Mid-Semester Study guide

October 6, 2022

This study guide is intended to help you to review for the midterm 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.

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
  - Similarity scoring. What are the required properties of simple similarity functions for sequence alignment? Which alignment problems can be solved with similarity scoring and which cannot? Why or why not?
  - What is edit distance? How does distance scoring differ from similarity scoring? Which alignment problems can be solved with edit distance and which cannot? Why or why not?
  - 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 type of alignment 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.
- n-step transitions in 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 calculate the stationary distribution of a Markov chain?
  - How can you verify that a given distribution is the stationary distribution?
  - What properties may prevent a Markov chain from having a stationary distribution?
  - What properties are required for a Markov chain to have a unique stationary distribution?

Markov models of nucleotide substitution

- What kinds of questions can be answered with sequence evolution models?
- What is the basic structure of a Markov model of DNA substitution?
  - States?
  - Meaning of transitions between states?
  - Underlying assumptions?
- The Jukes Cantor (JC) model
  - What are the underlying assumptions?
  - How are transitions modeled?
  - What is the stationary distribution?

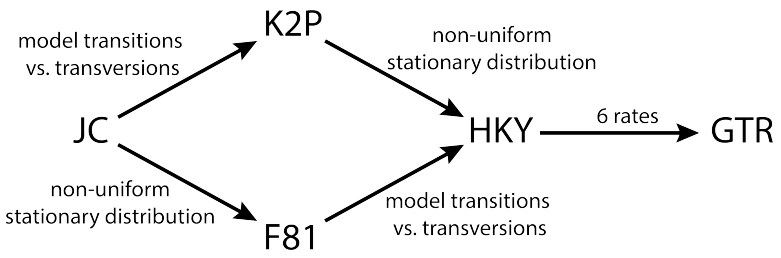
  - How is the rate parameter of the JC model related to the overall substitution rate?
  - The Jukes Cantor transition matrix gives the probability of a substitution occurring in a single time step. From this, we derived
    * the probability that nucleotide x at a given site has changed to nucleotide y after elapsed time, t, as well as the probability of observing the same nucleotide at a given site after elapsed time, t;
    * the probability of a mismatch at a given site in sequences that have been diverging independently from a common ancestor for time t;
    * the expected number of substitutions that occurred since the divergence of a pair of present-day sequences, given the number of mismatches observed in their alignment.

    You should understand each of these quantities and know how to apply them in simple scenarios. For the exam, you do not need to know how to derive these quantities.

- The Kimura 2 parameter (K2P) model
  - Kimura's model of DNA substitution distinguishes between are transitions and transversions. What are transitions and transversions? The word "transition" is also used to describe the progression from one state to another state in a Markov chain. It is also used to describe a class of nucleic acid substitions. You should understand both of these models.
  - What are the underlying assumptions?
  - What are the parameters of the model?
  - What is the stationary distribution of this model?
  - The K2P transition matrix gives the probability of a substitution occurring in a single time step. From this, we derived expressions for
    * the probability that a transition or transversion has changed nucleotide x at a given site after elapsed time, t, as well as the probability of observing the same nucleotide at a given site after elapsed time, t;
    * the probability of observing a mismatch, where the two nucleotides are transitions or transversions, at a given site in sequences that have been diverging independently from a common ancestor for time t.
    * the expected number of substitutions of each type that occurred since the divergence of a pair of present-day sequences, as a function of the number of observed transitions and transversions.

    You should understand each of these quantities and know how to apply them in simple scenarios. For the exam, you do not need to know how to derive these quantities.

- The DNA substitution model hierarchy: We discussed a hiearchy of increasingly complex models of DNA sequence evolution. In addition to the JC and K2P models, which we discussed in detail, we considered the Felsenstein (F81) model, the Hasegawa, Kishino, Yano (HKY) model, and the General Time Reversible (GTR) model.



- For each of the five models you should understand
  - What are the underlying assumptions of the model?
  - How many parameters does the model have? What do those parameters represent?
  - What is the meaning of transitions between states in the model?
  - What is the stationary distribution of the model?
- How are the different models related?
  - Non-uniform *transition probabilities*
    - The K2P, HKY, and GTR models all allow for different rates. The K2P and HKY models distinguish between transitions and transversions. The GTR model allows for a different substitution rate for each of the six possible pairs of nucleotides (rates are the same in both directions, i.e., A to G and G to A proceed at the same rate).
    - Both the JC and F81 models assume all substitutions proceed at the same rate.
  - 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.
    - The F81, HKY, and GTR models allow for different underlying base frequencies.
  - Transforming one model into another.
    - In contrast to the JC model, 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 HKY model combines the innovations of the K2P and Felsenstein models to give a matrix that has different 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 GTR model. The GTR allows for a different substitution rate for each of the six possible pairs of nucleotides and an arbitrary stationary distribution.
    - Given an instance of the JC model and a set of non-uniform base frequencies, could you turn it into an instance of the Felsenstein model?
    - More generally, 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 original transition matrix, but with the specified set of non-uniform base frequencies?
  - How do models compare in terms of complexity?
  - How can you decide which model to use?
  - Given the transition matrix for a nucleic acid substitution model, can determine which of the five models the matrix represents?

- 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,
    - insertions and deletions,
    - site-dependent rate variation (different rates at different sites), and
    - time-dependent rate variation (changes in rate over time).
  - What are the trade-offs associated with using a more complex models versus a less complex model?

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. 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?
- Why are the advantages of using log likelihood ratio, instead of simply the likelihood ratio?
- How is the log likelihood ratio used to constructing 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?