

Modeling indel hot-spots when inferring alignments

Benjamin Redelings

July 3, 2008

Indel Hotspots and Indel Information: Overview

1. Indel hotspots

- Why do we care?

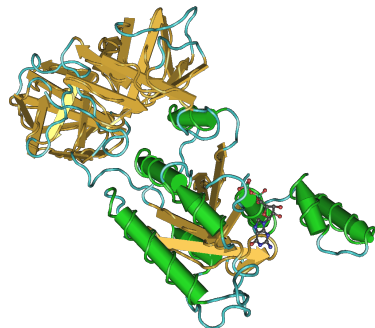
2. Background: Joint Estimation

- Effects of Ambiguity
- Sequential Estimation Paradigm
- Practical Benefits

3. MCMC sampling for indel hotspots

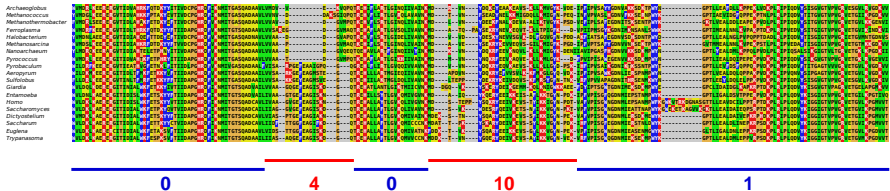
- Column Labels
- Transducers
- Sampling column labels

4. Future Directions



EF-Tu/EF-1 α

Insertions and deletions (indels) are often **clustered** in protein sequences:



One reason is because insertions are more common on the surface of proteins than in the hydrophobic core.

- 1 Most probabilistic models of insertion/deletion use a **spatially uniform** rate
 - **TKF91** and **TKF92** models have an insertion rate λ and deletion rate μ
 - The “**long indel model**” also assumes that indel rates do not vary spatially.
- 2 However, the biological reality seems to be better approximated by assuming that proteins are divided into **regions** with different indel rates.
 - **SAPF** model (Satija, et al, 2008) divides the sequence into regions, which alternate between **2 rates**.

Indel Hotspots: Why do we care?

Problem: If we assume **1 indel rate** when there are actually **2 indel rates**, then

- Indel rates in **hot-spots** are *under-estimated* → indel weight *over-estimated*
- Indel rates in **cold-spots** are *over-estimated* → indel weight *under-estimated*

Often most indels occur in hotspots → most indel weights are **overestimated!**

Use of indel information is unavoidable when averaging over **near-optimal alignments**.

Goal: Therefore, we wish to remove doubts about the use of indel information.

Additional benefits: Many other properties also vary spatially:

- Equilibrium frequencies: hydrophobic versus hydrophilic residues
- Substitution rates are also faster at the protein surface
- Idea: detect **cold-spots** based on hydrophobicity as well as lack of indels.

Alignment Ambiguity

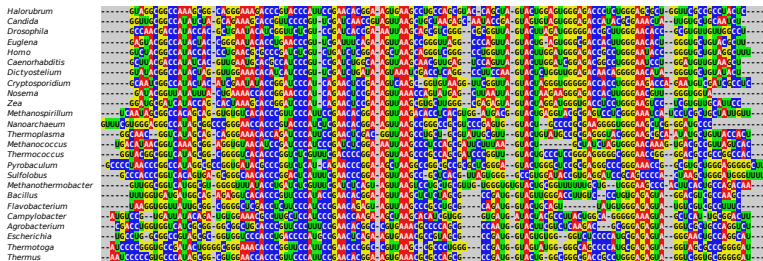
Bayesian, Parsimony, and Maximum Likelihood methods **rely on alignments**

- Residues in the same column have the same common ancestor
- But residue homology and gaps are *estimated*, not observed

Alignment ambiguity is a major problem in phylogenetic inference

Ambiguous alignments are easily affected by

- alignment method / program
- parameters: gap and mismatch costs



Clustal W alignment

Alignment Ambiguity

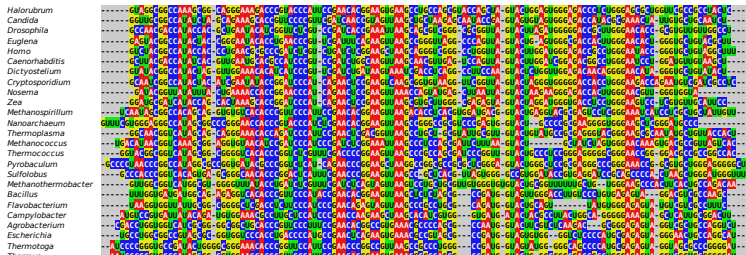
Bayesian, Parsimony, and Maximum Likelihood methods **rely on alignments**

- Residues in the same column have the same common ancestor
- But residue homology and gaps are *estimated*, not observed

Alignment ambiguity is a major problem in phylogenetic inference

Ambiguous alignments are easily affected by

- alignment method / program
- parameters: gap and mismatch costs



Muscle alignment

Alignment Ambiguity

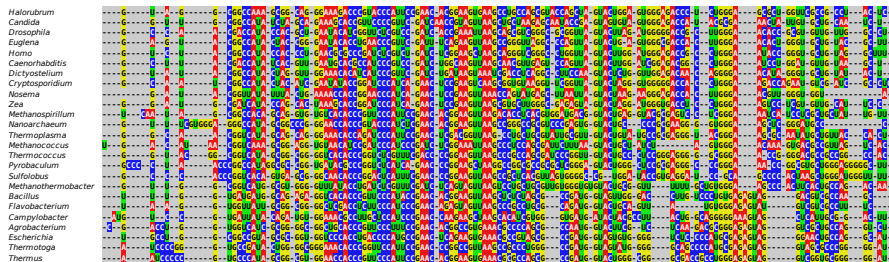
Bayesian, Parsimony, and Maximum Likelihood methods **rely on alignments**

- Residues in the same column have the same common ancestor
- But residue homology and gaps are *estimated*, not observed

Alignment ambiguity is a major problem in phylogenetic inference

Ambiguous alignments are easily affected by

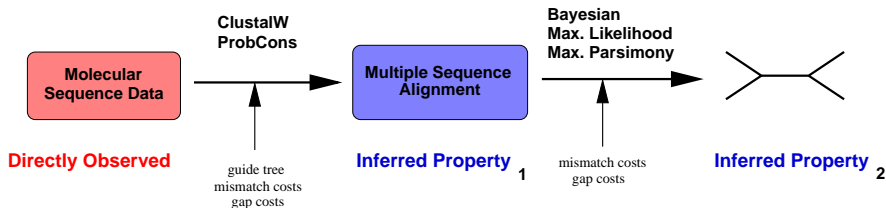
- alignment method / program
- parameters: gap and mismatch costs



BALI-Phy alignment (GTR + log-Normal₈ | RS07)

Sequential Estimation Pipeline

Current methods reconstruct phylogeny in **two stages**:

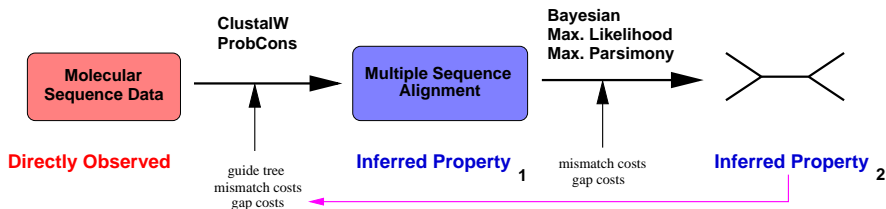


Problem #1a: guide tree is used to resolve **alignment ambiguity**

- alignment uncertainty \implies **bias toward the guide tree**

Sequential Estimation Pipeline


Current methods reconstruct phylogeny in **two stages**:



Problem #1a: guide tree is used to resolve **alignment ambiguity**

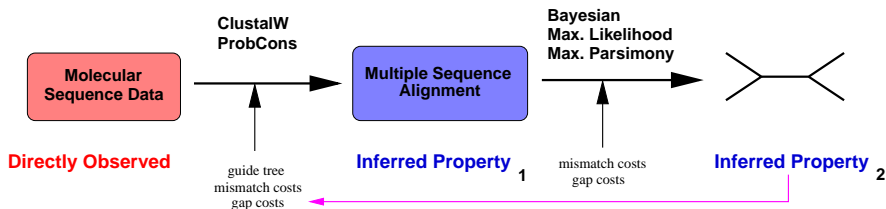
- alignment uncertainty \implies **bias toward the guide tree**

Problem #1b: alignment needed for tree estimate, and vice versa

- sequence data \rightarrow 

Sequential Estimation Pipeline

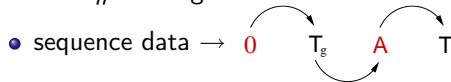
Current methods reconstruct phylogeny in **two stages**:



Problem #1a: guide tree is used to resolve **alignment ambiguity**

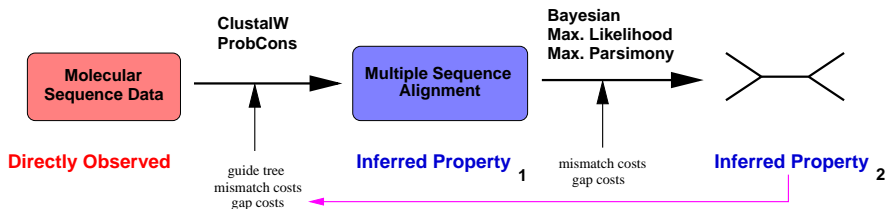
- alignment uncertainty \implies **bias toward the guide tree**

Problem #1b: alignment needed for tree estimate, and vice versa



Sequential Estimation Pipeline

Current methods reconstruct phylogeny in **two stages**:



Problem #1a: guide tree is used to resolve **alignment ambiguity**

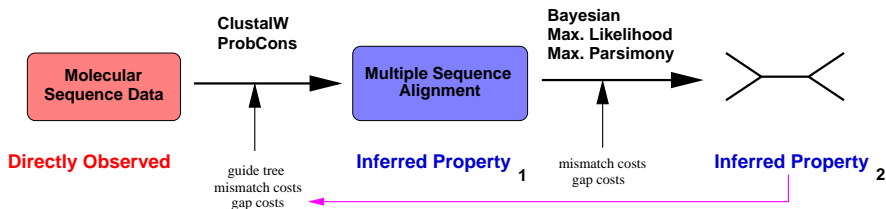
- alignment uncertainty \implies **bias toward the guide tree**

Problem #1b: alignment needed for tree estimate, and vice versa

- sequence data \rightarrow 0 \rightarrow T_{ig} \rightarrow A \rightarrow T \rightarrow A \rightarrow T \rightarrow ?

Sequential Estimation Pipeline

Current methods reconstruct phylogeny in **two stages**:



Problem #1a: guide tree is used to resolve **alignment ambiguity**

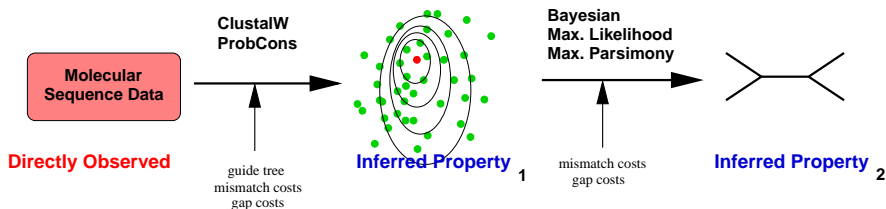
- alignment uncertainty \implies **bias toward the guide tree**

Problem #1b: alignment needed for tree estimate, and vice versa

- sequence data \rightarrow  \implies **chicken/egg problem**

Sequential Estimation Pipeline (2)

Current methods reconstruct phylogeny in **two stages**:

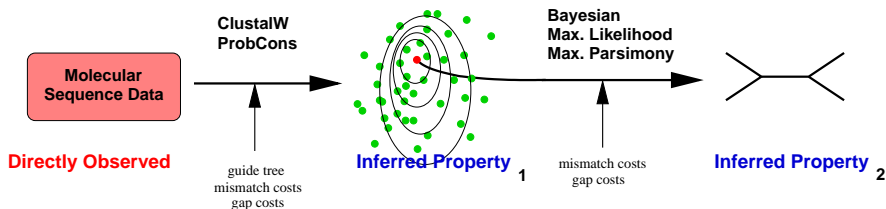


Problem #2: assumes complete certainty in a **single** alignment estimate

- alignment uncertainty is ignored \implies **exaggerated confidence in tree**

Sequential Estimation Pipeline (2)

Current methods reconstruct phylogeny in **two stages**:



Problem #2: assumes complete certainty in a **single** alignment estimate

- alignment uncertainty is ignored \implies **exaggerated confidence in tree**

Joint Bayesian Estimation: A Solution

Solution:

- 1 Joint estimation of \mathbf{A} and \mathbf{T} :

Joint Bayesian Estimation: A Solution

Solution:

① **Joint estimation** of \mathbf{A} and \mathbf{T} :

- sequence data $\rightarrow (\mathbf{A}, \mathbf{T})$

\Leftarrow no chicken/egg problem

Solution:

1 Joint estimation of \mathbf{A} and \mathbf{T} :

- sequence data $\rightarrow (\mathbf{A}, \mathbf{T})$
- *no guide tree*

- \Leftarrow no chicken/egg problem
- \Leftarrow no bias toward guide tree

Joint Bayesian Estimation: A Solution

Solution:

1 Joint estimation of \mathbf{A} and \mathbf{T} :

- sequence data $\rightarrow (\mathbf{A}, \mathbf{T})$
- *no guide tree*

\Leftarrow no chicken/egg problem
 \Leftarrow no bias toward guide tree

2 Weighted Sum over all alignments:

- consider near-optimal alignments
- alignments naturally weighted by posterior probability

\Leftarrow no exaggerated confidence

Joint Bayesian Estimation: A Solution

Solution:

1 Joint estimation of \mathbf{A} and \mathbf{T} :

- sequence data $\rightarrow (\mathbf{A}, \mathbf{T})$
- *no guide tree*

\Leftarrow no chicken/egg problem
 \Leftarrow no bias toward guide tree

2 Weighted Sum over all alignments:

- consider near-optimal alignments
- alignments naturally weighted by posterior probability

\Leftarrow no exaggerated confidence

Joint Probability Model

- Need a joint probability function $\Pr(\mathbf{A}, \mathbf{T})$ to weight alignments.
- Need a stochastic model of the insertion/deletion process.

Better alignments

- High-quality *substitution* models with rate heterogeneity *during alignment*
- High-quality *indel* model allows us to penalize *indels*, not *gaps*

Ambiguous regions are OK → deeper divergences can be analyzed

- Retain more phylogenetically informative characters.

No “cleaning” required → just throw in your sequences!

- Replaces subjective and *ad hoc* “visual inspection” method.

Avoids bias by summing over near-optimal alignments avoids

- The “best” alignment is atypical - has fewer gaps, more mismatches.
- Summing over alignments yields unbiased *indel rates* and *branch lengths*.

Improve power by using information in shared indels to infer phylogeny

- “Rare Genomic Changes” such as indels can help infer deep divergences
- When indel rates are high, indel information can improve shallow phylogenies.

Joint Probabilistic Model: Expression

\mathbf{Y} –data (τ, \mathbf{T}) –tree \mathbf{A} –alignment Θ –subst. params Λ –indel params

Current models implicitly **condition** on the alignment:

$$P(\mathbf{Y}, \tau, \mathbf{T}, \Theta | \mathbf{A}) = P(\mathbf{Y} | \mathbf{A}, \tau, \mathbf{T}, \Theta) \times P(\tau, \mathbf{T}) \times P(\Theta)$$

However, a joint model explicitly **includes** the alignment:

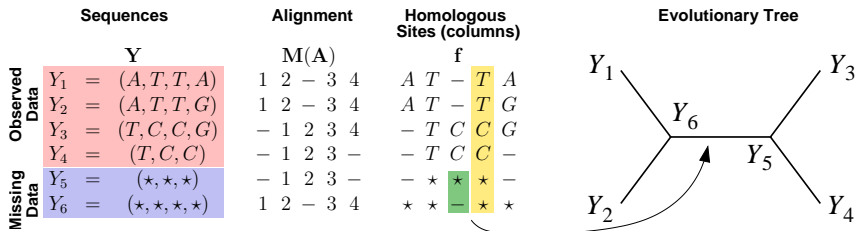
$$P(\mathbf{Y}, \mathbf{A}, \tau, \mathbf{T}, \Theta, \Lambda) = P(\mathbf{Y} | \mathbf{A}, \tau, \mathbf{T}, \Theta) \times P(\mathbf{A} | \tau, \mathbf{T}, \Lambda) \times P(\tau, \mathbf{T}) \\ \times P(\Theta) \times P(\Lambda)$$

Likelihood: $P(\mathbf{Y} | \mathbf{A}, \tau, \mathbf{T}, \Theta)$ — given by the substitution model.

Gap Prior: $P(\mathbf{A} | \tau, \mathbf{T}, \Lambda)$ — given by the indel model.

Alignments on Trees (part I)

- 1 We *augment* the alignment (**A**) by including unobserved sequences at internal nodes of the tree (τ , **T**) – Holmes and Bruno (2001).



- 2 We *separate* the indel process from the substitution process, and the alignment from the aligned data matrix:

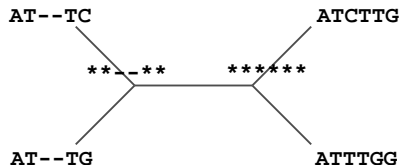
- A character cannot be deleted and re-inserted:
A column cannot contain: * \rightarrow - \rightarrow *
- We do not include '-' as a letter in the alphabet.

Alignments on Trees (part II)

Augmented alignments locate indels on the tree:

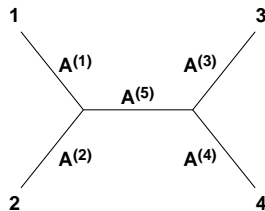
On **which branch** is the indel?

- 1 AT--TC
- 2 AT--TG
- 3 ATCTTG
- 4 ATTTGG



The multiple alignment $\mathbf{A} = (A^{(1)}, A^{(2)}, \dots, A^{(B)})$

- A is made up of **pairwise** alignments on each branch
- Pairwise alignment distribution is based on a pair-HMM:
- $\{A^{(i)}\}$ must agree on lengths of shared sequences



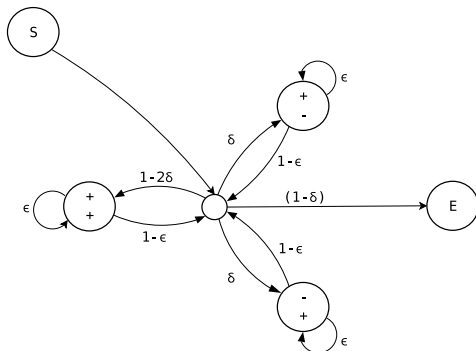
Gap Model: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

- **Pair HMM** model with 2 parameters:

- λ is the indel rate
- ϵ determines average indel length

indel probability $\delta \approx \lambda \times t$



- **Affine** gap penalty $\approx [\log \lambda t] + L \times [\log \epsilon]$

- separates gap *opening* from gap *extension* (different penalties)

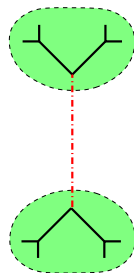
No guide tree → No progressive alignment!

Markov chain Monte Carlo (MCMC) for Sampling/Estimation:

- Specially constructed random walk on $(\mathbf{A}, \tau, \mathbf{T}, \Theta, \Lambda, \mu)$
- Equilibrium distribution is $P(\mathbf{A}, \tau, \mathbf{T}, \Theta, \Lambda, \mu | \mathbf{Y})$

Algorithm: Markov chain Monte Carlo (MCMC)

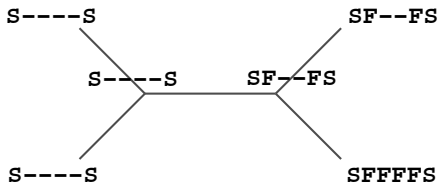
- Start from random tree and alignment
- Propose new trees: **NNI** + **SPR**
- Propose new alignments:
 - 1 Fix pairwise alignments on *most* branches
 - 2 Resample pairwise alignment on *one* branch
 - 3 Use *dynamic programming* with random traceback to re-align the *two sub-alignments*.



Idea: columns labels (F+S)

Extend the alignment: assign each column a label **S** (slow) or **F** (fast).

The form of the alignment **A** changes from



For now, we do not allow rate changes **S** → **F** or **F** → **S** over time.

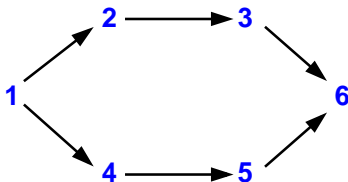
A simple auto-regressive model is not enough...

Felsenstein & Churchill (1996) model spatial structure in *substitution* rates

- The rate in each column depends on the rate in the previous column.
- A Markov chain is placed on substitution rate in each column.

However, in alignments with gaps, the “previous” column is not well-defined:

+	-	-	+	+	+
+	-	-	+	+	+
+	+	+	-	-	+
+	+	+	-	-	+
1	2	3	4	5	6

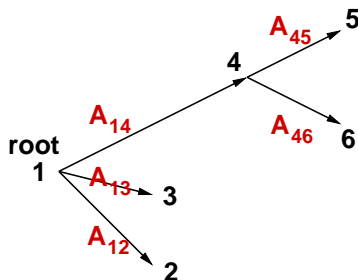


The columns in this alignment are only **partially ordered**:

- Columns (2,3) and (4,5) could be re-ordered so that (4,5) is before (2,3)
- It is unrealistic for the rate in column 4 to depend on the rate in column 3.

Model: Alignments on the tree

We now follow Holmes (2003) in choosing a root (ancestor) node



$$\begin{aligned} \Pr(A) = & \Pr(A_1) \times \\ & \Pr(A_{12}|A_1) \times \Pr(A_{13}|A_1) \times \Pr(A_{14}|A_1) \times \\ & \Pr(A_{45}|A_4) \times \Pr(A_{46}|A_4) \end{aligned}$$

We compute these conditional probabilities with **probabilistic transducers**.

Probabilistic Transducers

Probabilistic transducers have the same structure as pair-HMMs, but are normalized differently.

1. Pair HMMs yield a probability **distribution** $Q(x, y)$ on two sequences x and y

$$\sum_{x,y} Q(x, y) = 1. \quad \Pr(X, Y) = Q(X, Y)$$

2. Transducers yield a probability **measure** $Q(x, y)$ on two sequences x and y

$$\sum_{x,y} Q(x, y) = 1. \quad \Pr(Y|X) = Q(X, Y)$$

- x is the input sequence.
- y is the output sequence.

It is no longer the case that out-going edge probabilities sum to 1.

Transducers: Example

Consider a transducer with states $\begin{matrix} + \\ + \end{matrix}$, $\begin{matrix} + \\ - \end{matrix}$, $\begin{matrix} - \\ + \end{matrix}$, **S**, and **E**.

A sequence of states can represent a pairwise alignment:

$$\mathbf{S} \begin{matrix} + & + & + & - \\ + & - & + & + \end{matrix} \mathbf{E}$$

Probability is:

$$\Pr(\mathbf{S} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}) \Pr(\begin{matrix} + \\ - \end{matrix} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} - \\ + \end{matrix}) \Pr(\begin{matrix} - \\ + \end{matrix} \rightarrow \mathbf{E})$$

What does $\Pr\left(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}\right)$ mean?

Transducers: Example

Consider a transducer with states $\begin{matrix} + \\ + \end{matrix}$, $\begin{matrix} + \\ - \end{matrix}$, $\begin{matrix} - \\ + \end{matrix}$, **S**, and **E**.

A sequence of states can represent a pairwise alignment:

$$\mathbf{S} \begin{matrix} + & + & + & - \\ + & - & + & + \end{matrix} \mathbf{E}$$

Probability is:

$$\Pr(\mathbf{S} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}) \Pr(\begin{matrix} + \\ - \end{matrix} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} - \\ + \end{matrix}) \Pr(\begin{matrix} - \\ + \end{matrix} \rightarrow \mathbf{E})$$

What does $\Pr\left(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}\right)$ mean?

- Given that the previous letter in was **not deleted** in the **descendent...**

Transducers: Example

Consider a transducer with states $\begin{matrix} + \\ + \end{matrix}$, $\begin{matrix} + \\ - \end{matrix}$, $\begin{matrix} - \\ + \end{matrix}$, **S**, and **E**.

A sequence of states can represent a pairwise alignment:

$$\mathbf{S} \begin{matrix} + & + & + & - \\ + & - & + & + \end{matrix} \mathbf{E}$$

Probability is:

$$\Pr(\mathbf{S} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}) \Pr(\begin{matrix} + \\ - \end{matrix} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} - \\ + \end{matrix}) \Pr(\begin{matrix} - \\ + \end{matrix} \rightarrow \mathbf{E})$$

What does $\Pr\left(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}\right)$ mean?

- Given that the previous letter in was **not deleted** in the **descendent**...
- ... and given that there is **another letter** in the **ancestor**...

Transducers: Example

Consider a transducer with states $\begin{matrix} + \\ + \end{matrix}$, $\begin{matrix} + \\ - \end{matrix}$, $\begin{matrix} - \\ + \end{matrix}$, **S**, and **E**.

A sequence of states can represent a pairwise alignment:

$$\mathbf{S} \begin{matrix} + & + & + & - \\ + & - & + & + \end{matrix} \mathbf{E}$$

Probability is:

$$\Pr(\mathbf{S} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}) \Pr(\begin{matrix} + \\ - \end{matrix} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} - \\ + \end{matrix}) \Pr(\begin{matrix} - \\ + \end{matrix} \rightarrow \mathbf{E})$$

What does $\Pr\left(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}\right)$ mean?

- Given that the previous letter in was **not deleted** in the **descendent**...
- ... and given that there is **another letter** in the **ancestor**...
- ... there are no insertions in the **descendent** between the two letters...

Transducers: Example

Consider a transducer with states $\begin{matrix} + \\ + \end{matrix}$, $\begin{matrix} + \\ - \end{matrix}$, $\begin{matrix} - \\ + \end{matrix}$, **S**, and **E**.

A sequence of states can represent a pairwise alignment:

$$\mathbf{S} \begin{matrix} + & + & + & - \\ + & - & + & + \end{matrix} \mathbf{E}$$

Probability is:

$$\Pr(\mathbf{S} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}) \Pr(\begin{matrix} + \\ - \end{matrix} \rightarrow \begin{matrix} + \\ + \end{matrix}) \Pr(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} - \\ + \end{matrix}) \Pr(\begin{matrix} - \\ + \end{matrix} \rightarrow \mathbf{E})$$

What does $\Pr\left(\begin{matrix} + \\ + \end{matrix} \rightarrow \begin{matrix} + \\ - \end{matrix}\right)$ mean?

- Given that the previous letter in was **not deleted** in the **descendent**...
- ... and given that there is **another letter** in the **ancestor**...
- ... there are no insertions in the **descendent** between the two letters...
- ... and the second letter is deleted in the **descendent**.

Transducers: Normalization

The next state in the ancestor could be either $+$ or \mathbf{E} .

- We are conditioning on these options, so...
- Each option must have a total probability of 1.0.

If we ignore insertions, then putting a weight of 1.0 on $+$ gives:

$$\Pr\left(\mathbf{X} \rightarrow \begin{array}{c} + \\ - \end{array}\right) + \Pr\left(\mathbf{X} \rightarrow \begin{array}{c} + \\ + \end{array}\right) = 1.$$

If we ignore insertions, then putting a weight of 1.0 on \mathbf{E} gives:

$$\Pr(\mathbf{X} \rightarrow \mathbf{E}) = 1.$$

Therefore, the sum of transition probabilities from \mathbf{X} may be more than 1.

Transducers: Normalization

The next state in the ancestor could be either $+$ or \mathbf{E} .

- We are conditioning on these options, so...
- Each option must have a total probability of 1.0.

If we ignore insertions, then putting a weight of 1.0 on $+$ gives:

$$\Pr\left(\mathbf{X} \rightarrow \begin{array}{c} + \\ - \end{array}\right) + \Pr\left(\mathbf{X} \rightarrow \begin{array}{c} + \\ + \end{array}\right) = 1.$$

If we ignore insertions, then putting a weight of 1.0 on \mathbf{E} gives:

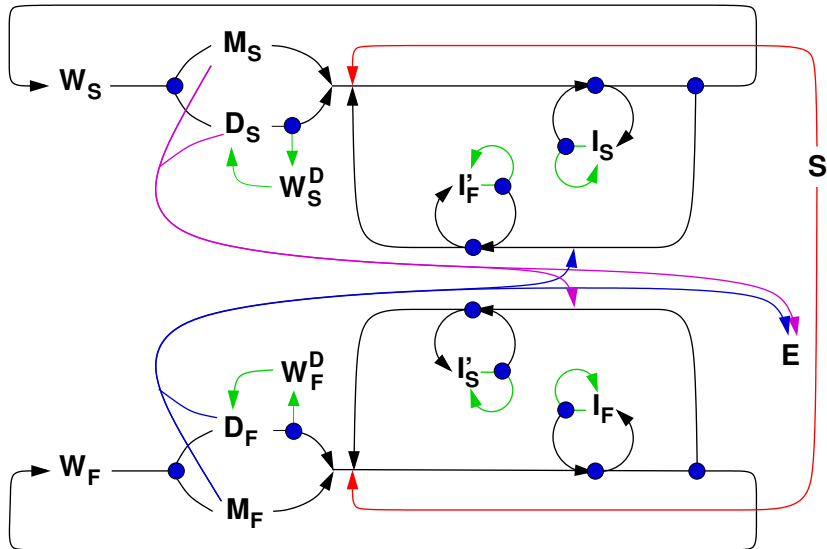
$$\Pr(\mathbf{X} \rightarrow \mathbf{E}) = 1.$$

Therefore, the sum of transition probabilities from \mathbf{X} may be more than 1.

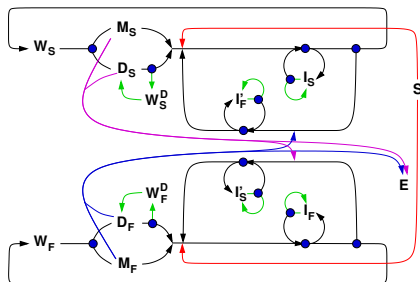
Issue: An insertion $\mathbf{F}[\text{insertion}]\mathbf{S}$ should depend on **both** left and right types

Model - F/S Transducer

Next ancestral letter may be **F**, **S** or **E** (no letter):



Model: Stochastic process for spatially clustered indels?



The transducer above may not represent a **stochastic process**:

- What stochastic process realistically depicts indel hot-spots/cold-spots?
- Can we find a **reversible** model? With a useable **equilibrium** distribution?
- Should we assume that child-insertions occur only on the **right** (or left)?

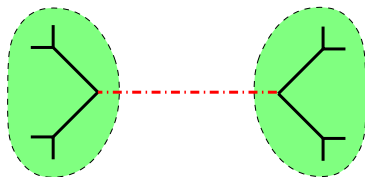
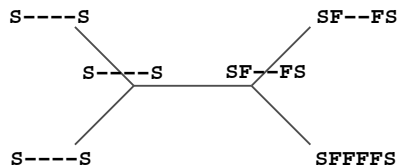
Resampling branch alignments with column F/S labels

Transition kernel HB01a:

Sample the (pairwise) alignment between two sequences connected by a branch.

Column labels add complexity to this transition kernel:

- 1 Fixed column labels inhibit resampling: we cannot align an **F** to an **S**.
- 2 “Unfixed” column labels introduce dependencies between distant branches.
- 3 Branch alignments are no longer independent, conditional on internal node sequence lengths.



Insertions on the fixed branches (black) give information about column labels.

Idea: Gibbs sample column labels

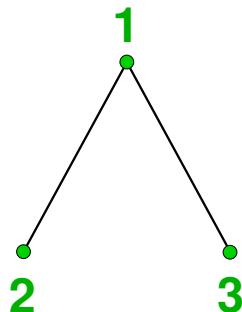
Consider the following alignment on the tree with branches $1 \rightarrow 2$ and $1 \rightarrow 3$:

1	+	-	-	-	-	+
2	+	+	+	-	-	+
3	+	-	-	+	+	+
	1	2	3	4	5	6

$$\mathbf{A}_1 = \sigma \rightarrow 1 \rightarrow 6 \rightarrow \epsilon$$

$$\mathbf{A}_{12} = \begin{array}{cccc} 1 & - & - & 6 \\ 1 & 2 & 3 & 6 \end{array}$$

$$\mathbf{A}_{13} = \begin{array}{cccc} 1 & - & - & 6 \\ 1 & 4 & 5 & 6 \end{array}$$

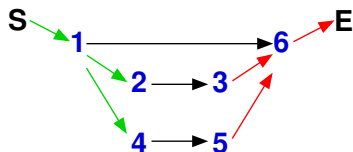


$$\Pr(\mathbf{A}) = \sum_{1,2,3,4,5,6} \Pr(\mathbf{A}_1) \times \Pr(\mathbf{A}_{12}|\mathbf{A}_1) \times \Pr(\mathbf{A}_{13}|\mathbf{A}_1).$$

Idea: Gibbs sample column labels

Consider the following alignment on the tree with branches $1 \rightarrow 2$ and $1 \rightarrow 3$:

1	+	-	-	-	-	+
2	+	+	+	-	-	+
3	+	-	-	+	+	+
	1	2	3	4	5	6

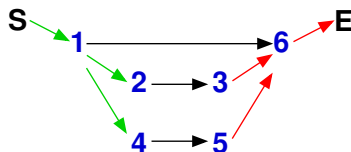


$$\begin{aligned} \Pr(\mathbf{A}) &= \sum_{1,2,3,4,5,6} T(\sigma \rightarrow 1)T(1 \rightarrow 6)T(6 \rightarrow \epsilon) \times \\ &Q(S \rightarrow M_1)Q(M_1 \rightarrow I_{26})Q(I_{26} \rightarrow I_{36})Q(I_{36} \rightarrow M_6)Q(M_6 \rightarrow E) \times \\ &Q(S \rightarrow M_1)Q(M_1 \rightarrow I_{46})Q(I_{46} \rightarrow I_{56})Q(I_{56} \rightarrow M_6)Q(M_6 \rightarrow E). \end{aligned}$$

Idea: Gibbs sample column labels

Consider the following alignment on the tree with branches $1 \rightarrow 2$ and $1 \rightarrow 3$:

1	+	-	-	-	-	+
2	+	+	+	-	-	+
3	+	-	-	+	+	+
	1	2	3	4	5	6

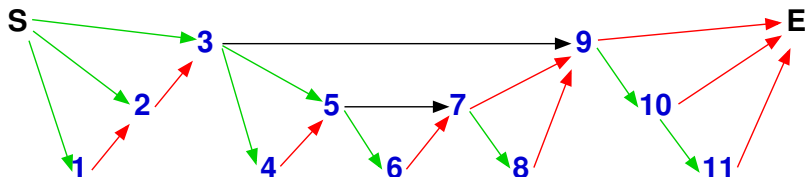


$$\begin{aligned}
 \Pr(\mathbf{A}) = & \sum_{1,6} T(\sigma \rightarrow 1) Q(S \rightarrow M_1)^2 \times \\
 & \left[\sum_3 \left[\sum_2 [Q(M_1 \rightarrow I_{26})]_{126} Q(I_{26} \rightarrow I_{36}) \right]_{136} Q(I_{36} \rightarrow M_6) \right]_{16} \times \\
 & \left[\sum_5 \left[\sum_4 [Q(M_1 \rightarrow I_{46})]_{146} Q(I_{46} \rightarrow I_{56}) \right]_{156} Q(I_{56} \rightarrow M_6) \right]_{16} \times \\
 & \times Q(M_6 \rightarrow E)^2 T(1 \rightarrow 6) T(6 \rightarrow \epsilon)
 \end{aligned}$$

Idea: Gibbs sample column labels

Consider the following alignment on the tree with branches $1 \rightarrow 2 \rightarrow 3$:

1	-	-	+	-	-	-	-	-	+	-	-
2	-	+	+	-	+	-	+	-	+	+	-
3	+	-	+	+	+	+	-	+	+	-	+
	1	2	3	4	5	6	7	8	9	10	11

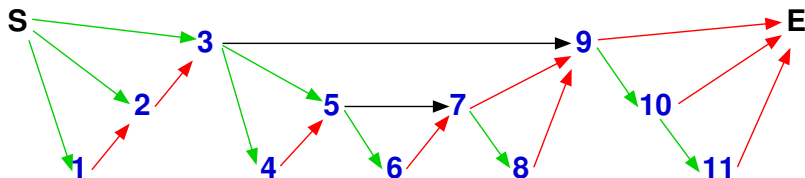


- 1 Each insertion has the **structure** $prev \Rightarrow i \rightarrow j \rightarrow k \Rightarrow next$
- 2 Insertions may be **nested**: $3 \Rightarrow 4 \Rightarrow 5$ and $5 \Rightarrow 6 \Rightarrow 7$ are nested within $3 \Rightarrow 5 \rightarrow 7 \Rightarrow 9$.

Idea: Gibbs sample column labels

Consider the following alignment on the tree with branches $1 \rightarrow 2 \rightarrow 3$:

1	-	-	+	-	-	-	-	-	+	-	-
2	-	+	+	-	+	-	+	-	+	+	-
3	+	-	+	+	+	+	-	+	+	-	+
	1	2	3	4	5	6	7	8	9	10	11



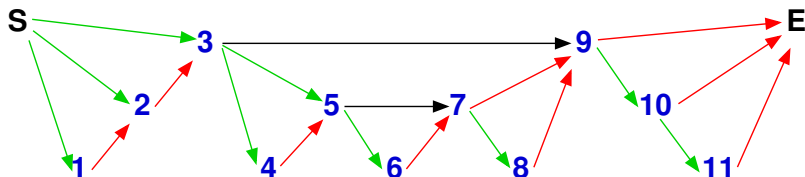
Forward summation to compute probability of insertion $prev \Rightarrow i \rightarrow j \rightarrow k \Rightarrow next$

- 1 We sum over the values of the **middle indices**.
- 2 The probability is **conditional on** the values of *prev* and *next* at each end.
- 3 Process nested insertions **first**: parent sums depend on child sums.

Idea: Gibbs sample column labels

Consider the following alignment on the tree with branches $1 \rightarrow 2 \rightarrow 3$:

1	-	-	+	-	-	-	-	-	+	-	-
2	-	+	+	-	+	-	+	-	+	+	-
3	+	-	+	+	+	+	-	+	+	-	+
	1	2	3	4	5	6	7	8	9	10	11



Backward sampling to compute probability of insertion [$prev \Rightarrow i \rightarrow j \rightarrow k \Rightarrow next$]

- 1 We sample rates **conditional on** the values of $prev$ and $next$ at each end.
- 2 Process nested insertions **last**: children depend on parents.
- 3 We sample rates for each middle index, going backwards from the end.

Creating the Graph

- 1 For each column c
 - 1 For each active branch b in the column
 - 1 What is the previous column p ?
 - 2 If $state(p, b) = I$ and $state(c, b) = I$ then extend the current insertion $p \rightarrow c$
 - 3 If $state(p, b) \neq I$ and $state(c, b) = I$ then start a new insertion $p \Rightarrow c$
 - 4 If $state(p, b) = I$ and $state(c, b) \neq I$ then end the insertion $p \Rightarrow c$

Forward algorithm: $O(N \times L \times R^4)$

- better than $e^{RN} \times L$ for using an evo-HMM
- we can do lookbacks, and so we don't need e^{RN} states to

Backward algorithm: $O(N \times L \times R)$.

Implement the algorithms... **not done**.

Future Work

- 1 Resample **pairwise alignments + labels**.
- 2 Compare to algorithms which hold column labels constant (Rajul)
 - 1 Less constrained mixing, but slower per iteration?
- 3 Stochastic process model? (Anybody have ideas?)
- 4 Decrease number of states in the transducer (Mealy machines vs Moore machines?)
- 5 Effects of prior on estimated rates and representation of each category?

Finally...

- 1 Joint sampling of **alignment + phylogeny + column labels!**
- 2 Does the weight of evidence for shared indels in hotspots decrease?
- 3 How does the alignment of hotspots change?
- 4 Does this result in inferring different trees?
- 5 Can we couple the indel and substitution process through column labels?

Acknowledgments

Postdoc Advisor



Dr. Jeff Thorne

Graduate Advisor



Dr. Marc Suchard

Grants:

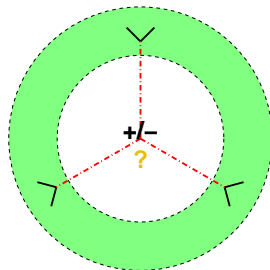
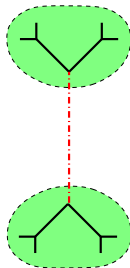
- NIH grant GM070806
- NSF grant DEB-0445180

The End

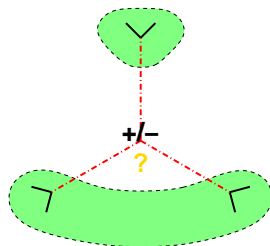
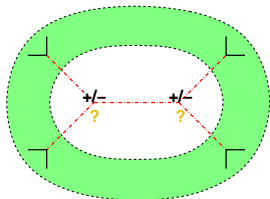
Questions?

Alignment Sampling

Transition kernels from Holmes & Bruno (2001):



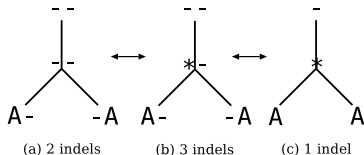
New transition kernels:



$O(L^2)$ sampling from 3-taxon tree

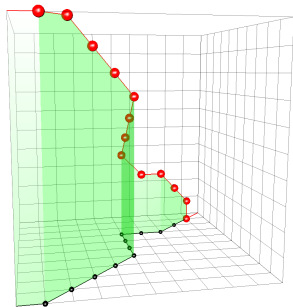
Sampling from a 3-star tree improves mixing:

- Removes a bad intermediate:
- Allows us to unalign/realign
- But its $O(L^3)$! Too expensive...



Solution:

- We constrain the “shadow” of the path, yielding a 2D slice
- Dynamic Programming changes from $O(L^3) \rightarrow O(L^2)$



Enables convergence starting from arbitrary trees & alignments.

Abstract:

Many common probabilistic models of the insertion-deletion process assume that insertions and deletions occur at a rate that is uniform throughout a gene. In protein coding genes, however, indel rates are often elevated in certain regions of the DNA sequence, such as those coding for amino acids that are exposed to solvent in the protein's 3-dimensional structure. Additionally, some viral genes, such as the HIV env gene, may contain indel hotspots as a result of selection for immune escape. The ability to consider spatially varying indel rates is an important biological feature for statistical models to capture. This is because the local indel rate affects both the number of indels that are inferred and the strength of evidence for common ancestry for each shared indel. When indels are clustered in a relatively small fraction of the protein, the unrealistic assumption of a uniform indel rate will lead to an underestimation of the local indel rate for most indels and will therefore lead to an exaggeration of the indel evidence for common ancestry. We therefore seek a model and an inference algorithm that can safely utilize the phylogenetic information in indel events because it is correctly weighted relative to the information in substitution events. At this initial stage, however, we propose to estimate only the indel rates for each column in a fixed alignment.

We consider models of indel rate heterogeneity that allow pooling of local indel events in order to infer spatially local indel rates. We note that inference under

such models is more complex than inference under models of even autoregressive models of substitution rate heterogeneity because a simple autoregressive model cannot be used. We therefore rely on the recently developed probabilistic transducer framework to construct the desired models. We assume that the indel rate in each column is a constant, and augment the alignment with this information. We additionally assume a fixed alignment that includes homology information for internal node sequences on a fixed tree. In this framework we seek to design an MCMC transition kernel that can directly sample the indel rates for all alignment columns from the full conditional distribution in about $O(N*L*R^3)$ time, where N is the number of sequences, L is the length of the alignments, and R is the number of possible indel rates. We will initially focus on models in which $R=2$, and therefore there are only two rates: fast and slow. We also hope to sample the parameters of the indel model including the relative rates of indels in each of the rate class. We anticipate that the prior on the mixture of indel rates may strongly influence the estimates of those rates, and therefore we hope to be able to investigate the effect of the prior on the estimated differences in indel rate between mixture components.

-BenRI