0003D_bio_1115-1130 1/26/05 12:17 PM Page 1115 Mac113 mac113:310_CM:

BIOINFORMATICS EXERCISES

DEVELOPED BY PAUL CRAIG
Department of Chemistry
Rochester Institute of Technology

INTRODUCTION

Over the last two decades, bioinformatics has become increasingly important in both teaching and learning biochemistry. The most obvious case is the sequencing of the human genome and those of many other species. 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 that generally refers to the use of computers and computer science approaches to the study of biological systems. This information is discussed in the text, mainly in 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 in order to introduce the techniques of bioinformatics that involve the use of computers, Internet-accessible databases, and the tools that have been developed to "mine" those databases.

As much as possible, the exercises are 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. The stable online resources that you will use most frequently include:

Genbank (http://www.ncbi.nlm.nih.gov/)

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

ExPASy Proteomics Server (http://us.expasy.org/)

European Bioinformatics Institute (http://www.ebi.ac.uk/)

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

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

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

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

The exercises include some questions that have definite answers, but many questions may be answered in a number of ways, depending on the approach you take or the topic you select. In most cases, the answer key includes the definite answers. For some open-ended questions, a typical correct answer may be presented. Please note that because biological databases are continually being expanded by the addition of new entries, the results of some database queries—particularly quantitative results—may appear to shift slightly over time.

These exercises are available on the website (www.wiley.com/college/voet). The online format offers you easy access to resources within the web environment by simply clicking on the relevant links. You can also use your computer's cut-and-paste functions to more conveniently perform tasks involving extensive sequences of text that would otherwise need to be typed in.

CHAPTER 3 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. The exercises below are designed to introduce you to some of the relevant databases and the tools they contain for examining and comparing different bits of information (see Sections 3-4C and 3-4D). Biological databases are an important resource for the study of biochemistry at all levels. These databases contain huge amounts of information about the sequences and structures 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—called web applications—can be used directly from a web browser. Other software—called freestanding applications—must be downloaded and installed on your local computer.

1. Finding Databases. We'll start with finding databases.

(a) What major online databases contain DNA and protein sequences?

(b) Which databases contain entire genomes?

(c) Using your textbook and online resources (http://www.google.com), make sure you understand the meaning of the following terms: BLAST, taxonomy, gene ontology, phylogenetic trees, and multiple sequence alignment. Once you have defined these terms, find resources on the Internet that enable you to study them.

2. TIGR (The Institute for Genomic Research). 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 are these?

(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? (See Section 3-4A for a discussion of restriction endonucleases.)

(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. Using high-throughput methods, scientists are now able to sequence entire genomes in a very

0003D_bio_1115-1130 1/26/05 12:17 PM Page 1116 Mac113 mac113:310_CM:

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:

TACGCAATGCGTATCATTCTGCTGGGCGCTCCGGGCGCAGGTAAA
GGTACTCAGGCTCAATTCATCATGGAGAAATACGGCATTCCGCAA
ATCTCTACTGGTGACATGTTGCGCGCCGCTGTAAAAGCAGGTTCT
GAGTTAGGTCTGAAAGCAAAAGAAATTATGGATGCGGGCAAGTT
GGTGACTGATGAGTTAGTTATCGCATTAGTCAAAGAACGTATCACA
CAGGAAGATTGCCGCGATGGTTTTCTGTTAGACGGGTTCCCGCGT
ACCATTCCTCAGGCAGATGCCATGAAAGAAGCCGGTATCAAAGTT
GATTATGTGCTGGAGTTTGATGTTCCAGACGAGCTGATTGTTGAG
CGCATTGTCGGCCGTCGGGTACATGCTGCTTCAGGCCGTGTTTATC
ACGTTAAATTCAACCCACCTAAAGTTGAAGATAAAGATGATGTTAC
CGGTGAAGAGCTGACTATTCGTAAAGATGATCAGGAAGCGACTGT
CCGTAAGCGTCTTATCGAATATCATCAACAAACTGCACCATTGGTT
TCTTACTATCATAAAGAAGCGGATGCAGGTAATACGCAATATTTTAA
ACTGGACGGAACCCGTAATGTAGCAGAAGTCAGTGCTGAACTGG
CGACTATTCTCGGTTAATTCTGGATGGCCTTATAGCTAAGGCGGTT
TAAGGCCGCCTTAGCTATTTCAAGTAAGAAGGGCGTAGTACCTACA
AAAGGAGATTTGGCATGATGCAAAGCAAACCCGGCGTATTAATGG
TTAATTTGGGGACACCAGATGCTCCAACGTCGAAAGCTATCAAGC
GTTATTTAGCTGAGTTTTTGAGTGACCGCCGGGTAGTTGATACTTC
CCCATTGCTATGGTGGCCATTGCTGCATGGTGTTATTTTACCGCTTC
GGTCACCACGTGTAGCAAAACTTTATCAATCCGTTTGGATGGAAG
AGGGCTCTCCTTTATTGGTTTATAGCCGCCGCCAGCAGAAAGCACT
GGCAGCAAGAATGCCTGATATTCCTGTAGAATTAGGCATGAGCTAT
GGTTCAC

(a) First, try to find an open reading frame in this segment of DNA. What is an open reading frame (ORF)? You can find the answer in your textbook (Section 3-4D) or online with a simple Internet search (http://www.google.com). You may also wish to try the bookshelf at PubMed (http://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 codes for the mRNA. According to the standard genetic code (Table 26-1), 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:

(1) It must start with ATG. In this exercise, the first ATG is the Start codon. In a real gene search, you would not have this information.

(2) It must end with TAA, TAG, or TGA.

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

(4) The ATG Start codon and the Stop codon must be in frame. This means that the total number of bases in the sequence from the Start to the Stop codon must be evenly divisible by 3 (see Section 26-1A).

Hints: Try this search by pasting the DNA sequence into a word processing program, then searching for the Start and Stop codons. Once you have found a pair, highlight the text of the proposed ORF and use the program's Word Count function to count the number of characters between (or including) the Start and Stop codons. This number must be evenly divisible by 3. You can also use a fixed-width font such as Courier, enlarge the size of the text, and adjust the margins so that each line holds just three characters (one codon). Once you find the first ATG, delete the characters that precede it. Then search for a Stop codon that fits all on one line (is in the same reading frame as the Start codon).

(b) Admittedly, Part (a) is a tedious approach. Here is an easier one: Highlight the entire DNA sequence again and copy it. Then go to the Translate tool on the ExPASy server (http://www.expasy.org/tools/dna.html). 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 the Output format and click on "Translate Sequence." The "Results of Translation" page that appears contains six different reading frames. What is a reading frame and why are there six? (Refer to Section 26-1A, the Internet, or the PubMed bookshelf for an answer.) Identify the reading frame that contains a protein (more than 100 continuous amino acids with no interruptions by a Stop codon) and note its name. Now go back to the Translate tool page, leave the DNA sequence in the sequence box, but select "Compact ("M", "-", no spaces)" as the 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 the Stop codon. Save this sequence to a separate text file.

(c) Now you will identify the protein and the bacterial source. Go to the NCBI BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/). What does BLAST stand for? You will do a simple BLAST search using your 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-protein BLAST." Enter your protein sequence in the "Search" box. Use the default values for the rest of the page and click on the "BLAST!" button. You will be taken to the "formatting BLAST" page. Click on the "Format!" button. You may have to wait for the results. Your protein should be the first one listed in the BLAST output. What is the protein and what is the source?

Note to instructors: You can do this exercise with any DNA sequence. You can also start from a DNA sequence directly in BLAST (use blastn) and find the genes that way. It is probably best to choose a DNA segment that encodes only one protein.

4. Sequence Homology. You will use BLAST to look at sequences that are homologous to the protein that you identified in Problem 3.

(a) First, some definitions: What do the terms "homolog," "ortholog," and "paralog" mean? Go to the NCBI BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/) and choose "Protein-protein BLAST." Paste your protein sequence into the "Search" box. Before clicking on the "BLAST!" button, narrow the search by kingdom. As you look down the BLAST page, you'll see an Options section. Under "Limit by entrez query" (followed by an empty box) or "select from:" (followed by a drop-down menu), select "Eukaryota." Now click on the "BLAST!" button. Click on

0003D_bio_1115-1130 1/26/05 12:17 PM Page 1117 Mac113 mac113:310_CM:

the "Format!" button on the next page. Can you find a homologous sequence from yeast?
(*Hint*: Use your browser's Find tool to search for the term "Saccharomyces.") Note the Score and E value given at the right of the entry.

Can you find a homologous sequence from humans?
(*Hint*: Search for the term "Homo.") Note its Score and E value.

Most biochemists consider 25% identity the cutoff for sequence homology, meaning that if two proteins are less than 25% identical in sequence, more evidence is needed to determine whether they are homologs. Click on the Score values for the yeast and human proteins to see each sequence aligned with the *Yersinia pestis* sequence and to see the percent sequence identity. Are the yeast and human sequences homologous to the *Yersinia pestis* sequence?

(b) Use the BLAST online tutorial (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html) to discover the meaning of the Score and E value for each sequence that is reported. What is the difference between an identity and a conservative substitution? Provide an example of each from the comparison of your sequence and a homologous sequence obtained from BLAST (see Section 5-4A for a discussion of conservative substitution).

(c) 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. Be sure you understand the meaning of the term "substitution matrix." What is 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 Problem 4(a) 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), which is supported by a number of commercial restriction enzyme suppliers (restriction enzymes are described in Section 3-4A). Go to the REBASE Enzymes page (http://rebase.neb.com/rebase/rebase.enz.html) and find a restriction enzyme from *Rhodothermus marinus* (it starts with the letters Rma). What is the abbreviation for this enzyme?

Click on the enzyme's abbreviation to be taken to the page for this enzyme. Follow the links there to answer the following two questions. 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 (http://www.ncbi.nlm.nih.gov/Entrez) and find the sequence of pBR322 by searching for the terms "pBR322, complete genome." You must select Nucleotide as your search option on the Entrez main page.

Look through the Entrez description of pBR322 and identify one gene encoded by pBR322 and name the antibiotic that it targets.

You can get Entrez to display your sequence in FASTA format by selecting this option next to the "Display" button. (Here are two of many sites that describe the FASTA format: http://ngfnblast.gbf.de/docs/fasta.html; http://bioinformatics.ubc.ca/resources/faq/?faq_id=1). Save the pBR322 sequence in FASTA format.

(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 program at http://www.adobe.com). What is the size of the pBR322 plasmid in number of base pairs?

How many cut sites are there for the restriction enzyme HaeIII on pBR322?

(d) Some restriction enzymes generate "blunt ends," and some generate "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). Enter your email address at the top, then input the pBR322 sequence file. Scroll down to the "Required section" and note that you have a Minimum recognition site length of four nucleotides and you have selected all the enzymes available in REBASE to digest pBR322 at the same time. Click on the "Run Restrict" button.

On the output screen, click on the "outfile.out" link. This takes you to a simple text page that lists all the cuts that were made in the pBR322 plasmid. How many pBR322 fragments did "all" the enzymes generate? (Look for the "HitCount" number on the output.out page).

What happens to the number of fragments when the minimum recognition site length is changed to six nucleotides? Why did the number change?

(f) Now change the enzyme name from "all" to "BamHI" in the enzymes box under the Required section on the RESTRICT page. How many fragments are generated? How many fragments are obtained using AvaI? What is the size of the restriction site for AvaI? How many fragments are obtained using Eco47III? What is the size of the restriction site for Eco47III?

(g) How many pBR322 fragments are produced when the three different enzymes are combined (separate the enzyme names by commas)? How large are the fragments?

(h) Use a mixture of the restriction enzymes BamHI, AvaI, and PstI to construct a restriction map of pUC18 similar to the one shown in Fig. 3-25. How does this procedure for restriction mapping differ from that used in Problem 10 at the end of Chapter 3?

(i) For the adventurous: Find an enzyme or combination of enzymes that will produce 10 fragments from pUC18. Draw a restriction map of your results.

CHAPTER 5 USING DATABASES TO COMPARE AND IDENTIFY RELATED PROTEIN SEQUENCES

1. Obtaining Sequences from BLAST. Triose phosphate isomerase is an enzyme that occurs in a central metabolic pathway called glycolysis (see Chapter 14). It is also known as an enzyme that demonstrates catalytic perfection (see Section 12-1B). For this problem, you'll start with the sequence of triose phosphate isomerase from rabbit muscle and look for related proteins in the online databases. Here is the sequence of rabbit muscle triose phosphate isomerase in FASTA format: