Announcements

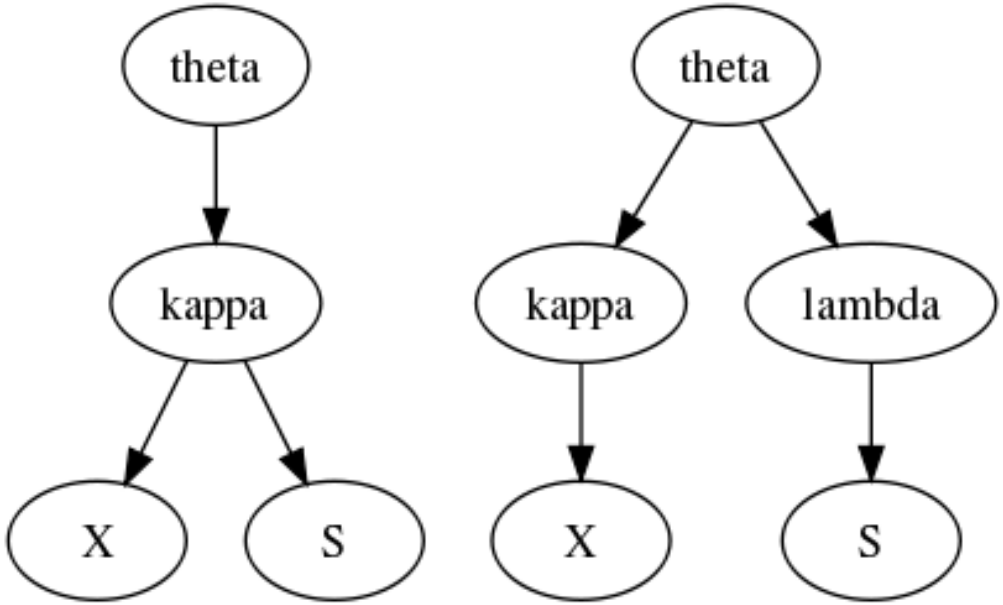
- HW 2 due today, please give it to me before end of class.
- **There is no discussion section this Friday, as the TA is out of town attending a conference**.
- Instead, we will use Monday's class as a discussion section where you will be able to get answers to questions you may have about implementation details of the project.
- After today, only one more class on basic theory; the rest of the course will focus on application examples, e.g. HMMs, alignment algorithms, phylogenetic tree algorithms.
- I think you'll find you use the basic theory in every bioinformatics problem you solve.

Programming Project 1

- the programming project (HW 3) is now available on the CourseWeb site.
- Due Oct. 19 in class.
- You may use any standard programming language, however if it requires compilation you must provide a `make` script (or `Makefile`) that will automatically compile it.
- See the detailed instructions in the project assignment.
- We will use automated testing (feeding your program different input files) so be careful to follow the detailed instructions.

Forensic Test Complexity

What is the computational complexity for computing $p(X, S|\theta)$ in the *match* vs. *mismatch* models? For the sake of simplicity, write your answer in big-O notation, assuming that all variables have N possible states.



Forensic Test Complexity Answer

For the match model, we are summing over a single variable

$$p(X, S|\theta) = \sum_{\kappa} p(\kappa|\theta)p(X|\kappa)p(S|\kappa)$$

so the computational complexity is $O(N)$.
For the mismatch model we are summing over two variables, but as *separate factors*:

$$p(X, S|\theta) = \left(\sum_{\kappa} p(\kappa|\theta)p(X|\kappa)\right)\left(\sum_{\lambda} p(\lambda|\theta)p(S|\lambda)\right)$$

so the computational complexity is still $O(N)$.

Half Your DNAs Is Belong to Us!

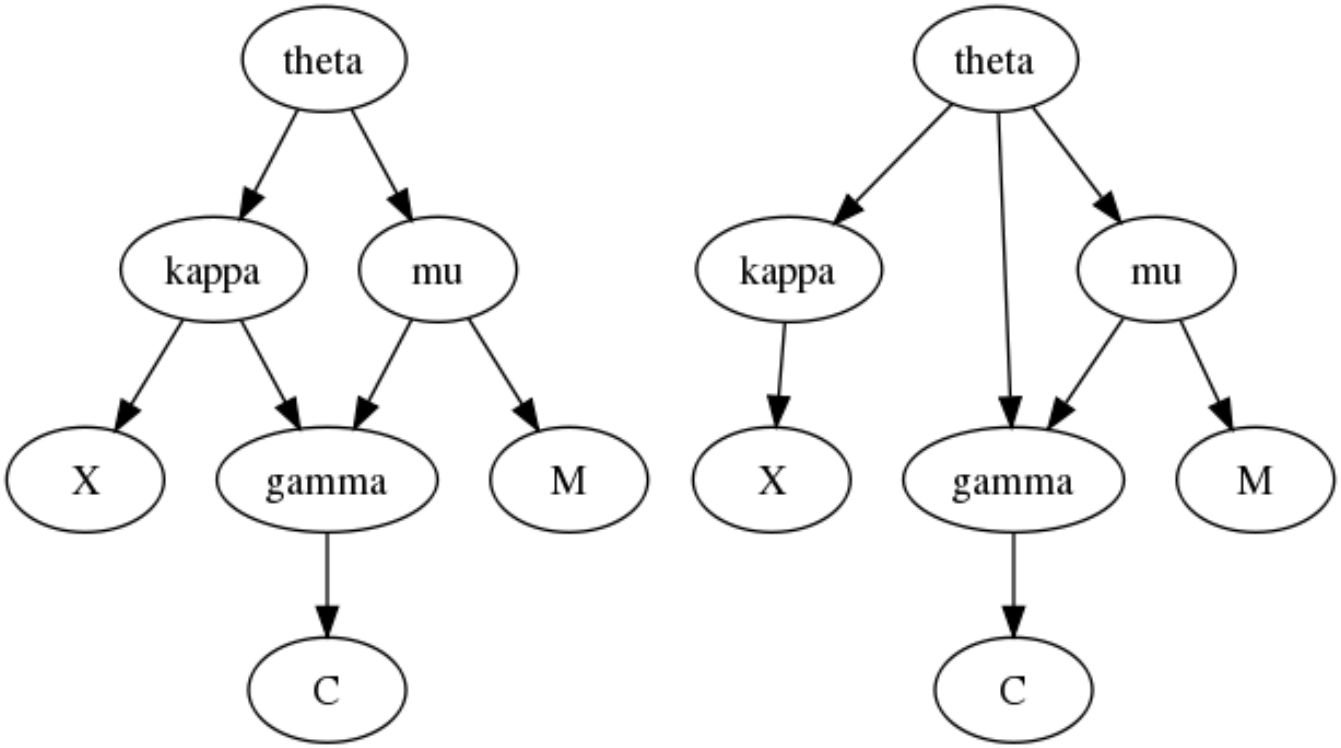
A Great Business Opportunity

Paternity Test Info Graph

After watching Maury Povitch you decide paternity testing would be more profitable. You decide to use the exact same microarray measurements of SNP markers, from three samples: measurement X from the candidate dad; measurement C from the child; measurement M from the mother.

Propose an information graph structure for the analysis of a single SNP, comparing two models: *dad*, which asserts that the candidate dad really is the father of the child; *not-dad*, that the candidate dad is unrelated to the child.

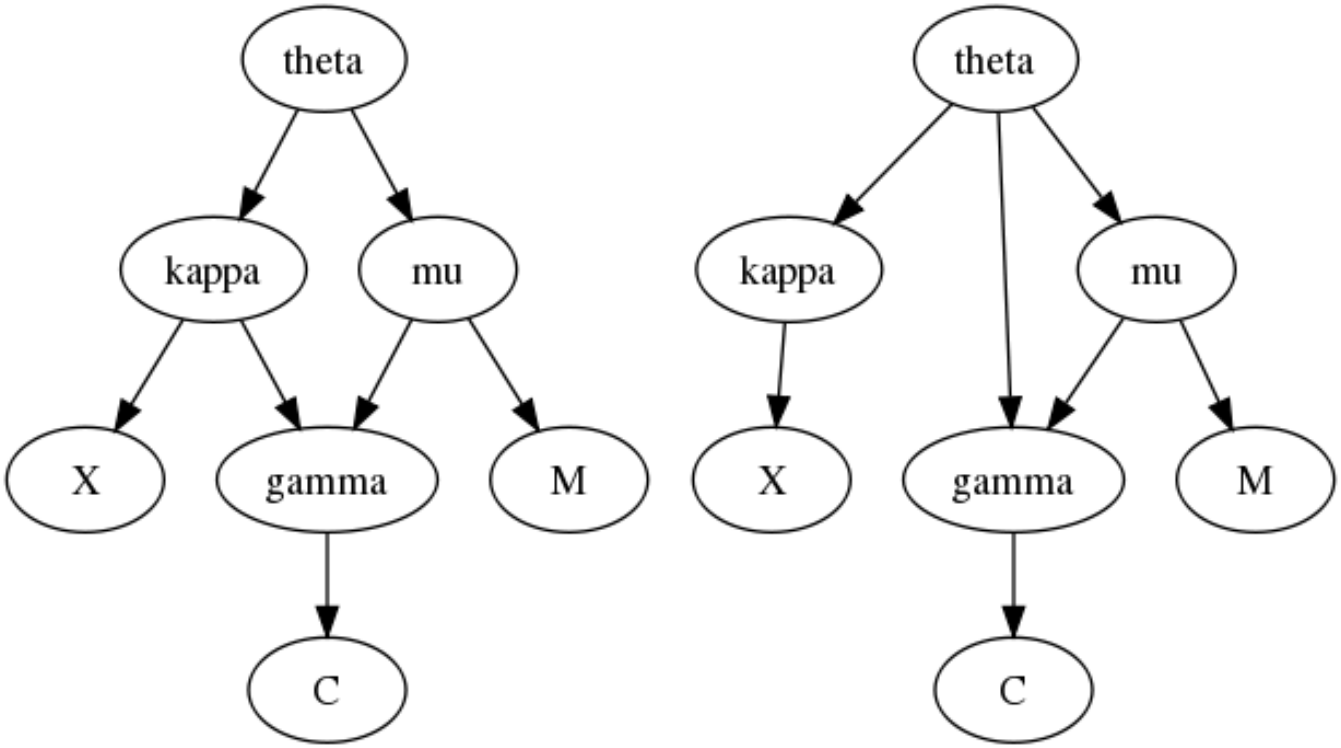
Paternity Test Info Graph Answer



Paternity Test Complexity

What is the computational complexity for computing $p(X, C, M|\theta)$ in the *match* vs. *mismatch* models? For the sake of simplicity, write your answer in big-O notation, assuming that all variables have N possible states.

Paternity Test Complexity Answer



- The *dad* model couples together all three hidden variables that we must sum over (κ, μ, γ), so the complexity is $O(N^3)$.
- The *not-dad* model still has three hidden variables that we must sum over, but only two of them are coupled together (μ, γ), so the complexity is $O(N^2)$.

$$P(X, G, M) = \sum_{\gamma} \left[P(\gamma) \left[P(G|\gamma) \right] \left[\sum_{K} P(K|\gamma) P(X|K) \right] \left[\sum_{\mu} P(\mu|\gamma) P(M|\mu) \right] \right]$$

$$P(X, C, M|\theta) = \sum_{K,\mu,\gamma} P(K|\theta) P(\mu|\theta) P(\gamma|K,\mu) P(X|K) P(C|\gamma) P(M|\mu)$$

Basic SNP Scoring

You are seeking to identify novel SNPs from short read sequencing of multiple people, pooled together. Assume that you are scoring a single genomic position from the set of base-calls at that position from different reads, and that the probability that a given base-call is wrong is a known value ε.

Propose a set of hidden variables and associated likelihood models for modeling this problem.

Basic SNP Scoring Answer

We are trying to find new SNPs so clearly we can't assume we know its frequency in the population beforehand: we must treat this as a hidden variable θ.
We observe some number of reads N, each of which has a basecall B for this position. Simplistically, the likelihood for B is just binomial (unmutated vs. mutated with probability θ), with a slight adjustment for the sequencing error rate e.g. $(1-\varepsilon)\theta + \varepsilon(1-\theta)$.
We should also consider the number of reads to be an (observable) variable, since it varies from site to site. A simple assumption is that the coverage is sampled randomly from some uniform density δ, a hidden variable; this implies a Poisson model for N.
Note that in this problem ε is a constant.

SNP Scoring

You are seeking to identify novel SNPs from short read sequencing of multiple people. Each read has a unique tag that identifies the individual person that it comes from. Assume that you are scoring a single genomic position from the set of base-calls at that position from different reads, and that the probability that a given base-call is wrong is a known value ε.

Propose a set of hidden variables and associated likelihood models for modeling this problem.

SNP Scoring Answer

We are trying to find new SNPs so clearly we can't assume we know its frequency in the population beforehand: we must treat this as a hidden variable θ. We don't know whether the putative SNP is actually present in a given person, so we treat its copy number in that person as a hidden variable κ. Assuming each copy is chosen independently, this implies a binomial likelihood.

Finally, we observe some number of reads N, each of which has a basecall B for this position. Simplistically, the likelihood for B is just binomial (unmutated vs. mutated with probability $\kappa/2$), with a slight adjustment for the sequencing error rate e.g. $(1-\varepsilon)\frac{\kappa}{2} + \varepsilon(1-\frac{\kappa}{2})$. We should also consider the number of reads to be an (observable) variable, since it varies from site to site. A simple assumption is that the coverage is sampled randomly from some uniform density δ, a hidden variable; this implies a Poisson model for N.

Note that in this problem ε is a constant.

SNP Scoring Info Graph

Write an information graph for the SNP scoring model. Make sure to explicitly represent the fact that there are multiple people by including variables for two people, and that there are multiple reads per person, again by including variables for two reads per person.

SNP Scoring Info Graph Answer