Exam 2 Study Guide

November 5, 2024

This study guide is intended to help you to review for the second in-class exam. This is not an exhaustive list of the topics covered in the class and there is no guarantee that these questions are representative of the questions on the exam. You should also review the notes you took in class, the notes and readings on the syllabus, and your homework assignments.

The second in-class exam will focus on material on substitution matrices and sequence motifs covered in lectures 9 through 17. This includes all material on amino acid substiution matrices, PSSMs, and the Gibbs sampler. You are responsible for the formal definition of a Hidden Markov model (HMM) and HMM design. You should understand how parameter values are estimated from labeled training data. You are not responsible for the HMM recognition algorithms covered on October 31st or motif discovery and parameter estimation with unlabeled data.

These subjects build on the material on Markov chains and nucleotide substitution models introduced in the first third of the semester, you should be familiar with that material as well. You may wish to take another look at the Study Guide for Exam 1 to remind yourself of the topics covered.

This exam is closed book. You may bring two pages (or one page, front and back) of your own notes. No electronic devices may be used during the exam. There is a clock in the class room. You will not need a calculator to answer the questions.

Amino acid substitution models and matrices

- Log-odds formulation.
  - A likelihood ratio compares the probability that an observation is the outcome of a process described by hypothesis H_A, and the probability that the observation is due to chance, described by the null hypothesis, H_0. You should understand the interpretation of a likelihood ratio in the context of a pairwise alignment. What are the alternate and null hypotheses, H_A and H_0, in this context?
  - What are the advantages of using the log likelihood ratio, instead of simply the likelihood ratio?
  - How is the log likelihood ratio used to construct a scoring function for an alignment?
  - What does it mean if the likelihood ratio is less than one? Greater than one?

  - What does it mean if the log-likelihood ratio is less than zero? Greater than zero?

- Deriving amino acid substitution matrices: overview
  - Desired properties for a substitution matrix
    - Substitution matrices should be parameterized by evolutionary divergence.
    - Substitution matrices should account, directly or indirectly, for multiple amino acid replacements at the same site.
    - Substitution matrices should reflect biophysical properties. Pairs of residues with similar properties represent conservative replacements and should have higher similarity scores than pairs of residues with different properties, which represent non-conservative replacements.
  - You should understand the similarities and difference between amino acid substitution matrices and DNA substitution models. Both formalisms are representations of the process of substitution at a single site in a sequence. Compared with DNA, the amino acid alphabet is larger and the properties of the amino acids are more varied. Amino acid substitution models rely more heavily on learning parameters from data than nucleotide models.
  - You should be familiar with two families of amino acid substitution matrices: the PAM matrices and the BLOSUM matrices. Both families were derived according to the following general approach, although the details of each step differ between the two methods.
    1. Use a set of "trusted" multiple sequence alignments (ungapped) to infer model parameters.
    2. Count observed amino acid pairs in the trusted alignments, correcting for various types of sample bias.
    3. Estimate substitution frequencies from amino acid pair counts.
    4. Construct a log odds scoring matrix from substitution frequencies.

- PAM matrices use the Dayhoff Markov model of amino acid replacement.
  - The unit of divergence used is the PAM or "percent accepted mutation". What is the precise definition of this unit?
  - The PAM matrices are derived from a Markov model of amino acid replacement. What is the basic structure of this model?
  - What are the properties of the data that Dayhoff used to obtain amino acid pair counts for her model? How are those properties related to the underlying assumptions of the modeling strategy that she used?
  - How did Dayhoff derive counts from that data set?
  - How did Dayhoff account for potential sample bias in her data?
  - How did Dayhoff use the amino acid counts to derive the PAM transition matrix? How does this derivation account for differences in amino acid frequency and amino acid mutability?

  - How did Dayhoff ensure that her basic model corresponds to exactly 1 PAM of divergence?
  - How is the PAM-N model derived from the PAM-1 model?
  - How are multiple substitutions accounted for in the PAM framework?
  - How are the PAM log odds substitution matrices derived from the Dayhoff Markov model transition matrices?
  - The transition matrices are not symmetric. The substitution matrices are symmetric. What is the biological intuition associated with these observations?

- BLOSUM matrices
  - What are the properties of the data that the Henikoffs used to obtain amino acid pair counts for the BLOSUM matrices?
  - Partitioning sequences into clusters based on percent identity is a key aspect of the BLOSUM method.
    - How are the clusters used in the process of counting amino acid pairs?
    - How does the use of clusters account for sample bias?
    - How does the use of clusters lead to a family of matrices parameterized by divergence?

- Log odds substitution matrices: Both the PAM and BLOSUM substitution matrices are log-odds matrices. You should understand and be able to work with the log odds substitution matrix framework.
  - When a log odds substitution matrix is used to score an alignment, the score of the alignment also corresponds to a log likelihood ratio; what does this mean?
  - How should a positive element in a substitution matrix be interpreted in this context?
  - How should a negative element in a substitution matrix be interpreted in this context?
  - When comparing the main diagonal elements of matrices representing different amounts of evolutionary divergence, what trends would you expect to see?
  - When comparing the off-diagonal elements of matrices representing different amounts of evolutionary divergence, what trends would you expect to see?

- What are the similarities and differences between the PAM and BLOSUM matrices?
  - What are the major differences between the data used for the BLOSUM matrices and the data used for the PAM matrices?
  - What are the major differences in how sequence divergence is represented in the BLOSUM matrices compared to the PAM matrices?
  - You should be able to rank levels of sequence divergence in the two models.

- What are the similarities and differences between models of DNA sequence evolution and amino acid substitutions matrices? What is the relationship between
  - the PAM and BLOSUM models/matrices?
  - the Jukes Cantor and PAM models?
  - the Jukes Cantor, Kimura 2 Parameter, Felsenstein, and HKY models?

Modeling Motifs and Patterns

- There are three major problems to solve in motif analysis. You are responsible for understanding this overarching framework. You are not responsible for the details of the HMM recognition and discovery algorithms on Exam 2.
  - Discovery: Given unlabeled sequences that share a conserved pattern or motif, discover the motif using unsupervised learning.
  - Modeling: Given labeled sequences that share a conserved pattern or motif, construct an abstract model that represents the frequencies of residues observed in the pattern.
  - Recognition: Given an abstract model of a motif and an unlabeled sequence, use the model to determine whether the unlabeled sequence contains the motif and/or predict the location of the motif in that sequence.
- Two major modeling approaches: Position Specific Scoring Matrices (PSSMs) and Hidden Markov models (HMMs).
  - PSSMs
    - Appropriate for ungapped, conserved motifs of fixed length, such as transcription factor binding sites.
    - Cannot model indels, variable length patterns, or positional dependences.
  - HMMs
    - Appropriate for modeling conserved motifs, as well as patterns in sequence composition, such as hydrophobic transmembrane regions.
    - Can model variable length patterns and positional dependences.

Position Specific Scoring Matrices and the Gibbs sampler

- Position specific scoring matrices (PSSMs)
  - A formalism for modeling ungapped multiple alignments
  - You should be familiar with each step in the calculation of a PSSM from an alignment, including the calculation of the
    1. Count matrix
    2. Frequency matrix
    3. Propensity matrix
    4. Log odds scoring matrix
  - Pseudocounts
    - What is a pseudocount?
    - What is the rationale for using pseudocounts?
    - Understand how to construct a PSSM using pseudocounts.
  - Recognition with PSSMs: You should know how to use a PSSM to score a sliding window in an unlabeled sequence to find new instances of the motif.
  - The score of a sequence segment is analogous to a log likelihood ratio. You should understand why this is true. What are the alternate and null hypotheses represented by this likelihood ratio?
  - How are PSSMs similar to amino acid substitution matrices? How do they differ from amino acid substitution matrices?

- The Gibbs sampler
  - In the context of biomolecular sequence analysis, the Gibbs sampler is a motif discovery method that uses the PSSM formalism as its fundamental data structure.
  - The Gibbs sampler simulates the stationary distribution of a Markov chain. You should have a basic understanding of this Markov chain. What do the states represent? How are states connected?
  - You should understand the basic structure of the Gibbs sampler algorithm.
  - The Gibbs sampler is guaranteed to find a globally optimal solution. What feature of the algorithm keeps it from getting trapped in local optima?
  - Even though the Gibbs sampler algorithm is guaranteed to converge to a global optimum, running the algorithm several times with different starting configurations is recommended. What is the rationale for this?
  - What is a probability mass function (pmf)? What is a cumulative mass function (cmf)? You should be able to calculate a cmf from a pmf.
  - You should know how to generate random numbers according to an arbitrary probability distribution, given the cmf of that distribution.

  - What are the underlying assumptions of the Gibbs sampler for biomolecular motif discovery? In what ways are they unrealistic?
  - What implementation decisions must the user make in order to apply the Gibbs sampler to a particular motif discovery problem?

- Limitations of PSSMs
  - You should understand the following limitations of PSSMs and be able to explain how these limitations result from the way in which PSSMs are defined.
    - PSSMs cannot model positional dependencies.
    - PSSMs are not well suited to modeling variable length patterns.
    - PSSMs cannot recognize pattern instances containing insertions or deletions.
    - Boundary detection: PSSMs are not well suited to determining the precise location of boundaries between distinct biological regions. Examples of such boundaries include the first membrane-bound amino acid in a transmembrane region, the first nucleotide in a binding site, the beginning of a gene, etc.

Hidden Markov models

- Definitions and terminology
  - A Hidden Markov model (HMM) has the following components:
    1. N states $E_1 \dots E_N$
    2. An alphabet, $\Sigma = \{\sigma_1, \sigma_2 \dots \sigma_M\}$
    3. Parameters, λ:
       (a) Initial state probability distribution vector $\pi = (\pi_i)$
       (b) Transition probability matrix a_{ij}
       (c) Emission probabilities: $e_i(\sigma)$ is the probability that state E_i emits $\sigma \in \sum$
  - An HMM is a generative model that emits a sequence $O = O_1, O_2, \dots O_T$ while passing through a sequence of states $Q = q_1, q_2, \dots q_T$. We refer to the sequence of states that emitted O as the "state path".
  - If multiple sequences are under consideration we use superscripts to distinguish them: $O^1, O^2, \dots O^k$, where $O^d = O^d_1, O^d_2, \dots O^d_{T_d}$. Similarly, multiple state paths are denoted $Q^1, Q^2, \dots$, where $Q^b = q^b_1, q^b_2, \dots q^b_{T_b}$.
  - Given a sequence $O = O_1, O_2, \dots O_T$ and a state path $Q = q_1, q_2, \dots q_T$, the joint probability of visiting the states in Q and emitting O is

$$P(O, Q|\lambda) = \pi_{q_1} \cdot e_{q_1}(O_1) \cdot a_{q_1 q_2} e_{q_2}(O_2) \cdot a_{q_1 q_2} \cdot e_{q_3}(O_3) \dots a_{q_{(T-1)} q_T} e_{q_T}(O_T).$$

  - The total probability that O was emitted by a given HMM, with parameters λ, is

$$P(O) = \sum_b P(O|Q^b, \lambda) \cdot P(Q^b|\lambda) = \sum_b P(O, Q^b|\lambda).$$

  - The sum of $P(O, Q|\lambda)$, over all sequences in Σ^* and all state paths is one:

$$\sum_d \sum_b P(O^d, Q^b) = 1.$$

  - What is meant by the "parameters" of an HMM?
  - What does λ usually refer to in HMM terminology?
  - What is "hidden" in a Hidden Markov model?

- Hidden Markov models (HMMs) are an extension of Markov chains.
  - What properties do HMMs have in common with Markov chains?
  - What features are unique to HMMs?
  - What are the advantages of using an HMM, compared to a Markov chain?

- HMM Design
  - Overview
    - HMM design involves two major tasks:
      1. designing the model topology and
      2. estimating the parameters.
    - If the pattern of interest is unknown, then parameter estimation also involves motif discovery.
    - HMM design involves a trade-off between model complexity, on the one hand, and overfitting and multiple local optima, on the other. More expressive models with more parameters can capture more complex biological phenomena, but require larger training sets to obtain accurate estimates of the parameters without overfitting.
  - HMM topology
    - The HMM topology is specified by the states, $E_1, \ldots, E_N$ and the transitions between them.
    - The state connectivity is specified by designating certain transitions to have zero probability, typically to reflect boundary conditions in the biological system that the model is intended to represent. For example, in the transmembrane model, $a_{CE} \equiv 0$, because a protein cannot jump from the cytosol to the extracellular matrix without passing through the membrane.
    - One could define the model to be fully connected and allow the parameter estimation process to discover which transitions have zero probability, but this is not done in practice. What are the disadvantages of that approach?
    - Alphabet of emitted symbols (Σ): For biomolecular sequences, the alphabet will typically be $\{A, C, G, T\}$ or the twenty amino acids. However, sometimes it is convenient to use a reduced alphabet. For example, amino acid sequences can be recoded

using a two letter alphabet, $\{H, L\}$, where H designates a hydrophobic amino acid and L designates a hydrophilic amino acid. A smaller alphabet reduces the number of emission probabilities to be inferred.

- Parameter estimation
    - Once the alphabet, states, and state connectivity have been chosen, the parameters of an HMM are estimated from training sequences, $O^1, O^2, ..., O^k$.
    - If the sequences are labeled, the transition and emission probabilities can be estimated from the observed transition and emission frequencies. If the sequences are unlabeled, we must first discover the conserved pattern using unsupervised learning.
    - Labeled sequences
        - If the sequences are labeled, the parameters are estimated by counting, for each state, the number of emissions and transitions observed in the data.
        - This is a form of maximum likelihood estimation (MLE).
        - You should understand the equations for estimating the initial, emission, and transition probabilities from labeled data and be able to apply them.
        - Pseudocounts can be used to account for emissions or transitions that are not observed in the training sequences. You should know how to incorporate pseudocounts in the estimation of both emission probabilities and transition probabilities.