Study guide

October 2, 2019

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.

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 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?
  - Under what properties is the stationary distribution the limiting distribution?
- What is the time reversibility property?
  - 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?

Markov models of nucleotide substitution

- What kinds of questions can be answered with sequence evolution models?
- What are transitions and transversions?
- What is the basic structure of a DNA substitution model?
  - States,
  - Meaning of transitions?
  - 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
  - What are the underlying assumptions?
  - How are transitions modeled?
  - What is the stationary distribution?
  - The K2P transition matrix gives the probability of a substitution occurring in a single time step. From this, we derived
    - Expressions for the probability of observing a transition, a transversion, or no change after time t has elapsed. (Given in the class notes.)
    - An expression for the expected number of substitutions of each type as a function of the number of observed transitions and transversions.

- The Felsenstein (F81) model
  - What are the underlying assumptions?
  - How are transitions modeled?
  - What is the stationary distribution?
- The Hasegawa, Kishino, Yano (HKY) model
  - What are the underlying assumptions?
  - How are transitions modeled?
  - What is the stationary distribution?
- The General Time Reversible (GTR) model.
  - What are the underlying assumptions?
  - How are transitions modeled?
  - What is the stationary distribution?
- How are the different models related?
  - Which use more complex models of nucleotide substitution? (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.
  - Which use more complex models with 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.
  - Which models are equivalent under certain conditions? How can you derive from one model to 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?

- 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,
    - modelling insertions and deletions,
    - different rates at different sites (site-dependent rate variation), and
    - changes in rate over time (time-dependent rate variation).
  - Pitfalls of using more and less complex models.