10-810 – Advanced Algorithms and Models for Computational Biology: a machine learning approach. Problem Set 3

This problem set is due on Monday, 03/27 in class.

Cubic splines
We have discussed the use of cubic splines to represent time series expression data with continuous curves.

1. a. In this part we will discuss interpolating splines, that is splines that go through each of the points. Assume we have a value of v for point i. Let $S_1 = ax^3 + bx^2 + cx + d$ be the spline to the left of this point and $S_2 = ex^3 + fx^2 + gx + h$ be the spline to its right. You may parameterize both splines as starting in 0 and ending at 1. How many equations are defined by point i ? Write all these equations and simplify each as much as you can.

1. b. In class we actually discussed approximating splines, that is splines that contain less control points than the number of actual measured points. One issue I did not mention is how to chose how many control points to assign. If we assign too many we will overfit the data but if we assign too few we might not be able to accurately reconstruct the underlying expression curve. Assume control points are uniformly spaced. Suggest a method for determining the number of control points we should use.

Synchronization
As mentioned in class, synchronization is an issue when working with a population of cells. In this question we will try to determine a synchronization loss model for yeast cells.

2. a. We assume that all cells have internal clocks which they follow. Clocks speeds are normally distributed with mean 1 (the real time) and an unknown variance. Describe a possible biological explanation to this assumption.

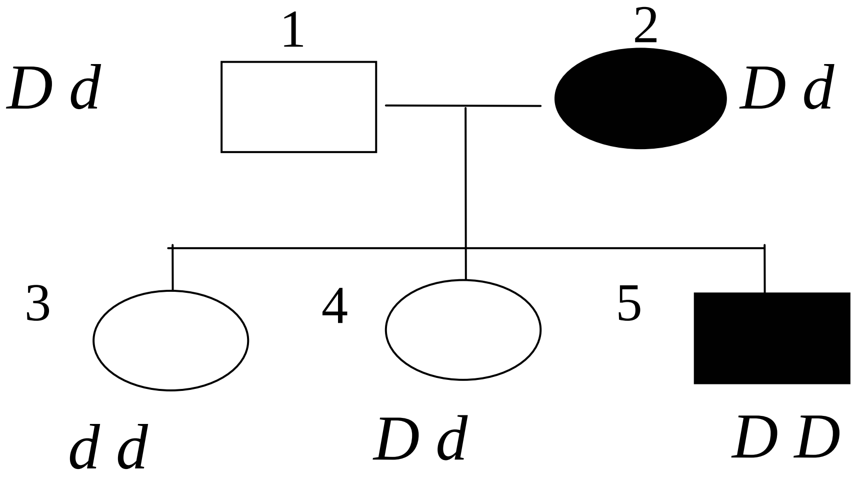
Our goal now is to estimate the variance for the speed of these cells. Assume we have a way to divide cells into three different groups (or phases): G1, S and G2/M (for example, using FACS data). Note that we know that all cells start in G1, move to S, then to G2/M and then to G1 again and so forth. What we do not know is the time they spend in each of these phases and the clocks variance so we have four parameters to fit. We measure the percentage of cells in each of these phases at 7 time points (every 15 minutes between 0 and 90). We know that none of the cells completed more than two cycles during these 90 minutes.

2. b. How many equations do we have for each of the measured time points (except for time point 0 where we assume that all cells are in G1) ? Write all equations for one of the time points.

2. c. How would you solve these set of equations (you should have more equations than unknowns since we have 6 time points beyond 0) ?

3. Some short proofs of the recombination process.
    a. Prove the Mether's formular: p(R(J) = 0.5*p(X(J)>0).
    b. Explain what is a first-division segregation (FDS) and a second-division segregation (SDS). Why they are intesting?
    c. Prove the inductive formula for the second-division segregation: $S_{k+1}=4F_k+2S_k$. (note that in class we proved that a similar formula for the FDS, mimic the technique we used there!)
    d. Discuss the relationship and difference between Haldane's mapping function p(R(d))=(1/2)(1-exp(-2d)) and the mapping function for the SDS frequency s(d)=(2/3)(1-exp(-3d)).

4. Genetic Linage Analysis
    a. What does "Hardy-Weinberg equilibrium" mean in terms of inheritance and mating outcome?
    b. Consider the following pedigree on slide 12, lecture 8:



What is the conditional probability of the genotype of the individual 2, i.e., $p(G_2=Dd|G_1=Dd,G_3=dd,G_4=Dd,G_5=DD)$? Use the allele frequencies given in the previous slides.

    c. What LODs of linkage are additive across independent pedigrees?
    d. Explain for *n* QTLs there are 2^n distinct genotypes for BC and 3^n distinct genotypes for IC?

5. A common strategy to infer the haplotype of multiple SNPs is to used a parsimony criteria to evaluate the results and devise algorithm that greedily optimize this criteria over possible solutions. Now suppose we have the genotype of 3 SNPs of individual 1: G_1={1/0, 1/0,1/0}, and we also know that the genotype of an individual 2 is: G_2={1/1, 0/0,1/1}. Can you guess what is a "parsimonious" solution of individual 1's haplotypes? What if G_2={0/0, 0/0,0/0}?
    a. Why haplotype is advantageous over single SNPs for linkage analysis?

b. Why is the computational and biological meaning for finding “blocks” in long sequences of SNPs?

c. Implement the EM algorithm for Haplotype inference and test it on the data to be posted.

To open the zipped file, use command “tar zxvf hap.data.tgz”. In the data directory, you will see a .genos and a .haplos file. The format is as follows:

Genotype:

```
1603 0  2 -1 1  1 0 0 1  1 1 1 2 1  0
2103 0 -1  2 1 -1 0 0 1 -1 1 1 0 1 -1
…
```

The first column is the sample id, starting from the second column, you see the genotype of 14 SNPs in each sample. “0” denote the genotype (0,0), “1” denote the genotype (1,1), “2” denote a genotype (0,1), i.e., a heterogeneous genotype, and finally, “-1” denote an artifact in the measurement of that SNP, and can be ignored.

Haplotype:

```
1603 0 0 1 -1 1 1 0 0 1  1 1 1 1 1 0
1603 1 0 0 -1 1 1 0 0 1  1 1 1 0 1 0
2103 0 0 0  1 1 1 0 0 1 -1 1 1 0 1 -1
2103 1 0 -1 0 1 0 0 0 1  1 1 1 0 1 0
…
```

The first column is the sample id; the second column is the index of paternal and maternal allele; starting from the third column, you see the haplotype of 14 SNPs in each of the two alleles of every sample. Note that “-1” denote an artifact in the measurement of that SNP, and can be ignored.

You need to infer the haplotype of the SNPs of all the samples using your program. Report both the haplotype and the overall error ratio (i.e., the number of wrongly inferred loci over all heterozygous loci).