

Spring 2008 10-810 / MSCBIO2070: Computational Genomics

Problem Set 4: Haplotype Inference and Bayesian Networks Due date: April 10, 2008, before class

The contact TAs for this assignment are Jacob Joseph (jmjoseph@andrew.cmu.edu) and Aabid Shariff (aabid@cmu.edu).

This assignment requires no implementation. If any machine code is submitted, supply source code in Python, Java, MatLab, C, C++, or R. Code should be portable; it should execute on a UNIX/Linux platform. Please include any auxiliary files and short description of how to execute your code. Submission of this source code may be by email to the contact TAs.

Collaboration is permitted, but solutions must be completed individually. Include a list of your collaborators. Refer to the course website for complete policies.

1. Haplotype Inference by Expectation Maximization (Jacob)

In his guest lecture, Prof. Schwartz described a variety of genome polymorphism mechanisms and techniques by which disease association may be ascertained. In particular, Single Nucleotide Polymorphism (SNP) refers to slight nucleotide variation among a population, whereby one or few nucleotide positions in an otherwise conserved chromosome segment achieve a distribution of bases. A common goal is then to examine correlation of disease with the values of blocks of such positions. Current sequencing technologies do not separately determine the sequence of a single chromosome (a *haplotype*), from that of its homolog from the other parent. As a result, disease association depends first upon haplotype inference. Without implementation, this question steps through such inference using expectation maximization.

Consider a collection of four pieces of DNA:

```
      *                               *
A C T T G G A C T G T T A C A
A C T T G G A C T G T T A A A
A C G T G G A C T G T T A A A
A C T T G G A C T G T T A C A
      *                               *
```

The sequence obtained from an individual would not resolve homologous copies, yielding

AA CC TG TT GG GG AA CC TT GG TT TT AA CA AA,

where the real pair of sequences was

```

      *                               *
    A C T T G G A C T G T T A C A
    A C G T G G A C T G T T A A A
      *                               *

```

or

```

      *                               *
    A C G T G G A C T G T T A C A
    A C T T G G A C T G T T A A A
      *                               *

```

Consider now the following formalism. Only the two sites which vary are informative; others may be ignored. That is, the four possible resolutions are GC, GA, TC, and TA. We wish to estimate the four haplotype frequencies:

$$\lambda = \{f_{GA}, f_{GC}, f_{TA}, f_{TC}\}$$

At each site, three possible pairs may be observed. For the first site, these include GG, TT, and GT/TG. These *genotypes* may be encoded as strings from the alphabet $\{0,1,2\}$, where $GG \rightarrow 0$, $TT \rightarrow 1$, and $GT/TG \rightarrow 2$. The genotypes of the second site may be similarly mapped: $AA \rightarrow 0$, $CC \rightarrow 1$, and $AC/CA \rightarrow 2$. Our input then consists of the counts of each of nine possible genotypes in the observed population:

$$x = \{n_{00}, n_{01}, n_{02}, n_{10}, n_{11}, n_{12}, n_{20}, n_{21}, n_{22}\}$$

From a maximum likelihood model based on Hardy-Weinberg equilibrium, the likelihood of a particular parameter set λ for an input set x is

$$Pr\{x|\lambda\} = \prod_{\text{genotypes } g_i} \sum_{(ab,cd) \text{ consistent with } g_i} f_{ab}f_{cd}$$

To formulate an EM approach, we will define a set of latent variables y . y is chosen as a piece of information not known to us, but one that would make our parameter estimation easy if it were known. Here, we select y as the counts of genotype pairs:

$$y = \{g_{GG,AA}, g_{GT,AA}, g_{TG,AA}, g_{TT,AA}, g_{GG,AC}, \dots\}$$

- (a) Using the above notation, describe any ambiguity of haplotype for observed genotypes 00, 02, 11, and 22.
- (b) (E-step) In terms of the four haplotype frequencies, λ , and genotype counts, x , derive the expected counts of the genotype pairs $g_{GG,AA}$, $g_{GT,AA}$, $g_{GT,CA}$, and $g_{GG,CC}$.
- (c) (M-step) Given any x and an the expected value of y , the optimal value of λ may be found. Derive this estimate.
- (d) Consider the input genotype counts and initial haplotypes frequencies as follows:

$$x = \begin{pmatrix} n_{00} = 1 & n_{01} = 5 & n_{02} = 4 \\ n_{10} = 8 & n_{11} = 17 & n_{12} = 22 \\ n_{20} = 2 & n_{21} = 19 & n_{22} = 22 \end{pmatrix}$$

$$\lambda = \begin{pmatrix} f_{GA} = 0.095 & f_{GC} = 0.22 \\ f_{TA} = 0.255 & f_{TC} = 0.43 \end{pmatrix}.$$

Perform an E-step by estimating $y = \{g_{GG,AA}, g_{GT,AA}, g_{TG,AA}, \dots\}$.

- (e) Perform an M-step by re-estimating $\lambda = \{f_{GA}, f_{GC}, f_{TA}, f_{TC}\}$.
- (f) Repeat another two EM iterations, and comment upon convergence of λ .

2. Bayesian Networks (Aabid)

You are a student enrolled in the Computational Genomics class and you decide to go for a Molecular Biology conference outside Pittsburgh. When listening to the talks you make the following interesting notes on Drosophila genetics research for neurodegenerative disorders

- The speaker concludes that if a fly has Alzheimers, the higher the chance of being blue-eyed. The speaker claims that the probability of seeing blue eyed flies among Alzheimers flies is 0.98.
- At another talk, having Alzheimers also shows curly wings. Again, like the earlier talk, this speaker too claims a probability that seeing curly winged among Alzheimers flies is 0.92.
- One of the graduate students of the speaker claims that eye color of Alzheimers flies is conditionally independent of wing appearance.
- Another student reveals to you that eye color can also relate to Huntington's disease. He says that the probability of seeing a blue eyed fly being healthy is 0.2, having Huntington's is 0.5 and having both disorders is 0.34. He claims that Huntington's and wing appearance are not related.

- Finally, a colorful poster shows that independent of any disorders, the probability of tufted bristles on flies with red eyes and flat wings is 0.4. Of those with red eyes and curly wings, this probability is 0.72. Of those with blue eyes and curly wings, it is 0.64. Finally, of those with blue eyes and flat wings, it is 0.44.

On your way back to Pittsburgh, you remember that in class you saw that Bayesian Networks (BNs) have been useful in Systems Biology and you would like to integrate what you noted at the conference to make more inferences. But you know that for the BN, you need to do Parameter learning and Structure learning before you make inferences.

- Identify the random variables for a BN. Generate a BN (Structure) based on the above notes. Write down an expression for the joint probability of the BN. The information given does not reveal all the entries of the CPTs (parameters) to describe the network you have drawn. List what can be found.
- (Inference) If an Alzheimers fly shows tufted bristles, what is the probability that one will see blue eyes in the fly?
- (Inference) If a non-tufted fly shows Alzheimers, do you think that eye color and wing appearance could be dependent? Why? If a tufted fly has a flat wing appearance, what is the probability that it is blue eyed?
- (Inference) If an Alzheimers fly shows tufted bristles, what is the probability that it will have curly wings?

3. Bayesian Networks Theory (Aabid)

We will now try to understand some theoretical aspects of Bayesian networks.

- What is the maximum and minimum number of edges in a BN with n nodes? Draw all possible BNs with 3 nodes representing variables X , Y , and Z .
- Derive a bound for the maximum number of BNs that can be constructed with n nodes (explain your answer).
- Derive a bound for the minimum number of BNs that can be constructed with n nodes (explain your answer).

Hint: You can get a valid lower bound if you can show that it is always possible to create some 'function of n ' number of networks with n nodes. Order the variables in some manner. Now think about how many BNs can be constructed.

Hint for another method: Use mathematical induction.