

# Galton-Watson Trees in Statistical Alignment

Dirk Metzler

Johann Wolfgang Goethe-Universität Frankfurt am Main  
Fachbereich Informatik und Mathematik

Workshop on Bayesian Phylogeny  
Budapest, 25.–29. June 2008

# Outline

- 1 Galton Watson Trees
- 2 Insertion-Deletion Models
- 3 Multiple Alignments
- 4 Challenges/Problems: Bayesian Sampling of Multiple Alignments

# Outline

- 1 Galton Watson Trees
- 2 Insertion-Deletion Models
- 3 Multiple Alignments
- 4 Challenges/Problems: Bayesian Sampling of Multiple Alignments

# Galton and Watson

Sir Francis Galton  
1822–1911



Henry William Watson  
1827–1903



# Galton and Watson

Sir Francis Galton  
1822–1911



Henry William Watson  
1827–1903



What's the probability for an English aristocratic family name to die out?

# Galton Watson Tree

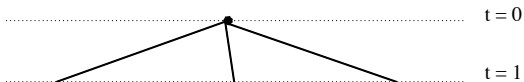
t = 0



$X_k$  := number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

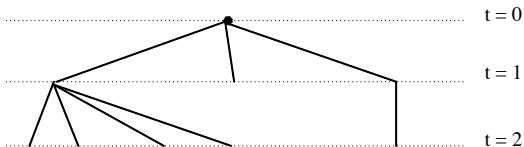
# Galton Watson Tree



$X_k$  := number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

# Galton Watson Tree

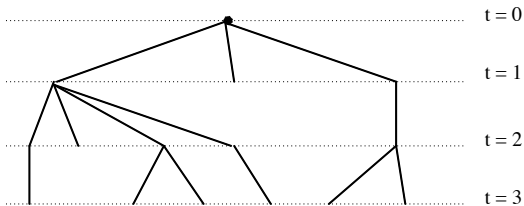


$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables



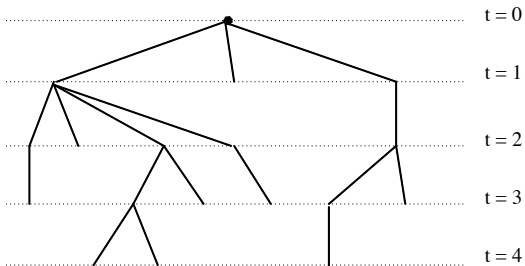
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

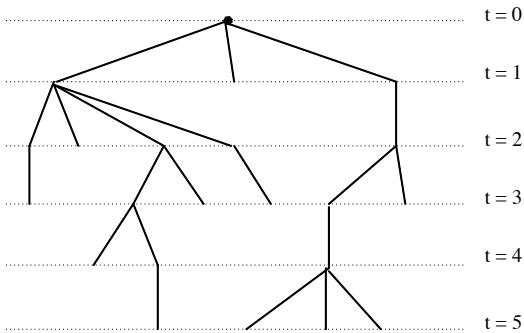
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

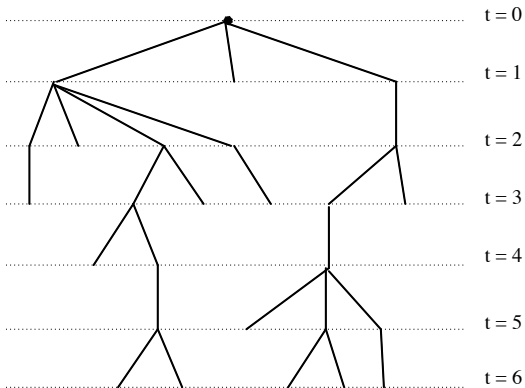
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

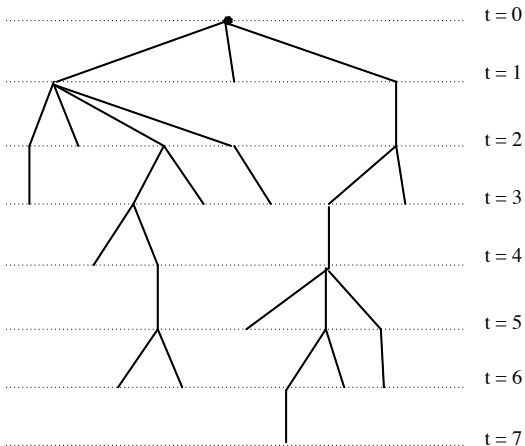
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

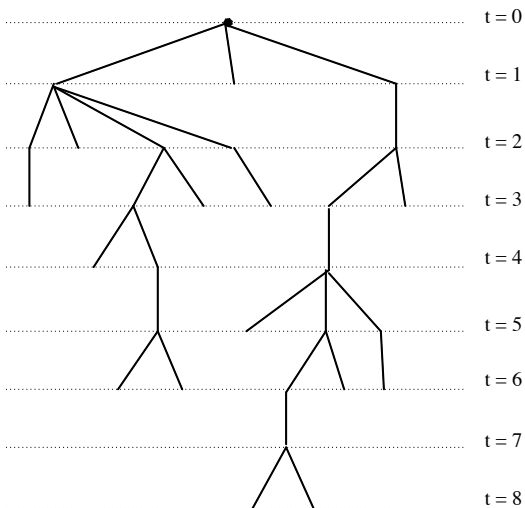
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

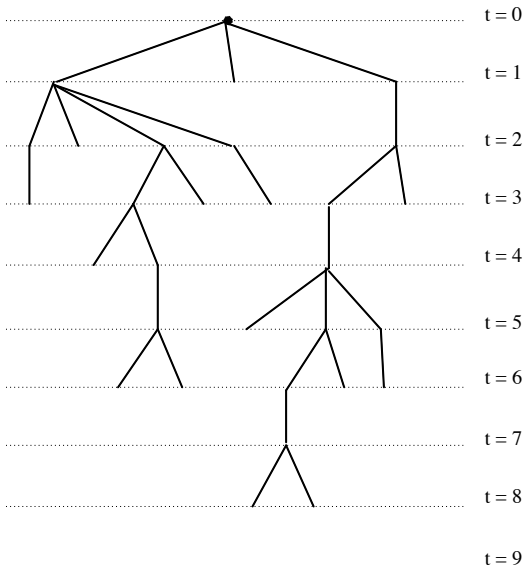
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

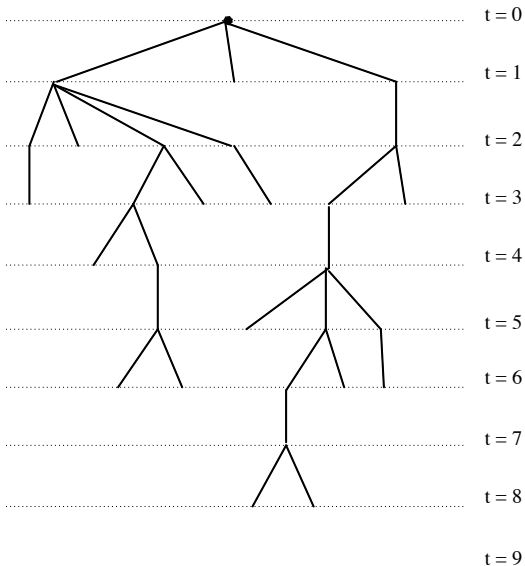
# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

# Galton Watson Tree



$X_k :=$  number of offsprings  
at node  $k$

$X_1, X_2, X_3, \dots$  i.i.d. random  
variables

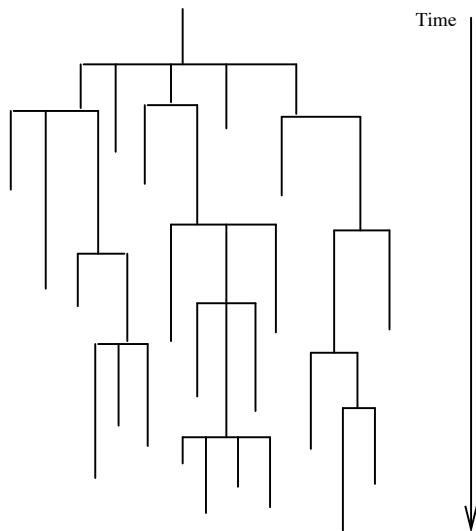
$\mathbb{E}X_k < 1$  : “subcritical”

$\mathbb{E}X_k = 1$  : “critical”

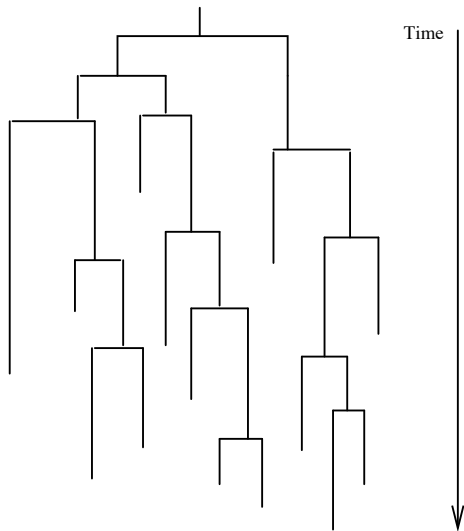
$\mathbb{E}X_k > 1$  : “supercritical”



# Galton-Watson Process in continuous time



# Binary Branching GW Process in cont. time



## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

on  $\{0, 1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^k \cdot p$

## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

$$\text{on } \{0, 1, 2, \dots\}: \quad \Pr(X = k) = (1 - p)^k \cdot p$$

$$\text{on } \{1, 2, \dots\}: \quad \Pr(X = k) = (1 - p)^{k-1} \cdot p$$

## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

on  $\{0, 1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^k \cdot p$ ,  $\mathbb{E}(X) = (1 - p)/p$

on  $\{1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^{k-1} \cdot p$

## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

on  $\{0, 1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^k \cdot p$ ,  $\mathbb{E}(X) = (1 - p)/p$   
on  $\{1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^{k-1} \cdot p$ ,  $\mathbb{E}(X) = 1/p$

## Theorem

*If a Galton-Watson process with binary branching or geometric offspring distribution (on  $\{0, 1, 2, \dots\}$ ) is still alive at time  $t$ , then the number of survivors at time  $t$  is geometrically distributed (on  $\{1, 2, 3, \dots\}$ ).*

on  $\{0, 1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^k \cdot p$ ,  $\mathbb{E}(X) = (1 - p)/p$

on  $\{1, 2, \dots\}$ :  $\Pr(X = k) = (1 - p)^{k-1} \cdot p$ ,  $\mathbb{E}(X) = 1/p$

The geometric distribution is the only one on  $\{(0, )1, 2, 3, \dots\}$  without memory:  $\Pr(X = n + k \mid X > n) = \Pr(X = k)$



# Outline

- 1 Galton Watson Trees
- 2 Insertion-Deletion Models**
- 3 Multiple Alignments
- 4 Challenges/Problems: Bayesian Sampling of Multiple Alignments

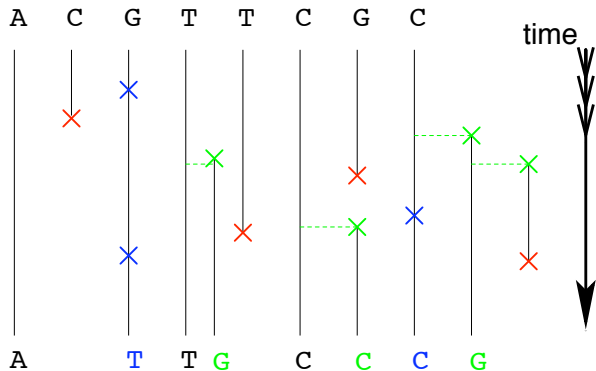
## Model of Sequence Evolution

Thorne, Kishino, Felsenstein (1991):

**Deletions** with rate  $\mu$  at each site.

**Insertions** with rate  $\lambda$  right of each site & at the very left.

**Substitutions** with Rate  $s$  at each site.



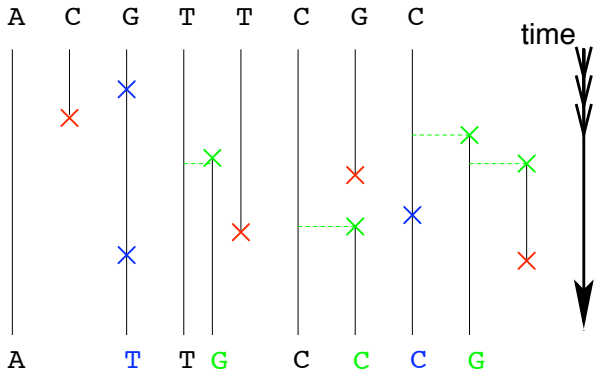
TKF alignment convention:

like this:

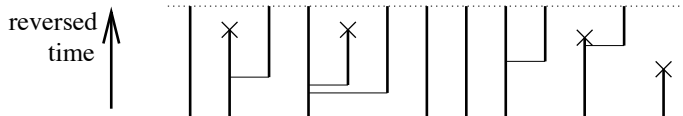
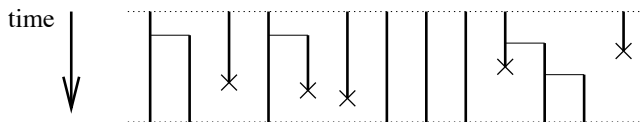
not like this:

ACGT\_TC\_GC\_  
A\_TTG\_CC\_CG

ACGT\_TCG\_C\_  
A\_TTG\_C\_CCG



# Reversibility?

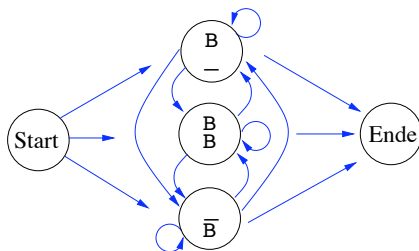


# Consequence of TKF convention

The **bare alignment**

```
BBBB_BB_BB_
B_BBB_BB_BB
```

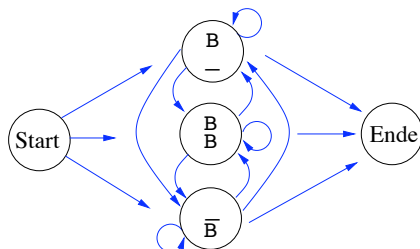
is generated by a **Markov chain**:



# Consequence of TKF convention

The **bare alignment**

BBBB\_BB\_BB\_  
 B\_BBB\_BB\_BB

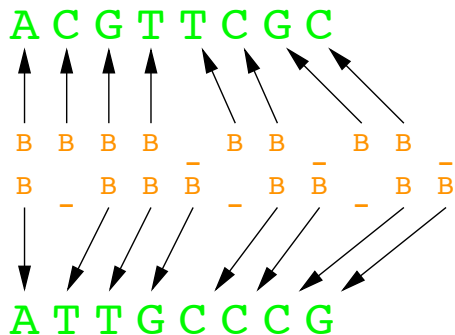


is generated by a **Markov chain**:

from \ to	$\overset{B}{B}$	$\overset{B}{-}$	$\bar{B}$
$\overset{B}{B}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} e^{-\mu}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} (1 - e^{-\mu})$	$\lambda\beta$
$\overset{B}{-}$	$\lambda\beta \frac{e^{-\mu}}{1 - e^{-\mu}}$	$\lambda\beta$	$\frac{1 - e^{-\mu} - \mu\beta}{1 - e^{-\mu}}$
$\bar{B}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} e^{-\mu}$	$(1 - \lambda\beta) \frac{\lambda}{\mu} (1 - e^{-\mu})$	$\lambda\beta$

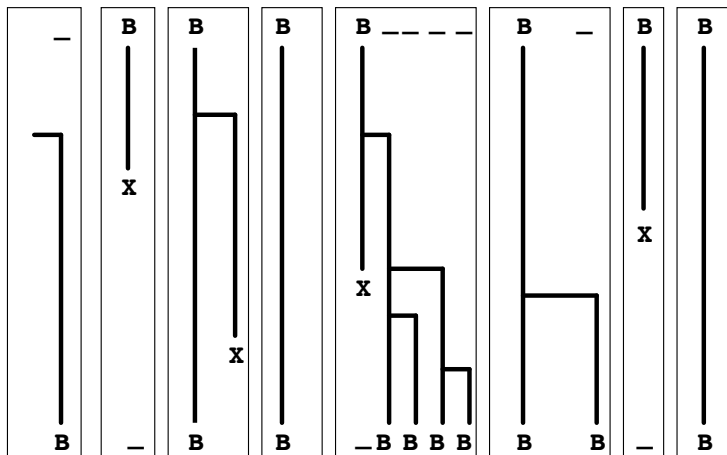
transition probabilities in (model: TKF'91),  $\beta = \frac{1 - e^{\lambda - \mu}}{\mu - \lambda e^{\lambda - \mu}}$

The **Markov chain (the alignment)** is hidden, **observable** is the **pair of sequences** emitted by the alignment.



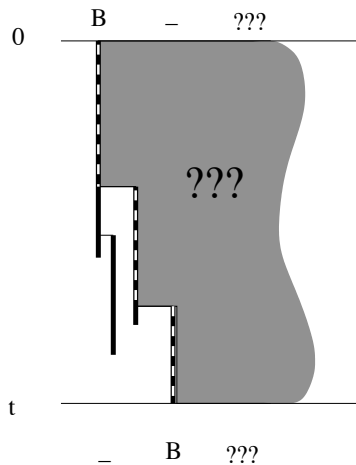
*pair Hidden Markov Model (pair HMM)*

# Why Markov?

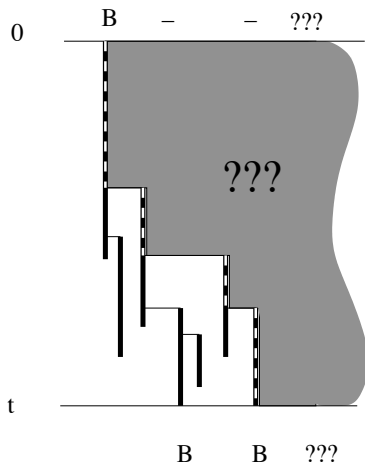




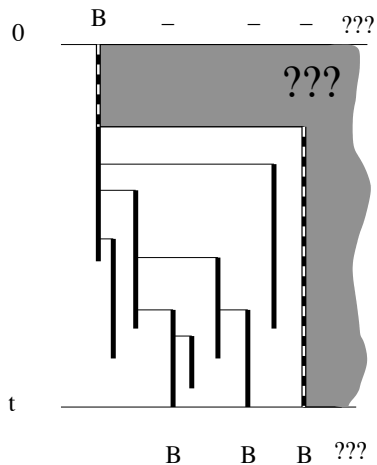
# Why Geometric Distribution?



# Why Geometric Distribution?

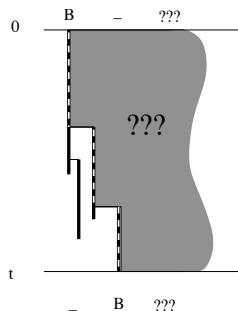


# Why Geometric Distribution?



# Computing transition probabilities

Simplification:  $\lambda = \mu$

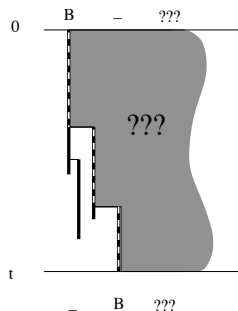


$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$

$$\mathbb{E}(X) = 1$$



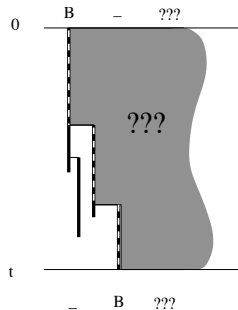
$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$

$$\mathbb{E}(X) = 1$$

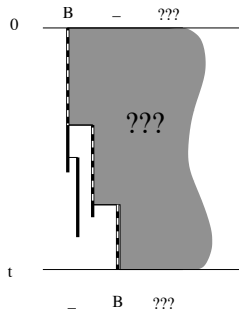
$$\Pr(X = k \mid X > 0) = (1 - p)^{k-1} \cdot p$$



$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$



$$\mathbb{E}(X) = 1$$

$$\Pr(X = k \mid X > 0) =$$

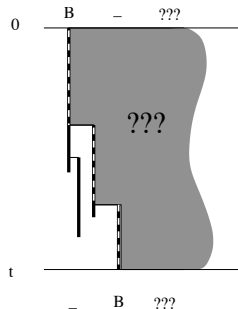
$$(1 - p)^{k-1} \cdot p$$

$$\frac{1}{p} = \mathbb{E}(X \mid X > 0) = 1 + t \cdot \lambda$$

$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$



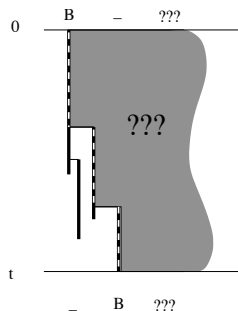
$$\begin{aligned} \mathbb{E}(X) &= 1 \\ \Pr(X = k \mid X > 0) &= \\ &= (1 - p)^{k-1} \cdot p \\ \frac{1}{p} &= \mathbb{E}(X \mid X > 0) = 1 + t \cdot \lambda \\ \Rightarrow p &= 1 / (1 + t \cdot \lambda) \end{aligned}$$

$X :=$  number of survivors at time  $t$



# Computing transition probabilities

Simplification:  $\lambda = \mu$



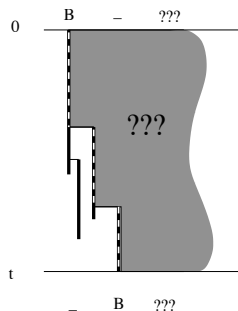
$$\begin{aligned}\mathbb{E}(X) &= 1 \\ \Pr(X = k \mid X > 0) &= \\ &= (1 - p)^{k-1} \cdot p \\ \frac{1}{p} = \mathbb{E}(X \mid X > 0) &= 1 + t \cdot \lambda \\ \Rightarrow p &= 1 / (1 + t \cdot \lambda)\end{aligned}$$

$$\begin{aligned}1 &= \mathbb{E}(X) \\ &= \Pr(X = 0) \cdot \mathbb{E}(X \mid X = 0) \\ &\quad + \Pr(X > 0) \cdot \mathbb{E}(X \mid X > 0)\end{aligned}$$

$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$



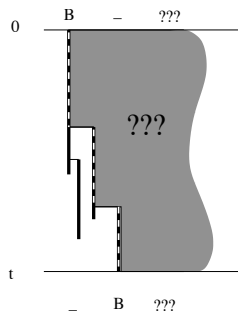
$$\begin{aligned}\mathbb{E}(X) &= 1 \\ \Pr(X = k \mid X > 0) &= (1 - p)^{k-1} \cdot p \\ \frac{1}{p} = \mathbb{E}(X \mid X > 0) &= 1 + t \cdot \lambda \\ \Rightarrow p &= 1 / (1 + t \cdot \lambda)\end{aligned}$$

$$\begin{aligned}1 &= \mathbb{E}(X) \\ &= \Pr(X = 0) \cdot \mathbb{E}(X \mid X = 0) \\ &\quad + \Pr(X > 0) \cdot \mathbb{E}(X \mid X > 0) \\ &= \Pr(X > 0) \cdot (1 + t \cdot \lambda)\end{aligned}$$

$X :=$  number of survivors at time  $t$

# Computing transition probabilities

Simplification:  $\lambda = \mu$

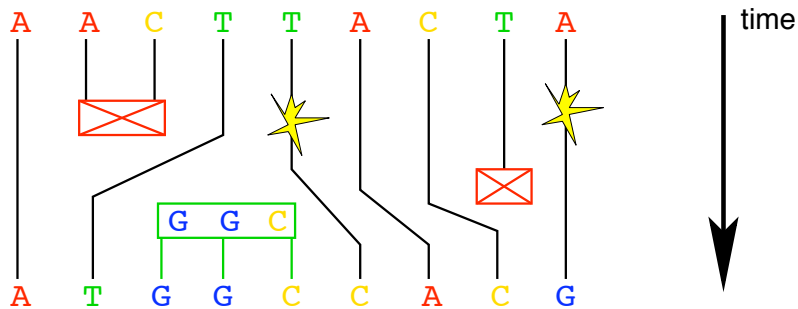


$$\begin{aligned} \mathbb{E}(X) &= 1 \\ \Pr(X = k \mid X > 0) &= \\ &= (1 - p)^{k-1} \cdot p \\ \frac{1}{p} = \mathbb{E}(X \mid X > 0) &= 1 + t \cdot \lambda \\ \Rightarrow p &= 1/(1 + t \cdot \lambda) \end{aligned}$$

$$\begin{aligned} 1 &= \mathbb{E}(X) \\ &= \Pr(X = 0) \cdot \mathbb{E}(X \mid X = 0) \\ &\quad + \Pr(X > 0) \cdot \mathbb{E}(X \mid X > 0) \\ &= \Pr(X > 0) \cdot (1 + t \cdot \lambda) \\ \Rightarrow \Pr(X > 0) &= 1/(1 + t \cdot \lambda) \end{aligned}$$

$X :=$  number of survivors at time  $t$

# InDels are usually longer than 1 position



J.L. Thorne, H. Kishino, J. Felsenstein (1992) Inching towards reality: an improved likelihood model for sequence evolution. *J. Mol. Evol.*, **34**, 3-16.

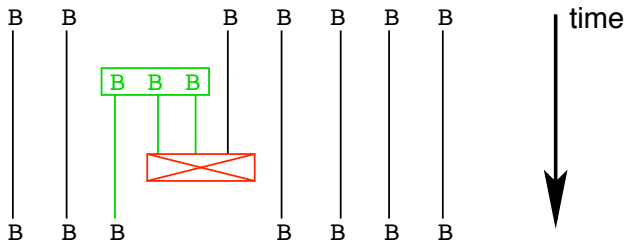
D. Metzler (2003) Statistical alignment based on fragment insertion and deletion models, *Bioinformatics* 19:490-499.

FID Model (also a pairHMM):

- instead of single nucleotides, fragments are inserted and deleted with rate  $\lambda$ .
- Length of the fragments: geometrically distributed, mean length:  $\gamma$ .

$$\Pr(L = k) = \frac{1}{\gamma} \left(1 - \frac{1}{\gamma}\right)^k$$

forbidden in TKF92 and FID:



GID Model:

- ↑ this is allowed
- **no hidden Markov structure**

Use GID to simulate data  
and test robustness of FID

How good are FID-based methods when GID/“Long Indel Model” is true?

- no problem for parameter estimations (Metzler, 2003)
- alignment accuracy can be decreased (Miklos, Lunter, Holmes, 2004)

How good are FID-based methods when GID/“Long Indel Model” is true?

- no problem for parameter estimations (Metzler, 2003)
- alignment accuracy can be decreased (Miklos, Lunter, Holmes, 2004)

Maybe generate mixed-geometric gap-length with different types of fragments.



How good are FID-based methods when GID/“Long Indel Model” is true?

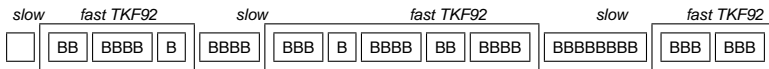
- no problem for parameter estimations (Metzler, 2003)
- alignment accuracy can be decreased (Miklos, Lunter, Holmes, 2004)

Maybe generate mixed-geometric gap-length with different types of fragments.

Along a tree fragmentation may change from edge to edge.

# InDel Model for detecting conserved regions

A. Arribas-Gil, D. Metzler, J.-L. Plouhinec (2007)

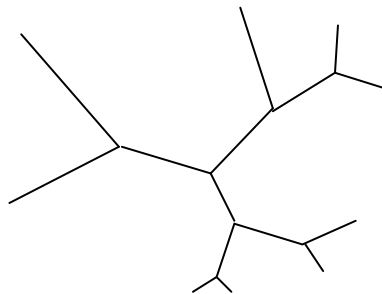


# Outline

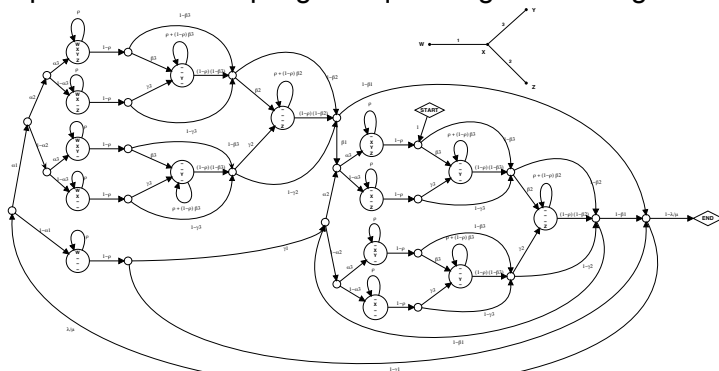
- 1 Galton Watson Trees
- 2 Insertion-Deletion Models
- 3 Multiple Alignments**
- 4 Challenges/Problems: Bayesian Sampling of Multiple Alignments

I. Holmes, W. J. Bruno (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment, *Bioinformatics* 17:803-820.

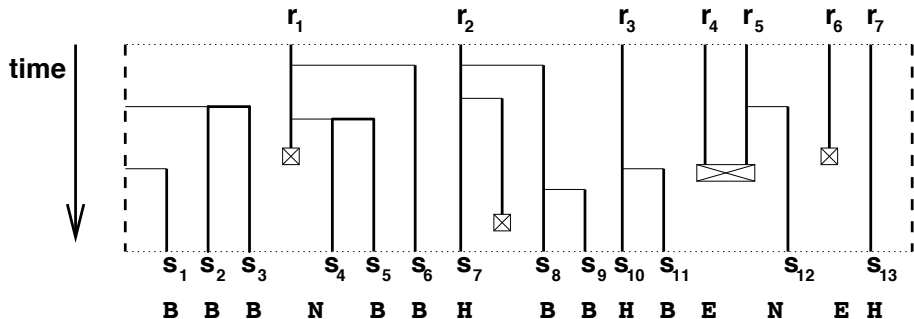
R. Fleißner, D. Metzler, A. von Haeseler (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology* 54(4):548-61.

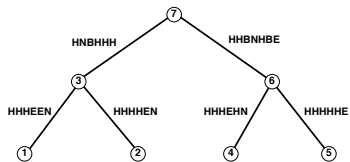
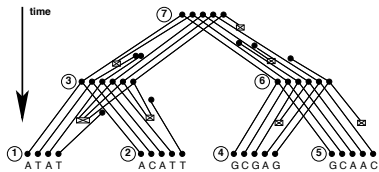


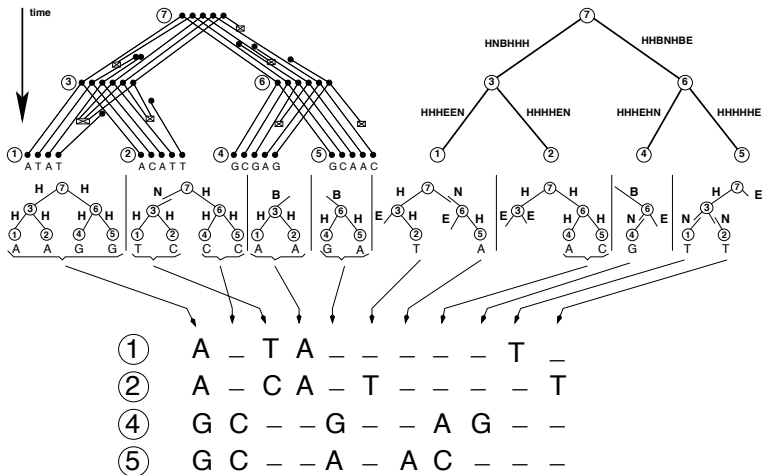
## multiple HMM for sampling a sequence given its neighbours



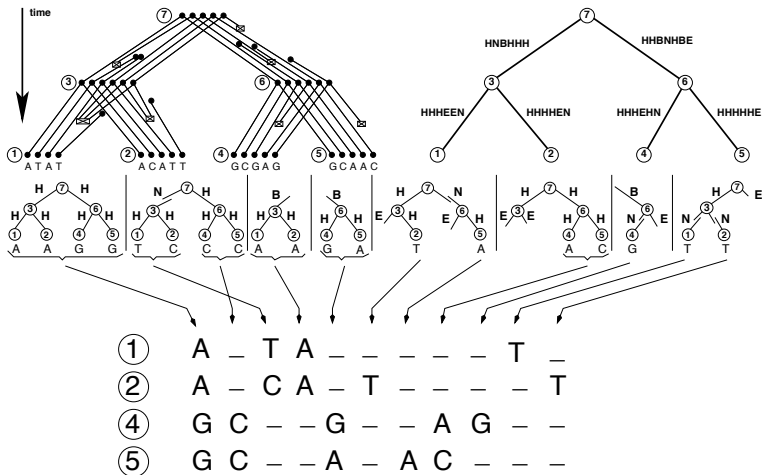
G.A. Lunter, I. Miklós, Y.S. Song, J. Hein (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.* 10(6):869-889.



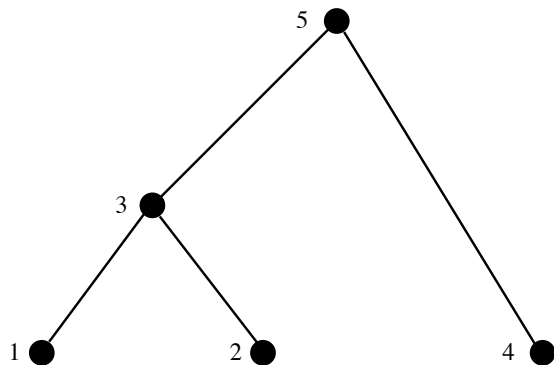


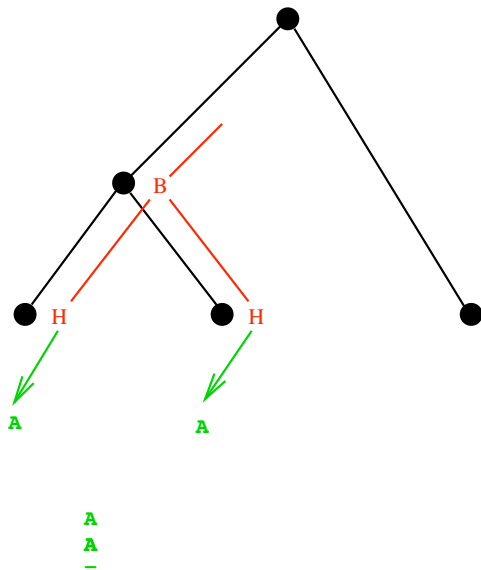


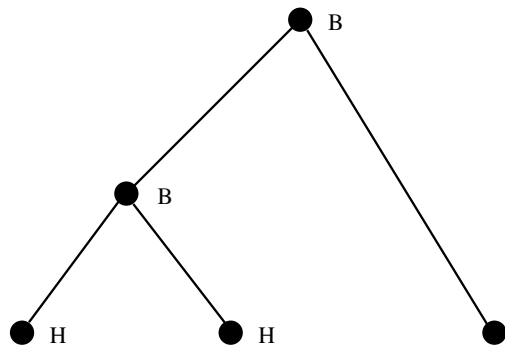




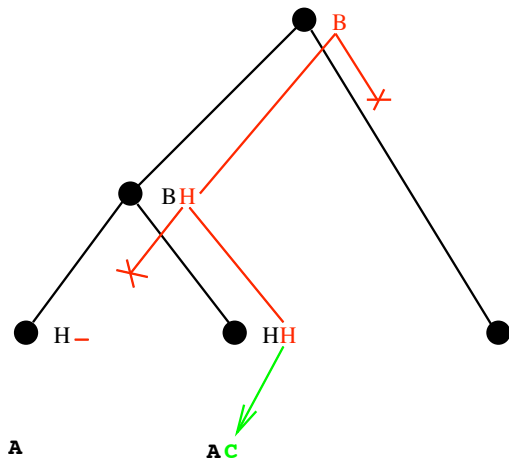
Tree-Indexed Heirs Line =: TIHL



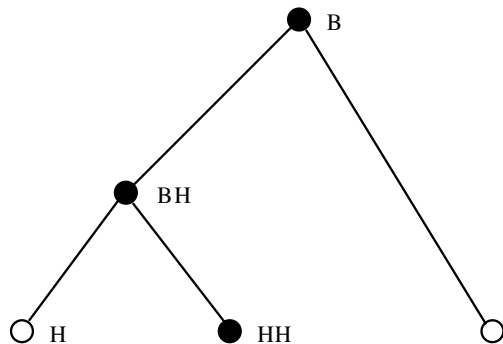


**A****A****A****A**

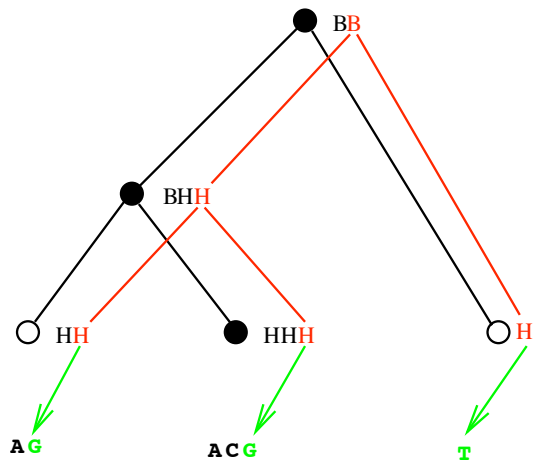
-



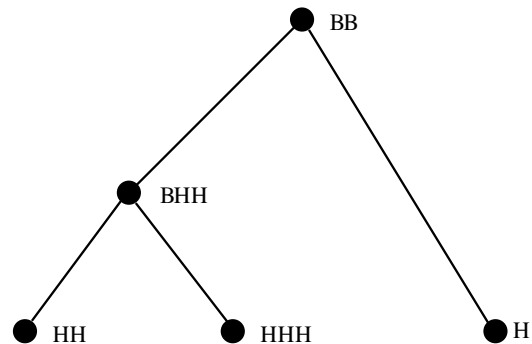
A -  
 A C  
 - -

**A****AC**

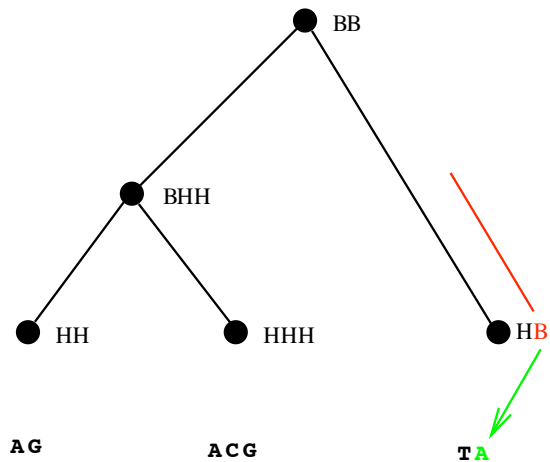
**A -**  
**A C**  
**- -**



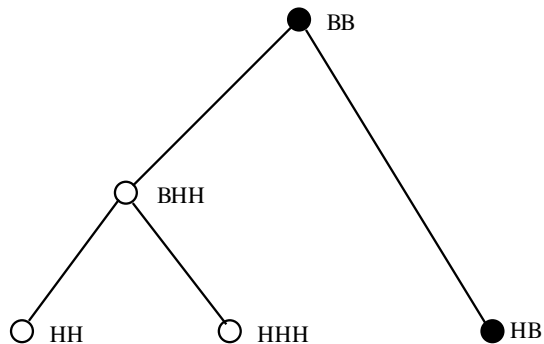
**A - G**  
**A C G**  
**- - T**

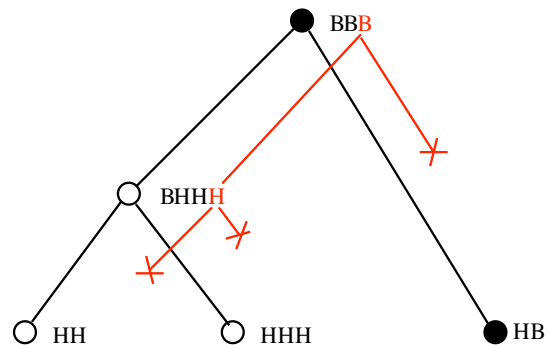
**AG****ACG****T****A - G****A C G****- - T**

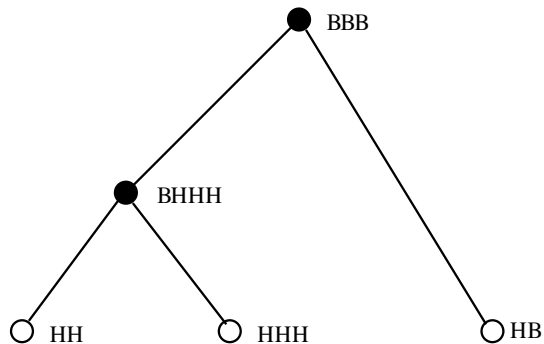


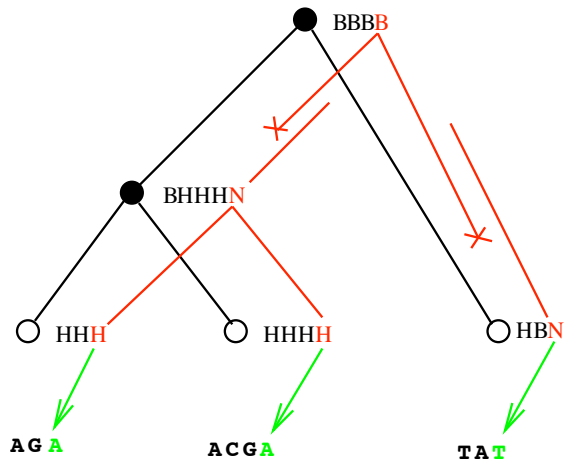


**A - G A**  
**A C G -**  
**- - T -**

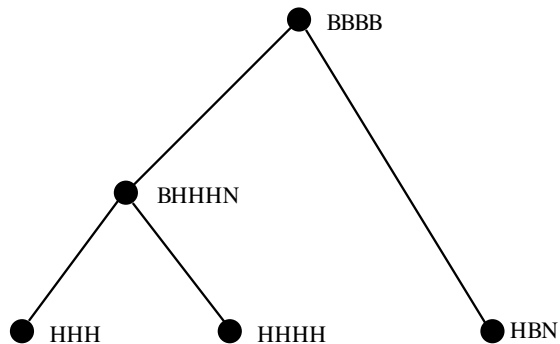
**AG****ACG****TA****A - G A****A C G -****- - T -**

**AG****ACG****TA****A - G A****A C G -****- - T -**

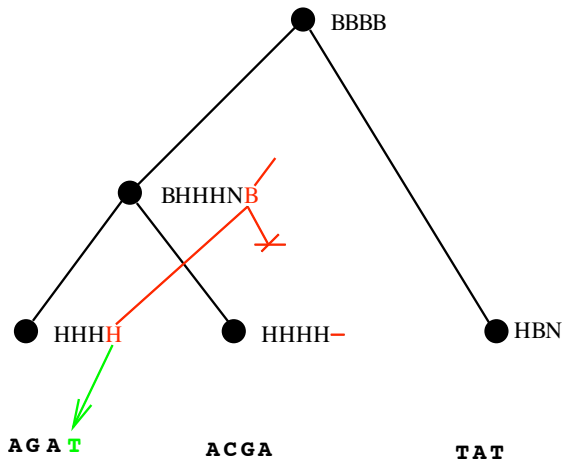
**AG****ACG****TA****A - G A****A C G -****- - T -**

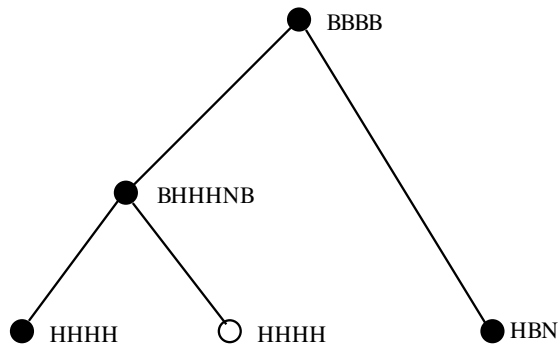


**A - G A A -**  
**A C G - A -**  
**- - T - - T**

**AG A****ACGA****TAT**

**A - G A A -**  
**A C G - A -**  
**- - T - - T**





**AGAT**

**ACGA**

**TAT**

**A - G A A \_ T**

**A C G - A - -**

**- - T - - T -**



TKF91: states of hidden Markov chain are the **Sets Of Active Nodes (soans)**.

$$P_S(k) = \sum_{(\mathcal{R}, e) : \mathcal{S}=[\mathcal{R}, e]} p(e)q(e)P_{\mathcal{R}}(k - v_e)\vartheta(e, k)$$

where

$k$  : Multi-index of Positions in sequences at leaves

$\mathcal{S} = [\mathcal{R}, e]$  : tihl  $e$  turns soan  $\mathcal{S}$  into soan $\mathcal{R}$

$P_S(k)$  : Pr(sequences up to  $k$  are generated and end there)

$p(e)$  = Pr(indel history of  $e$ )

$q(e)$  = Pr(no inserts at nodes in  $e$ )

$\vartheta(e, k)$  = Pr( $e$  emits base given in data types at  $k$ )

$v_e \in \{0, 1\}^n$  : indicates postions in leaf-sequences to which  $e$  emits

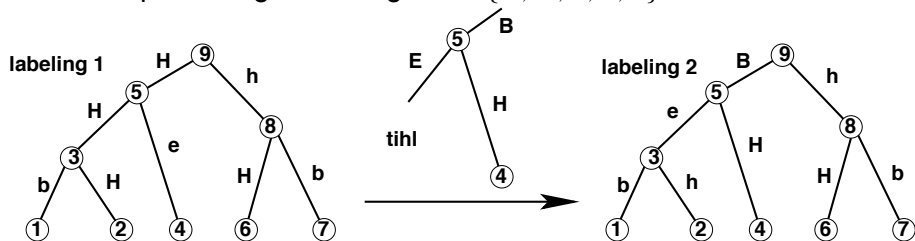
TKF91: states of hidden Markov chain are the [Sets Of Active Nodes \(soans\)](#).

TKF91: states of hidden Markov chain are the [Sets Of Active Nodes \(soans\)](#).

**Transfer this to FID or TKF92** (fragmentation may change from edge to edge)

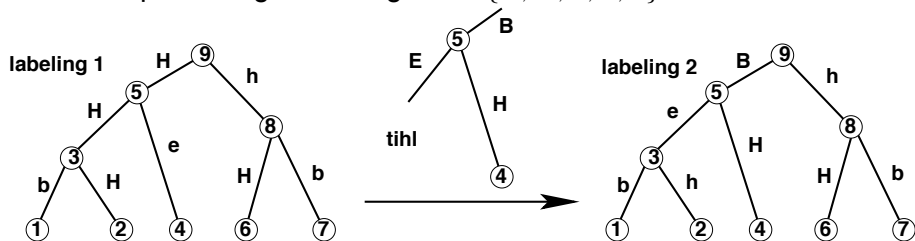
- D. Metzler, R. Fleißner, A. Wakolbinger, A. von Haeseler (2005) Stochastic insertion-deletion processes and statistical sequence alignment.
- D. Metzler, R. Fleißner (2007) Sequence Evolution Models for Simultaneous Alignment and Phylogeny Reconstruction.

state space: edge-labellings with  $\{B, H, e, b, h\}$ .



tihl = tree indexed heirs line

state space: edge-labellings with  $\{B, H, e, b, h\}$ .



tihl = tree indexed heirs line

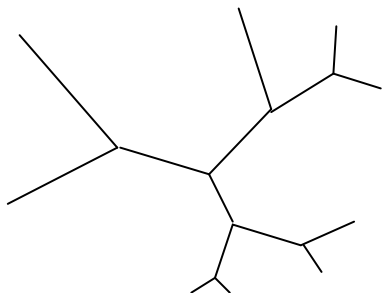
Example: 3-leaved tree

TKF91:  $2^3 = 8$  possible sets of active nodes

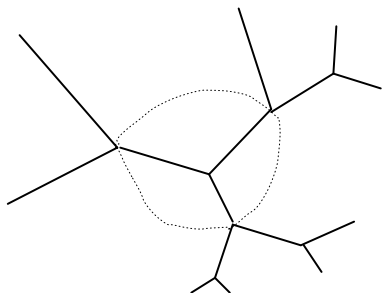
TKF92/FID:  $5^3 = 125$  possible labellings, 41 of them are relevant

# Outline

- 1 Galton Watson Trees
- 2 Insertion-Deletion Models
- 3 Multiple Alignments
- 4 Challenges/Problems: Bayesian Sampling of Multiple Alignments**



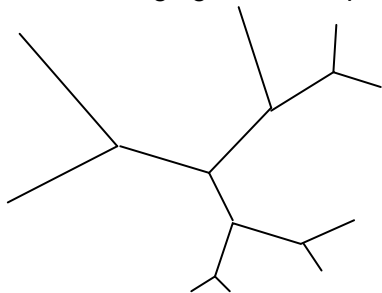
- re-sample alignments of 3-star subtrees (like J.L. Jensen and J. Hein, 2005, do for TKF91)
- do this only for limited parts of the sequences
- Can non-emitting tips be ignored?
- assigning sequences to internal nodes or use nucleotide (or AA) distributions?



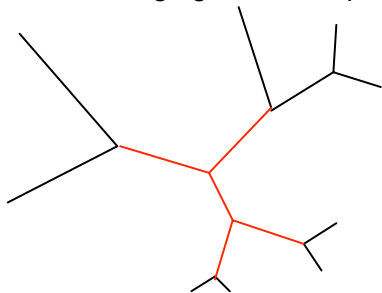
- re-sample alignments of 3-star subtrees (like J.L. Jensen and J. Hein, 2005, do for TKF91)
- do this only for limited parts of the sequences
- Can non-emitting tihs be ignored?
- assing sequences to internal nodes or use nucleotide (or AA) distributions?



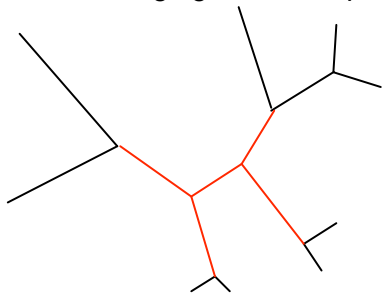
When changing the tree topology...



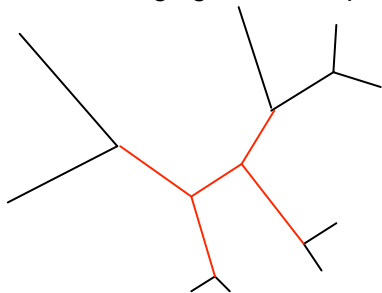
When changing the tree topology...



When changing the tree topology...



When changing the tree topology...



...keep alignments of exterior sequences fixed.  
(TKF91: 32 SOANS; FID: 437 relevant labellings)

# Conclusions

- We need multiple-alignment sampling to assess the full uncertainty in phylogeny estimation
- Let's write the software and try if it works!
- THANK YOU!

# Conclusions

- We need multiple-alignment sampling to assess the full uncertainty in phylogeny estimation
- Let's write the software and try if it works!
- THANK YOU!

# Conclusions

- We need multiple-alignment sampling to assess the full uncertainty in phylogeny estimation
- Let's write the software and try if it works!
- THANK YOU!