Introduction to Bioinformatics

Accompaniment to *Discovering Genomics…*

Reading in *Discovering Genomics, Proteomics, and Bioinformatics* by Campbell & Heyer
(henceforth referred to as DGPB by C&H)

The following (and similar lists throughout the semester) is advice, aimed at helping you prioritize your time. Suggestions to skip part of the text merely means that these sections won't play a big part in what will follow. By all means read anything you like! Much of what you read will require links available at the book's web-site (see our course web-site, **Resources and Links**). When the text refers to some mysterious program or object (e.g. “Perform a BLASTn search on the ORF sequence…”) look to an upper corner of the page to see where you might find it. The headers indicate the section of the DGPB web page that contains the desired link.

Chapter 1, Section 1.1 *Defining Genomics*

- What is Genomics? *[Skim]*
- How are Whole Genomes Sequenced?
  - Discovery Questions 1-5 (Sequencing)
  - Math Minute 1.1 (What is an E-value)
- *Skip Why Do the Databases Contain So Many Partial Sequences?*
  - *Skip Discovery Questions 6-12*
- How Do We Make Sense of All These Bases?
  - Discovery Questions 13-17a-c (*skip 17d*)
- Can We Predict Protein Functions?
  - Discovery Questions 18-20 (*skip 21*)
  - 3D Structures

What is Genomics?

You might read this section just for pleasure and a general overview.

How Are Whole Genomes Sequenced?

This section describes dideoxy sequencing or the Sanger method (see the DGPB web site, under links for a biography of Fred Sanger). The method is the basis for virtually all the sequence information we have. If you don't already know something about DNA replication and gel electrophoresis, then you'll probably be mystified by the discussion in the book. In that case, try going to the following animation of dideoxy sequencing (requires Flash):

http://smcg.cifn.unam.mx/enp-unam/03-EstructuraDelGenoma/animaciones/secuencia.swf

<u>Discovery Question 1 and 2</u>

Be sure you understand Fig. 1.2a, particularly the *directionality* of the DNA before you attempt to read the sequence of Fig. 1.2b. When you think you've got the sequence,… what is it? To find out, compare the sequence to all nucleotide sequences known, using a program called Blast (Basic Local Alignment Sequence Tool), the most widely used bioinformatics program on earth. There is a link to Blast at the DGPB web site under ***Links***. In fact, there are *two* links to Blast for some reason. They're identical, so take your pick. *Don't* go to Blast2 – that's something different.

The link will bring you to a menu of many flavors of Blast. You want to perform a search of *nucleotide sequences*. You'll find the basic nucleotide alignment program BlastN in the list. Click on it. Type your sequence into the big white box. There are lots of ways of modifying Blast's performance. For this time out, just accept all the default values, clicking on BLAST.

You'll be taken to a screen that promises you that the results will be ready in some number of seconds (depending on how heavily the server is being used). If you wait for the results to arrive, you'll wait forever. You need to click on FORMAT. Then the results will be sent to you when they've been calculated. On a bad day, it could take minutes.

Scroll down the results until you reach a box labeled ***Distribution of** n **Blast Hits….*** If all is well, you should see a pink line running from the left to the right, indicating that your sequence matched a sequence in the database from beginning to end. The color of the line relates to the quality of the match. More on that later. If you click on the pink line, you'll be taken to the details of the match, including a link to more information about the DNA your sequence matches.

The second part of DQ2 asks you to repeat the Blast but with the *bottom* strand, entered 5' to 3'. How do you get that? You should be able to do this by hand, but if the complementarity rules are new to you, it could be a chore. An alternative is to go to BioLingua and enter:

(OPPOSITE-STRAND-OF "*your sequence*")

How Are Whole Genomes Sequenced? (continued)

The description of automated sequencing may be entertaining, but the main point is to help you realize that machines aren't magical. The sequences produced by automated sequencing are sometimes good, sometimes bad, sometimes in between. The chromatogram (Chromat 1 listed under ***Sequences***) gives a good visual insight into the varying quality of a DNA sequence. Scroll through the sequence noting how easy it would be for you, a human, to judge what the sequence is at various points. Compare the end of the chromatogram with the top of the X-ray film available at the DGPB site (under ***Methods***). Both methods struggle with long fragments where a single base difference is a small fraction of the total.

Discovery Question 3

In answering this question consider that the top-left of the chromatogram comes out of the machine first and the lower-right comes out last.

Discovery Question 4 and 5

Question 5 doesn't say whether to submit the *first* 30 or the *last* 30. Try it both ways! Be sure to note the E-values.

Math Minute 1.1: What is an E-value?

In brief, the expect value (E-value) tells you how many matches of similar quality as yours you would expect if the database you're searching were composed of random sequences. If you got a hit with an E-value of 1, then you might chalk it up to random chance. If you got a hit with an E-value of 10^{-9} (written 1e-09), then you can feel pretty confident that the match is not fortuitous. There must be an connection between the two sequences.

For now, it is important to you see intuitively what an E-value is but I don't think you need to know right now how to calculate it (but if you take *Integrated Bioinformatics* BNFO 601, you'll spend a lot of time on this issue). To understand the nature of the Blast E-value, you need to know a little about how Blast works. Let's approach it by way of an example. Suppose you determine the sequence of a DNA sample fished out of the ocean. It's 800-nucleotides long, and you want to see if it is found in the genome of *Prochlorococcus* SS120. It may be that all of it is in the genome, some of it, or none of it. Maybe you happen to find a match to the genome of 10 nucleotides somewhere in the middle of your sequence:

```
...AAAAAGATTTAGAGTGGGATGATGCTGATGTACAAGG...  (800 nucleotides)
            ||||||||||
...NNNNNNNNNTAGAGTGGGANNNNNNNNNNNNNNNNNN...  (1751080  nucleotides)
```

What's the likelihood that a match this good could have arisen by chance? At a first approximation, you might think that each nucleotide match is a probability of 1/4, one chance out of {A, C, G, T}. If a match occurs 10 times, then the joint probability is:

$$1/4 * 1/4 * 1/4 * \ldots 1/4 \;=\; 1/4^{10} \;=\; \sim 1/1{,}000{,}000^{*}$$

and you'd expect to find the match one time in a million targets of this size. But this isn't fair. You would have been equally impressed if this match occurred *anywhere* in the target genome. So the probability should be:

10^{-6} (for possibly finding it in the first 10 nucleotides of the target)
+ 10^{-6} (for finding it at position 2 through 11 of the target)
+ 10^{-6} (for finding it at position 2 through 11 of the target)
. . .
+ 10^{-6} (for finding it at last 10 nucleotides of the target)
10^{-6} * (length of target sequence – 10)

But that's also unfair, because you'd also be equally impressed if the match occurred anywhere in the *query* sequence (the 800-nt sequence). That's another factor to multiply by. In general, then, you'd expect:

Expected number of matches = $m\, n\, (1/4)^{a}$

And this is the form of the equation to determine the E-value. The exact equation is different because the scoring is more complicated than one point per match and other reasons, but that's enough for now.

How Do We Make Sense of All These Bases?

The text asks you to select PubMed from the search menu. For those who can't find it, PubMed, as one of the most popular choices in the menu, was taken out of alphabetical order and placed at the top of the list. You can also find a link to PubMed on the blue horizontal bar. The Nucleotide database was also moved from its proper place in alphabetical order to near the top of the list.

[*] As a budding bioinformatician, you need to gain facility in working with powers of two. 1/4 = 1/2*1/2, so:

$$(1/4)^{10} \;=\; (1/2)^{10}\,(1/2)^{10} \;=\; \sim(1/1000)\,(1/1000) \;=\; \sim 1/1000000$$

The key point is that $(1/2)^{10}$ is approximately 1/1000, a good relationship to keep in front of you at all times.

When you search, put in the search terms *without* quotes, unless you want the terms to be adjacent to each other.

Once you've found the results of a search of the Nucleotide database, using *homo cyclooxygenase*, AMC advises you to look to the right for "terms that will allow you to search…" various databases. I suspect the interface has changed since the book was written. Now, a link called "Links" appears, which allows you to search the databases using terms in the item. Of the databases supplied, many of them are well worth browsing. You can spend hours in Online Mendelian Inheritance in Man (OMIM), investigating your favorite disease.

I used the Stanford mirror site of GeneCard, scrolled down and found the leptin receptor, clicked on its gene abbreviation and got the advertised treasure trove of information. I think the lesson to take away is that there is an overwhelming amount of information available. The trick is to somehow transform mostly computationally-derived information into human understanding.

ORFs and Translation and Discovery Questions 13 and 14

The ORF sequence mentioned in the text in paragraph 2 can be found at the DGPB web site under **Sequences**. Take a good look at the sequence.

SQ1. Is it an ORF, according to the definition given in the text?

If you copy and paste the sequence into the BlastN search window and go through the usual operations described above, you'll probably get results back in something less than a minute.

"Record the accession number..." What's an accession number? All databases give sequences an identification number as they are entered to enable the database to keep track of them. The first hit is identified by the following symbols (and many more besides):

gi|4506264|ref|NM_000963.1|.

- 4506254 is the GenBank ID (***gi***) number
- NM_000963 is the GenBank accession number. NM indicates that the sequence is from an mRNA.
- NM_000963.1 is the first (and probably only) version of NM_000963

Either the gi number or the accession number is recognized by ORF Finder, but it won't work to give both or to precede the first by "gi|" or the second by "ref|". Give only one field (symbols between two vertical lines).

However, before you go to ORF Finder, pause for a moment to smell the flowers. Click on the link of the first hit (i.e., "gi|4506…"). This will get you to the GenBank entry for the sequence. It says it's the mRNA. If so, it should contain the gene (by which I mean the part that encodes the protein, also called CDS or coding sequence). Shouldn't it?

Scroll down and look at the interesting part of the annotation (beyond the dozens of references).

SQ2. At what nucleotide does the coding sequence begin? What is the first triplet of the coding sequence? Does this make sense? At what nucleotide does the coding sequence end? What is the last triplet of the coding sequence? Does this make sense?

Now return to the BlastN report. What part of the mRNA did your short query sequence hit? Ask again, is the sequence part of an ORF, according to the definition given in the text? Go back to the GenBank entry and find the query sequence within the full sequence of the mRNA.

Now, on to ORF Finder. Enter the accession number as described. What do you get? Why six gray boxes with blue blobs in them? Why not one gray box? Why not three? What reading frame is the biggest? Click on it.

SQ3. Compare its amino acid sequence to that in the GenBank entry. Same? Different? Compare the from… to values given by ORF Finder to the gene boundaries given in the GenBank entry. Same? Different? Why? And while we're in the area, what do you think "mat_peptide" refers to?

SQ4. You've focused on the biggest of the ORFs found by ORF Finder. But it found 18 others. What's their significance?

Discovery Questions 15 – 17a-c

The purpose of these questions is to explore the similarity of two genes that encode proteins with similar biochemical function and the similarity of the proteins themselves.

When I did Discovery Question 15, I didn't get any patent-related sequences, but I did get several choices (including the gene I found with BlastN in DQ13). Ignore the several genes that don't encode cyclooxygenases or are not from *homo sapiens.* Of the remaining candidates, ignore the partial sequence. Take your pick of the two that remain (AMC used the one with the more similar accession number).

The link to Blast2 is broken. You can get to Blast2 by clicking on the Blast link, and clicking on "Align two sequences" in the **Special** section.

For Discovery Question 17, don't go back to the Blast2 Sequences window. Instead, open up a new Blast2 Sequences window and leave the old window open. This allows you to compare the Blast result for the nucleotide sequences with that of the protein sequencdes. The given protein accession numbers appeared by magic. You could get them yourself by clicking on the two GenBank entries you found above and scrolling down to the amino acid sequence. You'll see several ID numbers, including the GenBank protein ID's, beginning NP_… (P for protein).

The several X's in a row you find after the search suggested in DQ17c do not indicate mismatches but rather indicate that BlastP filtered out a region of the query protein before doing the comparison. It does this if it finds a region where there is significant repetition. To allow a full comparison to be made, go back to the Blast2 page, uncheck the Filter box, and rerun the Blast.

Why is it that there's so much more similarity in the protein sequences than in the nucleotide sequences? If the genes encodes the proteins and the proteins are similar, must not the genes be similar too? Actually, they're more similar than you think. Go back to the first Blast2 Sequences window (comparing the two nucleotide sequences), change the mismatch penalty from –2 to –1, and rerun the comparison. Now much more of the sequences align. What portion aligns? What portion does not align? Still the question remains:

SQ5. Why do the protein sequences align better than the nucleotide sequences?

Can We Predict Protein Functions?

Yes. Are the predictions correct? Sometimes. At present, the best predictors are similarities of sequences with sequences of proteins whose functions have been determined by biochemical means; supplemented by crude *ab initio* methods like Kyte-Doolittle hydropathy plots. This section takes you through these methods.

Run both protein sequences through the Kyte-Doolittle page. You should feel comfortable with the conclusion that one of them has a region that passes through a membrane. Be aware that often times results are not so clear-cut.

Discovery Questions 18 – 20

The Show Domain button is no longer top left. Now it's in the middle of the page.

You may have trouble finding the requested "gnl/CDD/7333". Look at the far right end of the match and mouse over a thin brown box.

3D Structures

C&H ask you to go to the Protein Data Bank or Entrez Structure. Take your pick. The interfaces are different, but either search will get you the same result. To see the 3D results (which are worth it), you'll need to download some software. I advise Protein Explorer, which does much more than let you see a pretty picture. The program requires the Chime plug-in. The downside is that Protein Explorer will not work on Netscape above 4.7, I think, and not with Internet Explorer on Macs. There is some fuss if you're using FireFox in Windows. Netscape 4.7 is OK everywhere. I confess, I haven't succeeded in installing it anywhere except on Netscape 4.7.

It's not easy to get this far, but if you have, you can rotate the protein within Protein Explorer to examine its structure. Get into the Quick View screen of PE, then select different classes of amino acids. Select hydrophobic amino acids and color them gray. Select polar (hydrophilic) amino acids and color them yellow. Select acidic (negatively charged) amino acids and color them red. Select basic (positively charged) amino acids and color them green. Then you'll see the protein as water sees it.

The suggested trip to PREDATOR only if you know something about secondary structure of proteins or if you look at the results while looking at the 3D structure of cyclooxygenase, available from the same page as the cyclooxygenase sequence. To see the structure, you'll need the Chime plug-in. In brief, **primary** structure is the sequence of the amino acids that compose a protein. **Secondary** structure is the local structure of a portion of a protein, determined by interactions between atoms of the protein backbone. The most important secondary structures are **alpha helices** and **beta pleated sheets**. **Tertiary** structure is the overall three-dimensional structure of a single polypeptide chain. Some proteins have multiple polypeptide chains that interact with each other. The structure resulting from that interaction is called the **Quaternary** structure. You can see pictures of an alpha helix and a beta pleated sheet at:

http://www.elmhurst.edu/~chm/vchembook/566secprotein.html

PREDATOR examines the amino acid sequence (the primary structure of a protein) and predicts those regions that are most likely to assume different secondary structures. To compare the regions predicted by PREDATOR with the 3D picture, rotate the picture until you've identified a chain-end that is near a red helix. I think this is the beginning of the protein. Now compare the relative positions of the red coils in the picture with the regions of *hhhh* given by PREDATOR.

SQ6. Is there a correspondence between the purported 3D structure and the secondary structure predicted by PREDATOR?