCCCGCATGCAGNCCCCCACCC
GCCGNCTTCTGCACCGAGAGAGATGGAATTAAACC
CTTGAACCCAGCANANAAAAAAAAGAAATAAAA
```

Next, supply an email address in the appropriate box, so that PipMaker can email the results. The box entitled "First sequence mask" will be left empty for this example. In fact, if the reference sequence is submitted to RepeatMasker, it responds that there are no repeats detected in this sequence. Feel free to try this, there is a link to RepeatMasker on the PipMaker Instruction page. In the box entitled "First sequence exons", paste in the following information, which specifies exons that correspond roughly to the Genescan Gene Predictions for the reference sequence (these were illustrated in BLAT in a previous module).

< 100 1307 PUTATIVE INSULIN PRECURSOR

```
100 304
1091 1307
```

Now, click on the submit button. Pipmaker will email several documents to you, two of which will be a "pip.pdf" and a "dot.pdf". View both of these figures (you may need to zoom in).

Exercise 2.17

What do you estimate the percent identities to be between (a) the query subsequence under Exon 1 and the reference subsequence, and (b) the query subsequence under Exon 2 and the reference subsequence?

Exercise 2.18

Which sequence is represented by the vertical axis of the dot plot, the query sequence (the EST) or the reference sequence (from chromosome 11)?

Pipmaker was designed to be used with sequences that can be much longer than these examples. Look under the input to PipMaker section of the "PipMaker Instructions" page.

Exercise 2.19

What is the maximum allowed length for the first (reference) sequence file?

Exercise 2.20

Is this the same as the maximum allowed length for the second (query) sequence file?

Feel free to try the PipMaker tool with other sets of related sequences. The PipMaker home page provides a set of advanced instructions for those who will be using this tool frequently for publications and reports.

Chapter 3

Multiple Sequence Alignment Lab

3.1 Multiple Sequence Alignment[1]

Sequence alignments can be used to study the relationship(s) between sequences in sets of more than two sequences. This application is particularly useful when studying the relationships between a similar type of gene product that is expressed by different organisms, like analyzing CFTR sequences from several different species, or when studying similar, yet divergent, sequences within the same organism, as the variance in troponin I isoforms in Homo sapiens.

Often, a primary focus of a multiple sequence alignment is to identify, within several related sequences, regions that are highly conserved in identity or similarity, and therefore probably have functional and/or structural significance. Many factors affect the analysis of conserved regions within related sequences, such as the number of sequences included in the analysis, and the ratio of the number of very similar (almost identical) sequences to the number of more distantly related sequences. Divergent sequences can cause problems in a multiple sequence alignment. It is more difficult to identify the correct alignment when two sequences that are related throughout part of the sequence also contain large sections that diverge. Therefore, the error rates in the alignment increase as divergence increases. These errors in the alignment can cause the related part of the sequences to show lower similarity than they actually have, and this sort of error is often amplified in subsequent steps. *ClustalW*[42] (1), a commonly used multiple sequence alignment program, addresses the problems associated with alignment of divergent sequences in several ways. Individual weights are assigned to each sequence in a partial alignment such that near-duplicate sequences are down-weighted and divergent sequences are up-weighted. Also, the amino acid substitution matrices at different alignment stages are chosen according to the divergence of the sequences to be aligned. Residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions increase the penalty for opening new gaps in regions of regular secondary structure. Therefore, it increases the likelihood that gaps will occur in loop regions than in highly structured regions such as alpha helices and beta sheets. Highly structured regions are commonly very important to the fold and function of a protein, and so divergence is often biologically less likely in these areas. For similar reasons, existing gaps receive locally reduced gap penalties to encourage the opening up of new gaps near the existing ones. These features have been designed into ClustalW to produce multiple sequence alignments that are biologically meaningful.

First, use the example of analyzing the variance in troponin I isoforms in Homo sapiens to get acquainted with multiple sequence alignments with ClustalW. ClustalW[2] is available at many bioinformatics web sites, but the EBI site is chosen here, for its nice interface and graphical display. Accept the default values on the submission form, and scroll down the page until the box for pasting in the sequences to be aligned appears. Now, the troponin I sequences must be obtained. For the sake of time, instead of searching with the string "troponin I human", the pertinent accession codes are supplied for this exercise. Open a new browser window

[1] This content is available online at <http://cnx.org/content/m11036/2.11/>.

[2] http://www.ebi.ac.uk/clustalw/

to the NCBI home page[3] . Locate the "Search" box at the top of the page and select "Protein" to search the protein database. Type in the accession code TNNI3_HUMAN in the box to the right, and click on "Go". The results of this search will become part of the search history. Select the history link from the menu under the query box to verify this. Now, search "Protein" again, using the accession code TNNI2_HUMAN, then perform one last search using the accession code TNNI1_HUMAN. As an illustration of the search history function in Entrez, combine the results of these three searches. Do this using the boolean operator "OR" in between the number assigned to each search in the history listing. For instance, if the Search History gave the following list,

```
#3 Search TNNI1_HUMAN  17:44:55
#2 Search TNNI2_HUMAN  17:44:49
#1 Search TNNI3_HUMAN  17:41:47
```

then "#1 OR #2 OR #3" would be the search string. The boolean operator OR is specifying that all results included in any of the above searches be combined. Using the boolean AND operator would specify only results that are common to all three searches, which in this case would yield zero results. To view the results, click on "GO" once you have entered the search string into the query box. Now, the combined search should return all three of the desired troponin I isoforms. Check the box to the left of each accession code for all three results. Just above the list of results, find the region on the left that states "Display Summary" and change "Summary" to "FASTA". This should yield the amino acid sequences in FASTA format for the 3 proteins that were selected. Copy and paste each sequence into the ClustalW window, pressing return after pasting each entry. There does not have to be a line space between entries, as long as the FASTA identifying line starts on a new line. However, it does not hurt to have a line space between entries, either. Note that the numbered search description line from Entrez is not part of the FASTA format. The numbered search description line will look something like this,

```
1:  P19237. Reports  Troponin I, slow ...[gi:1351298]
```

and should be omitted from the ClustalW search query. Also, check that the description header for the sequence in FASTA format does not take up more than one line. If it does, shorten the description so that it only occupies one line. Press the "Run" button, and click on the link for the browser window that displays the results, if this option is given. A text page will appear with the alignment scores.

Exercise 3.1
What is the score for the alignment of TNNI3 to TNNI2?

Exercise 3.2
What is the score for the alignment of TNNI3 to TNNI1?

Exercise 3.3
What is the score for the alignment of TNNI2 to TNNI1?

At the top of the page, a table contains several choices. There a link to the home page for the alignment editor Jalview,there is a button, labeled "Start Jalview", which is a link to the multiple sequence alignment results in the graphical format produced by the Jalview editor. Choose the "Start Jalview" button to view the alignment results. The sequences are labeled at the left and a bar graph below the sequences indicates which regions have high conservation or similarity. The cardiac isoform of troponin I is roughly 30 amino acids longer than the other two isoforms.

Exercise 3.4
Looking at the aligned sequences, in what region are the extra amino acids located in cardiac troponin I?

[3] http://www.ncbi.nlm.nih.gov/

Take a look at the color scheme of the alignment. Regions of the sequence with identical amino acids are colored the same, but notice that regions of the sequence with similar amino acids have also been colored alike. Usually these color schemes have groups consisting of hydrophobic amino acids, hydrophilic amino acids, positively-charged amino acids, and negatively-charged amino acids. Sometimes these large color groups are divided further to distinguish between sidechain lengths, the aromatic sidechains, glycines and prolines.

Try one more multiple sequence alignment, analyzing CFTR sequences from several different species. Return to the ClustalW[4] submission page. Cut and paste the CFTR sequences listed below from three species into the query submission box and run the multiple alignment.

```
>gi|203423|gb|AAA40918.1| cftr [Rattus norvegicus]
IKHSGRVSFSSQISWIMPGTIKENIIFGVSYDEYRYKSVVKACQLQEDITKFAEQDNTVLGEGGVTLSGG
QRARISLARAVYKDADLYLLDSPFGYLDVLTEEQIFESCVCKLMASKTRILVTSKMEQLKKADKILILHE
GSSYFYGTFSELQSLRPDFSSKLMGYDTFDQFTEERRSSILTETLRRFSVDDASTTWNKAKQSFRQTGEF
GEKRKNSILSSFSSVKKISIVQKTPLSIEGESDDLQERRLSLVPDSEHGEAALPRSNMITAGPTFPGRRR
QSVLDLMTFTPSSVSSSLQRTRASIRKISLAPRISLKEEDIYSRRLSQDSTLNITEEINEEDLKECFFDD
MVKIPTVTTWNTYLRYFTLHRGLFAVLIWCVLVFLVEVAASLFVLWLLKNNPVNGGNNGTKIANTSYVVV
ITSSSFYYIFYIYVGVADTLLALSLFRGLPLVHTLITASKILHRKMLHSILHAPMSTFNKLKAGGILNRF
SKDIAILDDFLPLTILT
>gi|192832|gb|AAA18903.1| cftr[Mus musculus]
MQKSPLEKASFISKLFFSWTTPILRKGYRHHLELSDIYQAPSADSADHLSEKLEREWDREQASKKNPQLI
HALRRCFFWRFLFYGILLYLGEVTKAVQPVLLGRIIASYDPENKVERSIAIYLGIGLCLLFIVRTLLLHP
AIFGLHRIGMQMRTAMFSLIYKKTLKLSSRVLDKISIGQLVSLLSNNLNKFDEGLALAHFIWIAPLQVTL
LMGLLWDLLQFSAFCGLGLLIILVIFQAILGKMMVKYRDQRAAKINERLVITSEIIDNIYSVKAYCWESA
MEKMIENLREVELKMTRKAAYMRFFTSSAFFFSGFFVVFLSVLPYTVINGIVLRKIFTTISFCIVLRMSV
TRQFPTAVQIWYDSFGMIRKIQDFLQKQEYKVLEYNLMTTGIIMENVTAFWEEGFGELLEKVQQSNGDRK
HSSDENNVSFSHLCLVGNPVLKNINLNIEKGEMLAITGSTGSGKTSLLMLILGELEASEGIIKHSGRVSF
CSQFSWIMPGTIKENIIFGVSYDEYRYKSVVKACQLQQDITKFAEQDNTVLGEGGVTLSGGQRARISLAR
AVYKDADLYLLDSPFGYLDVFTEEQVFESCVCKLMANKTRILVTSKMEHLRKADKILILHQGSSYFYGTF
SELQSLRPDFSSKLMGYDTFDQFTEERRSSILTETLRRFSVDDSSAPWSKPKQSFRQTGEVGEKRKNSIL
NSFSSVRKISIVQKTPLCIDGESDDLQEKRLSLVPDSEQGEAALPRSNMIATGPTFPGRRRQSVLDLMTF
TPNSGSSNLQRTRTSIRKISLVPQISLNEVDVYSRRLSQDSTLNITEEINEEDLKECFLDDVIKIPPVTT
WNTYLRYFTLHKGLLLVLIWCVLVFLVEVAASLFVLWLLKNNPVNSGNNGTKISNSSYVVIITSTSFYYI
FYIYVGVADTLLALSLFRGLPLVHTLITASKILHRKMLHSILHAPMSTISKLKAGGILNRFSKDIAILDD
FLPLTIFDFIQLVFIVIGAIIVVSALQPYIFLATVPGLVVFILLRAYFLHTAQQLKQLESEGRSPIFTHL
VTSLKGLWTLRAFRRQTYFETLFHKALNLHTANWFMYLATLRWFQMRIDMIFVLFFIVVTFISILTTGEG
EGTAGIILTLAMNIMSTLQWAVNSSIDTDSLMRSVSRVFKFIDIQTEESMYTQIIKELPREGSSDVLVIK
NEHVKKSDIWPSGGEMVVKDLTVKYMDDGNAVLENISFSISPGQRVGLLGRTGSGKSTLLSAFLRMLNIK
GDIEIDGVSWNSVTLQEWRKAFGVITQKVFIFSGTFRQNLDPNGKWKDEEIWKVADEVGLKSVIEQFPGQ
LNFTLVDGGYVLSHGHKQLMCLARSVLSKAKIILLDEPSAHLDPITYQVIRRVLKQAFAGCTVILCEHRI
EAMLDCQRFLVIEESNVWQYDSLQALLSEKSIFQQAISSSEKMRFFQGRHSSKHKPRTQITALKEETEEE
VQETRL
>gi|6862589|gb|AAF30300.1|AAF30300 cftr [Mus musculus]
MQKSPLEKASFISKLFFSWSTAILRKGYRQHLELSDIYQAPSADSADHLSEKLEREWDREQASKKNPQLI
HALRRCFFWRFLFYGILLYLGEVTKAVQPVLLGRIIASYDPENKVERSIAIYLGIGLCLLFIVRTLLLHP
AIFGLHRIGMQMRTAMFSLIYKKTLKLSSRVLDKISIGQLVSLLSNNLNKFDEGLALAHFIWIAPLQVTL
LMGLLWDLLQFSAFCGLGLLIILVIFQAILGKMMVKYRDQRAAKINERLVITSEIIDNIYSVKAYCWESA
MEKMIENLREVELKMTRKAAYMRFFTSSAFFFSGFFVVFLSVLPYTVINGIVLRKIFTTISFCIVLRMSV
TRQFPTAVQIWYDSFGMIRKIQDFLQKQEYKVLEYNLMTTGIIMENVTAFWEEGFGELLEKVQQSNGDRK
HSSDENNVSFSHLCLVGNPVLKNINLNIEKGEMLAITGSTGSGKTSLLMLILGELEASEGIIKHSGRVSF
```

[4] http://www.ebi.ac.uk/clustalw/

```
CSQFSWIMPGTIKENIIFGVSYDEYRYKSVVKACQLQQDITKFAEQDNTVLGEGGVTLSGGQRARISLAR
AVYKDADLYLLDSPFGYLDVFTEEQVFESCVCKLMANKTRILVTSKMEHLRKADKILILHQGSSYFYGTF
SELQSLRPDFSSKLMGYDTFDQFTEERRSSILTETLRRFSVDDSSAPWSKPKQSFRQTGEVGEKRKNSIL
NSFSSVRKISIVQKTPLCIDGESDDLQEKRLSLVPDSEQGEAALPRSNMIATGPTFPGRRRQSVLDLMTF
TPNSGSSNLQRTRTSIRKISLVPQISLNEVDVYSRRLSQDSTLNITEEINEEDLKECFLDDVIKIPPVTT
WNTYLRYFTLHKGLLLVLIWCVLVFLVEVAASLFVLWLLKNNPVNSGNNGTKISNSSYVVIITSTSFYYI
FYIYVGVADTLLALSLFRGLPLVHTLITASKILHRKMLHSILHAPMSTISKLKAGGILNRFSKDIAILDD
FLPLTIFDFIQLVFIVIGAIIVVSALQPYIFLATVPGLVVFILLRAYFLHTAQQLKQLESEGRSPIFTHL
VTSLKGLWTLRAFRRQTYFETLFHKALNLHTANWFMYLATLRWFQMRIDMIFVLFFIVVTFISILTTGEG
EGTAGIILTLAMNIMSTLQWAVNSSIDTDSLMRSVSRVFKFIDIQTEESMYTQIIKELPREGSSDVLVIK
NEHVKKSDIWPSGGEMVVKDLTVKYMDDGNAVLENISFSISPGQRVGLLGRTGSGKSTLLSAFLRMLNIK
GDIEIDGVSWNSVTLQEWRKAFGVITQKVFIFSGTFRQNLDPNGKWKDEEIWKVADEVGLKSVIEQFPGQ
LNFTLVDGGYVLSHGHKQLMCLARSVLSKAKIILLDEPSAHLDPITYQVIRRVLKQAFAGCTVILCEHRI
EAMLDCQRFLVIEESNVWQYDSLQALLSEKSIFQQAISSSEKMRFFQGRHSSKHKPRTQITALKEETEEE
VQETRL
>gi|1669377|gb|AAB46340.1| cftr [Homo sapiens]
LLLIVIGAIAVVAVLQPYIFVATVPVIVAFIMLRAYFLQTSQQLKQLESEGRSPIFTHLVTSLKGLWTLR
AFGRQPYFETLFHKALNLHTANWFLYLSTLRWFQMRIEMIFVIFFIAVTFISILTTGEGEGRVGIILTLA
MNIMSTLQWAVNSSIDVDSLMRSVSRVFKFIDMPTEGKPTKSTKPYKNGQLSKVMIIENSHVKKDDIWPS
GGQMTVKDLTAKYTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLLNTEGEIQIDGVSWDS
ITLQQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEVGLRSVIEQFPGKLDFVLVDGGCVL
SHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQIIRRTLKQAFADCTVILCEHRIEAMLECQQFLVI
EENKVRQYDSIQKLLNERSLFRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL
>gi|180332|gb|AAA35680.1| cftr [Homo sapiens]
MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLSEKLEREWDRELASKKNPKLI
NALRRCFFWRFMFYGIFLYLGEVTKAVQPLLLGRIIASYDPDNKEERSIAIYLGIGLCLLFIVRTLLLHP
AIFGLHHIGMQMRIAMFSLIYKKTLKLSSRVLDKISIGQLVSLLSNNLNKFDEGLALAHFVWIAPLQVAL
LMGLIWELLQASAFCGLGFLIVLALFQAGLGRMMMKYRDQRAGKISERLVITSEMIENIQSVKAYCWEEA
MEKMIENLRQTELKLTRKAAYVRYFNSSAFFFSGFFVVFLSVLPYALIKGIILRKIFTTISFCIVLRMAV
TRQFPWAVQTWYDSLGAINKIQDFLQKQEYKTLEYNLTTTEVVMENVTAFWEEGFGELFEKAKQNNNNRK
TSNGDDSLFFSNFSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMGELEPSEGKIKHSGRISF
CSQFSWIMPGTIKENIIFGVSYDEYRYRSVIKACQLEEDISKFAEKDNIVLGEGGITLSGGQRARISLAR
AVYKDADLYLLDSPFGYLDVLTEKEIFESCVCKLMANKTRILVTSKMEHLKKADKILILNEGSSYFYGTF
SELQNLQPDFSSKLMGCDSFDQFSAERRNSILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGEKRKNS
ILNPINSIRKFSIVQKTPLQMNGIEEDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVL
NLMTHSVNQGQNIHRKTTASTRKVSLAPQANLTELDIYSRRLSQETGLEISEEINEEDLKECLFDDMESI
PAVTTWNTYLRYITVHKSLIFVLIWCLVIFLAEVAASLVVLWLLGNTPLQDKGNSTHSRNNSYAVIITST
SSYYVFYIYVGVADTLLAMGFFRGLPLVHTLITVSKILHHKMLHSVLQAPMSTLNTLKAGGILNRFSKDI
AILDDLLPLTIFDFIQLLLIVIGAIAVVAVLQPYIFVATVPVIVAFIMLRAYFLQTSQQLKQLESEGRSP
IFTHLVTSLKGLWTLRAFGRQPYFETLFHKALNLHTANWFLYLSTLRWFQMRIEMIFVIFFIAVTFISIL
TTGEGEGRVGIILTLAMNIMSTLQWAVNSSIDVDSLMRSVSRVFKFIDMPTEGKPTKSTKPYKNGQLSKV
MIIENSHVKKDDIWPSGGQMTVKDLTAKYTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRL
LNTEGEIQIDGVSWDSITLQQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEVGLRSVIEQ
FPGKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQIIRRTLKQAFADCTVILC
EHRIEAMLECQQFLVIEENKVRQYDSIQKLLNERSLFRQAISPSDRVKLFPHRNSSKCKSKPQIAALKEE
TEEEVQDTRL
>gi|1809238|gb|AAB46352.1| cftr [Homo sapiens]
MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLSEKLEREWDRELASKKNPKLI
NALRRCFFWRFMFYGIFLYLGEVTKAVQPLLLGRIIASYDPDNKEERSIAIYLGIGLCLLFIVRTLLLHP
AIFGLHHIGMQMRIAMFSLIYKKTLKLSSRVLDKISIGQLVSLLSNNLNKFDEGLALAHFVWIAPLQVAL
LMGLIWELLQASAFCGLGFLIVLALFQAGLGRMMMKYRDQRAGKISERLVITSEMIENIQSVKAYCWEEA
```

```
MEKMIENLRQTELKLTRKAAYVRYFNSSAFFFSGFFVVFLSVLPYALIKGIILRKIFTTISFCIVLRMAV
TRQFPWAVQTWYDSLGAINKIQDFLQKQEYKTLEYNLTTTEVVMENVTAFWEEGFGELFEKAKQNNNNRK
TSNGDDSLFFSNFSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMVIMGELEPSEGKIKHSGRISF
CSQFSWIMPGTIKENIIFGVSYDEYRYRSVIKACQLEEDISKFAEKDNIVLGEGGITLSGGQRARISLAR
AVYKDADLYLLDSPFGYLDVLTEKEIFESCVCKLMANKTRILVTSKMEHLKKADKILILHEGSSYFYGTF
SELQNLQPDFSSKLMGCDSFDQFSAERRNSILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGEKRKNS
ILNPINSIRKFSIVQKTPLQMNGIEEDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVL
NLMTHSVNQGQNIHRKTTASTRKVSLAPQANLTELDIYSRRLSQETGLEISEEINEEDLKECFFDDMESI
PAVTTWNTYLRYITVHKSLIFVLIWCLVIFLAEVAASLVVLWLLGNTPLQDKGNSTHSRNNSYAVIITST
SSYYVFYIYVGVADTLLAMGFFRGLPLVHTLITVSKILHHKMLHSVLQAPMSTLNTLKAGGILNRFSKDI
AILDDLLPLTIFDFIQ
```

Before looking at the alignment in Jalview, look at the text results.

Exercise 3.5

Which 2 sequences align with the highest pairwise alignment score?

Look at the text display of the alignments, shown just below the scores. Try to determine the regions where the 6 sequences align most closely. Many people find this type of text display hard to interpret (a fact that is good to remember when putting together a figure for a presentation or an article). At the bottom of the text display is a cladogram. A cladogram is a branched phylogenetic-type tree where the branches are of equal length. Cladograms can show common ancestry, but because the branches have equal lengths, they do not provide an accurate indication of the evolutionary distance between the branches.

Exercise 3.6

Does the cladogram group together all the human CFTR sequences in the same branch?

Exercise 3.7

Now, look at the alignment in Jal view. Is it easier to identify the regions where all 6 sequences align most closely in the Jalview display?

Leave both of these windows open (closing the ClustalW text results browser page may cause the Jalview editor page to disappear automatically). Notice the file number assigned this Jalview sequence alignment. This is to help distinguish it for purposes of comparison from the next and last alignment in this tutorial, where the gap penalty will be manipulated from the default values.

Finally, repeat the CFTR alignment using the sequences above, but this time change "Gap Open" to 100. (Click on the term "Gap Open" to view the default value.) This will increase the penalty for opening a gap. For more information on gaps and gap penalties, view the EBI help page entitled "About Gaps"[5] . Compare the group alignment scores from the previous alignment with these group alignment scores. The group alignment scores can be found in the .output file (listed in the results section) under the section which starts:

```
Start of Multiple Alignment
There are 5 groups
Aligning...
```

Exercise 3.8

Does changing the gap open penalty affect the group alignment scores?

Exercise 3.9

Are the group scores higher or lower when the penalty is increased to 100?

View the Jalview display of the new alignment. Compare the Jalview displays of the first alignment, and the new alignment with the increased gap penalty, in the regions numbered 1220 to 1250.

[5] http://www.ebi.ac.uk/help/gaps_frame.html

Exercise 3.10

Describe the differences in the two alignments in this region, both in terms of the gaps, and in terms of the similarity bar graph displayed below the sequences.

A look at the ClustalW submission page is enough to see that there are many more parameters that can be manipulated by the user. However, the default values have been optimized and generally should be left alone, except in cases where the user has a definitive justification for making the change. It does frequently occur in biology, though, that the user knows some empirical information that conflicts with the results given by an alignment performed with the default values. In these cases, a scientific argument can be made for altering the parameters to force the alignment to reflect the empirical information.

Chapter 4

Substitution & Scoring Matrices

4.1 Scoring Matrices[1]

In bioinformatics, scoring matrices for computing alignment scores are often based on observed substitution rates, derived from the substitution frequencies seen in multiple alignments of sequences. Every possible identity and substitution is assigned a score based on the observed frequencies of such occurences in alignments of related proteins. The score is calculated from the frequency of occurrence of a match of the two individual amino acids in evolutionarily related sequences, and provides a measure of a chance alignment of the two amino acids. This score will also reflect the frequency that a particular amino acid occurs in nature, as some amino acids are more abundant than others. Higher scores indicate that the probability that those two amino acids aligned by chance is very small, and lower scores indicate a high probability the two amino acids aligned by chance, and are evolutionarily unrelated. Thus, identities are assigned the most positive scores, frequently observed substitutions also receive positive scores, but matches that are unlikely to have been a result of evolution, and are more likely indicative of unrelatedness at that position, are given negative scores. Matrices with scoring schemes based on observed substitution rates are superior to simple identity scores, or scores based solely on sidechain moiety similarity. The two most commonly used types of scoring matrices are the *PAM matrices*[41] and the *BLOSUM matrices*[28].

PAM (Percentage of Acceptable point Mutations per 10^8 years) matrices are based on global alignments of closely related proteins. The PAM 1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Scores are derived from a mutation probability matrix where each element gives the probability of the amino acid in column X mutating to the amino acid in row Y after a particular evolutionary time, for example after 1 PAM, or 1% divergence. A PAM matrix is specific for a particular evolutionary distance, but may be used to generate matrices for greater evolutionary distances by multiplying it repeatedly by itself. However, at large evolutionary distances the information present in the matrix is essentially degenerated. It is rare that a PAM matrix would be used for an evolutionary distance any greater than 256 PAMs.

Whereas the PAM matrices have been developed from global alignments, the BLOSUM (BLOcks SUbstitution Matrix) matrices are based on local multiple alignments of more distantly related sequences. For instance, BLOSUM 62, the default matrix in BLAST, is a matrix calculated from comparisons of sequences with no less than 62% identity. Unlike PAM matrices, new BLOSUM matrices are never extrapolated from existing BLOSUM matrices, but are always based on local multiple alignments. So, the BLOSUM 80 matrix would be derived from a set of sequences having 80% sequence identity.

The level of relatedness of a set of sequences, therefore, directly effects which scoring matrix is most appropriate for aligning the set, whether or not it is a PAM or a BLOSUM matrix. Comparisons of closely related sequences should use BLOSUM matrices with higher numbers and PAM matrices with lower numbers. Conversely, BLOSUM matrices with low numbers and PAM matrices with high numbers are preferable for

[1] This content is available online at <http://cnx.org/content/m11062/2.7/>.

comparisons of distantly related proteins. Nevertheless, a single matrix may be reasonably efficient over a relatively broad range of evolutionary change. The BLOSUM 62 matrix was chosen as the default for BLAST as a result of an analysis by *Henikoff and Henikoff*[29] wherein BLOSUM 62 detected more distant relationships in a BLAST search, and produced an alignment of diverged proteins more in agreement with three-dimensional structures, than did the corresponding PAM 60 matrix. The BLOSUM series does not include any matrices suitable for very short query sequences, so, in these cases, the PAM matrices may be used instead. Berkeley has a Matrix Information [2] website with a provisional table of recommended substitution matrices and gap costs for shorter sequences.

Now, take a look at some scoring matrices. A PAM Matrix[3] website sponsored by Wageningen University, in the Netherlands, allows online computation of PAM matrices. The default value is a PAM 250 matrix; calculate this matrix and look at the results. This PAM 250 matrix has a built-in gap penalty of -8, as seen in the * column. There are 24 rows and 24 columns. Of course, the first 20 are the amino acids, represented by the one letter code. B represents the case where there is ambiguity between aspartate or asparigine, and Z is the case where there is ambiguity between glutamate or glutamine. X represents an unknown, or nonstandard amino acid.

Exercise 4.1
In the PAM 250 matrix, where can the highest scores for each amino acid be found? Why?

Exercise 4.2
Would this be true for any scoring matrix?

Exercise 4.3
What row and column combination gives the highest score? (Specify the score value.)

Exercise 4.4
What is the second highest score? (Specify the score value.)

Exercise 4.5
Why are some scores for amino acid identities higher than others?

Exercise 4.6
Use the back button on the browser, and calculate a PAM 100 matrix. Are the two highest scoring matches the same combination of row and column as in the PAM 250 matrix? (Discuss with a sentence or two.)

Exercise 4.7
What is the gap penalty?

Exercise 4.8
Explain any differences in the gap penalties of the PAM 250 matrix versus the PAM 100 matrix.

To get an idea how the scoring matrix influences an alignment, perform the following exercise using the Biology Workbench[4] . The Workbench will require a password (it's free), but it will grant entrance immediately upon registration of a password. Enter the site, and scroll down the page until the five menu buttons are visible. The "Session Tools" button allows the naming of a session, so that different jobs in progress can be saved under distinct sessions. Select "Session Tools", then select "Start New Session" and click on "Run" to change the name of "Default Session" to a new name. Once the workbench has been exited, the session will remain. Subsequently, clicking on the dot to the left of the session name under the "Session Tools" menu, and then selecting "Resume Session", will recall the session. The Workbench policy at the time of this writing is that old jobs are deleted only when an account has not been accessed for 6 months. This tutorial will use sequences of hemoglobins (Hbs) from different organisms to illustrate the properties of scoring matrices. Choose the "Protein Tools" menu button, then choose the "Ndjinn Multiple Database Search" from the menu at the bottom of the page. Biology Workbench has a large number of databases to search, for this exercise, click in the box to left of the database description to choose the "PDBFINDER"

[2] http://mcb.berkeley.edu/labs/king/blast/docs/matrix_info.html
[3] http://www.bioinformatics.nl/tools/pam.html
[4] http://workbench.sdsc.edu/

database. Search the PDBFINDER database by typing in the PDB ID codes below into the search box at the top of the page. Import the sequences with the following PDB ID codes (use the OR operator between each PDB ID code to search for all of the records in the same search):

1. 1T1N from *trematomus newnesi* (antarctic fish)
2. 1SPG from *leiostomus xanthurus* (spot croaker)
3. 1QSI from *homo sapiens* (human)
4. 1IWH from *equus cabullus* (horse)
5. 1HV4 from *anser indicus* (goose)
6. 1HBR from *gallus gallus* (chicken)
7. 1H97 from *paramphistomum epiclitum* (trematode)
8. 1GVH from *escherichia coli* (enterobacteria)

The import function in the Workbench requires checking the boxes for all the PDB ID codes that were returned, then hitting the import button. There will be several subunits returned with most of these sequences, and some are duplicate sequences, so delete the following chains by clicking the box on the left of the ID code and selecting "Delete Protein Sequence(s)" from the pull-down menu at the bottom of the page:

1. 1HV4_C
2. 1HV4_D
3. 1HV4_E
4. 1HV4_F
5. 1HV4_G
6. 1HV4_H
7. 1HBR_C
8. 1HBR_D
9. 1H97_B
10. 1QSI_C
11. 1QSI_D

After the above sequences have been deleted, choose "Select All Sequence(s)" from the pull-down menu. Analyze the relatedness of this group of sequences by selecting "ClustalW" from the pull-down menu to perform a multiple alignment and draw a rooted cladogram. When the ClustalW page appears, before submitting the alignment, scroll down the page and change the "Guide tree display:" to "Rooted".

When the ClustalW results appear, first scroll down to the cladogram and observe which of these sequences are most closely related versus the more distant sequences. Notice there are three separate clusters of branches descending from the root. The two largest clusters are separated as a direct result of a structural characteristic of hemoglobins.

Exercise 4.9
What do each of these two clusters represent? (If the answer is not immediately clear, read this description of Hemoglobin[5] from the University of Brescia's on-line Biochemistry Course.)

Exercise 4.10
According to this cladogram, what is the sequence that is most closely related to human hemoglobin, ID code 1QSI?

Exercise 4.11
According to this cladogram, what is the sequence that is most closely related to the *E. coli* flavoHb, ID code 1GVH?

Exercise 4.12
According to this cladogram, what sequence is most closely related to the spot croaker Hb, ID code 1SPG?

[5] http://www.med.unibs.it/~marchesi/hemoglob.html#hemoglobin

It is not as clear from the cladogram which sequences are the most distantly related. However, scroll down past the cladogram to view the ClustalW pairwise alignment scores.

Exercise 4.13

Which two sequences yield the lowest pairwise alignment score?

At the very bottom of the alignment page, select "Import Alignments", to save this information for later reference, should that be necessary. The imported alignments can only be viewed through the "Alignment Tools" menu.

Apply the information elucidated by the multiple sequence alignment to test the impact of varying the scoring matrices in pairwise alignments. Return to "Protein Tools". Start with two sequences that are known to be closely related, the human Hb chain B, 1QSI_B, and the horse Hb chain B, 1IWH_B, by checking the box to the left of each of their codes. Choose "LALIGN" from the pull-down menu at the bottom of the page to compare two protein sequences to each other with BLAST. When the LALIGN page appears, next to select scoring matrix, choose PAM250 and run the alignment.

Exercise 4.14

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Now, return to "Protein Tools" and run LALIGN again with the same two sequences, 1QSI_B and 1IWH_B, except choose the "PAM120" matrix this time.

Exercise 4.15

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 4.16

(a) Which scoring matrix yielded the highest score for the alignment, and why is this matrix the best choice for this alignment? (b) List any regions where the two alignments differ.

Return to "Protein Tools", this time selecting the 1HV4_A and the 1TIN_B sequences, by checking the box next to their codes. Again, choose "LALIGN", and perform an alignment with the default PAM250 matrix.

Exercise 4.17

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 4.18

Run LALIGN again on the same two sequences, using the PAM120 matrix. What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 4.19

(a) Which scoring matrix yielded the highest score for the alignment, and why is this matrix the best choice for this alignment? (b) List any regions where the two alignments differ.

Exercise 4.20

Do the two different matrices always calculate the same value for percent identity when the same 2 sequences are being compared using each matrix? Why or why not?

Most bioinformatics tools available on the web have selected default scoring matrices that are based on a relatively exhaustive analysis of which scoring schemes work best over a wide range of query sequence characteristics. However, it is important to not only know which scoring matrix is used for a given alignment, but to consider the appropriateness of the default matrix for a given query as well. It is a recurring theme of bioinformatics that these computational tools should not be treated as "black boxes" where one can ignore the internal workings of the software, but instead require thoughtful interaction on the part of the user.

Chapter 5

Phylogenetics

5.1 Phylogenetic Trees[1]

A phylogenetic tree is a graphical representation of the evolutionary relationship between taxonomic groups. The term phylogeny refers to the evolution or historical development of a plant or animal species, or even a human tribe or similar group. Taxonomy is the system of classifying plants and animals by grouping them into categories according to their similarities. A phylogenetic tree is a specific type of cladogram where the branch lengths are proportional to the predicted or hypothetical evolutionary time between organisms or sequences. Cladograms are branched diagrams, similar in appearance to family trees, that illustrate patterns of relatedness where the branch lengths are not necessarily proportional to the evolutionary time between related organisms or sequences. Bioinformaticians produce cladograms representing relationships between sequences, either DNA sequences or amino acid sequences. However, cladograms can rely on many types of data to show the relatedness of species. In addition to sequence homology information, comparative embryology, fossil records and comparative anatomy are all examples of the types of data used to classify species into phylogenic taxa. So, it is important to understand that the cladograms generated by bioinformatics tools are primarily based on sequence data alone. Given that, it is also true that sequence relatedness can be very powerful as a predictor of the relatedness of species.

Cladograms cannot be considered completely true and accurate descriptions of the evolutionary history of organisms, because in any cladogram there are a number of possible evolutionary pathways that could produce the pattern of relatedness illustrated in the cladogram. The cladogram only illustrates the probability that two organisms, or sequences, are more closely related to each other than to a third organism, it does not necessarily clarify the pathway that created the existing relationships. However, the cladogram can be used in the formulation of new hypotheses and to cast new light on existing data. One of the most ambitious cladograms produced to date can be viewed at the Tree of Life[2] website, originated by *David and Wayne Maddison*[35] at the University of Arizona (1) . Please take a moment to view the "Root of the Tree" link on the Tree of Life web site. In this phylogenetic tree, the root is at the far left, termed the root of the cladogram because it is at the base of the cladogram, opposite the branches. Return to the home page and click on the link entitled "Popular Pages", then select "Mammals". At the right side of this cladogram are the terminal nodes, located at the tip of the branches in any cladogram. In the Mammalia cladogram illustrated here, there are six terminal nodes, labeled Triconodonts, Monotremata, Multituberculata, Marsupialia, Palaeoryctoids, and Eutheria. An internal node is a hypothetical common ancestor. The branching points between the root and the terminal nodes are internal nodes. Each internal node is also at the base of a clade, which includes the common ancestral node plus all its descendents. Sample a few more links on the Tree of Life. Be sure to read Darwin's quote on the home page and ponder how difficult it would be to get published in a scientific journal today, if it were necessary to write this beautifully in order to succeed.

[1] This content is available online at <http://cnx.org/content/m11052/2.8/>.

[2] http://tolweb.org/tree/phylogeny.html

The Tree of Life is an example of a cladogram illustrating the relationships between taxa, based on the collective evidence from many different fields of biology and bioscience. In contrast, the subject of this tutorial is the construction of cladograms through bioinformatics tools, where the cladograms are based on sequence data. First, use the Biology Workbench[3] *(2)*[40] to build a simple unrooted cladogram. The Workbench will require a password (it's free), but it will grant entrance immediately upon registration of a password. Enter the site, and scroll down the page until the five menu buttons are visible. The "Session Tools" button allows the naming of a session, so that different jobs in progress can be saved under distinct sessions. Select "Session Tools", then select "Start New Session" and click on "Run" to change the name of "Default Session" to a new name. Once the workbench has been exited, the session will remain. Subsequently, clicking on the dot to the left of the session name under the "Session Tools" menu, and then selecting "Resume Session", will recall the session. The Workbench policy at the time of this writing is that old jobs are deleted only when an account has not been accessed for 6 months.

Next, select "Protein Tools" from the menu buttons, highlight "Ndjinn Multiple Database Search", and click "Run". In the query box to the right of the term "Contains", type HSP70, for the molecular chaperone, heat shock protein 70 kDa. Scroll down the database list and check the box to the left of the database entitled "PDBFINDER" before hitting the "Search" button. Among the results, find 2BUP, chaperone, and check the box to the left. Then select the menu button entitled "Import sequence(s)". This will import the sequence in fastA format into the open session. Now, under the box of session options, there should be a listing for the 2BUP sequence, with a small box to the left. Notice that the main menu under "Protein Tools" allows more options such as "Delete Protein Sequence", "Copy Protein Sequence" and "Add New Protein Sequences". For now, select the "Ndjinn Multiple Database Search" again. Search the PDBFINDER Database again by scrolling down the page and selecting it, but this time, just search using the PDB ID codes 1HKB, 1ATN and 1DKG for hexokinase, actin and the molecular chaperone DnaK (use the OR operator between each PDB ID code to search for all three in the same search). Import all three sequences simultaneously by checking the box to the left of the PDB ID codes used in the query and clicking on "Import sequence(s)". 1DKG will return three chains, A, B and D. Only chain D is the molecular chaperone, chains A and B are nucleotide exchange factors that co-crystallized with DnaK. Delete chains A and B by checking the box to the left of 1DKG_A and 1DKG_B, highlighting "Delete Protein Sequence", and clicking on "Run". Actin (1ATN) returns two chains, but chain A is the actin, chain D should be deleted in the same manner. Hexokinase (also called phosphotransferase) will return two chains as well. They are both hexokinase, but two identical sequences are not desirable in the cladogram, so delete chain B. Four sequences should remain, 1DKG_D, 1ATN_A, 1HKB_A, and 2BUP_A; check the boxes to the left of each of these. Scroll down the protein tools menu and highlight "CLUSTALW - Multiple Sequence Alignment", then click "Run". The default parameters will be sufficient for our purposes, just select "Submit". When the sequence alignment is returned, scroll down the page and view the multiple alignment. The Workbench automatically returns an unrooted tree with the alignment. Look at the unrooted tree.

Exercise 5.1

Which two sequences appear to be most closely related by viewing the unrooted tree?

Look at the multiple sequence alignment scoring section. Notice the sequence list that assigns numbers to each sequence. The alignment scores are labeled by the assigned sequence numbers, so this list is necessary to interpret the scores.

Exercise 5.2

According to the pairwise scores, which two sequences are most similar?

Exercise 5.3

What is the score of the best pairwise alignment?

Return to the top of the page and select "Import Alignment(s)". This will save this sequence alignment under the "Alignment Tools" menu of this session. Select the alignment by clicking in the small box to the left of the listing. Choose "DRAWGRAM" from the options box, to view this alignment in a rooted tree. Accept all the default values on the drawgram page and click the submit button at the bottom of the page.

[3] http://workbench.sdsc.edu/

Drawgram will return a rooted tree using the program *"PHYLIP"*[24], a Phylogeny Inference Package. The PHYLIP suite includes packages that can infer phylogenies by parsimony, compatibility, distance matrix methods, and maximum likelihood methods. PHYLIP can also draw several types of tree diagrams, and allows editing of trees. The suite accepts an impressive number of input formats, including nucleotide and protein sequences in fastA format. Detailed information on the PHYLIP suite can be viewed at the PHYLIP website[4] .

Exercise 5.4

Do the same two sequences appear to be the most closely related by viewing the rooted tree in comparison to the unrooted tree?

Click on the "Return" button and this will yield the "Alignment Tools" menu. It will be necessary to return to "Protein Tools" to complete this tutorial, so select the "Protein Tools" icon with the mouse. Now, to put together a little more complex cladogram, search the PDBFINDER in the Ndjinn Database for the following sequences by copying and pasting the following into the query box:

```
1ECL OR 1BGW OR 1DUB OR 1AUX OR 1KAN OR 1BPE OR 2AAC
```

Import all of these sequences into this session. Once again, there are some duplicated sequences from multimeric proteins, so delete 2AAC_B, 1KAN_B, 1AUX_B and 1DUB_B, C, D, E, and F. This should leave 11 sequences, including the first four sequences imported in the example above. On the scroll down menu under "Protein Tools", highlight "Select all sequences" and click "Run". Next, select "ClustalW" and click "Run". On the ClustalW input page, change the "unrooted" tree option to "rooted and unrooted trees", then submit.

Exercise 5.5

Once the results are returned, click on the option "Download a PostScript version of the output" for the rooted tree. Send the postscript file as an attachment to your lab assignment.

Exercise 5.6

Which two sequences have the highest number of ancestral nodes as represented in the cladogram?

Exercise 5.7

Which sequence has the longest branch between the terminal node and the closest ancestral node, as represented in the cladogram?

Exercise 5.8

In this cladogram, what is the relationship between the two sequences that scored the highest pairwise alignment in the first example?

Attempt to construct another type of tree using sequences of personal interest, without explicit instructions.

Exercise 5.9

Find at least six related nucleotide sequences (*e.g.*, download the sequences for superoxide dismutase genes from six different species) and construct rooted and unrooted trees containing these sequences using the Biology Workbench[5] . Send the postscript files as attachments to your lab, and list the 6 (or more) PDB IDs for the chosen sequences, with brief descriptions (protein name).

[4] http://evolution.genetics.washington.edu/phylip.html

[5] http://workbench.sdsc.edu/

Chapter 6

Profiles and PSSMs

6.1 PSI-BLAST[1]

PSI-BLAST[20] (1) (Position-Specific Iterated BLAST) is a tool that produces a position-specific scoring matrix constructed from a multiple alignment of the top-scoring BLAST responses to a given query sequence. This scoring matrix produces a profile designed to identify the key positions of conserved amino acids within a motif. When a profile is used to search a database, it can often detect subtle relationships between proteins that are distant structural or functional homologues. These relationships are often not detected by a BLAST search with a sample sequence query.

For an oversimplified example of what a consensus sequence, or profile, looks like, consider that the EF-hand binding loop of the calmodulin family could be represented as follows:

```
Loop Position #         1   3 4  5  6   8       12
Profile                 D x D G D/N G x I x x x E
```

Here "x" stands for positions where there is variability in amino acid type, and therefore, that position is not heavily weighted in the alignment. Comparing the profile to some actual binding loop sequences from different calmodulins is the best way to illustrate the derivation of this profile.

```
POSITION #      1   3 4 5 6   8       12
CALM_HUMAN_1    D K D G D G T I T T K E
CALF_NAEGR_1    D K D G D G T I T T S E
CALM_SCHPO_1    D R D Q D G N I T S N E

CALM_HUMAN_2    D A D G N G T I D F P E
CALF_NAEGR_2    D A D G N G T I D F T E
CALM_SCHPO_2    D A D G N G T I D F T E

CALM_HUMAN_3    D K D G N G Y I S A A E
CALF_NAEGR_3    D K D G N G F I S A Q E
CALM_SCHPO_3    D K D G N G Y I T V E E

CALM_HUMAN_4    D I D G D G Q V N Y E E
CALF_NAEGR_4    D I D G D N Q I N Y T E
CALM_SCHPO_4    D T D G D G V I N Y E E
```

[1] This content is available online at <http://cnx.org/content/m11040/2.13/>.

The rules for deriving this simple profile are: 1) any position with 90% amino acid identity or greater is considered conserved in the profile, and thus a higher score would be given when the conserved amino acid is found at that position in the sequence, and 2) any position that always contains one of only two types of amino acids would be up-weighted to give a higher score whenever either of those two amino acids appears at that position. A program such as PSI-BLAST will employ more sophisticated rules to create a profile than this example, of course. It is easy to see even with these sequences that amino acid similarity could be taken into consideration in addition to amino acid identity, and exploited in the profile.

There are three common categories of homologues that are studied in relation to biological molecules, sequence homology, structural homology, and functional homology. Sequence homology is the easiest to identify, and is therefore the primary target of many bioinformatics methods. Sequence homology yields direct implications about the relatedness of proteins and their potential pathways of derivation. However, to help understand how a protein is implicated in a certain disease state, or how to design a pharmaceutical that interacts with a given protein, functional and/or structural information is necessary. Functional homologues are relatively easy to define, as they are any two proteins, or protein domains, that perform similar functions. Structural homologues contain similar "folds", which are localized regions of a molecule that comprise a structural feature such as a "beta barrel" or "four helical bundle" motif. The fold can encompass the entire protein, or just one domain of the protein. A good introduction to the topic of protein folds can be found at the website for the Internet Course on The Principles of Protein Structure [2] organized by *Birkbeck College*[3] (2). When considering sequence, functional, or structural homology, it is important to understand that one type of homology between proteins does not always infer another type of homology. Nevertheless, it is a reasonable assumption that proteins that are related through evolutionary pathways are likely to have some degree of all three types of homology. PSI-BLAST was engineered to identify distant relationships between sequences that are too subtle to discover with a regular BLAST search.

In the first round, PSI-BLAST is just like a normal BLAST; it finds sequence homologues. In the second round or "iteration" of PSI-BLAST, it figures out which residues tend to be conserved by creating a custom profile for each position of the sequence from a multiple alignment. Then another BLAST is performed, using the profile to produce a position-specific scoring matrix based on which positions evolution has conserved vs. which positions evolution has allowed to vary. The sequences found after the first round are added to the profile, allowing PSI-BLAST to detect more distant homologues in each iteration.

One of the known weaknesses of PSI-BLAST is that its ability to detect distant relationships between proteins is critically dependent on the choice of the query sequence. For this reason, a recommended strategy with PSI-BLAST is to query using individual functional domains. PSI-BLAST will then find other proteins that share this domain, even if they do not possess overall homology. To acquaint the new user with PSI-BLAST, this tutorial mimics an investigation performed by *Aravind and Koonin* [9] (3) in 1999, wherein new members of the HSP70/actin protein family were identified, except the analysis in our tutorial will be on a much smaller scale than that presented in the paper. The HSP70/actin family members were originally recognized to have a common evolutionary origin as a result of a study performed by *Bork et al.*[15] (4) in 1992. A structural superposition of the structures of actin, hexokinase, and the molecular chaperonin hsp70, and alignment of many sequences in each of the three families, uncovered a set of common conserved residues, distributed in five sequence motifs, that are involved in ATP binding and in a flexible interdomain hinge. Although each of these proteins performs very different functions, and their sequences are quite divergent, the similarity in the fold of the ATP-binding domain is visually recognizable. These are all ATP-dependent enzymes and the patterns discovered by Bork and associates could not be detected by traditional BLAST-type sequence searches. Therefore, Aravind and Koonin chose this family as a test of PSI-BLAST's ability to detect distant evolutionary relationships.

Aravind and Koonin chose actin from the PDB file with accession code 1atn as one of their query sequences. Begin the query by retrieving this sequence from the PDB[3] . Check the box for searching the PDB archive that says "PDB ID", then enter the accession code 1atn as the query. Notice that the crystal structure deposited in this entry contained DNase I complexed with actin. There will be a link in the menu

[2] http://www.cryst.bbk.ac.uk/PPS95/course/8_folds/

[3] http://www.rcsb.org/pdb/

in the blue border on the left entitled "FASTA Sequence", select this link to download the sequence file. This file will contain two sequences 1ATN:A (actin) and 1ATN:D (DNase I). Copy the sequence for 1ATN:A and paste it into the BLAST[4] query box that arises from choosing the Protein BLAST, then selecting "PSI-BLAST" under the algorithm section. Change the database from "nr" to "swissprot", but accept the default values for everything else. Click on BLAST, then view the results. NOTE THAT each time another iteration of PSI-BLAST runs, the results page will indicate the iteration number. This is very helpful for keeping track of the stage of the results.

Exercise 6.1

There is a statistical section at the very end of the BLAST report. BLAST query sequence windows will be lined up with similar regions in the database, then the algorithm will try to extend the alignment with the database sequence in both directions along the query sequence. What is the number of successful extensions for the first BLAST report?

Exercise 6.2

What is, by far, the most common protein shown as a hit in the list of scores?

Exercise 6.3

At the end of the list of scores, note the section entitled "Sequences with E-values WORSE than threshold". Why is it that PSI-BLAST includes this section, but BLAST does not?

Exercise 6.4

Do any members from the Actin-like ATPase domain superfamily, other than actin (or actin-like proteins), show up as hits in this section? If so, name one.

There are several buttons within the results document entitled "Run PSI-BLAST iteration 2"; if they are difficult to locate, there is one at the end of the list of descriptions and scores; click on one of these buttons to run a second iteration of PSI-BLAST. Look at the results window and above the graphical display, it should say "Results of PSI-Blast iteration 2". Keep track of what the next iteration number should be, and make sure it matches the iteration number that PSI-BLAST displays to avoid getting lost within the PSI-Blast search. When the results appear, look at the legend under the color alignment graph. It says that a yellow starburst with the word "NEW" inside it indicates a new sequence that was identified as a result of the most recent iteration, and a green dot indicates a sequence that was already present prior to the most recent iteration.

Exercise 6.5

What is the number of successful extensions for the second BLAST iteration?

Exercise 6.6

Did the second iteration locate any new sequences within (meaning less than or equal to) the threshold E-value? If so, how many?

Exercise 6.7

Are there new members of the Actin-like ATPase domain superfamily in the section entitled "Sequences with E-value WORSE than threshold"? If so, name one.

Exercise 6.8

How many new sequences are identified as a result of the third iteration?

Exercise 6.9

Are there new sequences returned that are within the threshold E-value that are NOT actin or actin-like proteins? If so, name one.

Finally, perform the first iteration of the PSI-BLAST search with the same query sequence as above, but with the database set to "nr".

Exercise 6.10

What is the number of successful extensions for the first BLAST report when searching the nr database?

[4] http://www.ncbi.nlm.nih.gov/BLAST/

The "swissprot" database is a curated database, and all the descriptions on the return list were informative. However, searching against the large "nr" database yields a list of returns that should contain some proteins described as "unknown", or "unnamed". The advantage of "nr" is that usually many more hits are returned. Sometimes this is desirable, and sometimes it is not, but it is one thing to consider when designing a PSI-BLAST search strategy.

A PSI-BLAST search can be forced to include more sequences in the profile at a given iteration by altering the threshold. The default value is to accept any results with E values less than 0.005. It is common to change this threshold to 0.05 if the PSI-BLAST converges too quickly at 0.005. We did not take our search to convergence, that would require continuing iterations until the last iteration did not return any new hits. As this tutorial illustrates, PSI-BLAST is a useful tool, but it often requires putting some thought into the search strategy before it will produce meaningful results.

Chapter 7

Independent Study Project

7.1 Bioinformatics Project[1]

The bioinformatics project is an opportunity to use the tools taught in previous modules to research an area of personal interest. The student will choose one biological sequence with some presumed functional or structural importance (it can be putative, as long as there is some evidence, preferably published evidence, that supports the putative significance of the sequence). The sequence can be nucleotide or protein, but it must be at least 140 residues in length. The sequence can represent only one domain of a protein or gene, or a regulatory region, as long as it meets the above requirements for significance and length and the student has sufficiently justified why it is appropriate to extract just one region from the entire protein or gene for the problem or question chosen. There will be five sections to the project:

1. Define the problem or question.
2. Materials and Methods.
3. Multiple sequence alignment figure.
4. Phylogenetic tree.
5. Discussion.

Examples of the types of problems that are appropriate for this project include:

- Detection of distantly related (divergent) sequences.
- Detection of sequence homologs in various species.
- Detection of homologous motifs in proteins of varied function.
- Generation of a functional domain profile from a set of sequences.

7.1.1 Section 1. Define the problem or question.

This section will define the sequence chosen as the starting or master sequence, from which the student will generate other sequences according to the question to be answered. Discuss the significance of the sequence and justify its selection as the topic for this project. Be sure it meets all the above-specified requirements. One well-written paragraph should be able to encompass this section.

7.1.2 Section 2. Materials and Methods.

Write a complete Materials and Methods section as for a journal article. Some journal articles may be selected and supplied by the instructor to serve as examples. Possible subsection titles could include:

[1] This content is available online at <http://cnx.org/content/m11881/1.2/>.

- Databases
- Database Search Tools
- Multiple Alignment Methods
- Generation of Phylogenetic trees

Most methods subsections can be sufficiently expressed with one paragraph.

7.1.3 Section 3. Multiple Sequence Alignment.

The student should choose a question to address that will generate at least 8 related sequences (this can be including the master sequence). Perform a multiple sequence alignment and generate a multiple sequence alignment figure that pertains to the question or problem of study. For example, a problem that deals with the detection of homologous motifs in proteins of varied function should be aligned to best illustrate similarity in the chosen motif or domain, which may not necessarily illustrate the best global alignment. Once aligned, sequences may require annotation to highlight specific regions or functional domains. The student may generate the multiple sequence alignment figure by a variety of methods, but the suggested approach is to use the Biology Workbench. Alignments can be performed under the PROTEIN TOOLS or NUCLEIC TOOLS menus. Under the alignment tools menu, options include editing, viewing, displaying (MVIEW) and drawing figures of (TEXSHADE, BOXSHADE) alignments. Note the warning under BOXSHADE that when the ruler is chosen for alignment numbering, boxshade can get stuck and never finish. The ruler is a nice addition to the figure, so try changing the font size or page orientation when this happens, as the Workbench recommends.

7.1.4 Section 4. Phylogenetic tree.

Generate a phylogenetic tree that illustrates the relationships of the sequences in the multiple alignment. The tree can be rooted or unrooted, but should demonstrate the sequence relationships clearly. Phylogenetic trees can be drawn using DRAWGRAM or DRAWTREE under the alignment tools menu. All figures in this project should be entitled, labeled and captioned (see journal articles for examples).

7.1.5 Section 5. Discussion.

Discuss the information discovered from the sequences generated and how this information addresses the problem or question of study. Be sure that the following questions are answered, if appropriate to the problem:

- How were functional domains determined or assigned?
- How were profiles generated or chosen, if used?
- Did the functional domains line up correctly in the alignments? If not, why not?
- Did you identify partial matches, or fringe homologs?

This project challenges the student to use the bioinformatics tools introduced in previous modules for independent study. The student gains experience in the proper design of a bioinformatics experiment, and in the proper presentation of bioinformatics methods, figures and results.

Chapter 8

Supplemental Material: Advanced Topics

8.1 RNA Secondary Structure Prediction[1]

The study of RNA structure calls for a distinct set of computational tools designed expressly for RNA applications. Recall there are three major categories of RNA, messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). Ribosomal RNA and ribozymes have catalytic functions, like proteins, while messenger RNA has an information storage function, like DNA. A good resource for reviewing the major types of RNA and their functions can be found in the online RNA Structure Primer [2] , available at the RNABase RNA Structure Database website *RNABase The RNA Structure Database,* [38]

RNA is usually thought of as a single stranded linear molecule, however, in a biological system this is not the case. Frequently, different regions of the same RNA strand will fold together via base pair interactions to make intricate secondary and tertiary structures that are essential for correct biological function. Common secondary structure motifs include hairpin loops, stems, and bulges. Diagrams of these motifs can be viewed on the IMB JENA Nucleic Acid Nomenclature and Structure[3] web page, under section 8 of the Table of Contents, secondary structural elements. To observe tertiary structure in an RNA molecule, view the structure of phenylalanine tRNA[4] from *the yeast crystal structure* [32], available from the Nucleic Acid Database (NDB) Project, Rutgers.

Though RNA is usually single stranded, in some RNA virus genomes it will form a double stranded helix. However, unlike DNA, RNA forms an A-form double helix. The RNA double helix differs from that of the DNA double helix because of the presence of ribose, rather than deoxyribose, in the sugar phosphate backbone of the molecule. The addition of a hydroxyl group at the C2 postion in the ribose sugar is responsible for the A-form geometry in double stranded RNA. The A-form makes a right-handed helix, like the B-form double helix, but is a shorter, wider helix than the B-form, and the major groove is deep, but narrow, making it virtually inaccessible to proteins. It is in the major groove that the chemical groups are sequence-specific and dependent on base identity, and therefore, this is where proteins tend to bind a DNA double helix. Because the RNA A-form double helix contains a major groove that is too narrow and deep for proteins to access, the minor groove becomes more important for protein interactions with RNA helices. Also, proteins that interact with specific RNA sequences commonly bind single-stranded RNA segments. For an example of an RNA/protein complex, view the *NDB entry*[8] for the Protein/Hepatitis Delta Virus Ribozyme Complex[5] . When a strand of DNA forms a double helix with a strand of RNA, this will also result in an A-form helix. An example of a *DNA/RNA complex*[_dna?] can be found by viewing the NDB entry for the RNA-DNA complex formed by the 10-23 DNA enzyme[6] .

[1] This content is available online at <http://cnx.org/content/m11065/2.4/>.

[2] http://www.rnabase.org/primer/

[3] http://www.imb-jena.de/ImgLibDoc/nana/IMAGE_NANA.html#sec_element

[4] http://ndbserver.rutgers.edu/atlas/xray/structures/T/trna04/trna04.html

[5] http://ndbserver.rutgers.edu/atlas/xray/structures/P/pr0010/pr0010.html

[6] http://ndbserver.rutgers.edu/atlas/xray/structures/U/uh0001/uh0001.html

RNA molecules sometimes contain the unusual nucleotides ribothymidine, dihydrouridine, pseudouridine and inosine. In addition, RNA somewhat commonly forms a G-U wobble base pair, and makes other non-canonical base pairs, a listing of which can be found on the George Fox group's Non-Canonical Base Pair Database[7] web page, University of Houston.

Exercise 8.1

List three non-canonical base pairs identified on the web page, besides a G-U wobble base pair.

It is clear to see that RNA molecules have many unique characteristics, distinct from the properties of DNA and protein structures. Many of these characteristics can be exploited to predict RNA structure from a nucleotide sequence.

One of the methodologies that is commonly used for RNA structure prediction is based on calculating free energy estimates for each possible fold, then choosing the fold that yields the lowest free energy. These free energy values are a combination of energy values calculated for each pair of adjacent base pairs, plus loop or bulge energies. The energy values are derived from melting studies of synthetically constructed oligoribonucleotides. For more information on the development of RNA free energy parameters, see the Development and References Page[8] of the Zuker-Turner RNA folding package.

Compute the free energy of an RNA structure using the efn server, an RNA free energy[9] web site authored by Michael Zuker, Rensselaer Polytechnic Institute. Copy and paste the following RNA sequence into the sequence query box.

```
G G C G C G G C A C C G U C C G C G G A A C A A A C G G
```

Just below the sequence query box, there is a box where the secondary structure can be defined by specifying base pairs. This is done using a triplet of numbers to define each stem region of consecutive base pairs. The first number in the triplet defines the sequence number of the first base in the pair from the 5' end of the sequence. The second number defines the sequence number of the opposing base in the pair. The third number defines how many consecutive bases are involved in the stem. In this example, use the following triplet:

```
10, 19, 3
```

The triplet in the above example means that the bases "C C G", number 10, 11 and 12, base pair with the bases "C G G", bases number 17, 18 and 19. Paste the above triplet into the secondary structure box and click on the box that says, "Send data for processing".

Exercise 8.2

What is the computed free energy for this RNA structure?

Exercise 8.3

Click on the link that says "png" to get a better picture of the structure. How many bases (non-paired) are in the loop in this structure?

Exercise 8.4

Save the png file to disk and send a copy to the course instructor.

Now, use the same sequence, but specify a different secondary structure. This time, paste the following triplet into the secondary structure box and send for data processing:

[7] http://prion.bchs.uh.edu/bp_type/all_record_page_by_type.html

[8] http://www.bioinfo.rpi.edu/~zukerm/rna/energy/node2.html#SECTION20

[9] http://www.bioinfo.rpi.edu/applications/mfold/old/rna/energy/form1.cgi

```
3, 18, 5
```

Exercise 8.5
What is the computed free energy for this RNA structure?

Exercise 8.6
Click on the link that says "png" to get a better picture of the structure. How many bases (non-paired) are in the loop in this structure?

Exercise 8.7
Save the png file to disk and send a copy to the course instructor.

Exercise 8.8
Which of these two structures is more likely to exist under physiological circumstances, given no additional constraints?

A second approach to RNA secondary structure prediction is to look for conserved stem regions in related sequences. This method involves looking for regions within sequences where stems have been conserved, even when the bases have mutated. For this to happen, it would require that if a G mutated to an A, then the opposing C in the base pair would mutate to a U. These regions are found by aligning related RNA sequences, and applying an algorithm that looks for these sorts of paired mutations in predicted stem regions. Align the RNA sequences of the following tRNAs using ClustalW[10] .

```
>1ASY:S ASPARTYL TRNA SYNTHETASE (ASPRS)
UCCGUGAUAGUUXAAXGGXCAGAAUGGGCGCXUGUCXCGUGCCAGAUXGGGGTXCAAUUC
CCCGUCGCGGAGCCA

>1EIY:C TRNA(PHE)
GCCGAGGUAGCUCAGUUGGUAGAGCAUGCGACUGAAAAUCGCAGUGUCCGCGGUUCGAUU
CCGCGCCUCGGCACCA

>1EFW:C ASPARTYL-TRNA
GGAGCGGXAGUUCAGXCGGXXAGAAUACCUGCCUXUCXCGCAGGGGXUCGCGGGXXCGAG
UCCCGXCCGUUCC

>1EHZ:A TRANSFER RNA (PHE)
GCGGAUUUAXCUCAGXXGGGAGAGCXCCAGAXUXAAXAXXUGGAGXUCXUGUGXXCGXUC
CACAGAAUUCGCACCA
```

IMPORTANT: After ClustalW alignment, the program puts asterisk below conserved residues. These must be removed before submitting the alignment to the RNA secondary structures prediction server.

Copy the multiple alignment and paste it into the query box at the RNA secondary structure prediction server[11] , Moscow State University. Click "submit query", and the results should appear within about 3 minutes. Scroll down the page and view the section where the stem regions were identified, and their free energies were computed.

Exercise 8.9
How many stems are predicted?

Exercise 8.10
List each of their computed free energy values.

[10] http://www.ebi.ac.uk/clustalw/

[11] http://www.genebee.msu.su/services/rna2_reduced.html

Exercise 8.11

Continue to scroll down the page and look at the predicted structure diagram. What is the total free energy of the structure?

Exercise 8.12

Does this structure that has been predicted from sequences agree well with the known structure of tRNAs?

RNA structure has some distinct differences from DNA structure that can be exploited to yield secondary structure predictions that are usually reasonably accurate. In addition, there are many on-line tools and databases that are specific to RNA. Here, the use of a few of these tools has been illustrated, but take some time to view more of the links that are available on the RNA World Website[12] , Institut fur Molekulare Biotechnologie, Jena, Germany.

_dna

8.2 Microarray Experiments[13]

Microarry chips are devices that enable the scientist to simultaneously measure the transcription level of every gene within a cell. Microarrays are commercially available from a number of companies, such as Affymetrix[14] , Invitrogen[15] and Sigma-Genosys[16] , to name a few. The chip is usually constructed by amplifying all the genes within the selected genome, yeast, for example, using polymerase chain reaction (PCR) methodology. The PCR products would then be "spotted" onto the chips by a robot, as single-stranded DNA that is linked by covalent bonds to the glass slide. The spots would be positioned in an array on a grid pattern, where each spot contains many identical copies of an individual gene. A discussion of the chemistry involved in creating a microarry can be found on the technology page[17] of the Affymetrix website. The position of the genes are recorded by spot location, so that the appropriate gene can be identified any time a probe hybridizes with, or binds to, its complementary DNA strand on the chip.

Microarray chips measure transcriptomes, which are the entire collection of RNA transcripts within a cell under the given conditions. To use the chip to measure an experimental transcriptome against a reference transcriptome requires cells grown under two different conditions, the experimental conditions and the reference conditions. The mRNA from the two different conditions are harvested separately, and *reverse transcriptase*[13] (1) is used to transcribe the mRNA into cDNA. The nucleotides used to synthesize the cDNA will be labeled with either a green or red dye, one color for the reference conditions and the other for the experimental conditions. The microarry chip is then incubated overnight with both populations of cDNAs, and a given cDNA will hybridize with the complementary strand from its gene that is covalently bound to a grid spot on the chip. The chips are washed to remove any unbound cDNAs and then two computerized images are produced by scanning first to detect the grid spots containing cDNAs labeled with green dye, and second to detect the spots contain red-labeled cDNAs. The computer also produces a merged image that will show a yellow spot for grid spots that contain both red- and green-labeled cDNAs, indicating transcripts that are expressed under both sets of conditions. A very nice on-line, animated demonstration of the entire protocol is offered by the Genomics Course[18] on the *Davidson College website*[2] (2).

In addition to producing a qualitative image that is easy visualize, a microarray experiment yields quantitative data for each spot, consisting of the measured fluorescence intensity of the red signal, the fluorescence intensity of the green signal, and the ratio of red signal to green signal. It is in storing and analyzing the quantitative data that bioinformatics really comes into play in microarray technology. These data sets are

[12] http://www.imb-jena.de/RNA.html

[13] This content is available online at <http://cnx.org/content/m11050/2.17/>.

[14] http://www.affymetrix.com/

[15] http://mp.invitrogen.com/

[16] http://www.sigma-genosys.com/epp.asp

[17] http://www.affymetrix.com/technology/manufacturing/index.affx

[18] http://www.bio.davidson.edu/courses/genomics/chip/chip.html

incredibly large. For instance, a typical mammalian cell is estimated to have between 10,000 to 20,000 different species of mRNA expressed at a given time.

As a demonstration, view the Stanford Microarray Database[19] website. Under the Public Data section, click on the "Public Login" link. Limit the data set search to the organism *Arabidopsis thaliana* and the author Gutierrez. Click on the button entitled "Display data", and a table of microarray datasets should be returned. Choose one of the experiments, making note of the Experiment ID number from the table and select the clickable image icon. (There is a legend for the icons at the top of the web page, if there is uncertainty as to which is the clickable image icon.) This yields the qualitative visualization of the microarray experiment, as described previously in this module. Take a look at the array image and note that it is difficult to draw many conclusions from this kind of visualization.

Exercise 8.13

What is the Experiment ID number for the viewed microarray image?

Exercise 8.14

Is it possible to get a feeling for which color dot, green, red, or yellow, is most predominant just by viewing the image? (If so, which color?)

Exercise 8.15

Are all of the dots, over the entire grid, well-shaped? (Give a brief explanation.)

Click on one of the individual spots in the microarray grid. This will open a new window that contains a close up of the individual spot and all the experimental information about that spot.

Exercise 8.16

Retrieve and list the following information about the chosen spot: a. spot number, b. description (under biological information), c. the Channel 1 intensity (mean), d. the Channel 1 background (median), e. the Channel 1 net intensity (mean), and f. the Log(base2) of R/G Normalized Ratio (Mean).

Return to the table of *Arabidopsis thaliana* data sets and for Experiment ID #11374, select the "Data" icon, which is the first icon under the "Options" column. Next to "Sort By", select "Log(base2) of R/G Normalized Ratio (Mean)", and "Descending". Under "Display:", click on "Spot", then scroll down and hold down the control key (or the apple key on macintoshes) while selecting "Log(base2) of R/G Normalized Ratio (Mean)". (The control key allows selection of additional choices without deselecting the previous choice.) Accept the default values for all remaining options and select "Display" at the bottom of the page. Recall that the data are converted to numbers representing the fluorescence intensity of red dye, green dye, and the ratios of red to green. Scientists commonly use a log transformation of the ratio data, because the logs are more mathematically tractable in reference to statistical analysis. The results page will show the top ranking spots from this chip, ranked from highest log red/green value to lowest.

Exercise 8.17

What are the spot numbers of the three highest ranking spots?

Exercise 8.18

What are the "Log(base2) of R/G Normalized Ratio (Mean)" values for the three highest ranking spots?

Use the browser's back button to go back and select new data for display. Chage the sorting selection to R/G Normalized (Mean), in descending order. under the "Display:" window select "Spot", then scroll down and hold down the control key while selecting "R/G Normalized (Mean)".

Exercise 8.19

What are the spot numbers for the three highest ranking spots?

Exercise 8.20

What are the "R/G Normalized (Mean)" values for the three highest ranking spots?

[19] http://genome-www5.Stanford.EDU/MicroArray/SMD/

Exercise 8.21

Does it change the ranking of the spots to use the log transformations of the ratios instead of the ratios?

To demonstrate a different method of visualizing and analyzing microarry data, take a look at the MIT Cancer Genomics Microarray Data Sets[20] . Scroll down to the section entitled "Gene Expression Correlates of Clinical Prostate Cancer Behavior". Click on the first data set, for Prostate tumor and normal samples, entitled "Prostate_TN_final0701_allmeanScale.res". This data set originates from Affymetrix chips. In this case, the signal is recorded as "A" for absent, "P" for present, and "M" for marginal, as determined by the Affymetrix GeneChip software. The numerical values are scaled average difference units for tumor vs. normal prediction, and these values are also generated by the Affymetrix software. A more complete discussion of gene expression data analysis for Affymetrix GeneChip Arrays can be found at the Affymetrix web site[21] .

So far, the discussion has been primarily about visualizing and quantifying the fluorescence signal from a microarray experiment. However, analysis of gene expression under experimental conditions versus reference conditions requires determining whether observed differences are significant or not. There are many sources of noise and variability in microarray data, including experimental sources such as image scanning inconsistencies, issues involved in computer interpretation and quantification of spots, hybridization variables such as temperature and time discrepancies between experiments, and experimental errors caused by differential probe labeling and efficacy of RNA extraction. In addition, as the size of the sample increases, so does the probability of finding some large differences due to chance. Therefore, statistical analysis is required to show that gene expression differences are real.

There are some complex problems underlying statistical analysis of microarray data, primarily related to the fact that the number of samples is very, very large, but the number of times that each measurement is repeated is comparatively very small. (This is due mostly to cost and time issues.) Also, the simplest statistical techniques commonly assume a normal distribution, which cannot necessarily be assumed in microarray experiments. For a detailed discussion, *D. K. Slonim*[19] (3) has authored a good review of the most current approaches to gene expression data analysis.

This tutorial will provide an oversimplified example of the type of statistical analysis that needs to be applied to microarray data, using the t-test. For a given gene, A, the gene will have two associated vectors: {a(ref)1, ..., a(ref)n} and {a(exp)1, ..., a(exp)n}, where a(ref) contains n measurements of expression levels under reference conditions and a(exp) contains n measurements of expression levels under experimental conditions.

The mean of each vector will be equal to:

$$<m> = [\sum_{i=1}^{n} a_i]/n$$

Figure 8.1

[20] http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi

[21] http://www.affymetrix.com/technology/data_analysis/index.affx

The standard deviation of each vector will be equal to:

$$\sigma = \sqrt{\left(\frac{[\sum_{i=1}^{n} (a_i - <m>)^2]}{(n - 1)} \right)}$$

Figure 8.2

The standard error of each vector will be equal to:

$$SE = \sigma / [\sqrt{(n - 1)}]$$

Figure 8.3

The formula for the t test is as follows:

$$t = \frac{<m>_{(exp)} - <m>_{(ref)}}{\sqrt{[(\sigma^2_{(exp)} / n_{(exp)}) + (\sigma^2_{(ref)} / n_{(ref)})]}}$$

Figure 8.4

The t-test is used to test the difference between the means of two test sets, as in before and after studies or matched-pairs studies. There is a confidence interval for the mean and a critical value for t for the chosen level of significance associated with the t-test. For instance, a level of significance equal to 0.05 means that 95% of the cases will be within the confidence range if there is no significant difference between the means of the two test sets, or experiments, being compared. The confidence limits set upper and lower bounds on an estimate of the mean for the chosen level of significance (0.05). The confidence interval is the range within the bounds of the confidence limits. The confidence interval can be computed, if you know the shape of your distribution. For normally distributed data, the confidence limits at the 0.05 significance level for an estimated mean are the sample mean plus or minus 1.96 times the standard error.

```
confidence interval (normal distribution):  mean +/- 1.96 * SE
```

For example, if the sample mean is 10 and the standard error is 1.2, then 95% of the cases will be within the range of 10 plus or minus 1.96 times 1.2, or 10 plus or minus 2.4, which is the range from 7.6 to 12.4. Thus, if the experimental mean is outside the limits of this range computed for the reference mean, then the difference between the means of the two test sets is considered to be significant within a probability of 95%. The critical value for t at a given significance level for a specific type of distribution can be looked up in a table; most statistics books contain them. In the case of microarray data, if the absolute value of t is greater than the critical value, this indicates a significant difference in the gene expression between the reference and experimental test sets. Because the t-test is a parametric test that assumes a normal distribution, the statistical tests that are commonly used to analyze microarray data are more complex variations that are used for distributions other than normal distributions.

Tutorial Data Set

	red:green ratio	**red:green ratio**	**red:green ratio**	**red:green ratio**	**red:green ratio**	**red:green ratio**	**red:green ratio**
Gene	**measurement 1**	**measurement 2**	**measurement 3**	**measurement 4**	**measurement 5**	**measurement 6**	**measurement 7**
A(ref)	0.97	1.54	1.32	0.89	1.06	1.21	
A(exp)	1.37	1.25	1.15	0.99	1.30	1.53	1.07
B(ref)	1.67	1.78	2.01	1.89	1.75	1.81	1.69
B(exp)	6.21	6.03	5.94	6.14	6.11		

Table 8.1

Assumptions for example problem:

1. Use the example data set to perform a t-test analysis.
2. Consider each row of the table a separate test set under either experimental or reference conditions.
3. Assume a normal distribution.
4. Choose 0.05 as the level of significance.
5. Compute answers to the second decimal place.
6. Assume the critical value for t is equal to an absolute value of t greater than 2.37.

Exercise 8.22
What are the means for each row of data?

Exercise 8.23
What are the standard deviations for each row of data?

Exercise 8.24

What is the standard error for each row of data?

Exercise 8.25

What is the value for t for the comparison between the reference and experimental test sets for Gene A?

Exercise 8.26

What is the 95% confidence interval computed for the Gene A reference set?

Exercise 8.27

Is there a significant difference between the mean values of the experimental versus the reference set for Gene A? (Explain the answer both in terms of the t value and the confidence interval.)

Exercise 8.28

What is the value for t for the comparison between the reference and experimental test sets for Gene B?

Exercise 8.29

What is the 95% confidence interval computed for the Gene B reference set?

Exercise 8.30

Is there a significant difference between the mean values of the experimental versus the reference set for Gene B? (Explain the answer both in terms of the t value and the confidence interval.)

Exercise 8.31

If there was a significant difference between the gene expression under experimental conditions versus the gene expression under reference conditions for either Gene A or Gene B, then estimate the significant increase or decrease observed.

There are many software packages available that have been designed expressly for microarray data analysis. In addition to testing gene expression under a set of experimental conditions versus reference conditions, it is possible to identified "clustered" genes that seem to have similar responses under similar conditions. Also, genes can be identified that show related responses under similar conditions, such as one gene's expression always increases when another's decreases. When two or more genes show this kind of clustered behavior, it can be an indication that they are part of the same pathway, or that they are regulating each other. Using this type of microarray data analysis, the scientist can combine the cluster analysis results with what is known through laboratory experiments and often come up with new hypotheses about biochemical pathways and regulation.

8.3 Expasy Proteomics Tools[22]

A proteome is the collection of all the proteins within a given organism, in the same way a genome is the collection of all the genes within a given organism. A proteome has some characteristics that are quite different from a genome, however. A principal difference is the fact that while a particular organism will have the same set of identical DNA in any undamaged, healthy cell throughout its lifetime, the organism's proteins will differ greatly from one tissue to another, and from one life stage to another. Furthermore, proteins commonly incur a variety of chemical modifications after they are made. These modifications are critical for proper protein functioning and/or regulation, and moreover, these modifications cannot be determined with certainty by looking at the DNA sequence alone. In a contempary high-throughput proteomics laboratory, the number of proteins identified and analyzed in one day can be on the order of hundreds.

The term "proteome" was originally coined by an Australian scientist, *Mark Wilkins*[22] (1), to describe the "PROTEin complement of the genOME". The term "proteomics" is used relatively loosely to describe any and all of the collection of high throughput techniques that have emerged to enable the scientist to analyze all the proteins expressed under a certain set of conditions within an individual cell or organism.

[22] This content is available online at <http://cnx.org/content/m11082/2.5/>.

The *ExPASy*[7] (Expert Protein Analysis System) website (2), Swiss Institute of Bioinformatics, offers the definition that "proteomics can be defined as the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes."

Common techniques for identifying the proteins within a proteome are 2D-PAGE (polyacrylamide gel electrophoresis) gels, amino acid (AA) composition analysis, peptide mass fingerprinting and other mass spectroscopy applications. A good starting point for becoming acquainted with 2D gels is the 2D PAGE tutorial[23] offered by the Institute of Biological Sciences, University of Wales at Aberystwyth. ExPASy offers a good synopsis of peptide mass fingerprinting and AA composition analysis[24] techniques, for those who are unfamiliar with these methods.

At the ExPASy Proteomics Tools server[25] , the first category of tools are for protein identification and characterization. Take a look at the tools listed in this section. These tools are designed to identify the proteins that make up the proteome of study, using the data received from gels, AA analysis and mass spectroscopy experiments.

Exercise 8.32

What tool from the ExPASy "protein identification and characterization" section would you use for identifying a protein for which you only know the amino acid composition?

Exercise 8.33

What is the name of at least one peptide mass fingerprint tool at the ExPASy site?

Exercise 8.34

Generally outline the underlying principles that allow the identification of a protein through peptide mass fingerprinting.

Scroll down on the ExPASy tools webpage to the section entitled "pattern and profile searches". The tools that populate this section are designed to identify proteins that belong to well characterized protein families, usually identified by conserved domains within family members. Also, well known protein motifs, or domains, are represented independently of their protein families in pattern databases that contain the conserved aspects of the domain sequence. Select the tool entitled *"InterPro Scan"*[37] (3) to perform an integrated search in PROSITE, Pfam, PRINTS and other family and domain databases. This tool is useful for identifying specific domains or motifs within a protein, once the sequence has been determined, and can sometimes recognize the protein as a member of an established protein family. Test the efficacy of this tool with the following sequences, one at a time, but make sure the interactive run button is selected. An email address will be required to submit the job, but the results can be viewed in the browser interactively.

```
>Seq1
MAGIAAKLAKDREAAEGLGSHERAIKYLNQDYEALRNECLEAGTLFQDPSFPAIPSALGFKELGPYSSKT
RGIEWKRPTEICADPQFIIGGATRTDICQGALGDCWLLAAIASLTLNEEILARVVPLNQSFQENYAGIFH
FQFWQYGEWVEVVVDDRLPTKDGELLFVHSAEGSEFWSALLEKAYAKINGCYEALSGGATTEGFEDFTGG
IAEWYELKKPPPNLFKIIQKALQKGSLLGCSIDITSAADSEAITFQKLVKGHAYSVTGAEEVESNGSLQK
LIRIRNPWGEVEWTGRWNDNCPSWNTIDPEERERLTRRHEDGEFWMSFSDFLRHYSRLEICNLTPDTLTS
DTYKKWKLTKMDGNWRRGSTAGGCRNYPNTFWMNPQYLIKLEEEDEDEEDGESGCTFLVGLIQKHRRRQR
KMGEDMHTIGFGIYEVPEELSGQTNIHLSKNFFLTNRARERSDTFINLREVLNRFKLPPGEYILVPSTFE
PNKDGDFCIRVFSEKKADYQAVDDEIEANLEEFDISEDDIDDGVRRLFAQLAGEDAEISAFELQTILRRV
LAKRQDIKSDGFSIETCKIMVDMLDSDGSGKLGLKEFYILWTKIQKYQKIYREIDVDRSGTMNSYEMRKA
LEEAGFKMPCQLHQVIVARFADDQLIIDFDNFVRCLVRLETLFKIFKQLDPENTGTIELDLISWLCFSVL

>Seq2
SGPRPVVLSGPSGAGKSTLLKRLLQEHSGIFGFSVSHTTRNPRPGEENGKDYYFVTREVM
QRDIAAGDFIEHAEFSGNLYGTSKVAVQAVQAMNRICVLDVDLQGVRNIKATDLRPIYIS
```

[23] http://www.aber.ac.uk/parasitology/Proteome/Tut_2D.html#Section%201

[24] http://us.expasy.org/ch2d/protocols/protocols.fm13.html

[25] http://us.expasy.org/tools/

```
VQPPSLHVLEQRLRQRNTETEESLVKRLAAAQADMESSKEPGLFDVVIINDSLDQAYAEL
KEALSEEIKKAQRTGA


>Seq3
MTEVISNKITAKDGATSLKDIDDKRWVWISDPETAFTKAWIKEDLPDKKYVVRYNNSRDE
KIVGEDEIDPVNPAKFDRVNDMAELTYLNEPAVTYNLEQRYLSDQIYTYSGLFLVAVNPY
CGLPIYTKDIIQLYKDKTQERKLPHVFAIADLAYNNLLENKENQSILVTGESGAGKTENT
KRIIQYLAAIASSTTVGSSQVEEQIIKTNPVLESFGNARTVRNNNSSRFGKFIKVEFSLS
GEISNAAIEWYLLEKSRVVHQNEFERNYHVFYQLLSGADTALKNKLLLTDNCNDYRYLKD
SVHIIDGVDDKEEFKTLLAAFKTLGFDDKENFDLFNILSIILHMGNIDVGADRSGIARLL
NPDEIDKLCHLLGVSPELFSQNLVRPRIKAGHEWVISARSQTQVISSIEALAKAIYERNF
GWLVKRLNTSLNHSNAQSYFIGILDIAGFEIFEKNSFEQLCINYTNEKLQQFFNHHMFVL
EQEEYMKEEIVWDFIDFGHDLQPTIDLIEKANPIGILSCLDEECVMPKATDATFTSKLDA
LWRNKSLKYKPFKFADQGFILTHYAADVPYSTEGWLEKNTDPLNENVAKLLAQSTNKHVA
TLFSDYQETETKTVRGRTKKGLFRTVAQRHKEQLNQLMNQFNSTQPHFIRCIVPNEEKKM
HTFNRPLVLGQLRCNGVLEGIRITRAGFPNRLPFNDFRVRYEIMAHLPTGTYVESRRASV
MILEELKIDEASYRIGVSKIFFKAGVLAELEERRVATLQRLMTMLQTRIRGFLQRKIFQK
RLKDIQAIKLLQANLQVYNEFRTFPWAKLFFNLRPLLSSTQNDKQLKKRDAEIIELKYEL
KKQQNSKSEVERDLVETNNSLTAVENLLTTERAIALDKEEILRRTQERLANIEDSFSETK
QQNENLQRESASLKQINNELESELLEKTSKVETLLSEQNELKEKLSLEEKDLLDTKGELE
SLRENNATVLSEKAEFNEQCKSLQETIVTKDAELDKLTKYISDYKTEIQEMRLTNQKMNE
KSIQQEGSLSESLKRVKKLERENSTLISDVSILKQQKEELSVLKGVQELTINNLEEKVNY
LEADVKQLPKLKKELESLNDKDQLYQLQATKNKELEAKVKECLNNIKSLTKELENKEEKC
QNLSDASLKYIELQEIHENLLLKVSDLENYKKKYEGLQLDLEGLKDVDTNFQELSKKHRD
LTFNHESLLRQSASYKEKLSLASSENKDLSNKVSSLTKQVNELSPKASKVPELERKITNL
MHEYSQLGKTFEDEKRKALIASRDNEELRSLKSELESKRKLEVEYQKVLEEVKTTRSLRS
EVTLLRNKVADHESIRSKLSEVEMKLVDTRKELNSALDSCKKREAEIHRLKEHRPSGKEN
NIPAVKTTEPVLKNIPQRKTIFDLQQRNANQALYENLKRDYDRLNLEKHNLEKQVNELKG
AEVSPQPTGQSLQHVNLAHAIELKALKDQINSEKAKMFSVQVQYEKREQELQKRIASLEK
VNKDSLIDVRALRDRIASLEDELRAA
```

View the results for Sequence 1. The first column of the results table identifies whether or not the match is of type "family" or of type "domain". The family and domain names appear at the top of each box in the second column of the results page, the same column that contains the diagrams which show the localization of the section of sequence that has been identified with the referenced family or domain.

Exercise 8.35

How many matches were of the type "family"?

Exercise 8.36

How many were domains?

Exercise 8.37

What are the names of the families identified with this sequence?

Exercise 8.38

List any domains that were identified within Sequence 1.

View the results for Sequence 2.

Exercise 8.39

How many families were returned as matches?

Exercise 8.40
How many domains?

Exercise 8.41
What families were identified with this sequence?

Exercise 8.42
List any domains that were identified within Sequence 2.

View the results for Sequence 3.

Exercise 8.43
How many families were returned as matches?

Exercise 8.44
How many domains?

Exercise 8.45
What families were identified with this sequence?

Exercise 8.46
List any domains that were identified within Sequence 3.

Return to the ExPASy Proteomics Tools server[26] . Now, scroll down to the section entitled "post-translational modification prediction". Use *NetPhos*[14] (4) to predict possible sites for serine, threonine and tyrosine phosphorylation on the three sequences above (all 3 sequences can be entered as one query). Accept the default values and select "submit". For help interpreting the results, view the NetPhos output format[27] .

Exercise 8.47
How many (a) serine, (b) threonine, and (c) tyrosine phosphorylation sites are predicted for Sequence 1?

Exercise 8.48
How many (a) serine, (b) threonine, and (c) tyrosine phosphorylation sites are predicted for Sequence 2?

Exercise 8.49
How many (a) serine, (b) threonine, and (c) tyrosine phosphorylation sites are predicted for Sequence 3?

Exercise 8.50
Are there any serine, threonine and tyrosine in the sequence that were not listed as a potential phosphorylation site? If so, explain why some of the residues were not listed as predicted phosphorylation sites. (Those uncertain about the answer to this question should view the above link explaining the output.)

Once a protein sequence has been determined through proteomics techniques, bioinformatics can be used to predict certain types of topology. Topology is the sequence of secondary structure elements within a protein. The most basic secondary structure elements within proteins are the alpha helix, the beta sheet and the random coil. However, some algorithms will predict topological features that are closely related to *in vivo* localization, such as signal sequences and transmembrane helices.

At the ExPASy Proteomics Tools server[28] , scroll down on the ExPASy tools webpage to the section entitled "topology prediction". This section contains tools that predict localization and sorting signals, as well as transmembrane regions within proteins. *PSORT*[31] (5) is a computer program for the prediction of protein localization. It requires input of an amino acid sequence and its source organism; and it searches for known, organism-specific protein sorting signals. It returns a list of candidate localization sites, accompanied by a score indicating the probability the protein encoded by the input sequence would be localized to that

[26] http://us.expasy.org/tools/
[27] http://www.cbs.dtu.dk/services/NetPhos-2.0/output.html
[28] http://us.expasy.org/tools/

site. To explore the use of PSORT, click on the PSORT link on the ExPASy tool page. Choose the "PSORT II" for eukaryotic sequences, and select the PSORT II Prediction. Cut and paste the following sequence for diacylglycerol kinase from Rattus norvegicus into the query box and click "Submit".

```
MEPRDPSPEARSSDSESASASSSGSERDADPEPDKAPRRLTKRRFPGLRLFGHRKAITKSGLQHLAPPPP
TPGAPCGESERQIRSTVDWSESAAYGEHIWFETNVSGDFCYVGEQYCVAKMLPKSAPRRKCAACKIVVHT
PCIGQLEKINFRCKPSFRESGSRNVREPTFVRHHWVHRRRQDGKCRHCGKGFQQKFTFHSKEIVAISCSW
CKQAYHSKVSCFMLQQIEEPCSLGVHAAVVIPPTWILRARRPQNTLKASKKKKRASFKRRSSKKGPEEGR
WRPFIIRPTPSPLMKPLLVFVNPKSGGNQGAKIIQSFLWYLNPRQVFDLSQGGPREALEMYRKVHNLRIL
ACGGDGTVGWILSTLDQLRLKPPPPVAILPLGTGNDLARTLNWGGGYTDEPVSKILSHVEEGNVVQLDRW
DLRAEPNPEAGPEERDDGATDRLPLDVFNNYFSLGFDAHVTLEFHESREANPEKFNSRFRNKMFYAGTAF
SDFLMGSSKDLAKHIRVVCDGMDLTPKIQDLKPQCIVFLNIPRYCAGTMPWGHPGEHHDFEPQRHDDGYL
EVIGFTMTSLAALQVGGHGERLTQCREVLLTTAKAIPVQVDGEPCKLAASRIRIALRNQATMVQKAKRRS
TAPLHSDQQPVPEQLRIQVSRVSMHDYEALHYDKEQLKEASVPLGTVVVPGDSDLELCRAHIERLQQEPD
GAGAKSPMCHPLSSKWCFLDATTASRFYRIDRAQEHLNYVTEIAQDEIYILDPELLGASARPDLPTPTSP
LPASPCSPTPGSLQGDAALPQGEELIEAAKRNDFCKLQELHRAGGDLMHRDHQSRTLLHHAVSTGSKEVV
RYLLDHAPPEILDAVEENGETCLHQAAALGQRTICHYIVEAGASLMKTDQQGDTPRQRAEKAQDTELAAY
LENRQHYQMIQREDQETAV
```

First, view the "k-NN" results by scrolling to the bottom of the page. The k-nearest neighbor (k-NN) algorithm takes the output of the many subprograms and determines a probability of localization at each candidate site within the cell using all of the predictions.

Exercise 8.51

What is the probability the sequence encodes a protein that is (a) secreted by vesicles? (b) localized to the endoplasmic reticulum? (c) cytoplasmic? or (d) localized to the nucleus?

Now, scroll through the results of the subprograms. Clicking on the links will reveal a brief description of the algorithm each individual subprogram utilizes.

Exercise 8.52

What is the localization prediction and reliability score produced by the NNCN subprogram, Reinhardt's methods for cytoplasmic/nuclear discrimination?

The first two subprograms, PSG and GvH, are tools that predict N-terminal signal peptide sequences. Just after their results are listed, there is a statement summarizing whether or not an N-terminal signal peptide has been predicted for the query sequence.

Exercise 8.53

Do these subprograms predict an N-terminal signal peptide for the diacylglycerol kinase query?

Exercise 8.54

After looking over all the results, what is the most likely localization of our query protein?

Read the title and abstract for this article[29] on the Rat diacylglycerol kinase used for the query sequence.

Exercise 8.55

Was PSORT able to predict the correct localization, using the sequence information alone?

Return to the ExPASy tools[30] , and scroll to the section entitled "primary structure analysis". Click on the link for the ProtParam tool. ProtParam is a suite of programs designed to predict various chemical and physical properties about a protein from its sequence. ProtParam will yield an estimated extinction coefficient at selected wavelengths based on protein sequence *(6)*[25], an estimation of the *in vivo* half-life of the protein (*7*[10] *8*[26] *9*[43] *10*[17]), an instability index *(11)*[27], an aliphatic index *(12)*[4], and an

[29] http://www.pnas.org/cgi/content/abstract/93/20/11196

[30] http://us.expasy.org/tools/

average value for hydropathicity *(13)*[34]. Cut and paste the Rat diacylglycerol kinase sequence above into the query box and click on "compute parameters".

Exercise 8.56
What is the molecular weight computed from the sequence?

Exercise 8.57
What does the amino acid composition analysis show as the most common amino acid in this protein? (Is that unusual?)

Exercise 8.58
What is the chemical formula for the query protein?

Exercise 8.59
What is the predicted extinction coefficient at 280 nm, in 6M guanidium HCl, 0.02M phosphate, pH6.5 buffer, assuming all cysteines appear as half cysteines?

Exercise 8.60
In what way could it be helpful to know the extinction coefficient?

Exercise 8.61
According to the instability index, is this protein classified as stable or unstable?

Return again to the ExPASy tools[31] . Notice there are two sections dealing with structure prediction, secondary structure prediction tools and tertiary structure prediction and visualization tools. The secondary structure prediction tools are designed to predict features such as the helical content, the beta sheet formations, and the turns, loops, and coil regions within a protein, given the sequence.

Exercise 8.62
Explore the secondary structure tools independently, and submit the diacylglycerol kinase sequence above to any of the available secondary structure prediction tools. Most of these tools will email the results, with at least a 20 minute delay between submission and receipt of results. Forward a results summary to the instructor, outlining the predictions created by the program of choice.

Tertiary structure prediction tools match the query sequence with sequences, or partial sequences, of proteins where the 3-D structure has been published in the Protein Data Bank (PDB). These tools will produce a model of the query protein by piecing together the structural regions from the best matches in the PDB, and threading the query sequence through the predicted structure. For more detailed explanations of available 3-D structure prediction software, view the Swiss-Model demo page[32] and the Geno3D reference page[33] . Although both of these tools are searching for templates from existing PDB entries, they are doing this in different ways.

Exercise 8.63
What program does Swiss-Model use to match the query sequence with sequences of known structures?

Exercise 8.64
What program does Geno3D use to match the query sequence with sequences of known structures?

Notice that the template selection process and the model structure refinement processes are different between these two programs as well.

Finally, in the tertiary structure section of the ExPASy tools page, Swiss PDB Viewer is a graphical tool for the visualization, comparison and analysis of 3-D coordinate files. Swiss PDB Viewer can superimpose 3-D structures by finding the rotation and translation that most closely aligns the two protein structures. Additionally, the Swiss PDB Viewer will perform amino acid mutations, prediction of hydrogen bonds, and calculation of angles and distances between atoms. Best of all, Swiss PDB Viewer is freeware and available for many different platforms, including Macintosh, PC, SGI IRIX, and Linux.

[31] http://us.expasy.org/tools/

[32] http://www.expasy.org/swissmod/SM_Demo_FA.html

[33] http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_references.html

Exercise 8.65

View this supplemental SPDBV web page[34] . What other function does Swiss PDB Viewer have, when used in conjunction with other applications such as OpenGL or POV-Ray?

ExPASy provides a very large library of tools, for proteomics as well as other bioinformatics applications.For those students interested in future research in the field of proteomics, this web server will be an important resource.

8.4 Protein Folding and Secondary Structure Prediction[35]

Proteins are the biological molecules that are the building blocks of cells and organs, and the biochemical processes required to keep living organisms alive are catalyzed and regulated by a particular category of proteins called enzymes. Proteins are linear polymers of amino acids that fold into complex conformations dictated by the physical and chemical properties of the amino acid chain. The biological function of a protein is dependent on the protein folding into the correct, or "native", state. Protein structure is described by biologists in terms of primary structure, which is the amino acid sequence, secondary structure, wherein the polypeptide backbone assembles into local regions of alpha-helices, beta-sheets, coils and turns, tertiary structure, which refers to the entire 3-dimensional structure of the protein, and quaternary structure, which describes interactions between separate polypeptide chains, called subunits, that exist in some large protein complexes. Computational methods have been developed that can predict protein secondary structure with a reasonable degree of accuracy. Prediction methods exist for predicting tertiary structure, but the accuracy of such methods is highly dependent on whether or not the protein in question is related in sequence to any members of the existing library of known protein structures. The development of *ab initio* tools to predict the complete structural fold of a protein from its amino acid sequence is a burgeoning field in computational biology, but true attainment of this goal is still pretty distant.

Protein folding is usually a spontaneous process, and often when a protein unfolds because of heat or chemical denaturation, it will be capable of refolding into the correct conformation, as soon as it is removed from the environment of the denaturant, meaning folding and unfolding under these circumstances are reversible. Protein folding can go wrong for many reasons. When an egg is boiled, the proteins in the white unfold and misfold into a solid mass of protein that will not refold or redissolve. In a similar way, irreversibly misfolded proteins form insoluble protein aggregates found in certain tissues that are characteristic of some diseases, such as Alzheimer's Disease.

Determining the process by which proteins fold into particular shapes, characteristic of their amino acid sequence, is commonly called "the protein folding problem". One approach to studying the protein folding process is the application of statistical mechanics techniques and simulations to the *study of protein folding*.[16] (1) These methods allow the investigation of larger systems than methods that try to represent atomic detail in their simulations of biological molecules, and have had success correlating the computational folding model with folding intermediates and transition states that have been experimentally measured for a limited test set of relatively large proteins.

An approach that uses an atomistic model for protein folding in a solvent environment is being taken by *The Stanford University* [44] (2) Folding@home[36] project, using large scale distributed computing that allows timescales thousands to millions of times longer than previously achievable with a model of this detail. Look at the menu on the left border of the Stanford Folding@home[37] web page. Click on the "Science" link to read the scientific background behind the protein folding distributed computing project.

Exercise 8.66

What are the 3 functions of proteins that are mentioned in the "What are proteins?" section of the scientific background?

[34] http://us.expasy.org/spdbv/text/gallery.htm

[35] This content is available online at <http://cnx.org/content/m11116/2.4/>.

[36] http://folding.stanford.edu/

[37] http://folding.stanford.edu/

Exercise 8.67

What are 3 diseases that are believed to result from protein misfolding?

Exercise 8.68

What are typical timescales for molecular dynamics simulations?

Exercise 8.69

What are typical timescales at which the fastest proteins fold?

Exercise 8.70

How does the Stanford group break the microsecond barrier with their simulations?

Return to the Stanford Folding@home[38] home page. Click on the "Results" link in the left border of the web page. Look at the information on the folding simulations of the villin headpiece.

Exercise 8.71

How many amino acids are in the simulated villin headpiece?

Exercise 8.72

How does this compare with the number of amino acids in a typical protein?

Exercise 8.73

Taking into consideration the size of the biological molecules in these simulations and the requirements that necessitated using large scale distributed computing methods for the simulations, what are the biggest impediments to understanding the protein folding problem?

Although attempts at predicting tertiary and quaternary structure from the amino acid sequence of proteins are relatively new, methods for predicting protein secondary structure have been in existence for some time. Depending on the method, secondary structure predictions can be performed with approximately 60 - 70% accuracy. Originally, empirical prediction methods were based on tables which listed each amino acid and the frequency with which that amino acid was found in alpha-helices, beta-sheets, turns and random coil. Currently, prediction methods usually employ machine learning in the form of neural networks that are trained with test sets consisting of sequences with known structure. In these cases, the selection of the test set is critically related to the accuracy of the method. However, given the ever increasing number of known structural folds, selecting a representative test set that includes many proteins of diverse structure has become easier.

Use the amino sequence below to explore some structure prediction tools. This is the sequence for lac repressor, a protein involved in gene regulation that is known to have both alpha-helical and beta-sheet structure:

```
>gi|33112645|sp|P03023|LACI_ECOLI Lactose operon repressor
MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQSLLIGVATSS
LALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAAC
TNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQI
QPIAEREGDWSAMSGFQQTMQMLNEGIVPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSS
CYIPPLTTIKQDFRLLGQTSVDRLLQLSQGQAVKGNQLLPVSLVKRKTTLAPNTQTASPRALADSLMQLA
RQVSRLESGQ
```

A quick and simple analysis of protein secondary structure can be performed by the nnpredict tool [39] at *UCSF.*[18] (3) Notice when pasting the above sequence into the query page that there is a separate line for the name of sequence, meaning that the first line in the above fasta format sequence should be entered here, separately from the rest of the sequence. Compare the nnpredict results to the actual secondary structure of lac repressor, known from the *crystal structure with PDBID 1LBI.*[36] (4)

[38] http://folding.stanford.edu/

[39] http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

```
Sequence and secondary structure - lac Repressor
Data is from PDB accession number 1LBI, RCSB Protein Data Bank.
The legend for the assignments are:
H=helix; B=residue in isolated beta bridge; E=extended beta strand;
G=310 helix; I=pi helix; T=hydrogen bonded turn; S=bend.

  1 MKPVTLYDVA EYAGVSYQTV SRVVNQASHV SAKTREKVEA AMAELNYIPN


 51 RVAQQLAGKQ SLLIGVATSS LALHAPSQIV AAIKSRADQL GASVVVSMVE
                  EEEEEEES S  HHHHHHH HHHHHHHHHH T EEEEEEE

101 RSGVEACKTA VHNLLAQRVS GLIINYPLDD QDAIAVEAAC TNVPALFLDV
    SSHHHHHHHH HHHHHHS  S EEEEES   S TTHHHHHHTS  SS EEESSS

151 SDQTPINSII FSHEDGTRLG VEHLVALGHQ QIALLAGPLS SVSARLRLAG
     TTSSS EEE E TTHHHHHH HHHHHHHT    EEEEE  SS SSHHHHTHHH

201 WHKYLTRNQI QPIAEREGDW SAMSGFQQTM QMLNEGIVPT AMLVANDQMA
    HHHHHTTTT    SEEEE  S SHHHHHHHHH HHHTTT   S EEEESSHHHH

251 LGAMRAITES GLRVGADISV VGYDDTEDSS CYIPPLTTIK QDFRLLGQTS
    HHHHHHHHTT TTTBTTTEEE E SB  TTGG GSSS   EEE   HHHHHHHH

301 VDRLLQLSQG QAVKGNQLLP VSLVKRKTTL APNTQTASPR ALADSLMQLA
    HHHHHHHHT   S  S EEE   EEE  TT   S TTS   HH HHHHHHHHHH

351 RQVSRLESGQ
    HHHHHH
```

Exercise 8.74

Look for regions identified as alpha-helical by nnpredict, but not identified as alpha-helical in the actual secondary structure features listed above, and vice versa. Are there any regions of 3 consecutive amino acids or more that differ? If so, how many alpha-helical regions differ and what residue numbers are involved?

Exercise 8.75

Look for regions identified as beta-sheet by nnpredict, but not identified as beta-sheet in the actual secondary structure features listed above, and vice versa. Are there any regions of 3 consecutive amino acids or more that differ? If so, how many beta-sheet regions differ and what residue numbers are involved?

Exercise 8.76

The PDB entry for the crystal structure of lac Repressor remarks that the N-terminal residues number 1 - 61 and the C-terminal residues number 358 - 360 are not seen in the electron density. How would this effect the assignment of actual secondary structure shown above?

A more complete sequence analysis tool that includes secondary structure prediction can be found at the PredictProtein server [40] at *Columbia University*.[39] (5) On the home page, select the tab for submission to enter a protein sequence for secondary structure prediction. The instructions indicate that you should only

[40] http://www.predictprotein.org/

enter the amino acid sequence, so omit the first line when you paste in the fasta format sequence above for lac Repressor. Click on the "Results on the site, not in email" option. The server will still send an email with a link to the results. Run the prediction tool for lac Repressor sequence. When the email arrives, click on the link for your results. The secondary structure prediction algorithms are within the section entitled PROF. This is the section needed to answer the following questions.

Exercise 8.77

Give a brief summary comparing the ProteinPredict results with the actual secondary structure from the PDB, listed above, and with the results from nnpredict.

With the publication of entire genomes that contain sequences to many unknown proteins, scientists would love to have the ability to predict the final folded structure of a protein based on its sequence. Although this is not yet a practical reality, tools exist that can predict secondary structure with some accuracy and inroads are being made toward solving the protein folding problem. Elucidating the mechanisms behind protein folding would provide important knowledge for fighting disease states where misfolded proteins are implicated.

Bibliography

[1] Dolan dna learning center. http://www.bioservers.org/.

[2] Genomics course, department of biology, davidson college, davidson, nc 28036. http://www.bio.davidson.edu/courses/genomics/genomics.html.

[3] Internet course on the principles of protein structure, a collaboration between birkbeck college and the virtual school of natural sciences(vsns) of the globewide network academy (gna). http://www.cryst.bbk.ac.uk/PPS95/.

[4] Ikai A. Thermostability and aliphatic index of globular proteins. *J. Biochem.*, pages 88:1895–1898, 1980.

[5] Miller W. Myers E.W. Lipman D.J. Altschul S.F., Gish W. Basic local alignment search tool. *J. Mol. Biol.*, pages 215:403–410, 1990.

[6] Miller W. Myers E.W. Lipman D.J. Altschul S.F., Gish W. Basic local alignment search tool. *J. Mol. Biol.*, pages 215:403–410, 1990.

[7] Hochstrasser D.F. Appel R.D., Bairoch A. A new generation of information retrieval tools for biologists: the example of the expasy www server. *Trends Biochem. Sci.*, pages 19:258–260, 1994.

[8] J.A. Doudna A.R. Ferre-D'Amare, K. Zhou. Crystal structure of a hepatitis delta virus ribozyme. *Nature*, 395:567–574, 1998.

[9] Aravind and Koonin. Gleaning non-trivial structural,functional and evolutionary information about proteins by interative database searches. *JMB*, pages 287:1023–1040, 1999.

[10] Varshavsky A. Bachmair A., Finley D. In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, pages 234:179–186, 1986.

[11] Lipman D.J. Ostell J. Benson D.A., Boguski M.S. Genbank. *Nucleic Acids Res.*, pages 22:3441–3444, 1994.

[12] Lipman D.J. Ostell J. Benson D.A., Boguski M.S. Genbank. *Nucleic Acids Res.*, pages 22:3441–3444, 1994.

[13] Puskas RS Eschenfeldt WH Berger SL, Wallace DM. Reverse transcriptase and its associated ribonuclease h: interplay of two enzyme activities controls the yield of single-stranded complementary deoxyribonucleic acid. *Biochemistry*, pages 22(10):2365–72, 1983.

[14] Gammeltoft S. Blom, N. and S. Brunak. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, pages 294(5): 1351–1362, 1999.

[15] Valencia A. Bork P, Sander C. An atpase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc Natl Acad Sci U S A*, pages 89(16):7290–4, 1992.

[16] H. Nymeyer C. Clementi and J.N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and 'on-route' intermediates for protein folding? an investigation for small globular proteins. *Journal of Molecular Biology*, pages 298: 937–953, 2000.

[17] Schwartz A.L. Ciechanover A. How are substrates recognized by the ubiquitin-mediated proteolytic system? *Trends Biochem. Sci.*, pages 14:483–488, 1989.

[18] F. E. Cohen D. G. Kneller and R. Langridge. Improvements in protein secondary structure prediction by an enhanced neural network. *JMB*, pages 214:171–182, 1990.

[19] Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet.*, pages 32 Suppl:502–8, 2002.

[20] Altschul et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, pages 25(17):3389–3402, 1997.

[21] Schwartz et al. Pipmaker-a web server for aligning two genomic dna sequences. *Genome Research*, pages 10:577–586, 2000.

[22] Wilkins et al. Progress with gene product mapping of the mollicutes. *Electrophoresis*, pages 16:1090–1094, 1995.

[23] G.J.B.Williams E.F.Meyer Jr M.D.Brice J.R.Rodgers O.Kennard T.Shimanouchi M.Tasumi F.C.Bernstein, T.F.Koetzle. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[24] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, pages 266:418–27, 1996.

[25] von Hippel P.H. Gill S.C. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.*, pages 182:319–326, 1989.

[26] Wunning I. Tobias J.W. Lane W.S. Varshavsky A. Gonda D.K., Bachmair A. Universality and structure of the n-end rule. *J. Biol. Chem.*, pages 264:16700–16712, 1989.

[27] Pandit M.W. Guruprasad K., Reddy B.V.B. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering*, pages 4:155–161, 1990.

[28] Henikoff JG. Henikoff S. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.*, pages 89(22):10915–9, 1992.

[29] Henikoff JG. Henikoff S. Performance evaluation of amino acid substitution matrices. *Proteins*, pages 17(1):49–61, 1993.

[30] Z.Feng G.Gilliland T.N.Bhat H.Weissig I.N.Shindyalov P.E.Bourne H.M.Berman, J.Westbrook. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[31] Paul Horton and Kenta Nakai. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Intelligent Systems for Molecular Biology*, pages 5:147–152, 1997.

[32] R.W. Warrant G.M. Church S.-H. Kim J.L. Sussman, S.R. Holbrook. Crystal structure of yeast phenylalanine t-rna. *J. Mol. Biol.*, 123:607–630, 1978.

[33] WJ. Kent. Blat–the blast-like alignment tool. *Genome Res.*, pages 12(4):656–64, 2002.

[34] Doolittle R.F. Kyte, J. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, pages 157:105–132, 1982.

[35] Maddison DR Maddison WP. Interactive analysis of phylogeny and character evolution using the computer program macclade. *Folia Primatol (Basel)*, pages 53(1–4):190–202, 1989.

[36] N.C.HORTON M.A.KERCHER H.C.PACE M.A.SCHUMACHER R.G.BRENNAN P.LU M.LEWIS, G.CHANG. Crystal structure of the lactose operon repressor and its complexes with dna and inducer. *SCIENCE*, page 271:1247, 1996.

[37] Attwood T.K. Bairoch A. Barrell D. Bateman A. Binns D. Biswas M. Bradley P. Bork P. Bucher P. Copley R.R. Courcelle E. Das U. Durbin R. Falquet L. Fleischmann W. Griffiths-Jones S. Haft D. Harte N. Hulo N. Kahn D. Kanapin A. Krestyaninova M. Lopez R. Letunic I. Lonsdale D. Silventoinen V. Orchard S.E. Pagni M. Peyruc D. Ponting C.P. Selengut J.D. Servant F. Sigrist C.J.A. Vaughan R Zdobnov E.M. Mulder N.J., Apweiler R. The interpro database, 2003 brings increased coverage and new features. *Nucl. Acids. Res.*, pages 31:315–318, 2003.

[38] V. L. Murthy. Rnabase, the rna structure database. *http://www.rnabase.org*, Copyright 2000-2002.

[39] B Rost. Phd: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology*, pages 266:525–539, 1996.

[40] Subramaniam S. The biology workbench–a seamless database and analysis environment for the biologist. *Proteins*, pages 32(1):1–2, 1998.

[41] Dayhoff MO Schwartz RM. *Atlas of Protein Sequence and Structure, 5 suppl.*, volume 3:353-358. Nat. Biomed. Res. Found., Washington D.C., 978.

[42] Gibson T.J. Thompson J.D., Higgins D.G. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, pages 22:4673–4680, 1994.

[43] Rocap G. Varshavsky A. Tobias J.W., Shrader T.E. The n-end rule in bacteria. *Science*, pages 254:1374–1377, 1991.

[44] Sorin E. Zagrovic B. and Pande V. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *JMB*, pages 313:151–169, 2001.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

A amino acid sequence, § 2.1(9)
atomic coordinates, § 1.3(4)

B BCM Search Launcher, § 1.4(7)
bioinformatics, § 1.1(1), § 1.2(2), § 1.3(4), § 1.4(7), § 2.1(9), § 3.1(17), § 4.1(23), § 6.1(31), § 7.1(35), § 8.3(45)
BLAST, § 2.1(9)
BLAT, § 2.1(9)
BLOSUM, § 4.1(23)

C cladogram, § 5.1(27)
clustalw, § 3.1(17)
computational biology, § 8.4(51)

D data mining, § 1.4(7)
database, § 1.2(2)
database search, § 7.1(35)
DNA, § 8.2(40)

E EBI, § 1.4(7)
Entrez, § 1.1(1), § 1.2(2)
evolutionary distance, § 5.1(27)
ExPASy, § 1.4(7), § 8.3(45)
expression levels, § 8.2(40)

G GenBank, § 1.2(2)
gene expression, § 8.2(40)

M microarray, § 8.2(40)
misfolding, § 8.4(51)
mRNA, § 8.2(40)
multiple sequence alignment, § 3.1(17), § 7.1(35)

N NCBI, § 1.1(1), § 1.2(2)
nucleic acid, § 8.1(37)
nucleic acid structure, § 1.3(4)
nucleotide sequence, § 1.1(1), § 2.1(9), § 2.2(13)

P pair-wise sequence alignment, § 2.1(9)
PAM, § 4.1(23)
percent identity plot, § 2.2(13)
phylogenetic tree, § 7.1(35)
phylogeny, § 5.1(27)
pip, § 2.2(13)
project, § 7.1(35)
protein database search, § 6.1(31)
protein folding, § 8.4(51)
protein motif profile, § 6.1(31)
protein secondary structure, § 8.4(51)
protein sequence, § 1.1(1), § 3.1(17)
protein structure, § 1.3(4)
proteome, § 8.3(45)
proteomics, § 8.3(45)
PSI-BLAST, § 6.1(31)

R related sequences, § 5.1(27)
RNA free energy, § 8.1(37)
RNA secondary structure, § 8.1(37)

S scoring matrix, § 4.1(23)
search, § 1.2(2)
sequence alignment, § 2.1(9), § 2.2(13), § 4.1(23), § 6.1(31)
sequence analysis, § 1.4(7)
sequence comparison, § 3.1(17)
sequence database, § 1.1(1)
sequence homology, § 5.1(27)
simulations, § 8.4(51)
structure files, § 1.3(4)
structure prediction, § 8.1(37), § 8.4(51)
substitution matrix, § 4.1(23)

Attributions

Collection: *Bios 533 Bioinformatics*
Edited by: Susan Cates
URL: http://cnx.org/content/col10152/1.16/
License: http://creativecommons.org/licenses/by/1.0

Module: "NCBI: National Center for Biotechnology Information"
By: Susan Cates
URL: http://cnx.org/content/m11789/1.3/
Pages: 1-2
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Entrez"
By: Susan Cates
URL: http://cnx.org/content/m10996/2.7/
Pages: 2-4
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "PDB"
By: Susan Cates
URL: http://cnx.org/content/m10997/2.4/
Pages: 4-7
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Tour of Bioinformatics Sites"
By: Susan Cates
URL: http://cnx.org/content/m10998/2.2/
Pages: 7-8
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Introduction to Sequence Alignment"
By: Susan Cates
URL: http://cnx.org/content/m11026/2.13/
Pages: 9-13
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "PipMaker"
By: Susan Cates
URL: http://cnx.org/content/m11027/2.5/
Pages: 13-15
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Multiple Sequence Alignment"
By: Susan Cates
URL: http://cnx.org/content/m11036/2.11/
Pages: 17-22
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Scoring Matrices"
By: Susan Cates
URL: http://cnx.org/content/m11062/2.7/
Pages: 23-26
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Phylogenetic Trees"
By: Susan Cates
URL: http://cnx.org/content/m11052/2.8/
Pages: 27-29
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "PSI-BLAST"
By: Susan Cates
URL: http://cnx.org/content/m11040/2.13/
Pages: 31-34
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Bioinformatics Project"
By: Susan Cates
URL: http://cnx.org/content/m11881/1.2/
Pages: 35-36
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "RNA Secondary Structure Prediction"
By: Susan Cates
URL: http://cnx.org/content/m11065/2.4/
Pages: 37-40
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Microarray Experiments"
By: Susan Cates
URL: http://cnx.org/content/m11050/2.17/
Pages: 40-45
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Expasy Proteomics Tools"
By: Susan Cates
URL: http://cnx.org/content/m11082/2.5/
Pages: 45-51
Copyright: Susan Cates
License: http://creativecommons.org/licenses/by/1.0

Module: "Protein Folding and Secondary Structure Prediction"
By: Susan Cates
URL: http://cnx.org/content/m11116/2.4/
Pages: 51-54

Bios 533 Bioinformatics
Computer laboratory modules for the Introduction to Bioinformatics course. This course is designed for the beginning graduate student or advanced undergraduate in the biosciences. The goal is to introduce the student to various biologically relevant databases, methods to effectively search the databases, and an overall view of the various aspects of computational biology.

About Connexions
Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities.

Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai. Connexions is part of an exciting new information distribution system that allows for **Print on Demand Books**. Connexions has partnered with innovative on-demand publisher QOOP to accelerate the delivery of printed course materials and textbooks into classrooms worldwide at lower prices than traditional academic publishers.