Study guide

This study guide is intended to help you to review for exams. 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 your class notes, the notes and readings on the syllabus, and your homework assignments.

Pairwise sequence alignment

• Terminology: Alphabet, sequence, string, subsequence, substring.

• Dynamic programming algorithms for local, global and semiglobal alignment.
  – Be familiar with the basic components of these algorithms: initialization, recursion, optimal score, traceback. What is the computational complexity of alignment with dynamic programing?
  – How do the basic algorithmic components differ for local, global and semiglobal alignment? What types of scoring functions are (un)suitable for each of these? Do any of the three types of alignment impose more restrictive criteria on the scoring function used? If so, what is the rationale for these criteria?

• Scoring functions
  – Edit distance. What are the required properties of distance functions for sequence alignment?
  – Similarity scoring. What are the required properties of simple similarity functions for sequence alignment?
  – You should be able to explain how changing a scoring function will influence the nature of optimal alignments obtained with respect to that scoring function.

• Applications: Given a particular sequence analysis scenario (e.g., sequence assembly, identifying introns, etc.), you should be able to state which alignment algorithm is most appropriate and why.

Markov chains

• Definitions and terminology
  – States, the state probability distribution at time \( t \), the initial state probability distribution.
  – The transition probability matrix. What requirements must a matrix satisfy to be a valid transition probability matrix?
  – What is the Markov property?
• Absorbing states, reflecting states, periodic states.

• We discussed finite-state, discrete-time, time-homogeneous Markov chains. You should understand each of these terms.

• Higher-order Markov chains: Given a transition matrix for 1 time step, you should understand how to construct a transition matrix for n time steps.

• Stationary state distributions. What is the formal definition of a stationary distribution? How can you verify that a given distribution is the stationary distribution? What Markov chain properties are required for a Markov chain to have a (unique) stationary distribution? What properties prevent a Markov chain from having a stationary distribution?

• What is the time reversibility property and how do you test whether a Markov chain is time reversible? Why is this property important for working with sequence evolution models?

• Simple random walk models. What are they? How are they related to sequence analysis?

Markov models of nucleotide substitution

• What kinds of questions can be answered with sequence evolution models?

• The Jukes Cantor (JC) model
  – What is the basic structure of the JC model (i.e., states, transition matrix)? What are the underlying assumptions?
  – What is the stationary distribution of the Jukes Cantor model?
  – How is the rate parameter of the JC model related to the overall substitution rate?
  – The Jukes Cantor transition matrix gives the probability of a substitution occurring in a single time step. From this, we derived
    * the probability that nucleotide $x$ at a given site has changed to nucleotide $y$ after elapsed time, $\Delta t$, as well as the probability of observing the same nucleotide at a given site after elapsed time, $\Delta t$;
    * the probability of a mismatch at a given site in sequences that are separated by $\Delta t$,
    * the expected number of substitutions that occurred since the divergence of a pair of present-day sequences, given the number of mismatches in their alignment.

You should understand each of these quantities and know how to apply them in simple scenarios.

• More complex models of nucleotide substitution
  – Non-uniform transition probabilities
The Kimura 2 parameter (K2P) assumes that transitions and transversions occur at different rates. (What are transitions and transversions?) Like JC, the K2P model also has a uniform stationary distribution. Expressions for the probability of observing a transition, a transversion, or no change following elapsed time $\Delta t$ are given in the class notes. An expression for the expected number of substitutions as a function of the number of observed transitions and transversions is also given.

Models with three or more rates are also possible. The maximum is six, one for each nucleotide pair.

- Non-uniform stationary distributions
  - Both the JC and the K2P models have uniform stationary distributions. This distribution is an implicit consequence of the symmetric structure of the transition matrices of these models.
  - In contrast, the Felsenstein model assumes all substitutions proceed at the same rate, but allows for different underlying base frequencies. How is the transition matrix in the Felsenstein model modified to achieve this?

  - The Hasegawa, Kishino, Yano (HKY) model combines the innovations of the K2P and Felsenstein models to give a matrix that has separate rates for transitions and transversions and allows for non-uniform base frequencies.

  - More complex models allow three or more rates. The most complex of the models within this framework is the General Time Reversible (GTR) model. The GTR allows for a different rate for each of the six possible substitutions and an arbitrary stationary distribution.

  - Given a set of non-uniform base frequencies and a transition matrix that implies uniform base frequencies, can you construct a new model that has the same rate structure as the transition matrix, but with the specified set of non-uniform base frequencies?

- Limitations: Properties of sequence evolution that are not captured by the models we learned in class include interactions between different sites in the same sequence, different rates at different sites (site-dependent rate variation) and changes in rate over time (time-dependent rate variation).

Amino acid substitution models and matrices

- Deriving amino acid substitution matrices: overview
  - Substitution models should reflect biophysical properties. Residues with similar properties represent conservative replacements and should have higher similarity scores than residues with different properties that represent non-conservative replacements.
  - Substitution matrices should be parameterized by evolutionary divergence.
  - Given the greater number and variety of the amino acids, compared with nucleotides, amino acid substitution models rely more heavily on learning parameters from data than nucleotide models.
We considered 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 sample bias.
3. Estimate substitution frequencies from amino acid pair counts.
4. Construct a log odds scoring matrix from substitution frequencies.

**The PAM model**

- Dayhoff’s PAM matrices are derived from a Markov model of amino acid replacement. What is the basic structure of this model?
- The unit of divergence used is the PAM or “percent accepted mutation”. How is the PAM defined?
- 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 Markov chain strategy that she used?
- How did Dayhoff derive counts from that data set and how did she account for potential sampling 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 1 PAM divergence?
- How is the PAM-n model derived from the PAM-1 model?
- How are multiple substitutions accounted for in the PAM framework?

**The PAM substitution matrices**

- 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 this observation?

**BLOSUM matrices**

- What are the properties of the data that the Henikoffs used to obtain amino acid pair counts for the BLOSUM matrices? What are the major differences between the data used for the BLOSUM matrices and the data used for the PAM matrices?
- Partitioning sequences into clusters based on percent identity is a key aspect of the BLOSUM method.
* Broadly speaking, 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 alignment score 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 levels of divergence, what trends would you expect to see?
  – When comparing the off-diagonal elements of matrices representing different levels of divergence, what trends would you expect to see?

• What are the similarities and differences
  – between the Jukes Cantor, Kimura 2 Parameter, and Felsenstein models?
  – between the Jukes Cantor and PAM models?
  – between the PAM and BLOSUM models/matrices?

BLAST

• Karlin Altschul statistics
  – You should understand the equation
    \[ E = K m n e^{-\lambda S} \] (1)
    and be able to explain each of the variables in the equation. How does \( E \) vary if one of the independent variables increases (or decreases)? How does this make sense in terms of the behavior of a data base search?
  – You should understand the equation
    \[ E = m n 2^{-S_b} \]
    and be able to explain each of the variables in the equation. How is this equation related to equation 1?
  – What are bit scores? What are raw scores? How are they related?
What is an E value? How does it differ from a p value?
What is meant by a “random sequence” in this context?
Karlin Altschul statistics provide an estimate of the probability of observing a test statistic under a null hypothesis. What is this null hypothesis? What is the alternate hypothesis?
Karlin Altschul statistics were derived based on the assumption that the scoring matrix satisfies certain criteria. What are those criteria?
The BLAST output includes “effective lengths”. What does this mean?
What are target frequencies?

The BLAST heuristic
You should understand the role of each of the BLAST parameters and how the parameters influence the performance of the heuristic.
What is a “hit”? How were hits found in the 1990 BLAST heuristic?
How would increasing or decreasing $w$, $T$, $A$, or the E value threshold influence each of the following?
* the speed of the heuristic
* the number of false negatives
* the number of false positives
What problems were Gapped Blast and Two-Hit Blast designed to address?
The 1990 version of BLAST did not consider alignments with gaps. What are the pros and cons of including gaps in the model? Consider running time, the sensitivity of the search, and the statistical model.

Information theoretic aspects of BLAST
What is the relative entropy of a matrix?
What are target frequencies?
How is the evolutionary divergence of a matrix related to the evolutionary divergence between a query sequence and the matches retrieved during a data base search?
Which matrix will give the best discrimination between true and false positives? What is meant by true and false positives in this context?
What happens if you don’t use this “ideal” matrix? How could you work around this problem?
What is the relationship between the length of the query matrix and the scoring matrix used?
You should be able to calculate the minimum information needed to retrieve meaningful matches.
– How much information is there in an alignment?
– How is scoring an alignment like flipping a coin? How is scoring an alignment like a random walk?

Local multiple sequence alignment

• Overview
  – Applications of these methods
  – Three major problems to solve
    * Discovery
    * Modeling
    * Recognition

• 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:
    1. Frequency matrix
    2. Propensity matrix
    3. Log odds scoring matrix
  – Pseudocounts
    * What are they?
    * What is the rationale for using pseudocounts?
    * How to construct a PSSM using pseudocounts.
  – Recognition with PSSMs: You should know how to use a PSSM to score each position in an unlabeled sequence to find new instances of the motif?
  – The score of a window is analogous to a log likelihood ratio. You should understand why this is true. What is the alternate hypothesis? What is the null hypothesis?
  – How are PSSMs like amino acid substitution matrices? How do they differ from amino acid substitution matrices?
  – PSSMs are suitable for modeling some kinds of biomolecular motifs. What are the properties of those motifs?
  – What are the limitations of PSSMs? What motif features are not captured by the PSSM formalism?

• The Gibbs sampler
  – In the context of biomolecular sequence analysis, the Gibbs sampler is a motif discovery method based on the PSSM formalism.
- The Gibbs sampler simulates the stationary distribution of a Markov chain. You should have a basic understanding of this Markov chain: What are the states? How are they 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 it is guaranteed to converge to a globally optimal, running the algorithm several times with different starting configurations is recommended. What is the rationale for this?
- What is a probability density function?
- What is a cumulative density function? You should be able to calculate a cdf from a pdf?
- You should know how to generate random numbers according to an arbitrary probability 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 discover problem?