**Bioinformatics Exercises**

Over the last two decades, information has been gaining increasing importance in both  teaching and learning biochemistry.  The most obvious case is the sequencing of the human genome and many other complete genomes.  In 1990, the determination of the sequence of a protein was often the topic of a full publication in a peer-reviewed journal such as *Science, Nature,*or *The Journal of Biological Chemistry.*  Now entire genomes are the topic of individual research papers.  The term "bioinformatics" is a catch-all phrase which generally refers to the use computers and computer science approaches to the study of biological systems.  The main chapters where this information is discussed in the text are chapters 3 (Nucleotides, Nucleic Acids and Genetic Information), 5 (Proteins: Primary Structure), 6 (Proteins: Three-Dimensional Structure), 12 (Enzyme Kinetics, Inhibition and Regulation) and 13 (Introduction to Metabolism).  Here we provide exercises appropriate to these chapters aimed at introducing the techniques of bioinformatics that involve the use of computers, Internet-accessible databases and the tools that have been developed to “mine” those databases.

**General principles**

1.       *Open ended questions*.  The exercises may include some questions that have definite answers, but in many cases there will also be questions which may be answered in a number of ways, depending on the approach you take or the topic you select.

2.       *Stable Internet Resources*.  As much as possible, the exercises will be based on well established, stable web sites.  If it is necessary to use less reliable sites and/or resources, attempts have been made to provide multiple sites that perform similar functions.

3.       Here are the stable online resources that will be used most frequently:

a.       Genbank  (//www.ncbi.nlm.nih.gov/)

b.       Protein Data Bank  (http://www.rcsb.org)

c.        Expasy Proteomics Server  (http://us.expasy.org/)

d.       European Bioinformatics Institute  (//www.ebi.ac.uk/)

e.        Pfam  (//www.sanger.ac.uk/Software/Pfam/)

f.        SCOP  (http://scop.mrc-lmb.cam.ac.uk/scop/)

g.        CATH  (http://www.biochem.ucl.ac.uk/bsm/cath/)

h.       PubMed  (//www.ncbi.nlm.nih.gov/entrez/query.fcgi)

i.         PubMed Central  (http://www.pubmedcentral.nih.gov/)

4.       *Answer key*.  Where a definite answer is known, it will be provided in an answer key.  For more open-ended questions, a typical correct answer will be presented.

5.       *Historical perspective*.   If historical resources are available online (including PubMed), there may be questions designed to help students identify some of the historical roots of biochemistry and molecular biology.

**Project 1: Databases for the Storage and “Mining” of Genome Sequences**

Chapter 3 is an introduction to nucleotides, nucleic acids (DNA and RNA), and the processes of transcription and translation.

1.       **Finding Databases.**One of the major bioinformatics tools is the biological database.  These databases are an important resource for the study of biochemistry at all levels.  These databases contain huge amounts of information about the sequence and structure of nucleic acids (DNA and RNA) and proteins.  They also contain software tools that can be used to analyze the data.  Some of the software you can use directly from a web browser - these tools are called web applications.  Other software must be downloaded and installed on your local computer - these are called freestanding applications.  We'll start with *finding databases*

a.       What major databases are available online that contain DNA and protein sequences?

b.       Which database contains entire genomes?

c.        Using your textbook and online resources (http://www.google.com), make sure you understand the meaning of the following terms.

                     i.            BLAST

                    ii.            Taxonomy

                  iii.            gene ontology

                  iv.            phylogenetic trees

                   v.            multiple sequence alignment

d.       Once you have defined these terms, find resources on the Internet which enable you to study them.

2.       **TIGR (The Institute for Genomic Research) Exercise**.  Open the TIGR site  (http://www.tigr.org).  Find the Comprehensive Microbial Resource.

a.       What 2001 publication describes the Comprehensive Microbial Resource at TIGR?

b.       How many completed genomes from *Pseudomonas* species have been deposited at TIGR?

c.        Which *Pseudomonas* species?

d.       Identify the primary reference for *Pseudomonas putida* KT2440.

e.        Find the link on the Comprehensive Microbial Resource home page for restriction digests.  Perform a computer-generated restriction digest on*Pseudomonas putida* KT2440 with BamH1.  How many fragments form and what is the average fragment size?

f.        In addition to microbial genomes, TIGR also contains the genomes of many higher organisms.  Identify five eukaryotic genomes that are available at TIGR.

3.       **Analyzing a DNA Sequence.**Throughthe use of high throughput methods, scientists are now able to sequence entire genomes in a very short period of time.  Sequencing a genome is quite an accomplishment in itself, but it is really only the beginning of the study of an organism.  Further study can be done both at the wet lab bench and on the computer.  In this problem, you will use a computer to help you identify an open reading frame, determine the protein that it will express and find the bacterial source for that protein.  Here is the DNA sequence.

|  |
| --- |
|  |

TACGCAATGCGTATCATTCTGCTGGGCGCTCCGGGCGCAGGTAAAGGTACTCAGGCTCAATTCATCATGGAGAAATACGGCATTCCGCAAATCTCTACTGGTGACATGTTGCGCGCCGCTGTAAAAGCAGGTTCTGAGTTAGGTCTGAAAGCAAAAGAAATTATGGATGCGGGCAAGTTGGTGACTGATGAGTTAGTTATCGCATTAGTCAAAGAACGTATCACACAGGAAGATTGCCGCGATGGTTTTCTGTTAGACGGGTTCCCGCGTACCATTCCTCAGGCAGATGCCATGAAAGAAGCCGGTATCAAAGTTGATTATGTGCTGGAGTTTGATGTTCCAGACGAGCTGATTGTTGAGCGCATTGTCGGCCGTCGGGTACATGCTGCTTCAGGCCGTGTTTATCACGTTAAATTCAACCCACCTAAAGTTGAAGATAAAGATGATGTTACCGGTGAAGAGCTGACTATTCGTAAAGATGATCAGGAAGCGACTGTCCGTAAGCGTCTTATCGAATATCATCAACAAACTGCACCATTGGTTTCTTACTATCATAAAGAAGCGGATGCAGGTAATACGCAATATTTTAAACTGGACGGAACCCGTAATGTAGCAGAAGTCAGTGCTGAACTGGCGACTATTCTCGGTTAATTCTGGATGGCCTTATAGCTAAGGCGGTTTAAGGCCGCCTTAGCTATTTCAAGTAAGAAGGGCGTAGTACCTACAAAAGGAGATTTGGCATGATGCAAAGCAAACCCGGCGTATTAATGGTTAATTTGGGGACACCAGATGCTCCAACGTCGAAAGCTATCAAGCGTTATTTAGCTGAGTTTTTGAGTGACCGCCGGGTAGTTGATACTTCCCCATTGCTATGGTGGCCATTGCTGCATGGTGTTATTTTACCGCTTCGGTCACCACGTGTAGCAAAACTTTATCAATCCGTTTGGATGGAAGAGGGCTCTCCTTTATTGGTTTATAGCCGCCGCCAGCAGAAAGCACTGGCAGCAAGAATGCCTGATATTCCTGTAGAATTAGGCATGAGCTATGGTTCAC

a.       First we will find an open reading frame in this segment of DNA.  What is an open reading frame (ORF)?  You can probably find the answer in your textbook or online with a simple Internet search (http://www.google.com).  You may also wish to try the bookshelf at PubMed (//www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books).  In bacteria, an open reading frame on a piece of mRNA almost always begins with AUG, which corresponds to ATG in the DNA segment that coded for the mRNA.  According to the standard genetic code (see your textbook), there are three stop codons on mRNA:  UAA, UAG, and UGA, which correspond to TAA, TAG, and TGA in the parent DNA segment.  Here are the rules for finding an open reading frame in this piece of bacterial DNA:

                     i.            It must start with ATG

                    ii.            It must end with TAA, TAG or TGA

                  iii.            It must be at least 300 nucleotides long (coding for 100 amino acids)

                  iv.            The ATG start codon and the stop codon must be in frame.  This means that if you count the total number of bases in the sequence from the start to the stop codon, it must be evenly divisible by 3.  
  
Hints:  I suggest you do this search by pasting the DNA sequence into a Word processor, then searching for the start and stop codons. Once you have found a pair, you need to count the number of bases between them (including the 3 bases each for the start and stop codon).  If you use Microsoft Word, you can highlight the text of your proposed ORF, then use the Tools..Word Count option from the drop-down menu and check the number for "Characters."  This number must be divisible by 3.

b.       Admittedly, part a is a tedious approach.  Here is an easier one.  Highlight the DNA sequence again and copy it.  Then go to Translate tool on the Expasy server (http://web.expasy.org/translate/).  Paste the sequence into the box entitled, "Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).”  Then select "Verbose ("Met”, "Stop", spaces between residues)” as your Output format and click on Translate sequence.  The "Results of Translation" page which appears will contain 6 different reading frames. What is a reading frame and why are there 6?  Once again, you can refer to your textbook, the Internet or the PubMed bookshelf for the answer.  Select the reading frame that contains a protein (more than 100 continuous amino acids with no interruptions by a stop codon).  Which reading frame is that? Write that down.  Now go back to the Translate tool page, leave the DNA sequence in the sequence box, but select "Compact ("M", "-", no spaces)" as your Output format.  Go to the same reading frame as before and copy the protein sequence (by one letter abbreviations) starting with "M" for methionine and ending in "-" for stop codon.  Save this sequence to a separate text file.

c.        Now we're going to identify the protein and the bacterial source.  Go to the NCBI BLAST page (//www.ncbi.nlm.nih.gov/BLAST/).  What does BLAST stand for?  We're going to do a simple BLAST search using our protein sequence, but you can do much more with BLAST.  You are encouraged to work the Tutorials on the BLAST home page to learn more.  On the BLAST page, select protein BLAST.  Enter your search sequence in the "Search" box.  Use the default values for the rest of the page and click on the BLAST! button. You may have to wait a while before you get your results.  What is the protein and what is the source?  It should be the first one that is listed in the BLAST output.    
  
  
For instructors:  You can do this exercise with any DNA sequence you choose.  It is probably best to choose a DNA segment that encodes only one protein.

4.       **Sequence Homology**.  We are going to use the protein that you identified in problem 3 to look at sequence homology with BLAST.

a.       First - some definitions: what do the terms "homolog", "ortholog", and "paralog" mean?

b.       Go to the NCBI BLAST page (//www.ncbi.nlm.nih.gov/BLAST/).  Find the text file of your protein sequence and paste it into the BLAST in the "Search" box.  Before you click on the "Search" button, we're going to narrow the search by kingdom.  As you look down the BLAST page, you'll see a Choose Search Set section.  Enter “Eukaryota” in the box next to Entrez query.  Now click on the "BLAST!" button. Can you find a homologous sequence from yeast (hint:  use the Find tool in your browser with the term Saccharomyces)?  From humans (use the Find tool in your browser with the term Homo)?   Most biochemists consider 25% identical amino acid sequence the cutoff for sequence homology, meaning that if two proteins have less than 25% identical sequence, more evidence is needed to determine if they are homologs.  Can you find any sequences that are above the 25% identity mark in your BLAST search? You can get the listing for an individual protein by searching for the word “Identities” on the web page.

c.        Use the BLAST Glossary on the BLAST help page (//blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\_TYPE=BlastDocs) to discover the meaning of the Score and E Value for each sequence that is reported.  Also, what is the difference between an identity and a conservative substitution?  Provide an example of each from the comparison of your sequence and a search sequence obtained from BLAST.

d.       BLAST uses a substitution matrix to assign values in the alignment process, based on the analysis of amino acid substitutions in a wide variety of protein sequences. Search the BLAST glossary to find the default substitution matrix on the BLAST page. What other matrices are available?  What is the source of the names for these substitution matrices?  Repeat the BLAST search in step 3.b. using a different substitution matrix.  Do you find different answers?

5.       **Plasmids and Cloning**

a.       REBASE is the Restriction Enzyme Database ([http://rebase.neb.com/rebase/rebase.html](http://rebase.neb.com/rebase/rebase.html" \t "_blank)) which is supported by a number of commercial restriction enzyme suppliers.  Search by organism name on their main page (a drop-down menu and a box) to find a restriction enzyme from*Rhodothermus marinus*. What is the abbreviation for this enzyme?  What is the recognition sequence for this enzyme?  What are the expected and actual frequencies of restriction enzyme recognition sites for this enzyme in *Bacillus halodurans* C-125?

b.       What is a plasmid?  pBR322 was one of the first plasmids to be developed for experimental work.  Go to the Entrez site (//www.ncbi.nlm.nih.gov/Entrez) and find the sequence of pBR322 using the terms, "pBR322, complete genome."  Save the sequence of pBR322 to a file in FASTA format (here are two sites that describe the FASTA format, but there are many more: http://ngfnblast.gbf.de/docs/fasta.html; http://bioinformatics.ubc.ca/resources/faq/?faq\_id=1).  You can also get Entrez to display your sequence in FASTA format.  Once you find the pBR322 sequence page, select "FASTA" as the Display option.  Plasmids normally contain genes that encode enzymes that confer bacterial resistance.  Identify one enzyme encoded by pBR322 and name the antibiotic that it targets.

c.        Go to PubMedCentral and search for a 1978 article in Nucleic Acids Research about restriction mapping of pBR322.  Download the article in pdf format (use Adobe Acrobat to read it - you can get this at [http://www.adobe.com](http://www.adobe.com/" \t "_blank)).  What is the total size of the pBR322 plasmid in bases? How many cut sites are there for the restriction enzyme HaeII on pBR322?

d.       When using restriction enzymes, some digests result in "blunt ends" and some in "sticky ends."  Explain the meaning of those terms and provide an example of each.

e.        Go to the RESTRICT site at the Pasteur Institute (http://bioweb.pasteur.fr/seqanal/interfaces/restrict.html).  Input your pBR322 sequence file.  Make sure you enter your email address properly.  Scroll down to the "Required Section" and notice that you have a Minimum recognition site length of 4 nucleotides and you have selected all the enzymes available in REBASE to digest pBR322 at the same time.  Click on the "Run Restrict" button.

f.        After you hit the "Run Restrict" button, you'll see an output screen.  Click on the link "output.out."  This will take you to a simple text page that lists all the cuts that were made in the pBR322 plasmid.  Select all the text on this page, copy it and save it to a simple text file on your computer.  How many fragments did the pBR322 get cut into by "all" enzymes?  (look for the "HitCount" number on the output.out page).

g.        Now let's play with the web site a bit.   What happens to the number of fragments you get if you change the Minimum recognition site length to 6 nucleotides?  Why did the number change?

h.       Return to REBASE (http://rebase.neb.com/rebase/rebase.html) and click on the icon labeled "REBASE enzymes."  The goal here is to find a restriction enzyme that will cut pBR322 into 3 or 4 fragments.

i.         Try BamHI (You must spell it exactly correctly for RESTRICT to execute properly.  How many fragments do you get from BamHI?  What is the size of the restriction site for BamHI?

j.         Now try AvaIII.  Again, how many fragments do you get from AvaIII?  What is the size of the restriction site for AvaIII?

k.       Now I'll give you one more example to try:  Eco47III.  How many fragments do you get from Eco47III?  What is the size of the restriction site for Eco47III?

l.         What happens if you combine the three different enzymes (separate the enzyme names by commas)?  Study the output.out file to find the following information?

                   v.            How many fragments are generated?

                  vi.            Where are the restriction sites on pBR322?

                vii.            What is the size of the fragments that are generated?

m.     Use the NEBcutter V2.0 site ([http://tools.neb.com/NEBcutter2/](http://tools.neb.com/NEBcutter2/" \t "_blank)) to create a restriction map of pBR322, showing the relative positions of the restriction enzymes BamH1, AvaI and Pst1. Save an image of the restriction map.

n.       For the adventurous:  Search REBASE to find an enzyme or combination of enzymes that will produce 10 fragments from pBR322.