Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Insertions and deletions
Probabilistic models of indel evolution

I. Holmes

Department of Bioengineering
University of California, Berkeley

Spring semester

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Outline

1. Mechanisms of sequence mutation

2. The "links model" and string transducers

3. Multiple alignment with the links model

4. More realistic evolutionary models

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- Point substitution: the canonical models
- Context-dependent and multiple-nucleotide substitutions
  - e.g. CpG $\rightarrow$ 5-methyl-cytosine (5mC) $\rightarrow$ C-deamination $\rightarrow$ 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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- Motivating question: how to handle gaps?
- Gaps as a fifth nucleotide in standard point substitution model: advantages (simplicity), drawbacks (irreversibility, "ghost" bases)

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- Continuous-time Markov chains where the state space is the set of sequences Ω^*
- String-rewriting system (semi-Thue system) + rates = *rate grammar*
  - Point substitution with rate matrix **Q**: rule is $AxB \rightarrow AyB : Q_{xy}$ where $A, B \in \Omega^*$ and $x, y \in \Omega$
  - Insertion: $AC \rightarrow ABC : \lambda(B)$
  - Deletion: $ADC \rightarrow AC : \mu(D)$
- 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.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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 $\to \infty$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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 $\kappa/(1 - \kappa)$.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

"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."

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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).

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- A specific consequence of fixing the zone co-ordinates is that the alignment $\begin{smallmatrix} X- \\ -X \end{smallmatrix}$ is distinct from the alignment $\begin{smallmatrix} -X \\ X- \end{smallmatrix}$, 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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Differential equations:

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

Boundary conditions:

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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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

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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Introduce generating function $G(s, t) = \sum_{n=0}^{\infty} s^n r_n(t)$. The r_n are recoverable as $r_n(t) = \frac{1}{n!} \left. \frac{\partial^n G}{\partial s^n} \right|_{s=0}$

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

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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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

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 $G(s,0) = 1$.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Use method of characteristics to rewrite this p.d.e. as o.d.e.'s.

Suppose $s = s(u)$ and $t = t(u)$. Then $\dfrac{dG}{du} = \dfrac{\partial G}{\partial t}\dfrac{dt}{du} + \dfrac{\partial G}{\partial s}\dfrac{ds}{du}$

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 $s = e^{-u} + \mu/\lambda$.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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

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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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.

I. Holmes   Indels

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

TKF91 as a transducer and a Pair HMM

- Transducer: stochastic finite state machine that "absorbs" 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*(des|anc) rather than joint *P*(anc, des)
    - Emission probabilities $\exp(\mathbf{Q}t)_{xy}$ (MATCH state), π_y (INSERT state), 1 (DELETE state)

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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)

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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)

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Now consider two TKF91 transducers in series:

$$\text{Ancestor} \xrightarrow{T(t)} \text{Intermediate} \xrightarrow{T(t)} \text{Descendant}$$

Can we marginalize the intermediate sequence, and write this as a single transducer?

$$\text{Ancestor} \xrightarrow{T(t)^2} \text{Descendant}$$

For TKF91, we know the transducer models the finite-time transition probabilities of a stochastic process, so $T(t)^2 = T(2t)$. However we can also think of the state space of the two coupled TKF91 machines as a single ensemble transducer.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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	.	.	.	.	.	.	.	.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

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*.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

The letter transducer is a tuple $T = (\Omega_I, \Omega_O, \Phi, \phi_S, \phi_E, \tau, \mathcal{W})$ where

- Ω_I is an input alphabet,
- Ω_O is an output alphabet,
- Φ is a set of states,
- $\phi_S \in \Phi$ is the start state,
- $\phi_E \in \Phi$ is the end state,
- $\tau \subseteq \Phi \times (\Omega_I \cup \{\epsilon\}) \times (\Omega_O \cup \{\epsilon\}) \times \Phi$ is the transition relation,
- $\mathcal{W} : \tau \to [0, \infty)$ is the transition weight function.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Transition paths: The transitions in τ correspond to the edges of a labeled multidigraph over states in Φ. Let $\Pi \subset \tau^*$ be the set of all labeled transition paths from ϕ_S to ϕ_E.

Transduction weight: For a transition path π from ϕ_S to ϕ_E, define the path weight $\mathcal{W}(\pi)$ and (for sequences $x \in \Omega_I^*, y \in \Omega_O^*$) the transduction weight $\mathcal{W}(x : [T] : y)$

$$
\mathcal{W}(\pi) = \prod_{\tau \in \pi} \mathcal{W}(\tau)
$$

$$
\mathcal{W}(x : [T] : y) = \sum_{\pi \in \Pi, S_I(\pi) = x, S_O(\pi) = y} \mathcal{W}(\pi)
$$

where S_I and S_O denote the input and output labelings of a path.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Equivalence: If for transducers T, T' it is true that $\mathcal{W}(x : [T] : y) = \mathcal{W}'(x : [T'] : y) \; \forall x, y$ then the transducers are equivalent in weight, $T \equiv T'$.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Types of state and transition: If there exists a function,

- type $: \Phi \to \mathcal{T}$

mapping states to types in the set

$$\mathcal{T} = \{S, M, D, I, N, W, E\}$$

and, furthermore, we can find functions

- $\mathcal{W}_{\text{trans}} : \Phi^2 \to [0, \infty)$
- $\mathcal{W}_{\text{emit}} : (\Omega_I \cup \{\epsilon\}) \times (\Omega_O \cup \{\epsilon\}) \times \Phi \to [0, \infty)$

such that the labeled transition weights separate into transition
and emission factors (see next slide), then the transducer is in
(weak) normal form.
If, additionally, there are no null states ($\Phi_N = \emptyset$), then the
transducer is in *strict normal form*.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

$$\Phi_U = \{\phi : \phi \in \Phi, \text{type}(\phi) \in U \subseteq \mathcal{T}\}$$
$$\Phi_S = \{\phi_S\}$$
$$\Phi_E = \{\phi_E\}$$
$$\Phi \equiv \Phi_{SMDINWE}$$
$$\tau_M \subseteq \Phi_W \times \Omega_I \times \Omega_O \times \Phi_M$$
$$\tau_D \subseteq \Phi_W \times \Omega_I \times \{\epsilon\} \times \Phi_D$$
$$\tau_I \subseteq \Phi_{SMDIN} \times \{\epsilon\} \times \Omega_O \times \Phi_I$$
$$\tau_N \subseteq \Phi_{SMDIN} \times \{\epsilon\} \times \{\epsilon\} \times \Phi_N$$
$$\tau_W \subseteq \Phi_{SMDIN} \times \{\epsilon\} \times \{\epsilon\} \times \Phi_W$$
$$\tau_E \subseteq \Phi_W \times \{\epsilon\} \times \{\epsilon\} \times \Phi_E$$
$$\tau = \tau_M \cup \tau_D \cup \tau_I \cup \tau_N \cup \tau_W \cup \tau_E$$
$$\mathcal{W}(\phi_{\text{src}}, \omega_{\text{in}}, \omega_{\text{out}}, \phi_{\text{dest}}) \equiv \mathcal{W}_{\text{trans}}(\phi_{\text{src}}, \phi_{\text{dest}}) \mathcal{W}_{\text{emit}}(\omega_{\text{in}}, \omega_{\text{out}}, \phi_{\text{dest}})$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Interpretation: A normal-form transducer can be thought of as associating inputs and outputs with states, rather than transitions. (Thus, it is like a Moore machine.) The state types are start (S) and end (E); wait (W), in which the transducer waits for input; match (M) and delete (D), which process input symbols; insert (I), which writes additional output symbols; and null (N), which has no associated input or output. All transitions also fall into one of these types, via the destination states; thus, τ_M is the set of transitions ending in a match state, etc. The transition weight ($\mathcal{W}$) factors into a term that is independent of the input/output label ($\mathcal{W}_{\text{trans}}$) and a term that is independent of the source state ($\mathcal{W}_{\text{emit}}$).

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Universality: For any weak-normal form transducer *T* there exists an equivalent in strict-normal form which can be found by applying the state-marginalization algorithm to eliminate null states. For any transducer, there is an equivalent letter transducer in weak normal form, and therefore, in strict normal form.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Mealy machines are transducers with I/O occurring on transitions, as with our general definition of the letter transducer.
Moore machines are transducers whose I/O is associated with states, as with our normal form.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Transducer composition: Given letter transducers
$T = (\Omega_X, \Omega_Y, \Phi, \phi_S, \phi_E, \tau, \mathcal{W})$ and
$U = (\Omega_Y, \Omega_Z, \Phi', \phi'_S, \phi'_E, \tau', \mathcal{W}')$, there exists a letter transducer
$TU = (\Omega_X, \Omega_Z, \Phi'' \ldots \mathcal{W}'')$ such that $\forall x \in \Omega_X^*, z \in \Omega_Z^*$:

$$\mathcal{W}''(x : [TU] : z) = \sum_{y \in \Omega_Y^*} \mathcal{W}(x : [T] : y) \mathcal{W}'(y : [U] : z)$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Example construction of TU: Assume without loss of generality that T and U are in strict normal form. Then $\Phi'' \subset \Phi \times \Phi'$, $\phi''_S = (\phi_S, \phi'_S)$, $\phi''_E = (\phi_E, \phi'_E)$ and

$$
\mathcal{W}''((t, u), \omega_x, \omega_z, (t', u')) =
\begin{cases}
\delta(t = t')\delta(\omega_x = \epsilon)\mathcal{W}'(u, \epsilon, \omega_z, u') & \text{if } \text{type}(u) \neq W \\
\mathcal{W}(t, \omega_x, \epsilon, t')\delta(\omega_z = \epsilon)\delta(u = u') & \text{if } \text{type}(u) = W, \text{type}(t') \notin \{M, I\} \\
\displaystyle\sum_{\omega_y \in \Omega_Y} \mathcal{W}(t, \omega_x, \omega_y, t')\mathcal{W}'(u, \omega_y, \omega_z, u') & \text{if } \text{type}(u) = W, \text{type}(t') \in \{M, I\} \\
0 & \text{otherwise}
\end{cases}
$$

The resulting transducer is in weak-normal form (it can be converted to a strict-normal form transducer by eliminating null states).

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Transducer intersection: Given letter transducers
$T = (\Omega_X, \Omega_T, \Phi, \phi_S, \phi_E, \tau, \mathcal{W})$ and
$U = (\Omega_X, \Omega_U, \Phi', \phi'_S, \phi'_E, \tau', \mathcal{W}')$, there exists a letter transducer
$T \circ U = (\Omega_X, \Omega_V, \Phi'' \dots \mathcal{W}'')$ where
$\Omega_V \subseteq (\Omega_T \cup \{\epsilon\}) \times (\Omega_U \cup \{\epsilon\})$ such that
$\forall x \in \Omega_X^*, t \in \Omega_T^*, u \in \Omega_U^*$:

$$\mathcal{W}(x : [T] : t)\mathcal{W}'(x : [U] : u) = \mathcal{W}''(x : [T \circ U] : (t, u))$$

where the term on the right is defined as follows

$$\mathcal{W}''(x : [T \circ U] : (t, u)) = \sum_{v \in \Omega_V^*, S_1(v)=t, S_2(v)=u} \mathcal{W}''(x : [T \circ U] : v)$$

Here Ω_V is the set of all possible pairwise alignment columns,
$v \in \Omega_V^*$ is a pairwise alignment and $S_1(v)$ and $S_2(v)$ are the
sequences in (respectively) the first and second rows of v.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Example construction of T ∘ U: Assume without loss of generality that *T* and *U* are in strict normal form. Then $\Phi'' \subset \Phi \times \Phi'$, $\phi''_S = (\phi_S, \phi'_S)$, $\phi''_E = (\phi_E, \phi'_E)$ and

$$\mathcal{W}''((t, u), \omega_x, (\omega_y, \omega_z), (t', u')) =$$
$$\begin{cases} \delta(t = t')\delta(\omega_x = \omega_y = \epsilon)\mathcal{W}'(u, \epsilon, \omega_z, u') & \text{if } \text{type}(u) \neq W \\ \mathcal{W}(t, \epsilon, \omega_x, t')\delta(\omega_x = \omega_z = \epsilon)\delta(u = u') & \text{if } \text{type}(u) = W, \text{type}(t) \neq W \\ \mathcal{W}(t, \omega_x, \omega_y, t')\mathcal{W}'(u, \omega_x, \omega_z, u') & \text{if } \text{type}(t) = \text{type}(u) = W \\ 0 & \text{otherwise} \end{cases}$$

The resulting transducer is in weak-normal form (it can be converted to a strict-normal form transducer by eliminating null states).

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Identity: There exists a transducer $\mathcal{I} = (\Omega, \Omega \dots)$ that copies its input identically to its output. An example construction (not in normal form) is

$$\begin{aligned}
\mathcal{I} &= (\Omega, \Omega, \{\phi\}, \phi, \phi, \tau_{\mathcal{I}}, 1) \\
\tau_{\mathcal{I}} &= \{(\phi, \omega, \omega, \phi) : \omega \in \Omega\}
\end{aligned}$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Bifurcation: There exists a transducer $\Upsilon = (\Omega, \Omega^2 \dots)$ that duplicates its input in parallel. That is, for input $x_1 x_2 x_3 \dots$ it gives output $\begin{pmatrix} x_1 \\ x_1 \end{pmatrix} \begin{pmatrix} x_2 \\ x_2 \end{pmatrix} \begin{pmatrix} x_3 \\ x_3 \end{pmatrix} \dots$. An example construction (not in normal form) is

$$\Upsilon = (\Omega, \Omega^2, \{\phi\}, \phi, \phi, \tau_\Upsilon, 1)$$
$$\tau_\Upsilon = \left\{ \left(\phi, \omega, \begin{pmatrix} \omega \\ \omega \end{pmatrix}, \phi \right) : \omega \in \Omega \right\}$$

It can be seen that $\Upsilon \equiv \mathcal{I} \circ \mathcal{I}$.

An intersection $T \circ U$ may be considered a parallel composition of Υ with T and U. We write this as $\Upsilon(T, U)$ or, diagrammatically, $(\, .. \, (\, .\Upsilon. \, (\, .T. \,) \, (\, .U. \,) \,) \,)$

We use the notation $\Upsilon(T, U)$ in several places, when it is convenient to have a placeholder transducer Υ at a bifurcating node in a tree.

I. Holmes   Indels

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Recognition profiles: A transducer *T* is a *recognizer* if it has a null output alphabet, and so generates no output except the empty string.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Exact-match recognizer: For $S \in \Omega^*$, there exists a transducer $\nabla(S) = (\Omega, \emptyset \dots \mathcal{W})$ that accepts the specific sequence S with weight one, but rejects all other input sequences

$$\mathcal{W}(x : [\nabla(S)] : \epsilon) = \delta(x = S)$$

Note that $\nabla(S)$ has a null output alphabet, so its only possible output is the empty string, and it is a recognizer.
In general, if $T = (\Omega_X, \Omega_Y \dots \mathcal{W}')$ is any transducer then $\forall x \in \Omega_X^*, y \in \Omega_Y^*$

$$\mathcal{W}'(x : [T] : y) \equiv \mathcal{W}(x : [T\nabla(y)] : \epsilon)$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

An example construction (not in normal form) is

$$
\begin{aligned}
\nabla(S) &= (\Omega, \emptyset, \mathbb{Z}_{\mathsf{len}(S)+1}, 0, \mathsf{len}(S), \tau_\nabla, 1) \\
\tau_\nabla &= \left\{ (n, \mathsf{symbol}(S, n+1), \epsilon, n+1) : n \in \mathbb{Z}_{\mathsf{len}(S)} \right\}
\end{aligned}
$$

where $\mathbb{Z}_N$ is the set of integers modulo N, and $\mathsf{symbol}(S, k)$ is the k'th position of S (for $1 \leq k \leq \mathsf{len}(S)$). Note that this construction has $\mathsf{len}(S) + 1$ states.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Generative transducers: A transducer T is generative (or "a *generator*") if it has a null input alphabet, and so rejects any input except the empty string. Then T may be regarded as a state machine that generates an output, equivalent to a Hidden Markov Model. Define the probability (weight) distribution over the output sequence

$$P(x|T) \equiv \mathcal{W}(\epsilon : [T] : x)$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Chapman-Kolmogorov equation:
If T_t is a transducer parameterized by a continuous time parameter t, modeling the evolution of a sequence for time t under a continuous-time Markov process, then the Chapman-Kolmogorov equation can be expressed as a transducer equivalence

$$T_t T_u \equiv T_{t+u}$$

The TKF91 transducers, for example, have this property. Furthermore, for TKF91, T_{t+u} has the same state and transition structure as T_t, so this is a kind of self-similarity.

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

We can write the instantaneous form of Chapman-Kolmogorov in transducer form,

$$\frac{d}{dt} p_t = p_t Q$$

where p_t is the generator transducer for the sequence at time t and Q is an instantaneous rate transducer, a generalized transducer whose weight is negative for on-diagonal elements (i.e. when the inputs and outputs are equal):

$$Q = Q_+ - (\mathcal{I} \circ \mathcal{P}(Q_+))$$

where Q_+ is a positive (i.e. nonnegatively-weighted) transducer representing the rates and $\mathcal{P}(Q_+)$ is the projection of Q_+ to a recognizer (marginalizing the output tape).

This further implies a transducer matrix exponential

$$T_t = \exp(Qt)$$

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Felsenstein pruning recursion for transducers

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

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).

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- Phylogenetic transducers/HMMs
    - Exhaustive DP (c.f. Hein 2001)
- MCMC
    - Gibbs-sampling kernels for local tree neighborhoods
    - Chaining programs together via Metropolis-Hastings propose/accept/reject schemes
    - Redelings and Suchard's kernel
    - Bouchard-Côté, Jordan, and Klein's kernel

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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 leading terms computed 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)

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

- 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

Mechanisms of sequence mutation
The "links model" and string transducers
Multiple alignment with the links model
More realistic evolutionary models
Summary

Summary

- Indels