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 to calculate conditional probabilities (matrix exponential) or multiple alignment likelihoods?
  - Infinite-dimensional (albeit sparse) rate matrix; total mutation rate scales with sequence length.
- Under point sub. model, likelihood for whole alignment factorizes into product of column likelihoods. Each column is independently evolving zone. Extension to indel models uses same idea of independent zones.

- 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 $P_t(n') = P(n(t) = n')$

$$\frac{d}{dt}P_t(n) = \lambda n P_t(n-1) + \mu(n+1)P_t(n+1) - (\lambda n + \mu(n+1))P_t(n)$$

Same as equation for immortal zone fates (see below)

- Clearly individual nucleotides are distributed according to π at equilibrium (since newly-inserted nucleotides are also sampled from this distribution). So equilibrium probability distribution over sequences X is

$$P(X) = \kappa^{|X|}(1-\kappa)\prod_{i=1}^{|X|}\pi_{X_i}$$

NB this is also the likelihood for generating X from a single-state HMM with self-loop transition probability κ and emit vector π.

The mean sequence length is $\kappa/(1-\kappa)$.

- TK&F (following Bishop & Thompson, 1986)
  - A biological sequence is a chain 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.

"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 any other link. Both types of links can be associated with births. The birth rate per normal link is equal to the birth rate per immortal link (λ). A newborn link is always a normal link. We adopt the convention that it appears immediately to the right of its parent link. Accompanying the birth of a normal link is the birth of a DNA base immediately to the left of the newborn link. The probabilities that the newborn DNA base will be A, G, T, or C are π_A, π_G, π_T and π_C, respectively. Normal links are subject to death (μ is the death rate per normal link) but immortal links are not."

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

- 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 $\begin{matrix}\mathrm{X}- \\ -\mathrm{X}\end{matrix}$ is distinct from the alignment $\begin{matrix}-\mathrm{X} \\ \mathrm{X}-\end{matrix}$, since the former implies a deletion followed by an insertion in the same zone, whereas the latter does not imply any chronological ordering on the insertion & deletion events, since they occurred in different zones. (This asymmetry relates to the apparent contradiction to reversibility mentioned above. Both can, in fact, be “fixed” by swapping the order of such alignment columns when reversing the arrow of time.)
- The zone corresponding to the immortal link can acquire and lose new links, but never dies off

- 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:

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

Boundary conditions:

$$\begin{aligned}
p_n(0) &= \delta(n=1) \\
q_n(0) &= 0 \\
r_n(0) &= \delta(n=0)
\end{aligned}$$

Solutions:

$$\begin{aligned} p_n(t) &= \alpha\beta^{n-1}(1-\beta) \\ q_0(t) &= (1-\alpha)(1-\gamma) \\ q_n(t) &= (1-\alpha)\gamma\beta^{n-1}(1-\beta) \quad \text{for } n > 0 \\ r_n(t) &= \beta^n(1-\beta) \end{aligned}$$

where

$$\begin{aligned} \alpha(t) &= \exp(-\mu t) \\ \beta(t) &= \frac{1-\alpha(t)\exp(\lambda t)}{\kappa^{-1}-\alpha(t)\exp(\lambda t)} \\ \gamma(t) &= 1-\frac{\beta(t)}{\kappa(1-\alpha(t))} \end{aligned}$$

- Thorne *et al* quote these results without derivation (they were obtained by comparison with formulae for similar generic birth-death processes). They can readily be verified.
- Metzler argument: if n is the no. of surviving links at time t, then $P(n \geq k+1 | n \geq k)$ must be independent of k, since $n \geq k$ implies that there has been an available insertion site for time t
  - Holmes used this to derive numerical estimates of posterior expectations for number of indels
- Feller's generating function approach can be used to solve for $r_n(t)$ and guess forms of p_n, q_n
  - This also yields posterior summary statistics: Minin *et al*, `http://arxiv.org/abs/1009.0893`
- Karlin and McGregor analysed birth-death processes in detail, and found a series of orthogonal polynomials associated with transition probabilities of the process

Introduce generating function $G(s,t) = \sum_{n=0}^{\infty} s^n r_n(t)$. The r_n are recoverable as $r_n(t) = \left.\frac{\partial^n G}{\partial s^n}\right|_{s=0}$

Let $D = \frac{\partial}{\partial s}$ and use the following operator table:

	Operator L	Coefficient of s^n in LG
	1	r_n
sDs	$= s(1+sD)$	nr_{n-1}
	D	$(n+1)r_{n+1}$
Ds	$= 1+sD$	$(n+1)r_n$
	sD	nr_n

$$\begin{aligned}\frac{\partial G}{\partial t} &= [\lambda s(1+sD) + \mu D - \lambda(1+sD) - \mu sD]\,G \\ &= \lambda(s-1)G + (\lambda s - \mu)(s-1)\frac{\partial G}{\partial s}\end{aligned}$$

$$\begin{aligned}\frac{\partial G}{\partial t} &= [\lambda s(1+sD) + \mu D - \lambda(1+sD) - \mu sD]\, G \\ &= \lambda(s-1)G + (\lambda s - \mu)(s-1)\frac{\partial G}{\partial s}\end{aligned}$$

Can rewrite this as

$$\frac{1}{\lambda(s-1)}\frac{\partial G}{\partial t} + (\mu/\lambda - s)\frac{\partial G}{\partial s} = G$$

Boundary condition is $G(s,0) = 1$.

Use method of characteristics to rewrite this p.d.e. as o.d.e.'s. Suppose $s = s(u)$ and $t = t(u)$. Then $\frac{dG}{du} = \frac{\partial G}{\partial t}\frac{dt}{du} + \frac{\partial G}{\partial s}\frac{ds}{du}$ which looks like our p.d.e. if

$$\begin{aligned} \frac{dt}{du} &= \frac{1}{\lambda(s-1)} \\ \frac{ds}{du} &= \mu/\lambda - s \\ \frac{dG}{du} &= G \end{aligned}$$

The general solution for G is

$$G(s,t) = g(v)e^u$$

where $g(\cdot)$ is any function and v is constant along a characteristic, so $dv/du = 0$. Solving the o.d.e. for $s(u)$ (e.g. using an integrating factor) we obtain $s = e^{-u} + \mu/\lambda$.

Furthermore, on a characteristic curve, the following is true

$$\frac{dt}{ds} = \frac{dt/du}{ds/du} = \frac{1}{\lambda(s-1)(\mu/\lambda - s)}$$

and hence

$$\log\left|\frac{s-\mu/\lambda}{s-1}\right| + (\mu-\lambda)t = \text{const.}$$

The general solution for G can thus be written

$$G(s,t) = g\left(\frac{s-\mu/\lambda}{s-1}e^{(\mu-\lambda)t}\right)(s-\mu/\lambda)^{-1}$$

where $g(\cdot)$ is an arbitrary function, to be determined by the boundary condition – which is $G(s,0) = 1$, so

$$g\left(\frac{s-\mu/\lambda}{s-1}\right) = s - \mu/\lambda$$

Boundary condition $G(s,0) = 1$ leads to

$$\begin{aligned}
g\left(\frac{s-\mu/\lambda}{s-1}\right) &= s-\mu/\lambda \\
g(v) &= (\mu/\lambda - 1)\left(\frac{1}{1-v} - 1\right) \\
G(s,t) &= \frac{\mu/\lambda - 1}{\mu/\lambda - e^{(\lambda-\mu)t} - s\left(1 - e^{(\lambda-\mu)t}\right)} \\
r_n(t) &= \left.\frac{\partial^n G}{\partial s^n}\right|_{s=0} \\
&= \left.\frac{(\mu/\lambda - 1)\left(1 - e^{(\lambda-\mu)t}\right)^n}{\left(\mu/\lambda - e^{(\lambda-\mu)t} + s\left(e^{(\lambda-\mu)t} - 1\right)\right)^{n+1}}\right|_{s=0} \\
&= (1-\beta(t))\,\beta(t)^n
\end{aligned}$$

as expected.

Integrating factors. Consider the equation

$$y' + P(x)y = Q(x)$$

Multiply by integrating factor $M(x)$ to yield

$$M(x)y' + P(x)M(x)y = Q(x)M(x)$$

If we choose $M(x)$ such that $M'(x) = M(x)P(x)$, then

$$(M(x)y)' = Q(x)M(x)$$

Equivalently $M'/M = P$ and so

$$\begin{aligned} M(x) &= \exp\left[\int P(x)dx\right] \\ y(x) &= \frac{\int Q(x)M(x)dx + C}{M(x)} \end{aligned}$$

where C is a constant of integration.

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	$1-\beta$	.	.	.
W	.	.	.	α	$1-\alpha$	1
M	.	β	$1-\beta$	.	.	.
D	.	γ	$1-\gamma$	.	.	.
E	.	.	.	.	.	.

(dots represent zeroes)

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	.	.	.	$\kappa\alpha$	$\kappa(1-\alpha)$	.	.	$1-\kappa$
IM	.	.	.	.	.	β	$1-\beta$	.
ID	.	.	.	.	.	γ	$1-\gamma$	.
II	.	.	.	.	.	β	$1-\beta$	.
IW	.	.	.	$\kappa\alpha$	$\kappa(1-\alpha)$	.	.	$1-\kappa$
EE	.	.	.	.	.	.	.	.

Here, in constructing transitions for the expanded transducer, we have used the general rule: *“Update the last transducer that is not in a* `WAIT` *state. If this transducer emits an output symbol, then update any* `WAIT`*-ing transducers that receive a symbol on their input, repeating this last step recursively until no more transducers have input symbols to process.”*
All transitions update one transducer only, except those from {`SW`,`IW`} to {`IM`,`ID`,`EE`
Note redundancy: can eliminate `SW` and `IW`, collapse `SI` and `II` together... in fact, for this model [TKF91], can collapse DP right down to one variable; see e.g. Miklòs, Song *et al*.

- We just described transducer *composition*
  - Output of transducer A connected to input of B
  - Analogous to matrix multiplication AB
- Also useful to consider transducer *intersection*
  - Input duplicated, fed to inputs of A and B
  - Analogous to pointwise (Hadamard) product, $A \circ B$
- Generators, recognizers
  - Unit generator $\Delta(S)$ and recognizer $\nabla(S)$
- Felsenstein pruning: Westesson *et al*, `http://arxiv.org/abs/1103.4347`

`http://en.wikipedia.org/wiki/Finite_state_transducer`

Felsenstein pruning recursion for transducers

$$F_n = \begin{cases} \left(M^{(l)} F_l\right) \circ \left(M^{(r)} F_r\right) & \text{if } n \text{ internal} \\ \nabla(y_n) & \text{if } n \text{ leaf} \end{cases}$$

The Felsenstein likelihood is given by summing all paths through RF_{root}, where

- R is generator for root distribution,
- $M^{(n)}$ is transducer on branch above node n,
- y_n is sequence at leaf node n,
- AB is transducer composition (matrix multiplication),
- $A \circ B$ is transducer intersection (pointwise product).

- 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
  - Motivation: many indels appear, empirically, to be miniature local duplications
- Mutation-selection models
  - Description of model
- Rearrangement models: combinatorial explosion in histories, MCMC slow
  - Most authors write the genome in bigger units (e.g. identifiable genes or blocks of synteny, rather than individual nucleotides)
  - Hannenhalli & Pevzner, 1999; Siepel, 2002; Miklòs, Bioinformatics 2003

Summary

- Indels