Bi 1x, Spring 2009

Week 6 Prelab

Problem 1: thinking about genomes

Genome science is fast becoming one of the most important facets of modern biology. The idea is to examine the content of genomes as a window on biological function, the evolutionary history of organisms and a host of other issues. In this problem, use a few simple rules of thumb e.g. that a "typical" protein is 300 amino acids long.

A. Given that the *E. coli* genome has roughly 5 million basepairs, make a simple estimate of the number of genes in the genome of *E. coli*.

B. Compare and comment on your result from part 1A with the roughly 4400 genes actually observed in the *E. coli* genome.

"GeneSweep was an informal gene-count betting pool that began at the 2000 Cold Spring Harbor Laboratory Genome Meeting."
http://www.ornl.gov/sci/techresources/Human_Genome/faq/genenumber.shtml
This betting pool was aimed at guessing how many genes would be found in the human genome.

C. Try out a similar estimate on the human genome with its 3 billion basepair genome to that you tried for a bacterium in part A. Explain the rationale behind your estimate and then comment on what bet you would make and why?

D. No matter what the ultimate gene count turns out to be, this estimate is way off! **Give at least two possible reasons why this naïve reasoning failed in the human case.**

Problem 2: the basics

The most famous polymer in the world, DNA, can be thought of as just a simple sequence of 4 letters: A, T, C and G, and yet put a few billion or so of these letters together and we can have the code for making a tree, a frog or Jack Black. Modern initiatives like the Human Genome project have generated reams of genetic sequence data that have become one of biologists' most treasured resources. Some examples of research questions arising from this sequence data are: Which parts of the DNA sequence code for genes? How has evolution changed sequences over time and between species? How do we determine the evolutionary relationship between living things based on their DNA? Bioinformatics is the field of science that attempts to tackle these questions among many others concerned with how information is stored, used and passed on in living things.

Both DNA and proteins are linear polymers that can be thought of usefully in terms of their sequences: DNA as a sequence of nucleotides and proteins as a sequence of amino acids. Most of you are familiar with the alphabet of DNA; the (perhaps) less familiar alphabet of amino acids is shown in Table 1 below.

One-letter code	Three-letter code	Amino acid	One-letter code	Three-letter code	Amino acid
A	Ala	alanine	M	Met	methionine
C	Cys	cysteine	N	Asn	asparagine
D	Asp	aspartic acid	P	Pro	proline
E	Glu	glutamic acid	Q	Gln	glutamine
F	Phe	phenylalanine	R	Arg	arginine
G	Gly	glycine	S	Ser	serine
H	His	histidine	T	Thr	threonine
I	Ile	isoleucine	V	Val	valine
K	Lys	lysine	W	Trp	tryptophan
L	Leu	leucine	Y	Tyr	tyrosine

Table 1 – One and three letter codes for the 20 commonly occurring amino acids.

One useful way to compare two sequences (DNA or protein) is to align them so that the conserved parts of the sequence (the parts that have stayed the same) line up with one another. Methods to do this quickly and accurately are an active area of research in bioinformatics, but most of them rely on one of two main schemes for comparing sequences: global alignment[1] and local alignment[2]. In a global alignment we seek to align every letter in the sequences, whereas in a local alignment we only seek the best matching subsequences. This can be illustrated by a

[1] Needleman and Wunsch (1970), *J Mol Biol*. **48**:443-453.

[2] Smith and Waterman (1981), *J Mol Biol*. **147**:195-197.

simple artificial example (taken from the excellent book “BLAST” by Korf, Yandell and Bedell and published by O'Reilly). Suppose we want to align the sequences PELICAN and COELICANTH. In a global alignment we must use each letter from both sequences, but we allow ourselves to use gaps. One such global alignment is:

```
P-ELICAN--
COELICANTH
```

where a “-” character indicates a gap. With a local alignment we are only concerned with subsequences that match well; in this case the best matching subsequences are:

```
ELICAN
ELICAN
```

Local alignments lend themselves better to various computational tricks so that they can be executed quickly even when searching databases with millions of sequences. A typical situation is that you have a sequence for your favorite protein, XYZ, and you would like a tool that will compare your sequence with an entire database of proteins and quickly find similar sequences. That's where BLAST comes in.

Introducing BLAST

BLAST (Basic Local Alignment Search Tool) is built to do just exactly that: search a database quickly for proteins sequences that are closely related to a query sequence with local alignments. BLAST has become such a popular and useful tool that people talk about "blasting" sequences in just the same way that they talk about "googling" web searches. BLAST is a bit like the Google of biology.

What makes BLAST so useful is that it is very fast. It uses some computational shortcuts to speed up searches so that even if the database being queried contains millions of sequences you can get results back in a matter of a few seconds. However there are at least a couple of important concepts to keep in mind when using this marvelous tool. The first is that BLAST must have a lookup table or matrix which tells it the score (positive or negative) for lining up letters in the two sequences being compared; this table is symmetric and will be at least 20 by 20 in the case of proteins since there are 20 naturally occurring amino acids. BLAST calculates the score for the entire alignment by adding up the score for each pair of aligned letters. These matrices are called substitution or scoring matrices and they tell us the likelihood that one amino acid will be substituted for another as evolution proceeds. Different methods have been applied to generate scoring matrices and two of the most popular are the PAM and BLOSUM matrices. For most purposes the default matrix used by BLAST, BLOSUM62, does a fine job of finding related proteins. But you should be aware that the numbers in these matrices were arrived at by heuristic methods and if you are going to use programs like BLAST regularly it would behoove you to learn how these matrices are generated so that you know their limitations.

The second key concept to understand is the statistics of sequence alignment. There is always a possibility that two proteins completely unrelated by evolution will happen to have subsequences that are similar just by chance. Understanding the likelihood that a particular alignment reflects random chance rather than a genuine evolutionary relationship is a major key to correctly interpreting BLAST's results. Of course it always pays to know the biology of the proteins that are returned in BLAST's results as well. BLAST quantifies the probability of getting a result by chance with a number called an "E value" or "Expect value". This number, roughly speaking, is the number of alignments with a given score or better that would be expected by pure random chance given the size of the database searched. If you have a hit with an expect value of 1e-10, there is a good bet that the query sequence and the hit sequence are related to one another. If the expect value is 10 or 100, then that hit can't be distinguished from matches due to random chance and BLAST is not telling you anything useful.

The BLAST homepage is found at:

http://www.ncbi.nlm.nih.gov/blast

There are many resources listed on the BLAST page; besides the web interface for submitting queries there are links to download genomic data and one can download the entire BLAST package (source and binaries). BLAST is really a whole suite of programs that can compare protein and nucleotide sequences in various ways.

Finally you should know that while BLAST is popular and efficient it is by no means the only game in town... many other packages are available. Each of them has their individual strengths and weaknesses.

Sequence data resources

There are many places where protein sequence data can be retrieved; here are a few useful sites:

- UniProt (http://www.pir.uniprot.org/)
- Entrez Protein (http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein)
- Entrez Genome (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome)

Using BLAST

HIV virions wrap themselves with a lipid bilayer membrane as they bud off from infected cells; in this viral membrane envelope are "spikes" composed of two different proteins (actually glycoproteins), gp41 and gp120. The gp denotes glycoprotein, and the number indicates their molecular weight in kilodaltons. gp120 and gp41 together form the trimeric envelope spike on the surface of HIV that functions in virus entry into a host cell. The primary receptor for gp120 is CD4, a protein found mainly on the white blood cells known as T lymphocytes. gp120 avoids detection by the host immune system through a number of strategies, including rapid changes in sequence due to mutations. See Figure 1 below to see a schematic drawing of HIV with the gp120 and gp41 proteins labeled.

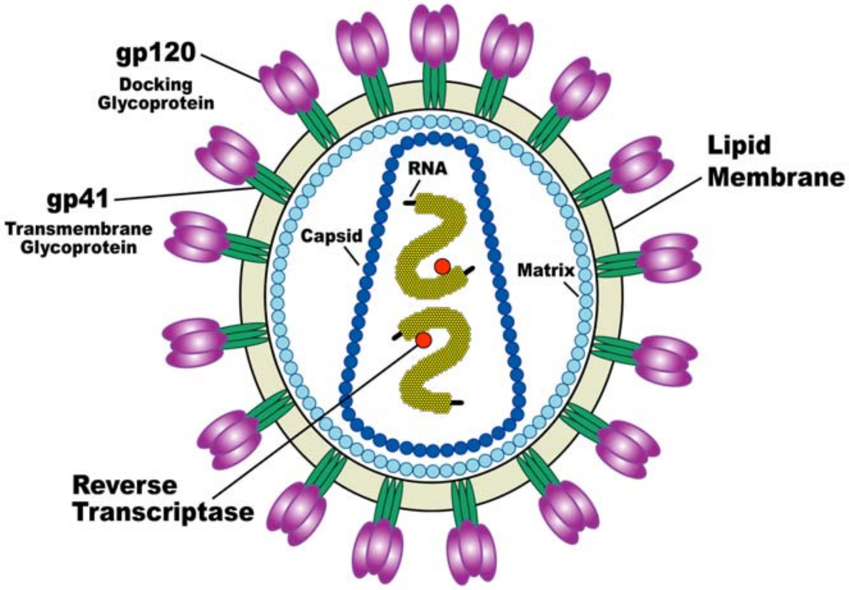

Figure 1 – Diagram of an HIV virus. Source: NIH

In this exercise we will grab a sequence for gp120 and blast it to find related proteins. Use the LANL HIV Sequence Database site to find a sequence for gp120:

http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html

From this webpage, click on "Search interface" under the “Programs and Tools” heading and then type "gp120" into the "Sequence name" field and execute the search to find some gp120 sequences. The complete gp120 molecule has about 500 amino acids so a complete DNA sequence will have about 1500 bases (plus or minus about 100 bases). Make sure that you use an HIV sequence and not SIV (Simian Immunodeficiency Virus)—SIV is a virus related to HIV that infects monkeys. Click on the Accession code for the sequence you select from your search and scroll to the bottom to find your gp120 nucleotide sequence.

Open a new window with the BLAST website and select "blastx" under the "Basic BLAST" heading. Copy and paste your ~1500 nucleotide sequence of gp120 into the top box labeled "Enter Query Sequence"; BLAST will only pay attention to letters so it's OK if there are extra numbers in your pasted sequence. For the database select "Protein Data Bank proteins(pdb)". The PDB is a worldwide repository for the structures of proteins that have been determined in atomic detail (i.e. the positions of every atom have been determined by crystallography or NMR). What we are doing is having BLAST translate the gp120 nucleotide sequence into an amino acid sequence and then compare it with amino acid sequences of proteins in the PDB. Note that there are some proteins which appear many times in the PDB (because their structures have been analyzed and determined more than once or in different contexts). Finally, push the "BLAST" button; BLAST will give you updates on your query until it is finished and the final results are displayed.

Now that you have your list of BLAST results, take a look at the structure of the results. There is a colored graph showing the quality of alignments that BLAST was able to generate, going from black (low score, low significance) to red (high score, high significance). The red bar next to

“Query” represents the sequence you searched with, and the thinner bars below it represent sequences BLAST found in the PDB that aligned with the query sequence. The length and position of the bar represents where the PDB sequence aligned with the Query you found in the LANL database.

Below this is a ranked list with the same sequences that were represented as bars on the graph. Finally below the ranked list you will find the actual alignments that BLAST generated, with the “Query” sequence showing the translation of the LANL nucleotide sequence you pasted into the search form and “Sbjct” being a sequence of a protein from the PDB that aligned with it. There is a line between “Query” and “Sbjct” that helps guide the eye with the alignment; if “Query” and “Sbjct” agree identically, the matching letter is repeated in the middle; if they do not match exactly but the amino acids are compatible in some sense (a favorable mismatch), then a “+” is displayed on the middle line to indicate a positive score. Where there is no letter on the middle line there is either an unfavorable mismatch or a gap in the alignment. The numbers at the beginning and end of the “Query” and “Sbjct” lines tell you the position in the sequence.

One last caveat: be aware that your hits from the PDB represent only that small subset of gp120 sequences that were convenient to study with crystallography.

A. How does BLAST determine the ranking of the results from your search?

B. For your top search result, identify the percentage of sequence identity with your query sequence. Explain how this number is determined.

Cells have evolved sophisticated machinery to copy DNA with great fidelity but occasionally mutations occur. These mutations can range in size from a single change to a base pair in the DNA to crossing over large regions of an entire chromosome. There are also processes that can result in insertions or deletions in the DNA.

C. What evidence in your alignment do you see for insertions and deletions?

D. For your top hit, how many alignments with this high of a score or better would have been expected by chance?

Print out the first page of your BLAST results and turn it in along with your answers to the above questions when you submit the homework. No credit will be given if you do not turn in your BLAST results, and it may not be a photocopy.

Problem 3: Multiple alignments

Even more information can be gleaned by the natural next step in sequence alignment: aligning more than 2 sequences in a "multiple alignment". Tools like ClustalW align many sequences simultaneously and help us discern relationships between related proteins and therefore between their parent organisms. This is illustrated by an example with hemoglobin.

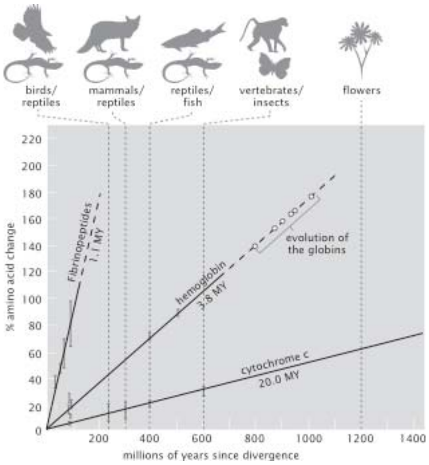

Figure 2 – Examples of three molecular clocks. Fibrinopeptides, hemoglobin and cytochrome c have all been used to characterize divergence of species over evolutionary time. (Adapted from R. E. Dickerson, Sci. Am., 226:58, 1972.)

The function of hemoglobin as an oxygen carrying protein is similar for almost all vertebrates. However, not all of the amino acid residues in the hemoglobin sequence are critical for its three-dimensional structure or its function. These residues are subject to alteration through mutation. Interestingly, even before the advent of modern sequence analysis, molecular comparisons between hemoglobins were used as the basis for deductions about molecular (and thereby, organismal) evolution. In particular, by comparing the mobility of parts of hemoglobins from different species on two-dimensional gels, it was possible to relate species ranging from humans to sharks. From this and many similar studies, molecular phylogenists have inferred the historical relatedness of modern species. In most cases, these kinds of molecular analyses are consonant with what was already known about relatedness on the basis of anatomical comparisons. One fascinating outcome of these kinds of sequence detective work in conjunction with independent dating using the fossil record is the ability to use sequence similarity as a molecular clock. Figure 2 shows some classic studies on the molecular clock idea using hemoglobin and several other molecules. The outcome of this and other studies is the recognition that molecular similarity can be used as the basis for deducing how long ago species diverged from their common ancestor.

In this problem you will examine the changes in hemoglobin over evolutionary time from the point of view of modern sequence analysis with a comparison of the hemoglobins from eight different organisms (humans, chimps, gorillas, cattle, horses, donkeys, rabbits and carp). The concept of the problem is to take the known sequences of hemoglobin from all of these organisms, do a multiple alignment using ClustalW and then explore the differences and similarities between them. ClustalW can be found at:

http://www.ebi.ac.uk/clustalw

You can obtain the necessary hemoglobin sequences using the links below.

Human	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=57013850
Chimp	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=110835747
Gorilla	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=122407

Cow	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=13634094
Horse	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=122411
Donkey	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=62901528
Rabbit	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=229379
Carp	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&id=122392

Cut and paste the sequences above into a text file; before each sequence, make a line with a greater-than symbol and the name of the organism, like so:

```
> Human
mvlspadktn…
```

What you are doing is constructing a FASTA file containing all the sequences you downloaded. In FASTA format, the “>” symbol is recognized as a comment that gives information about the sequence and is used to make your results more readable. Go to the ClustalW site, fill in your email but leave the other options as they are and click the “Choose File” button. Navigate to your FASTA file with the sequences, select it and then run the alignment. After a few moments the results will appear; the parts of highest interest to us in this problem are the Alignment, shown after the Scores Table, and the Cladogram, a graph which shows the degree of similarity among each of the sequences. Very similar sequences are connected by short branches; increasing branch length indicates decreasing sequence similarity.

A. Hand in your alignment for the eight hemoglobin sequences. Examine the variability in the amino acids between the sequences and comment on the relatedness of the species in question. Do the results jibe with your intuition?

B. Consider the ‘Scores’ table. **From here, which two organisms are the most closely related? Most distantly related? Rationalize these relationships (in terms of phylogeny).**

C. Choose a position in the alignment that has at least three different amino acids in it among the eight different species. **Make a list of all of the different amino acids used at this position and comment on their similarities or differences (i.e. hydrophobic or polar, size, etc.).**

D. Now look at the Cladogram in your results that ClustalW generated based on the alignment. **Does the tree correspond to your intuition about relatedness of species?**