Insertions and deletions

Probabilistic models of indel evolution

I. Holmes

Department of Bioengineering
University of California, Berkeley

Spring semester

Outline

- Point substitution: the canonical models
- Context-dependent and multiple-nucleotide substitutions
  - e.g. CpG → 5-methyl-cytosine (5mC) → C-deamination → TpG
- Insertions and deletions; mechanisms
  - DNA foldback, cruciforms, etc; polymerase stutter; microsatellites
  - Transposition (DNA cut’n’paste; retrotransposition; rolling-circle); horizontal transfer
- Duplications (local, nonlocal); inversions; whole-chromosome and -genome duplications, polyploidy
- Rearrangement
- Recombination; gene conversion

- Motivation: how to handle gaps? Gaps as a fifth nucleotide in standard point substitution model: advantages (simplicity), drawbacks (irreversibility, “ghost” bases)
- Whole-sequence models: state space Ω^*. Rate grammar notation: write mutation rules
  - Point substitution with rate matrix $\mathbf{Q}$: rule is $AxB \rightarrow AyB : Q_{xy}$ where $A, B \in \Omega^*$ and $x, y \in \Omega$
  - Insertion: rule can be written $AC \rightarrow ABC$, etc.
- How can we calculate pairwise likelihoods (i.e. matrix exponential) or multiple alignment likelihoods for such a model? At first glance, appears impossible (or very difficult): infinite-dimensional (albeit sparse) rate matrix where total mutation rate (and number of mutations) scales with sequence length.
- Under point substitution model, likelihood for whole alignment factorises into product of column likelihoods.

- Thorne, Kishino, Felsenstein 1991. **Links model.**
  - First consider the following rate grammar:

Event	Rule			Rate
Substitution	AxB	$\rightarrow$	AyB	Q_{xy}
Insertion	AB	$\rightarrow$	AyB	$\lambda\pi_y$
Deletion	AxB	$\rightarrow$	AB	μ

  - Here $A, B \in \Omega^*$, $x, y \in \Omega$, $\mathbf{Q}$ is a reversible Markov process on Ω with equilibrium π, λ is a point insertion rate and μ is a point deletion rate

- Reversibility and equilibrium of the model
  - Number of links in sequence evolves independently from actual nucleotides themselves
  - Let n be sequence length (i.e. number of mortal links)
    - n evolves according to a classical "linear birth-death process with constant immigration" (in fact immigration rate = birth rate)
    - Total insertion rate (i.e. rate of $n \to n+1$) is $\lambda(n+1)$
    - Total deletion rate (i.e. rate of $n \to n-1$) is μn
    - If reversibility holds, then $P(n) \times \text{Rate}(n \to n+1) = P(n+1) \times \text{Rate}(n+1 \to n)$ where $P(n)$ is equilibrium length distribution
    - Thus $P(n)\lambda(n+1) = P(n+1)\mu(n+1)$ so that $P(n) = \kappa^n(1-\kappa)$ where $\kappa = \frac{\lambda}{\mu}$ i.e. geometric
    - NB for normalisation, $\kappa < 1 \Rightarrow \lambda < \mu$
    - Expected sequence length at equilbrium is $\frac{\kappa}{1-\kappa}$. As $\lambda \to \mu$, equilibrium sequence length $\to \infty$
    - Equation of state for $n(t)$. Here $P_t(n') = P(n(t) = n')$

$$\frac{d}{}$$

- TK&F introduced the following construction to help analyse this model (following Bishop & Thompson, 1986)
  - A biological sequence is modeled as a sequence of links: one “immortal link” followed by zero or more normal or “mortal links”
  - Mortal and immortal links spawn new “child links” to their right (rate λ). Mortal links can also die (rate μ).
  - Each mortal link corresponds to an observed, independently evolving nucleotide (or amino acid). The immortal link is invisible. (Without the immortal link, if the sequence ever reached zero length it would get irreversibly stuck. This might or might not be realistic, but one goal of Thorne *et al* was a reversible model.)
  - From the TKF 1991 paper:

> “The insertion-deletion process is framed in terms of a birth-death process of these links. Each link evolves independently from all other link; a birth or death of one link does not affect the probability of a birth or death of

Solving for conditional probabilities $P(\text{descendant}|\text{ancestor})$ in TKF91

- Consider an ancestral sequence with n mortal links (and one immortal link)
- Each ancestral link defines a zone that evolves independently from all other zones
- Zones corresponding to mortal ancestral links can acquire new links, lose links (including the original mortal link) and even die off (i.e. lose all its links, and become inert)
  - NB this implies the process w.r.t. zones is irreversible. Apparent paradox arises because we’ve introduced new information in the form of the zone (alignment) co-ordinates, and our original model did not promise to be reversible with respect to zones (alignments).
  - A specific consequence of fixing the zone co-ordinates is that the alignment $\begin{matrix} X- \\ -X \end{matrix}$ is distinct from the alignment

- Differential equations for zone fates. Suppose time t has elapsed since ancestral sequence observed. Let
  - $p_n(t)$ be the probability that n mortal links are descended from a mortal link **and that one of them is the original**;
  - $q_n(t)$ be the probability that n mortal links are descended from a mortal link **and that the original died**;
  - $r_n(t)$ be the probability that n mortal links are descended from an immortal link.

  Differential equations:

$$\begin{aligned}
\frac{d}{dt}p_n(t) &= \lambda(n-1)p_{n-1}(t) + \mu n p_{n+1}(t) - (\lambda+\mu)np_n(t) \\
\frac{d}{dt}q_n(t) &= \lambda(n-1)q_{n-1}(t) + \mu(n+1)q_{n+1}(t) + \mu p_{n+1}(t) - (\lambda \\
\frac{d}{dt}r_n(t) &= \lambda n r_{n-1}(t) + \mu(n+1)r_{n+1}(t) - (\lambda(n+1)+\mu n)r_n(t)
\end{aligned}$$

  Boundary conditions:

TKF91 as a transducer and a Pair HMM

- Transducer: stochastic finite state machine that "eats" input symbols and "emits" output symbols, representing the action of a finite time interval t (ancestor=input, descendant=output)
  - States `START`, `INSERT`, `WAIT`, `MATCH`, `DELETE`, `END`
    - `WAIT` is a null state introduced for later convenience; it means "wait for input symbol"
  - A transducer is similar to a Pair HMM, but normalised differently
    - Forward likelihood is conditional $P(\text{des}|\text{anc})$ rather than joint $P(\text{anc}, \text{des})$
  - Emission probabilities $\exp(\mathbf{Q}t)_{xy}$ (`MATCH` state), π_y (`INSERT` state), 1 (`DELETE` state)
  - Transition matrix

From/To	S	I	W	M	D	E
S	.	β	$1-\beta$	.	.	.
I		β	$1-\beta$			

Can obtain joint Pair HMM for likelihood $P(\text{anc}, \text{des})$ by "left-multiplying" transducer by single-state HMM for ancestor

- Ancestor states `S`, `I`, `E`; transition matrix

From/To	S	I	E
S	.	κ	$1-\kappa$
I	.	κ	$1-\kappa$
E	.	.	.

- Joint anc-des states `SS`, `SI`, `SW`, `IM`, `ID`, `II`, `IW`, `EE`
- Emission probabilities $\pi_x \exp(\mathbf{Q}t)_{xy}$ (`IM` state), π_y (`SI`, `II` states), π_x (`ID` state)
- Transition matrix

From/To	SS	SI	SW	IM	ID	II	IW	EE
SS	.	β	$1-\beta$	.	.	.	.	
SI	.	β	$1-\beta$	.	.	.	.	.
SW	.	.	.	$\kappa\alpha$	$\kappa(1-\alpha)$	.	.	$1-\kappa$

Consider also multiplying transducer by itself and summing out missing states

- Sequences $X \xrightarrow{t_1} Y \xrightarrow{t_2} Z$
- Markov property:
  $P(Z|X, t_1 + t_2) = \sum_Y P(Y|X, t_1)P(Z|Y, t_2)$
- Can write an expanded transducer for $P(Y|X, t_1)P(Z|Y, t_2)$
  - Joint $XY - YZ$ states: `SS, SI, SW, IM, ID, II, IW, WW, MM, MD, MI, DW, EE`
  - Exercise: write out transition matrix & emit probabilities

- Evolutionary HMMs
  - Exhaustive DP (c.f. Hein 2001)
- MCMC and evolutionary HMMs
  - Branch sampling; node sampling; aunt displacement; (parent sampling); sequence sampling
  - Chaining programs together via MCMC; `tkfalign` and propose/accept/reject

- The long indel model
  - Knudsen-Miyamoto transducer
  - Miklòs-Lunter-Holmes transducer
- Short-range context-sensitive substitution models
  - Siepel-Haussler approach: model k-mer as Ω^k-state stochastic process $x_1(t), x_2(t) \ldots x_k(t)$
    - Conditional probability of next aligned pair, given previous $k-1$ aligned pairs: $P(x_k(0), x_k(t) | x_1(0) \ldots x_{k-1}(0), x_1(t) \ldots x_{k-1}(t))$
  - Lunter-Hein approach: model full Ω^L process. Rate matrix is sum of k-mer terms, some of which commute, some don't. Taylor series for matrix exponential can be factorised and solved by DP.
  - Short-range context-dependence: relevant for protein? Probably not for substitution (eg recent Brenner *et al* study) but maybe for indels (which might look like local duplications)
- Short-range context-sensitive indel models

Summary

- Indels