Reading for Lecture 13

Release v10

Christopher Lee

October 03, 2015

Contents

Read Jones & Pevzner 10.1-10.6, and the following material.

1 Evolutionary Trees

One of the greatest success stories in science is how the theory of evolution has been confirmed and elaborated in almost incredible detail by modern genome sequencing. Data that biologists in the past could barely have even imagined, are now available for download on the Internet with a single click. It will take decades for science and medicine to absorb all the lessons and new capabilities that these data are beginning to make possible.

The funny thing about evolution is its conservatism: it makes changes very slowly, and it takes eons to genuinely "forget" anything (in the sense of all traces of that feature completely disappearing from the genome). For example, even though the human and mouse genomes diverged approximately 100 million years ago, coding regions of the human and mouse genomes are about 87% identical at the nucleotide level. Thus, about nine-tenths of the coding sequence is unchanged from its proto-mammalian ancestor 100 million years ago. Furthermore, there are many other separate mammal lineages (e.g. dog, cow, bear, cat etc.) and more distant relatives (e.g. marsupials, reptiles, birds), all of which contribute information for infering the sequence of the mammalian ancestor. Although no one is trying to recreate a dinosaur a la Jurassic Park, researchers will undoubtedly be working out in considerable detail what the genome of the mammalian ancestor looked like (and thus learning a lot about how that animal lived). Closer to home, the DNA of the human population can be thought of as millions of detailed individual recordings of the history of humanity, including who invaded who, who went where when, etc. etc. In effect, the history of life is written in detail

in modern genomes; now that we can read genomes at will, quickly and cheaply, that book is open for everyone to read.

As an introduction to this huge and fascinating subject, we begin by exploring some basic models of how sequences evolve.

1.1 Evolution as a Markov Process

The entire information for producing the next generation of an animal or plant passes through a tiny "eye of the needle" indeed: a microscopically small egg cell, and an even tinier sperm cell. This represents the complete "state" that the "daughter" animal or plant depends on (i.e. the *genetic* component of its phenotype as opposed to the *environmental* component). By definition this is a Markov process: the "next variable" depends only on the "previous variable". Extending this Markov chain to greater and greater lengths (e.g. 1 million years corresponds to about 50,000 human generations) does not alter the fact that it is a Markov process. Thus we model sequence evolution as a Markov chain, in the following simple ways:

- *descent from a common ancestor* is modeled as a graph structure whose nodes represent individual sequences, and whose edges represent some period of evolutionary change between a pair of sequences (i.e. a set of "edit operations" or mutation events that transform one sequence into the other). This graph structure represents the information graph for the set of sequences.
- The simplest models assume that each sequence depends on only a single predecessor sequence (a simple first-order Markov chain) with only a few allowed edit operators (e.g. single letter substitution / insertion / deletion). Note that more complex models are possible (and in some cases necessary); for example, in animals with two copies of each chromosome, *recombination* can occur between the two copies to produce a hybrid or "chimeric" sequence that contains pieces of *both* ancestral copies. We will restrict ourselves to the simple first-order model here. In this case, the graph structure cannot contain any cycles, so it is a *tree*.
- For convenience (and mostly without loss of generality), we typically assume that each sequence has at most two descendants, which restricts the graph structure to a *binary tree*. This assumption is in a sense fully general, in that it can actually represent the case where one sequence has more than two descendants, by simply creating a linear chain of "copies" of it in the form of a chain of descendants connected to it through zero-length edges (i.e. zero mutational change).
- Leaf nodes of the tree represent *modern sequences* (which we can directly observe), whereas internal nodes represent *ancestral sequences* (which ordinarily we cannot directly observe, but instead can only infer).

1.2 Rooted vs. Unrooted Trees

- If we draw the tree using *directed edges* it explicitly represents a specific temporal sequence (i.e. each directed edge points from an earlier sequence (ancestor) to a later, child sequence). In this case, for a specific set of sequences $s_1, s_2, ...$, the *closest* sequence in the tree that has a directed path to all of the designated sequences is referred to as their *most recent common ancestor* (MRCA). By definition, the root of the tree is the MRCA of the other sequences in the tree. Because the directed edges define a unique root for the tree, it is referred to as a *rooted tree*.
- If we draw the tree using *undirected edges* it gives no explicit information about the temporal order of the sequences or the location of the root (MRCA). For this reason, such a tree is called an *unrooted tree*. Because inferring the location of the root can be challenging, the most commonly used tree construction algorithms produce unrooted trees.

1.3 Evolutionary Distances

- Each edge in a tree has an associated *branch length*, proportional to the length of evolutionary time represented by that edge. That is, a single mutation likelihood (rate) model is assumed to hold over the entire period represented by that edge, in which case only a single "total time" parameter is needed to predict all the mutation likelihoods associated with that edge.
- For substitution models, the branch length parameter is typically expressed in terms of the *average number of mutation events per site* (which for clearly related sequences is a fraction substantially less than 1). Note that this is not necessarily the same as the average number of *observed* differences per site between a pair of sequences, since two or more mutation events could occur at the *same* site, resulting in only a single observed letter difference. Thus the number of *observed* differences could under-estimate the number of actual mutation events that occurred.
- *The "Molecular Clock" Hypothesis*: one important class of models assumes that the rate of mutation μ is constant over all time periods, i.e. that a single mutation rate model applies at all times.
  - In this case, the expected number of mutation events per site on each edge is just equal to μt where t is time-length of that edge.
  - Furthermore, since the rate is constant over any sequence of consecutive edges $t_1, t_2, ...$, the expected number of mutations per site over that sequence of edges is just $\mu(t_1 + t_2 + ...)$ i.e. just proportional to the sum of the time-lengths of those edges.
  - *Reversibility*: if the mutation rate matrix is time-reversible, i.e. it obeys the *detailed balance* equation, then the above assertion holds true regardless of whether some (or all) of the specified edges are traversed in forward or backward orientation. In other words, the only thing that affects the distance between any two sequences is just the sum of the time-lengths on the edges that connect them, regardless of the orientations of those edges.
- *Ultrametric distances*: this assumption is especially meaningful on *rooted trees*, because by definition the total time from an MRCA to any of its modern descendants must be the same. This implies that the distance metric between any pair of sequences should depend only on how long ago they diverged (i.e. the length of time since their most recent common ancestor (MRCA)).

2 Continuous Time Markov Chains

Evolution introduces one change compared with our previous study of discrete Markov processes: evolution takes place over continuous time, so it is appropriate to model it using a *continuous time Markov process*. We define the Markov property in continuous time as follows. Just as a discrete first-order Markov chain obeys the joint probability rule that each variable X_t is conditionally independent of the starting variable X_0 given any intermediate variable X_s where $0 < s < t$, the same holds true for a continuous-time Markov chain:

$$p(X_t|X_0, X_s) = p(X_t|X_s)$$

We now consider the practical aspects of how to use and compute continuous time transition probabilities. We first define the unit-time transition matrix as

$$T^1 = e^{\Lambda}$$

and the time-dependent transition matrix for a time interval t as

$$T^t = e^{\Lambda t}$$

where $e^{\Lambda t}$ is the *matrix exponential*

$$e^{\Lambda t} = I + t\Lambda + \frac{t^2\Lambda^2}{2!} + \frac{t^3\Lambda^3}{3!} + ...$$

Note that t is a scalar value, I is the identity matrix, and Λ is the *instantaneous transition rate matrix* representing the rate of flow per unit time from each possible state to the other states. For example, for a two-state system that starts with $p(\theta_t = s_1|t = 0) = 1, p(\theta_t = s_2|t = 0) = 0$, the rate of growth of the probability of state s_2 at time $t = 0$ will be

$$\frac{dp(\theta_t = s_2)}{dt} = \lambda_{12}$$

where λ_{12} is matrix element $i = 1, j = 2$ of Λ. Note also that raising a matrix to an integer power (e.g. Λ^k) simply means multiplying that matrix by itself (using the standard definition of matrix multiplication) that number of times (i.e. $\Lambda^2 = \Lambda \times \Lambda$).

Note that in general, the unit-time transition matrix and the instantaneous transition rate matrix are not the same thing, i.e. $T^1 \neq \Lambda$. This is because T^1 takes into account the possibility that *multiple* mutation events could occur, whereas Λ represents the probabilities associated with a *single* mutation event. However, if the off-diagonal elements of matrix Λ become very small, i.e. $\lambda_{ij} \to 0$, then the two become approximately equal, i.e. $T^1 \to \Lambda$ (because the probability of multiple mutation events becomes insignificant).

The main consequence of the matrix exponential form of the continuous time transition matrix is that we expect the rows of T^t to converge to the stationary distribution $\vec{\pi}$ via an *exponential decay* process. Recall that for a discrete Markov process with transition matrix T with non-zero elements, the transition matrix T^k after k time steps also converges such that all of its rows converge to the stationary distribution $\vec{\pi}$. The continuous time Markov process follows the same behavior, but gives a smooth convergence trajectory across all real-numbered values of $t \geq 0$. Concretely, let $\tau_{ij}(t)$ be matrix element i, j of T^t.

- at time $t = 0$, the transition matrix $T^0 = I$ is just the identity matrix, i.e. its diagonal elements are $\tau_{ii}(0) = 1$ and all other elements are zero.
- as $t \to \infty$, $\tau_{ij}(t) \to \pi_j$.
- this convergence should follow an exponential decay of the general form $e^{-\lambda t}$.

3 The Jukes-Cantor Model of Nucleotide Evolution

3.1 Assumptions

As a simple example of this exponential decay process, we consider the Jukes-Cantor model, which makes the following simplifying assumptions:

- the transition rate λ_{ij} depends only on the *target* nucleotide j, not the source nucleotide i.
- specifically it is assumed to be just proportional to the stationary distribution for that nucleotide, times a scalar rate parameter λ,

  $$\lambda_{ij} = \lambda \pi_j$$

  for $i \neq j$, and consequently the rate of decrease for state i is just

  $$\lambda_{ii} = -\sum_{j \neq i} \lambda \pi_j = -\lambda(1 - \pi_i)$$

- the stationary distribution is assumed to be equal over all four nucleotides, i.e.

  $$\pi_A = \pi_C = \pi_G = \pi_T = \pi = \frac{1}{4}$$

Then the Jukes-Cantor instantaneous rate matrix is just

$$\Lambda = \lambda \begin{bmatrix} -(1-\pi) & \pi & \pi & \pi \\ \pi & -(1-\pi) & \pi & \pi \\ \pi & \pi & -(1-\pi) & \pi \\ \pi & \pi & \pi & -(1-\pi) \end{bmatrix}$$

and the continuous time transition matrix T^t is given by

$$\tau_{ij}(t) = (1 - e^{-\lambda t})\pi$$

for $i \neq j$, and

$$\tau_{ii}(t) = \pi + e^{-\lambda t}(1-\pi)$$

3.2 Jukes-Cantor Distance Metric

It is convenient to estimate the distance between two sequences in terms of their *time of divergence* t. Specifically, we wish to estimate the total number of mutations per site that have occurred during this time of divergence. Note that under the Jukes-Cantor model this can be treated as a Poisson distribution where mutation events occur at a uniform density $\lambda(1-\pi)$. The expectation value for the total number of events over a time interval t is just $\lambda(1-\pi)t$. We adopt this as our distance metric D. We can estimate this from the *observed* fraction of mutated sites f, which for a large number of independent sites should converge to the expectation value

$$f \to E(f) = \sum_i \pi_i \sum_{j \neq i} \tau_{ij}(t)$$

$$= (1 - e^{-\lambda t})(1-\pi)$$

$$e^{-\lambda t} = 1 - \frac{f}{1-\pi}$$

$$\lambda t = -\log\left(1 - \frac{f}{1-\pi}\right)$$

$$D = \lambda(1-\pi)t = -(1-\pi)\log\left(1 - \frac{f}{1-\pi}\right)$$

Note that for small f (i.e. $f \to 0$), $D \to f$, but as f increases, D becomes much larger than f. (Indeed, as f approaches its maximum possible value $1-\pi$, D diverges to ∞). This is closely analogous to the difference between T^1 vs. Λ: whereas f counts each observed sequence letter difference as if only a *single* mutation event occurred, D takes into account the possibility that *multiple* mutation events occurred at that site.