18.417 - Introduction to Computational Molecular Biology **PS 2**
Bonnie Berger and Manolis Kamvysselis October 2, 2001

Problem Set 2
Due Date: Tuesday, October 16

1. **Using BLAST:** Open up a web browser and go to http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast Using the NR database (NR stands for Non-Redundant - this is the default), enter the amino acid sequence GLSPETRRLVRQRQ. Note that if you just type in those 14 letter in upper case, the sequence will be in FASTA format. How many database sequences are returned? Click on the first 14-mer to find out more about it. What is the accession ID? Now do a BLAST search (from the same page as before) on this accession ID. What family have you found? Family is not meant here as a technical term - just explain what biological function all the most closely aligned proteins share.

2. **Understanding FASTA:**

   (a) In class we said that FASTA does not compute an exact alignment. In this problem we will investigate one way in which FASTA can fail. Take the sequences

   $ACCGTAGAAGAA$
   $ACCTACAACAAC$

   and consider the effect of running FASTA to align the two sequences using only 3-mers. What is the only offset value at which we have a non-zero score? Now consider the alignment depicted below.

   $ACCGTAGAAGAA-$
   $-ACCTACAACAAC$

   Using the scoring metric that matches receive a +1, mismatches a -1, and gaps a -2, what is the score of the alignment produced by FASTA? What is the score of the alignment depicted in the second picture? What does this show about FASTA?

   (b) Compute the optimal FASTA alignments of ACTGTACGTA and ACTGCGTACG using a window of length 3 and a band of width 2 on either side of the diagonal. Only consider the single optimal offset produced by FASTA. Turn in your scoring matrix.

3. **The Gibbs Sampling algorithm.** Choose a set of orthologous sequences from `ftp://ncbi.nlm.nih.gov/pub/COG/COGs/`, and apply the code in

/mit/18.417/share/problem_code/gibbs.py to find a motif in it. You may find the code in .../problem_code/tests/test_gibbs.py has helpful examples of how to use it. Don't forget to mention which group of proteins you ran the code on!

Which residues have the greatest frequency at each position? Does this seem like a genuine motif?

The code in gibbs.py does not take into account the *background frequency* of the residues. Find a way to fix this. (Hint: personally, I found the description of Gibbs Sampling at

http://bayesweb.wadsworth.org/gibbs/content.html very helpful.)

4. **Physical Mapping:** Suppose you are studying a new plasmid (with circular DNA) that is 2500 bases long, whose restriction map you wish to construct. You treat the plasmid DNA with a set of restriction endonucleases and measure the size of the resulting fragments by gel electrophoresis to obtain the following results.

EcoRI	–	2500
HindIII	–	2500
PstI	–	2500
MboI	–	1300, 800, 400
MboI	+	EcoRI – 1300, 600, 400, 200
MboI	+	HindIII – 1300, 800, 300, 100
MboI	+	PstI – 1000, 800, 400, 300
EcoRI	+	HindIII – 2000, 500
EcoRI	+	PstI – 1600, 900
HindIII	+	PstI – 2100, 400

Construct a restriction map based on this information. To break the circularity, place base pair 1 at the HindIII cleavage site.

5. **Genetic mapping:** The family described below has been genotyped for three linked markers, A, B, and C. The pair of alleles found (each represented by a digit) is listed below for each marker and each person in the family. All the alleles come from genes located on a single chromosome.

Father:	I–1
Mother:	I–2
Offspring:	II–1 through II–10

		marker	
Person	**A**	**B**	**C**
I–1	1,5	2,3	6,8
I–2	1,9	4,7	3,6
II–1	1,5	2,7	6,6
II–2	1,9	3,4	3,8
II–3	1,5	2,7	6,8
II–4	5,9	3,4	3,6
II–5	1,9	3,7	3,8
II–6	5,9	2,4	6,6
II–7	1,1	3,4	6,8
II–8	1,5	2,7	6,6
II–9	5,9	2,4	3,6
II–10	1,9	3,4	6,8

a. For each of the offspring, list which alleles were inherited from the mother and which from the father.

b. Count the frequency of each parternally- and maternally-derived haplotype. What are the parents' haplotypes most likely to be?

c. For each pair of markers, count the total number (both maternal and paternal) of crossovers that occurred between just those markers. (I.e., for markers A and C, ignoring the data for marker B entirely, how many crossovers do you see?) Use this information to build a map of the markers.

d. Verify your map by determining the haplotypes corresponding to double crossovers. How often do they occur?