

Problem Set 8: Bioinformatics and Evolution

Due Tuesday, June 2 at 4:00 PM in the Bi 1 closet

HOMEWORK INSTRUCTIONS

- 1) Turn in your homework stapled to this cover page.
- 2) Use separate sheets of paper for your answers.
- 3) Write or type your answers neatly.
- 4) Put your name on each page of your answers.
- 5) Box your answers, please, so that the grader can find them.

Points may be deducted if you don't follow these instructions!

Name: _____

Section #: _____

Mail Code: _____

TA Names: _____

Date and Time turned in: _____

Number of pages including this one: _____

AFTER YOU FINISH:

How many hours did you spend on this set?: _____

Please also go to the Bi1 moodle site at <http://courses.caltech.edu/> and take the homework survey

There are **2** questions. The number of parts to each question is listed at the beginning of each; be sure to answer all the parts!

Grade:

Problem 1 _____

Problem 2 _____

TOTAL: _____

Problem 1: Bioinformatics and Sequence Comparisons (50 points – 14 Parts)

Bioinformatics is the application of information technology to biological problems. An example is the comparison of different sequences (DNA or protein) to determine evolutionary relationships between various organisms.

One useful way to compare two sequences (DNA or protein) is to align them so that the conserved parts of the sequences (the parts that have stayed the same) line up with one another. Most methods that do this quickly and accurately rely on one of two main schemes for comparing sequences: global alignment¹ and local alignment². 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 (within the given global sequence). For example, in a global alignment of the sequences PELICAN and COELICANTH, 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. A typical situation is that you have a sequence of interest, and you would like a tool that will compare your sequence with an entire database of proteins and find similar sequences.

Introducing BLAST

BLAST (Basic Local Alignment Search Tool) searches a database quickly (often within seconds) for protein sequences that are closely related to a query sequence with local alignments. For each alignment, BLAST returns statistics that measure how “good” the alignment is:

- **Score.** BLAST calculates the score for a particular local alignment by adding up the score for each pair of aligned letters based on a chosen symmetric lookup table (or matrix). These so-called substitution or scoring matrices were originally constructed based on the likelihood that one amino acid will be substituted for another as evolution proceeds.
- **E-value (“Expect value”).** There is always a possibility that two proteins completely unrelated by evolution will happen to have subsequences that are similar just by chance. BLAST quantifies the probability of getting a result by chance with a number called an “E value” or “Expect value”. Roughly speaking, this is the number of alignments with a given score or better that would be expected by pure random chance given the size of the

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

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

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.

The BLAST homepage, which contains many resources and tutorials that you are welcome to explore, is found at:

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

Using BLAST

In this exercise we will use BLAST to follow how one of the most fundamental discoveries in cancer biology was made. This discovery was made in the late 1970s, when whole genome sequences and tools like BLAST were not available, which meant the experiments were much more time-consuming.

Peyton Rous won the Nobel Prize in 1966 for the discovery of the eponymous Rous sarcoma virus (RSV), a retrovirus infecting chickens and other birds that causes cancer in their connective tissue (bone, muscle, fat) after infection. It was found that a single gene called *src* (short for sarcoma) in the RSV genome was necessary for inducing the unchecked division and proliferation of cells (a phenomenon known as neoplasia) that is the hallmark of cancer. In other words, when *src* was deleted from the genomes of RSV virions, they were still able to infect chickens, but they did not cause cancer. Later biochemical studies on Src, the protein encoded by the RSV *src* gene, showed that Src is a tyrosine kinase, a type of enzyme that can transfer a phosphate group from ATP onto the tyrosine residues of other proteins in a process known as phosphorylation. Tyrosine phosphorylation is a signal transduction mechanism that is thought to have evolved in multicellular eukaryotes, as this phenomenon has not been found in prokaryotes. In eukaryotic cells, phosphorylation of a protein acts as a switch that can turn on or off the activity of that protein. Eukaryotes use phosphorylation and numerous phosphorylating enzymes (called kinases) to regulate their own processes even in the absence of viral infection. Many cellular decisions such as whether to proliferate (divide) can be enacted simply by modifying the set of proteins that is phosphorylated.

A. (4 points) Speculate on how the neoplastic changes that result from avian RSV infection might benefit the virus. (< 5 sentences)

Two biologists, Harold Varmus and J. Michael Bishop, wondered about the evolutionary origin of RSV *src*. One possibility is that RSV *src* evolved independently of eukaryotic tyrosine kinases to hijack the avian cells' signal transduction pathways.

B. (4 points) Assuming that RSV evolved *src* independent of eukaryotic tyrosine kinase genes, would you expect to *src*-like sequences in viruses that are evolutionarily closely-related to RSV? Explain. (< 5 sentences)

Another possibility is that RSV *src* and eukaryotic tyrosine kinases evolved together and that there was some exchange between the viral genomes and the eukaryotic genomes. This possibility predicts that some eukaryotic genomes should have a single protein that has high sequence similarity and a similar function to viral Src. This eukaryotic protein would be referred to as the "ortholog" of viral Src. Varmus and Bishop tested this possibility using molecular hybridization techniques. We will recapitulate their results using BLAST.

Open a new window with the BLAST website and select "protein blast" under the "Basic BLAST" heading. Copy and paste the amino acid sequence of RSV Src protein (from the file: "RSV p60 src.txt") into the top box labeled "Enter Query Sequence"; it's OK if there are extra numbers in your pasted sequence. For the database select "Reference proteins (refseq_protein)." For the organism, type without quotes: "human (taxid:9606)". Click on "Algorithm parameters" and change "Max target sequences" to 10. Leave the rest of the options unchanged, and hit the BLAST button at the bottom of the page. Wait until the results are displayed. The "blastp" program will compare your input RSV Src sequence to sequences of human proteins; if there is a human ortholog of RSV Src, this program will find it.

Now that you have your list of BLAST results, take a look at the structure of the results:

- A colored graph shows 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 show the lengths and positions of the found protein sequences that aligned with the query sequence.
- A ranked list gives the actual identities and scores of the aligned sequences.
- Below the ranked list you will find the actual alignments that BLAST generated, with the "Query" sequence showing the pasted viral Src amino acid sequence and "Sbjct" being the sequence of an aligned human protein. There is a line between "Query" and "Sbjct:" if "Query" and "Sbjct" agree identically, the matching letter is repeated here; if they do not match exactly but the amino acids are compatible in some sense (a favorable mismatch), then a "+" is displayed to indicate a positive score. Where there is no letter 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.

- C. (3 points) How does BLAST determine the ranking of the results from your search? (< 3 sentences)**
- D. (2 points) For your top search result, identify the percentage of sequence identity with your query sequence.**
- E. (2 points) For your top hit, how many alignments with this high of a score or better would have been expected by chance?**
- F. (3 points) Based on the scores and E-values of the returned results, is there a human ortholog of RSV *src*? If so, identify which one BLAST "hit" represents the ortholog. Explain how you obtained your answer. (< 5 sentences)**
- G. (4 points) What region of RSV Src protein is highly conserved among human tyrosine kinases? Give approximate residue numbers. Explain how you obtained your answer. (< 5 sentences)**
- H. (5 points) 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.**

The discovery of viral *src*-like genes in human cells was exciting because it implies that there are genes within our bodies that, when mutated, could potentially cause cancer even in the absence of viral infection. These genes are called “proto-oncogenes” because in their unaltered state they serve essential functions in maintaining normal cell signaling and proliferation. The gene versions possessed by cancer-causing viruses represent mutational paths that can convert these “normal” genes into cancer-causing genes. Since this discovery, numerous other proto-oncogenes have been identified, whose normal functions range from serving as kinases to directly regulating the initiation of transcription.

How did proto-oncogenes like *src* arise in eukaryotic cells? One hypothesis is that the RSV genome was incorporated into infected cells and that the eukaryotic cells managed to mutate viral *src* into a gene that had an undisruptive, essential cellular function. Another hypothesis is that RSV “picked up” eukaryotic *src* and mutated into an oncogenic (cancer-causing) form. To examine these hypotheses we will use a tool called ClustalW.

Introducing ClustalW

ClustalW is bioinformatics tool that performs global alignments on a defined set of input nucleic acid or amino acid sequences. ClustalW will perform all pairwise alignments (that is, take all combinations of 2 sequences and align them to each other) and cluster the sequences by their similarity to one another. This result is represented as a phylogenetic tree, which shows the relationship between entities that are believed to have a common evolutionary ancestor. For more information on phylogenetic trees, please refer to Chapter 27 and page B-3 in Freeman (3rd ed). Below is a sample phylogenetic tree showing how panda bears (*Ailuropoda melanoleuca*, circled in red) are believed to have evolved:

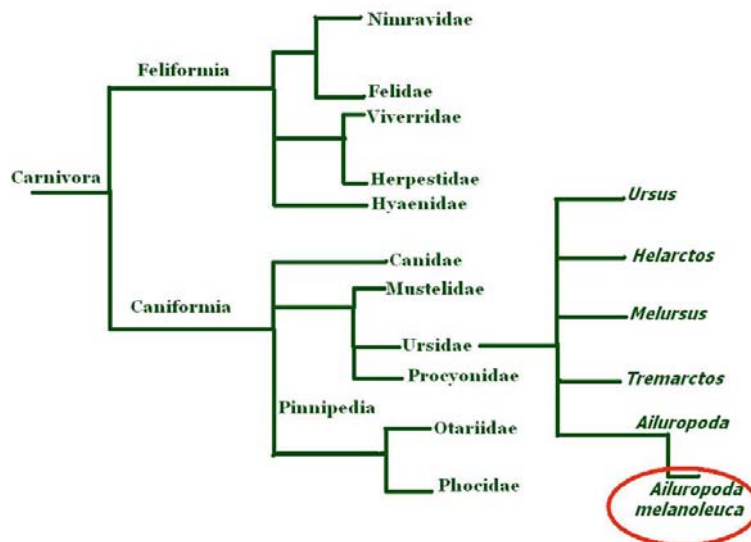


Figure 1: Phylogenetic tree showing the evolutionary history of panda bears (taken from http://bioweb.uwlax.edu/bio203/s2007/barger_rach/)

The lengths of the lines on a phylogenetic tree indicate evolutionary distance (a.k.a. time). Panda bears and brown bears (*Ursus*) have a common ancestor *Ursidae*. Bears and wolves (*Canidae*) have a common canine ancestor (*Caniformia*). This canine ancestor shares a common carnivorous ancestor (*Carnivora*) with the feline branch of the tree (*Feliformia*). Considered in a different way: carnivores diverged into feline (cat) and canine (dog) forms; the canines diverged into wolves, bears, and others; the bears diverged to become brown bears, spotted bears, and panda bears. The *Ailuropoda* genus, which contains many types of panda bears, is most similar to spectacled bear genus (*Tremarctos*).

How does all of this relate to *src*? We will ask ClustalW to create a phylogenetic tree based on the Src protein sequences of various organisms. If the sequences for viral Src occur at the root of the tree, this suggests that all eukaryotic Src sequences are derived from viral Src and that RSV evolved an oncogenic Src first before eukaryotes mutated it to a non-disruptive form. However, if RSV Src occurs in the middle of the tree, this suggests that the virus picked up Src from eukaryotic genomes and then mutated it into an oncogenic form.

Using ClustalW

Go to the ClustalW webpage:

<http://www.ebi.ac.uk/Tools/clustalw2/index.html>

Input your e-mail address and title your alignment “Src.” At the bottom of the page, upload the file “src proteins.txt”. Then hit “Run” and wait for your results to be displayed. When the results are displayed, go to the bottom of the page and hit the “Show as phylogram tree” button. The results of interest to us are:

- Scores table. The scores table shows the pairwise alignment similarity score (out of 100) of the sequences in the input sequence set. Higher scores mean that the two compared sequences are more similar.
- Alignment. The alignment results are shown with consensus residues marked by an asterisk.
- Phylogram. ClustalW clusters the input sequences together based on their similarity and organizes the clusters in a phylogenetic tree.

- I. (3 points) Consider the Scores table. From here, which Src sequences are most closely related? Most distantly related?**
- J. (4 points) Now look at the phylogenetic tree. Does the tree correspond to your intuition about the relatedness of species? Explain. (< 5 sentences)**
- K. (4 points) Based only on the alignment, which mutated region of viral Src might you guess to be responsible for conferring its oncogenic properties? Give approximate residue numbers. Justify how you obtained your answer. (< 3 sentences)**
- L. (3 points) Which eukaryotic Src protein does viral Src most resemble? Does this jibe with what you know about the infectivity of RSV? Explain. (< 3 sentences)**

M. (4 points) Based on the phylogenetic tree, did RSV evolve Src and then pass it to eukaryotes, or vice-versa (*i.e.*, which hypothesis is correct)? Explain. (< 3 sentences)

N. (5 points) Print out *just your phylogenetic tree* and hand it in with your assignment.

Problem 2: Evolution and Sequence Space (50 points – 8 parts)

A protein undergoes amino acid substitutions during evolution. It is a useful conceptual tool to think of a “sequence space” consisting of all possible protein sequences. This is a biological design space. For the following, assume that 20 amino acids are possible at each position and that the proteins are all 200 amino acids in length.

A. (3 points) How many total sequences are there in protein sequence space?

B. (5 points) For a given protein sequence, calculate the number of possible single, double, and triple mutants. Then derive a general formula for calculating $N(m)$, the number of mutants with m mutations.

There are many points in sequence space. However, sequence space is highly connected in that you can get anywhere in the space by a path that never exceeds 200 steps (in the case of proteins that are 200 amino acids long). A single step is defined as a change (mutation) in a single amino acid in the protein.

C. (5 points) For a given protein sequence, how many possible shortest-length single-mutation paths are there to go to another protein sequence that differs at 2, 10, and m positions?

Evolution corresponds to motion across sequence space. Each point in sequence space can be associated with a “fitness,” which for molecular evolutionists means the ability of the organism to survive and reproduce. The “landscape” is the plot of fitnesses across sequence space. Evolution can thus be conceptualized as a random walk on a fitness landscape in sequence space. If evolution by natural selection is to occur, mutations in protein sequences must not result in intermediates that deleteriously affect the reproductive capabilities (fitness) of the organism.

Consider a constantly-evolving protein X that has subsequence TIME (single letter code for threonine, isoleucine, methionine, glutamic acid) in residues 100-103. Suppose that there is a version of X with higher fitness that has the subsequence LAND (single letter code for leucine, alanine, asparagine, aspartic acid) in residues 100-103 (but does not differ in any other residues). In this example, only versions of protein X that have an English word in residues 100-103 do not deleteriously affect the organism (we will call this the “fitness criterion”).

D. (4 points) Taking into account the fitness criterion, write out two possible shortest-length single-mutation paths to go from ...TIME... to ...LAND...

E. (3 points) What fraction does your answer in part D represent of the theoretical (i.e., removing the fitness criterion) number of possible shortest-length single-mutation paths?

F. BONUS: (5 extra points) Can you think of a version of X that has high fitness with m mutations from ...TIME... but is not accessible by a shortest-length single-mutation path? Write it down. Prove it.

To observe evolution in action, one needs to look no further than HIV. It is known that there is considerable person-to-person variability in the duration between initial HIV infection and the onset of AIDS. Some of this variability is explained by the presence of particular MHC class I alleles that are more effective at presenting HIV peptides and mediating the killing of HIV-infected cells. One such allele, called HLA-B*51, commonly presents the HIV-derived peptide TAFTIPSI (listed in single letter code), and has been associated with delaying the onset of AIDS.

G. (3 points) From what protein is the HIV epitope TAFTIPSI derived? Choose from: i) gp120, ii) gp41, iii) reverse transcriptase, iv) Vpu, v) p17, vi) p24. Describe how you obtained your answer in fewer than 3 sentences. Hint: use one of the techniques already introduced to you on this problem set.

An international research group led by Philip Goulder analyzed how the sequences of HIV proteins differed between HIV-positive individuals who possessed the HLA-B*51 allele and those who did not. In one group of individuals who were recently infected with HIV-1 and also positive for HLA-B*51, the HIV-derived TAFTIPSI peptide was mutated to the so-called “escape sequence” TAFTIPSX (X denotes a non-isoleucine, non-valine amino acid) in 205/213 (96%) individuals!

H. (5 points) Propose a mechanism based on natural selection to explain the high prevalence of TAFTIPSX in HIV-positive, HLA-B*51-positive individuals. For this part, you do not have to go into structural details of the MHC:peptide complex. (< 8 sentences)

I. (4 points) If evolutionary forces are indeed shaping the high prevalence of TAFTIPSX in HIV-positive, HLA-B*51-positive individuals, what prevalence of this mutated epitope is expected in HIV-positive, HLA-B*51-negative individuals? Explain.

If you remember from lecture, peptides that bind to a particular MHC molecule share structurally-related anchor residues. The anchor residues that contribute to binding a particular MHC allele need not be identical, but are always related. The figure below highlights this:

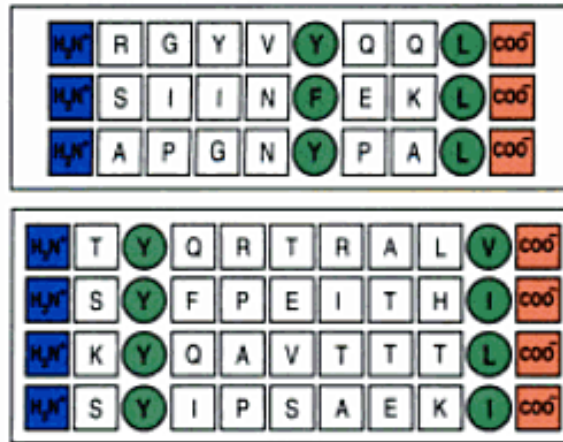


Figure 2: Peptides eluted from two different MHC class I molecules are shown in the upper and lower panels, respectively. Sequences are presented using the single letter code. Anchor residues (green) differ for peptides that bind different alleles of MHC class I molecules but are similar for all peptides that bind to the same MHC molecule. For example, phenylalanine (F) and tyrosine (Y) are both aromatic amino acids. Peptides also bind to MHC class I molecules through mainchain atoms in their amino (blue) and carboxy (red) termini.

Of the HIV TAFTIPSX mutations in HIV-positive, HLA-B*51-positive individuals, the TAFTIPST variant was the most common and the TAFTIPSV variant was often observed to convert to another mutation shortly after infection.

J. (5 points) Give two plausible reasons why the TAFTIPST variant was the most commonly observed HIV variant in HLA-B*51-positive individuals. For the first reason, think about the accessibility of different amino acids in DNA sequence space. For the second reason, consider the structural basis of MHC class I molecule-to-peptide binding and what role the last amino acid in this peptide sequence might play in mediating MHC:peptide interactions.

K. (3 points) Give one plausible reason why the TAFTIPSV variant was commonly observed to convert to another mutation after infection.

As you have learned in lecture, the HIV reverse transcriptase is highly error-prone because it lacks the proofreading ability found in other nucleotide polymerases. This error rate, reported to be as high as 1 in 30,000 nucleotides, accelerates the evolution of HIV.

L. (5 points) What is the probability that a reverse transcription reaction with the HIV reverse transcriptase results in a point mutation of TAFTIPSI to an escape sequence? Assume that: (i) we do not care about the rest of the protein that contains the TAFTIPSI subsequence; (ii) the only errors made by HIV reverse transcriptase are nucleotide substitution errors; (iii) each of the 3 isoleucine codons has equal probability of being used; and (iv) the probability that two or more nucleotide substitution mutations in this subsequence is miniscule and can be neglected.

M. (5 points) How many T cells would have to be infected by HIV in order for one to expect to find an HIV virion with an HLA-B*51 escape sequence? How does this number compare to the number of T cells ($\sim 10^9$) that are in circulation in the human body?
Assume that: (i) on average there are 4 integrations of HIV DNA per infected T cell; (ii) newly-budded HIV virions do not re-infect the T cell that produced them; and (iii) all HIV integration events are of unique HIV DNA sequences.

Please remember to enter the number of hours you spent on the set on the cover page and submit your comments to the Bi1 moodle site.

<http://courses.caltech.edu>