6.874/7.90/6.915 Computational Functional Genomics — Handout 4
Spring 2005 — May 6, 2004

Potential Final Quiz

Problem 1:

We demonstrated how to model the expression of genes using a multivariate Gaussian model. A multivariate Gaussian can be created by combing independent Gaussians with the formulation Y = AZ + M, where A is a transformation matrix, Z is a vector of observations from independent normal Gaussians, and M is a vector that contains the mean of Y.

a) Express how to approximate the transformation matrix A in terms of the eigenvectors and eigenvalues of a covariance matrix from observations of Y.

b) Explain what one could learn from expressing the model of the data in this way.

c) Assume that you are given a set of (multidimensional) points, and told to cluster them using a mixture of multivariate Gaussians. What additional unobserved data must we assume to accomplish this clustering? What algorithm could we use to perform this clustering? What value does that algorithm maximize?

Problem 2:

You observe that a motif is in 20 of the 6000 intergenic regions in yeast, but the motif is only in 12 of the 15 intergenic regions that are bound by your favorite transcription factor.

a) Under the "null hypothesis" that bound regions and motif-bearing regions are both drawn independently and uniformly at random from the total set of genomic regions, how would you approximate the probability that this motif occurred in 12 of the 15 sequences "drawn at random?"

b) Now, assume that the motif sequences are not independent at all from the bound sequences (regions), but rather were discovered by using those bound sequences *as input* to some form of motif discovery tool. Assume that this is a "special" motif discovery tool which, when given 15 intergenic regions, returns a (learned) motif model *nearly instantaneously.* Furthermore, assume you have a function that chooses regions uniformly at random from the set of total genomic regions.

   Devise a method to calculate a an *empirical* significance score for your motif, using these tools. Explain, in one ore two sentences, your null hypothesis for this method.

Problem 3:

We have seen that there is an important difference in how we interpret graphical models of gene regulatory function when edges are causal. In particular, the graphs A -> B and B -> A are equivalent under a non-causal interpretation, but are not identical under a causal interpretation.

(a) Assume that A and B are observed binary variables, and that we have observed that P(A = 1) = .5, P(B = 1) = .9. Give one conditional probability table for A->B and one conditional probability table for B->A that fits this observed data.

(b) If we know A causes B, why do A->B and B->A no longer have equivalent explanatory power? Explain in detail.

Problem 4:

A large-scale low-resolution ChIP-Chip binding experiment has determined that a transcription factor A binds the upstream intergenic region of gene B. A series of systematic genetic deletion experiments were performed, to narrow that binding down to a small genomic locus with known sequence:

TGCCTGAGGCT

Previous genome-wide comparisons and alignments have been performed between the species in which the experiments were performed, S, and a closely-related species S*. The ortholog to gene B in species S* is called B*, and the corresponding region of DNA in B*'s upstream region has the sequence

AGTCTGGCTA

A previous study showed that a reasonable scoring matrix for substitutions between these two species is approximated by:

-	A	T	G	C
A	1	-1	-3	-3
T	-1	1	-3	-3
G	-3	-3	2	-1
C	-3	-3	-1	2

a) Using the scoring matrix as your pairwise match-function on nucleotides, and a gap penalty of -2, use the Smith-Waterman algorithm to find the optimal local alignment between the two sequences. Show your alignment matrix, the trackback pointers, and circle the optimal score of the best alignment.

b) Now, use this alternate matrix to perform the same alignment.

-	A	T	G	C
A	2	-1	-3	-3
T	-1	2	-3	-3
G	-3	-3	1	-1
C	-3	-3	-1	1

How do your results change?

Problem 5:

In this problem, we ask you several basic questions about Bayesian networks, the assertions of independence they encode, and the factorization of the joint probability distribution that they imply. We also ask you to read off independence assertions from a given Bayesian network.

We start with a canonical set of variables: A, B, C, D, E, and F. You may assume, if you wish, that each variable is binary.

a) Draw a network that respects the following independence assertions:
   a. E is independent of C *if both* D and B are known.
   b. D and B are independent, *unless* C is known.
   c. D and A are independent, *if* C is known, *unless both* C *and* F are known.
   d. C and F are independent, *if both* D and A are known.
   e. If D is known, then E is independent of every other variable.

b) Write down the joint probability distribution, in terms of the conditional probability distributions of each variable.

c) Answer the following questions:
   a. If F and A and known, are B and D asserted to be independent?
   b. What about if just F is known?
   c. If C is known, are A and B asserted to be independent?
   d. If B is known, are A and C asserted to be independent?

d) We'd like to modify the assumption (a:a) above: We like to assume that E and C are independent if D=1, but *not* if D=0. How could we modify the network from (a) above to encode this new assumption?

Problem 6:

In our class lectures and the second problem set, we outlined a probabilistic form of motif discovery known as "OOPS" EM – Expectation Maximization under the "(exactly) one occurrence per sequence" assumption. In this model, each sequence *S* (that is, each region of DNA which is believed to be separately bound by the DNA-binding protein in question) is considered an independent sample. Furthermore, each sequence *S* is associated with a hidden variable *L*, the "location" of the motif within that sequence.

If we knew, for a sequence *S*, the value of *L* then we could compute the likelihood of each sample <S, L>:

P(S, L) = P(S|L) * P(L)

(and from that, we could compute the total data likelihood, P(S*, L*) for all S and L).

In this problem, we will ask you to adapt the OOPS-EM setup in order to take into account a simple form of sequence conservation information.

In the old OOPS-EM, each sequence *S* was considered to be a string of letters, s_1…s_N. In our new version of OOPS-EM, which we will call cOOPS-EM, each s_i will instead be a *pair* of DNA letters, corresponding to a pair of aligned bases in homologous regions from two related species (as in a sequence alignment). Therefore, s_i will be a pair representing the aligned column at position i of a sequence, and the letters from each sequence will be denoted s_i0 (the letter at position i of the 'reference' sequence) and s_i1 (the letter at position i of the aligned sequence).

Note that our alphabet is no longer just the standard DNA bases, but also now includes a "gap" character: {A, T, G, C, -}. However, for simplicity we will assume that s_i0 is always a non-gap character.

The model parameters of cOOPS-EM are a superset of those in OOPS-EM. As before, we have a PWM that represents our motif model, and a background model (you may assume a 0^{th} order Markov model) that represents the "background" sequence. However, we will also add two additional components: the 'conservation matrices'.

A 'conservation matrix' has a form similar to that of a substitution matrix in sequence alignment. Formally, a conservation matrix is a conditional distribution on the letter in sequence 1 (s_i1, for some i), *given* the letter in the corresponding position in sequence 0 (s_i0, for the *same i*). We assume that we have one such conservation matrix, C_M, for the positions which are "in a motif," and a second matrix C_B for all those positions which are not in a motif (in "the background"), for a given value of *L*.

Therefore, if position i is "in the motif" (that is, L <= i <= L+w), P(s_i1 | s_i0) = C_M(s_i1, s_i0).

a) How many parameters are required to completely specify C_M?

b) In OOPS-EM we wished to calculate P(S, L). Now, in cOOPS-EM, we want to calculate the new total data likelihood, P(S_1, S_2, L). Assuming that individual columns are independent (given the value of *L*) in much the same way as individual letters are independent in OOPS-EM (again, when the value of *L* is known), derive the formula for P(S_1, S_2, L) in terms of M, B, C_M, and C_B.

c) Using Bayes Rule, use the expression you derived in (b) to derive an expression for P(L=i | S_1, S_0), for all i.

d) In an iterative algorithm, such as EM, we wish to update the values of C_M and C_B (and all the other parameters) to maximized the total expected log likelihood of the data, where the expectation is taken with respect to the P(L=i|S_1, S_0) which we calculated in (c) (for convenience, let us call this quantity e_i). Considering only the C_M matrix for a moment (you may ignore C_B for this problem), derive the equations for the components of this which maximize this expected log likelihood.

Problem 7:

Suppose we entertain only the following two causal models:

M1 : f (+)→ g
M2 : f ←(-) g

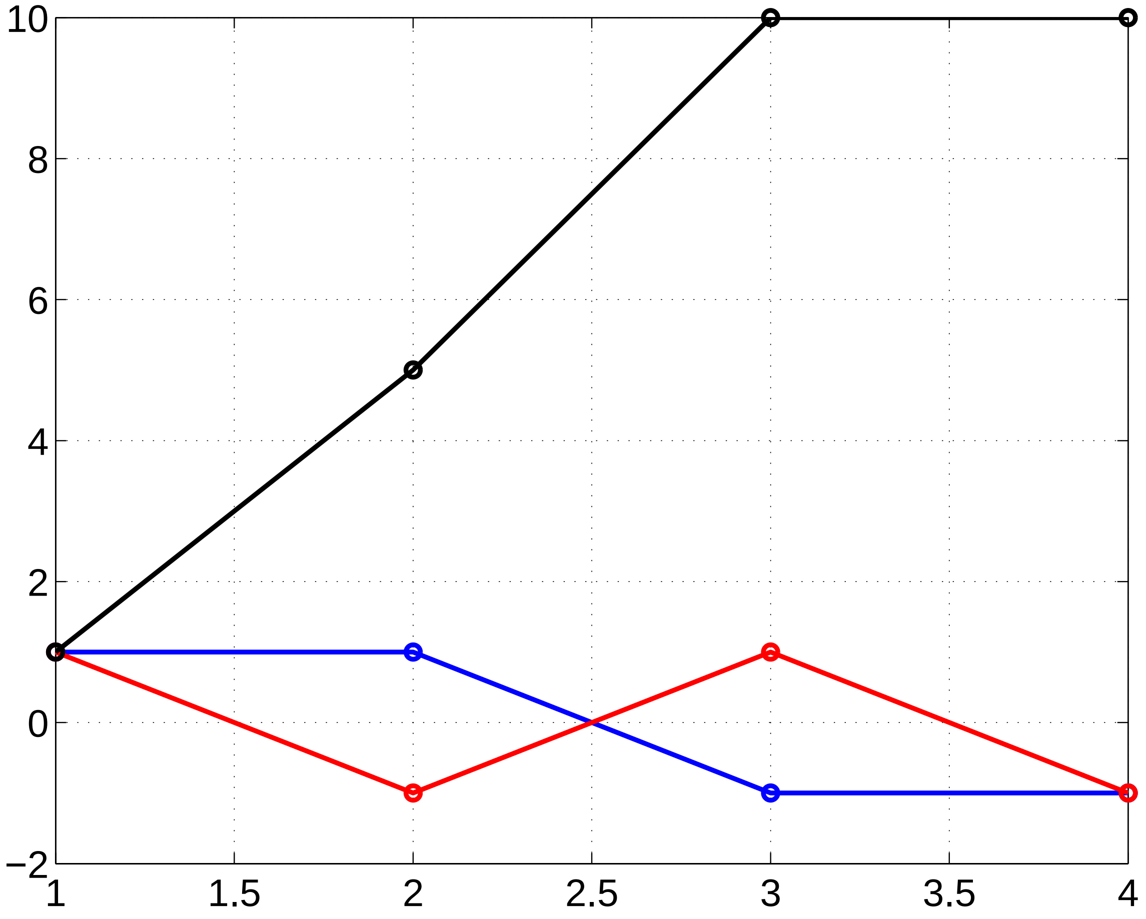
Based on the experiments we have carried out so far we believe that P(M1) = 0.1 and P(M2) = 0.9. We can delete either f or g. Our goal is to find out which model is correct.

a) Does it matter which experiment we carry out in this case?

b) Provide a brief justification (we don't expect you to need to do any numerical calculations)

Problem 8:

The plot below provides four expression measurements for three genes (black, blue, andred). We wish to use a decision tree to predict the expression response of the "black" gene on the basis of hypothesized transcription factors "red" and "blue". Provide a possible (simple) decision tree, indicate the decisions at the branches, and the mean responses at the leafs.


10
8
6
4
2
0
−2
1
1.5
2
2.5
3
3.5
4