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 the notes you took in class, the notes and readings on the syllabus, and your homework assignments.

- Applications of DNA substitution models
  - 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, Δt, as well as the probability of observing the same nucleotide at a given site after elapsed time, Δt;
    - the probability of a mismatch at a given site in sequences that are separated by Δt,
    - 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.
  - For the Kimura 2 Parameter model, expressions for the probability of observing a transition, a transversion, or no change following elapsed time Δt 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.

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

Amino acid substitution models and matrices

- Deriving amino acid substitution matrices: overview
  - Substitution models 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.
  - Substitution matrices should be parameterized by evolutionary divergence.
  - Amino acid substitution matrices implicitly correct for multiple substitutions.
  - Given the greater number and variety of the amino acids, compared with nucleotides, amino acid substitution models rely more heavily on learning parameters from training 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
  - The Dayhoff Markov model of amino acid replacement.
    - 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 one 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?
  - The PAM substitution matrices
    - How are the PAM 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 each of 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? 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.
    - 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.
  - What is a likelihood ratio? Frequently we take the logarithm of a likelihood ratio. What are two advantages of using log likelihood ratios (i.e., log odds ratios), as opposed to likelihood ratios, alone?
  - 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

- 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 was the Blast heuristic modified in 1997? What problems were Gapped Blast and Two-Hit Blast designed to address?
  - How would increasing or decreasing w, T, A, or the reporting threshold influence each of the following?
    - the speed of the heuristic
    - the number of false negatives
    - the number of false positives
  - 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.

- Karlin Altschul statistics
  - You should understand the equation

$$E = Km'n'e^{-\lambda \mathcal{S}} \tag{1}$$

    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 makes sense in terms of the behavior of a database search?
  - You should understand the equation

$$E = m'n'2^{-\mathcal{S}_b}$$

    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 "effective lengths" of the query and the database and why are effective lengths used instead of actual lengths in Karlin Altschul statistics?
  - What is meant by a "random sequence" in this context?
  - Karlin Altschul statistics provide an estimate of a the probability of observing a test statistic under a null hypothesis. What 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?
- 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 database 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 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?