BMI/CS 776
Lecture #29:
Insertion/deletion models

Colin Dewey
May 6, 2008
Substitution models

- Substitution models only describe change in state of nucleotide positions.
- How can we model the deletion and insertion of nucleotide positions?

\[ P[X|Z] = P_{AG}(t)P_{CC}(t)P_{TT}(t)P_{AT}(t) \]

\[ P[X|Z] = ? \]
Importance of indel models

- PHMMs are great for statistical alignment
- Are PHMMs a good model of evolution?
  - Emission probabilities: easily derived from evolutionary model (substitution model)
  - Transition probabilities: not so clear (e.g., how do they depend on time?)
TKF91 Model

- Thorne, Kishino, and Felsenstein 1991
  - “An evolutionary model for maximum likelihood alignment of DNA sequences”
- Joint probability distribution on pairs of sequences $(X,Y)$, related via a common ancestor $Z$ \( P(X,Y) \)
- Superposition of two stochastic processes:
  - Substitution (continuous-time Markov model)
  - Insertion and deletion (birth-death process)
Time reversibility

- TKF model is for evolution of ancestral sequence to extant sequence
- With time reversibility, can model two extant sequences by considering one to be the ancestor
Substitution in TKF

- The same as substitution processes we’ve seen before

\[ P(t) = e^{Q\lambda t} \]

- Can use any rate matrix \( Q \)
TKF Insertion-deletion process

- Represent sequences as characters separated by “links”
  
  Sequence: GCAGAT
  
  With links: ≈G~C~A~G~A~T~
  
  Links only: ≈~~~~~~

- Leftmost link (≈) is special: “immortal”

- All other links (~) are “mortal”

- Insertion-deletion modeled by
  
  - links giving “birth” to new links (always immediately to the right of parent link)
  
  - link “deaths” (delete character to the left of link)
Birth-death process

• Continuous time Markov model
• State space: \{0, 1, 2, ... \}
• Transitions:
  • from \(i\) to \(i-1\) (death) with rate \(\mu_i\)
  • from \(i\) to \(i+1\) (birth) with rate \(\lambda_i\)

[Diagram of birth-death process with state transitions labeled accordingly.]
Linear birth-death process

• Rates proportional to state: $\lambda_i = \lambda_i$, $\mu_i = \mu_i$

• In TKF, state is number of nucleotides

  • Links give birth at rate $\lambda$

  • Mortal links die at rate $\mu$

  • If there are $i$ nucleotides, there are $i$ mortal links and $i+1$ total links

  • Rates are very close to linear birth-death process: $\lambda_i = \lambda(i+1)$, $\mu_i = \mu_i$
Length distribution

• Immortal link keeps sequences from dying over infinite length of time

• \( \mu > \lambda \) to keep sequences from going to infinite length

• From detailed balance:
  \( \mu_k q_k = \lambda_{k-1} q_{k-1} \rightarrow \mu k q_k = \lambda_k q_{k-1} \)

• Derive that equilibrium length distribution is
  \( q_k = \left( 1 - \frac{\lambda}{\mu} \right) \left( \frac{\lambda}{\mu} \right)^k \) (geometric)
Animation

(http://biowiki.org/, created by Ian Holmes)
Evolutionary histories

• TKF breaks the evolution of the ancestral sequence to the extant sequence into the histories of individual links in the ancestor

• There are three classes of histories for an ancestral link:

\[
\begin{align*}
  p_n(t) & \quad \sim-\ldots- \\
  & \hspace{1cm} \sim\ldots\sim \\
  p'_n(t) & \quad \sim-\ldots- \\
  & \hspace{1cm} \sim\ldots\sim \\
  p''_n(t) & \quad \approx-\ldots- \\
  & \hspace{1cm} \approx\ldots\approx
\end{align*}
\]

- Link survives, gives rise to $n-1$ new links
- Link gives rise to $n$ new links, then dies
- Immortal link gives rise to $n-1$ new links
TKF birth-death
differential equations

• For $n > 0$

\[
\frac{dp_n(t)}{dt} = \lambda(n-1)p_{n-1}(t) - (\lambda + \mu)np_n(t) + \mu np_{n+1}(t)
\]

\[
\frac{dp_n'(t)}{dt} = \lambda(n-1)p_{n-1}'(t) - (\lambda + \mu)np_n'(t) + \mu(n+1)p_{n+1}'(t) + \mu p_{n+1}(t)
\]

\[
\frac{dp_n''(t)}{dt} = \lambda(n-1)p_{n-1}''(t) - (\lambda n + \mu(n-1))p_n''(t) + \mu np_{n+1}''(t)
\]

• Initial conditions

\[
\frac{dp_0'(t)}{dt} = \mu p_1'(t) + \mu p_1(t)
\]

\[
p_0(t) = p_0''(t) = 0
\]

\[
p_1(0) = p_1''(0) = 1
\]

\[
p_n(0) = p_n''(0) = 0, \quad n > 1
\]

\[
p_n'(0) = 0
\]
Transition probabilities

- Solving the differential equations gives:

\[ p_n(t) = \alpha(t) \beta(t)^{n-1}(1 - \beta(t)) \]

\[ p'_0(t) = (1 - \alpha(t))(1 - \gamma(t)) \]

\[ p'_n(t) = (1 - \alpha(t))\gamma(t)\beta(t)^{n-1}(1 - \beta(t)), \quad n > 0 \]

\[ p''_n(t) = (1 - \beta(t))\beta(t)^{n-1} \]

\[
\alpha(t) = e^{-\mu t}
\]

\[
\beta(t) = \frac{\lambda(1 - e^{(\lambda-\mu)t})}{\mu - \lambda e^{(\lambda-\mu)t}}
\]

\[
\gamma(t) = 1 - \frac{\mu(1 - e^{(\lambda-\mu)t})}{(1 - e^{-\mu t})(\mu - \lambda e^{(\lambda-\mu)t})}
\]
Probability of a history

\[ p''_2(t) \quad p'_0(t) \quad p_1(t) \quad p'_1(t) \quad p_2(t) \]
Probability of an alignment

- Probability combines link fate and substitutions

alignment (h)
- TGT-C-
G-C-ACA

alignment links
≈-～～～～～～～
≈～～～～～～～

\[ P[h] = q_4 p_2'(t) p_0'(t) p_1(t) p_1'(t) p_2(t) \pi_G \pi_T \pi_G P_{GC}(t) \pi_T \pi_A \pi_C P_{CC}(t) \pi_A \]

links only
nucleotide states
TKF91 PHMM

\[ g = \frac{\lambda}{\mu} \]
TKF91 summary

\[ P[h] = \left( \frac{\alpha(1 - \beta)}{(1 - \alpha)p_M(1 - \gamma)} \right)^{h_M} \left( \frac{\alpha(1 - \beta)}{(1 - \alpha)p_X(1 - \gamma)} \right)^{h_X} (\beta \pi)^n (g \pi)^m (1 - g)(1 - \beta) \left( \frac{\gamma(1 - \beta)}{\beta(1 - \gamma)} \right)^{DI} \]

Less than one if \( \lambda < \mu \)

- With a simple substitution model, TKF91 alignments use a simple summary: (matches, mismatches)
- D→I transitions matter also, but any alignment with summary (m,x) can be transformed into another with the same summary and no D→I transitions.
TKF in action

(http://biowiki.org/, created by Ian Holmes)
How good a model is TKF91?

- **Good**
  - Continuous time Markov model for entire sequences
  - Computing likelihood, ML alignment is tractable
- **Bad**
  - Unrealistic indel model - insertion/deletion events often involve more than one nucleotide
  - This was somewhat corrected for in TKF92