Models for molecular evolution

How do we model evolution at the sequence level?

We've talked about probabilistic alignment scores, but we'd like to generalize to multiple events (a base going through intermediate states before reaching the final state) and have an explicit dependence on the time since the sequences diverged.

We want a way to answer $\Pr(\text{base} = \text{A} | \text{base} = \text{C}, \text{dt} = 5)$, that is, the probability that a base mutated from C to A in 5 time units.

How does this differ from alignment scores that we've seen? They assumed a fixed time since divergence. They can be considered a special case of these more general models.

Let's consider a discrete-time Markov chain. What is a Markov chain? What is the Markov property?

Roughly, a Markov chain is a random process that depends only on the current state (not all past history). This dependence structure is the Markov property.

Does this modeling assumption make sense for evolution? Note that with this approximation we are only allowing one change per time step.

Consider the state diagram:

![State Diagram](image)

Add transition probabilities, the simplest case being the Jukes-Cantor model (stay with $p=0.7$, change with $p=0.1$ to each of the other nucleotides).

Example:

Start with $\text{C}$: $[0 \ 1 \ 0 \ 0]$.

After one time step, what probability does it have of being each other nucleotide?

Easy: $[0.1 \ 0.7 \ 0.1 \ 0.1]$

What about one more time step?

$[0.16 \ 0.52 \ 0.16 \ 0.16]$ \ (notice that $0.52 = 0.1*0.1*3 + 0.7*0.7*1$)

What about 4 time steps? Harder. $[0.22 \ 0.35 \ 0.22 \ 0.22]$

What about infinite time steps? $[0.25 \ 0.25 \ 0.25 \ 0.25]$
How are we doing these calculations? Let's formalize as a finite discrete-time Markov chain: need a transition matrix.

Set up a transition matrix $Q$.

- e.g. Jukes-Cantor model:

$$
\begin{bmatrix}
1-3u & u & u & u \\
u & 1-3u & u & u \\
u & u & 1-3u & u \\
u & u & u & 1-3u
\end{bmatrix}
$$

Instantiate with $u=0.1$:

Note that this is the transition matrix, not the rate matrix.

$$
\begin{bmatrix}
0.7 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.7 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.7 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.7
\end{bmatrix}
$$

Note that we can simplify this (by symmetry) into two states, staying with the same nucleotide or changing to another nucleotide.

$\Pr(\text{Staying on the same nucleotide in one time step}) = 1-3u$

$\Pr(\text{Change to a different nucleotide in one time step}) = 3u$

Each entry $i,j$ in the transition matrix corresponds to $\Pr(\text{base}=i \mid \text{base}=j)$, letting us form the following expressions with matrix algebra:

$$
\Pr(\text{base}=i) = \sum_{\text{all } j} \Pr(\text{base}=i, \text{base}=j) = \sum_{\text{all } j} \Pr(i|j) \cdot \Pr(j) = \\
\Pr(i|A) \cdot \Pr(A) + \Pr(i|C) \cdot \Pr(C) + \Pr(i|G) \cdot \Pr(G) + \Pr(i|T) \cdot \Pr(T)
$$

Here we used marginalization and Bayes' theorem, as covered in the first recitation, to evaluate a probabilistic expression based on expressions that we know.

We can use matrix algebra to generate the distribution over nucleotides from any starting distribution and any number of time steps. Let $x_i$ be the column vector describing the probability of being in each state (nucleotide value).

$$
x_1 = Q \cdot x_0 \\
x_2 = Q \cdot x_1 \\
\text{thus: } x_2 = Q \cdot Q \cdot x_0 \\
\text{and similarly: } x_t = Q^t \cdot x_0
$$

How do we calculate this? We can diagonalize the transition matrix $Q$. For a complete reference, see a linear algebra textbook. Briefly, we can write the transition matrix $Q$ as:

$$
Q = P \cdot D \cdot P^{-1}
$$

...where $D$ is a diagonal matrix. Then we can calculate matrix powers in closed form because the
P*P^-1 terms cancel and raising a diagonal matrix to a power is easy to calculate.

Now, what if we want time to be continuous? We need a continuous-time Markov process

We need to convert Q from a transition matrix into a rate matrix. We can think about each entry in the transition matrix as 1-3u*t or u*t, which denote the fraction of nucleotides that change from one state to another in a fixed amount of time. Let's take the derivative with respect to t in order to form the rate matrix. Roughly, the rate matrix is simply Q-I (where I is the identity matrix) because we didn't include the t term in our original definition of the rate matrix.

Let X be the column vector describing the distribution over the nucleotides. The rate matrix lets us definite a matrix differential equation:

\[ \frac{dX}{dt} = (Q-I)X \]

...which has a simple solution:

\[ X(t) = e^{((Q-I)t)} * X(0) \]

We will omit the derivation of the specific terms, but note that we can calculate the entries of the matrix exponential in closed form. This gives us:

- \[ \text{Pr(Staying on the same nucleotide, after t time units)} = 0.25 + 0.75 * \exp(-4ut) \]
- \[ \text{Pr(Change to a different nucleotide, after t time units)} = 0.25 – 0.25*\exp(-4ut) \]

Does this have the right properties (in the limits)?

Extensions: the Kimura model allows for different penalties for transitions (A-G, C-T) and transversions. Its rate matrix is:

\[
\begin{bmatrix}
-2b-a & b & a & b \\
b & -2b-a & b & a \\
a & b & -2b-a & b \\
b & a & b & -2b-a
\end{bmatrix}
\]

You can get more complicated (but still closed-form) answers for the three types of entries.

What are the limitations of these models?
- => stationary distributions are equiprobable for each base
- => evolution is uniform (though unclear on how to avoid this)

What are the strengths?
- => quick to compute
- => can calculate for any time or starting nucleotide identity

How do these relate to (empirical) substitution matrices?
What is a key assumption of a particular substitution rate (what does 62 stand for in BLOSUM62)?
Can we use these techniques to change the time using a particular substitution matrix?

Later lectures will show how to use models like this on a phylogeny (on each edge) and perform complicated inference tasks.