# Insertion-deletion models for sequence evolution and Bayesian sampling methods for multiple alignment

Dirk Metzler

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

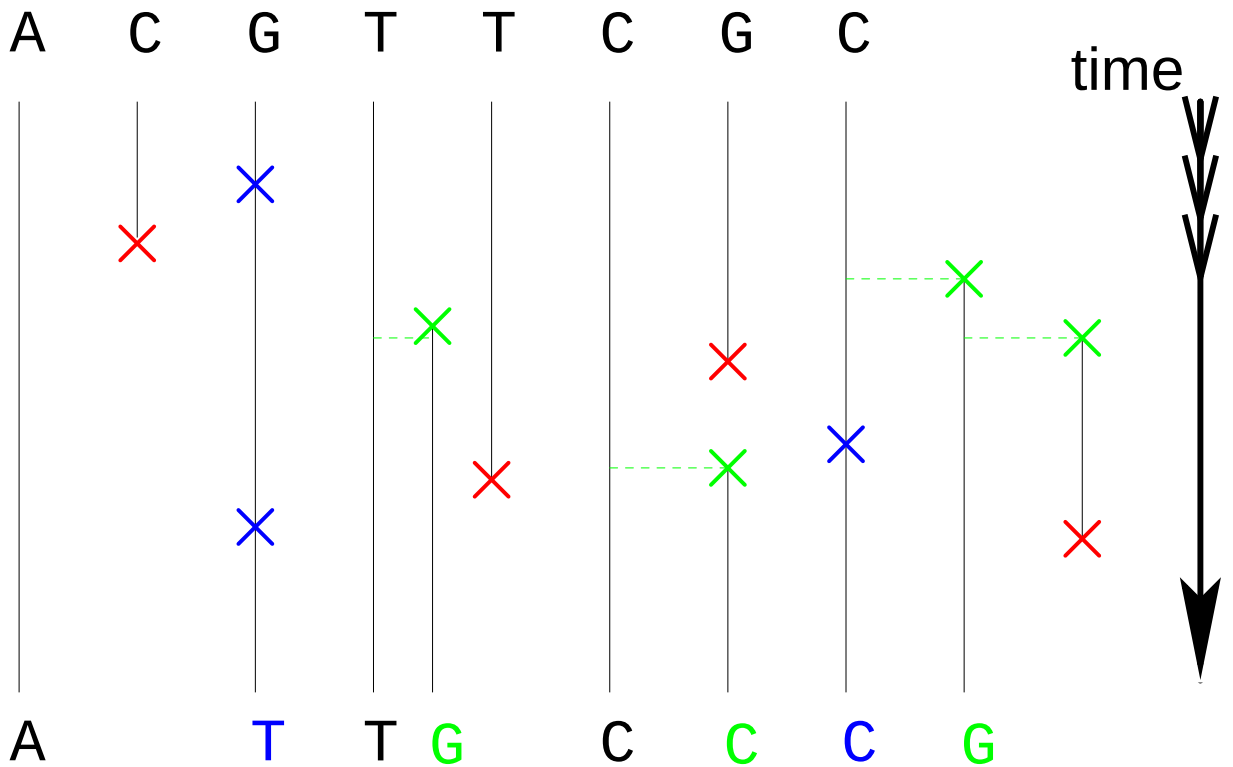
Newton Institute, Cambridge, Sept. 2007

# Outline

# Outline

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.



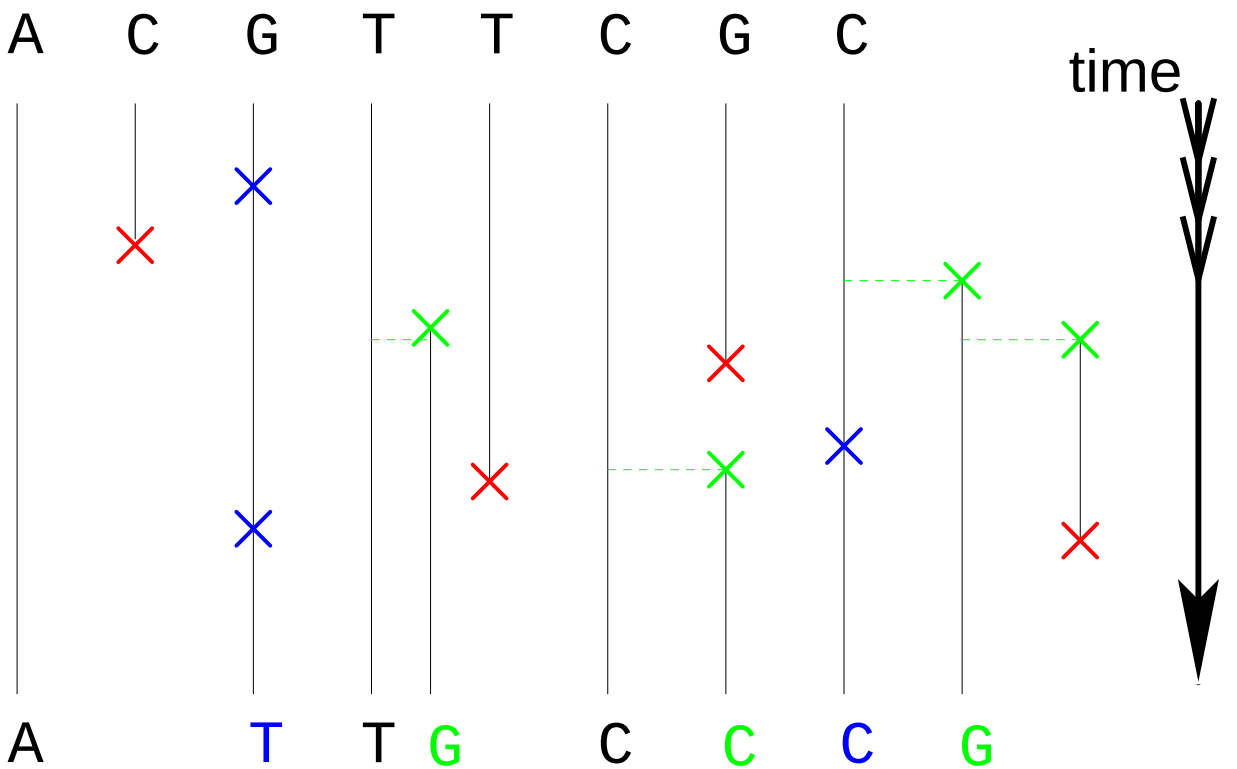
TKF alignment convention:

like this:

```
ACGT_TC_GC_
A_TTG_CC_CG
```

not like this:

```
ACGT_TCG_C_
A_TTG_C_CCG
```

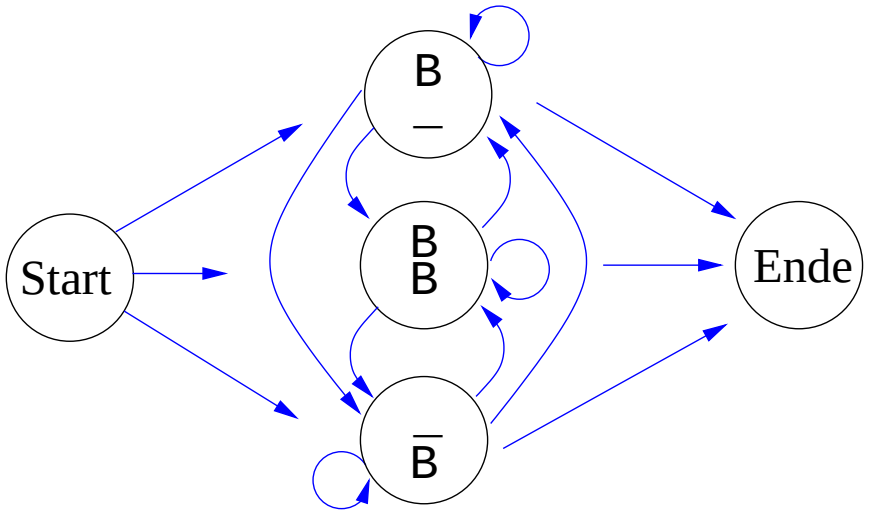


# Consequence of TKF convention

The bare alignment

BBBB_BB_BB_
B_BBB_BB_BB

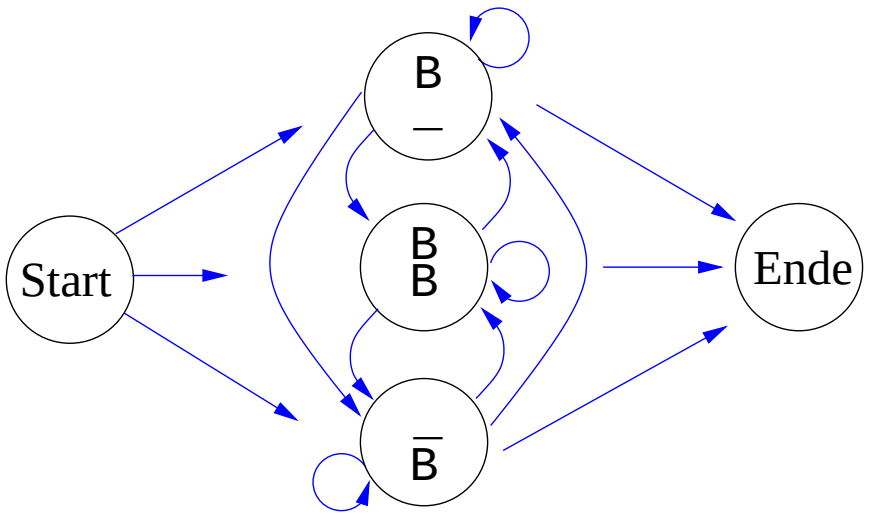
is generated by a Markov chain:



# Consequence of TKF convention

The bare alignment

BBBB_BB_BB_
B_BBB_BB_BB

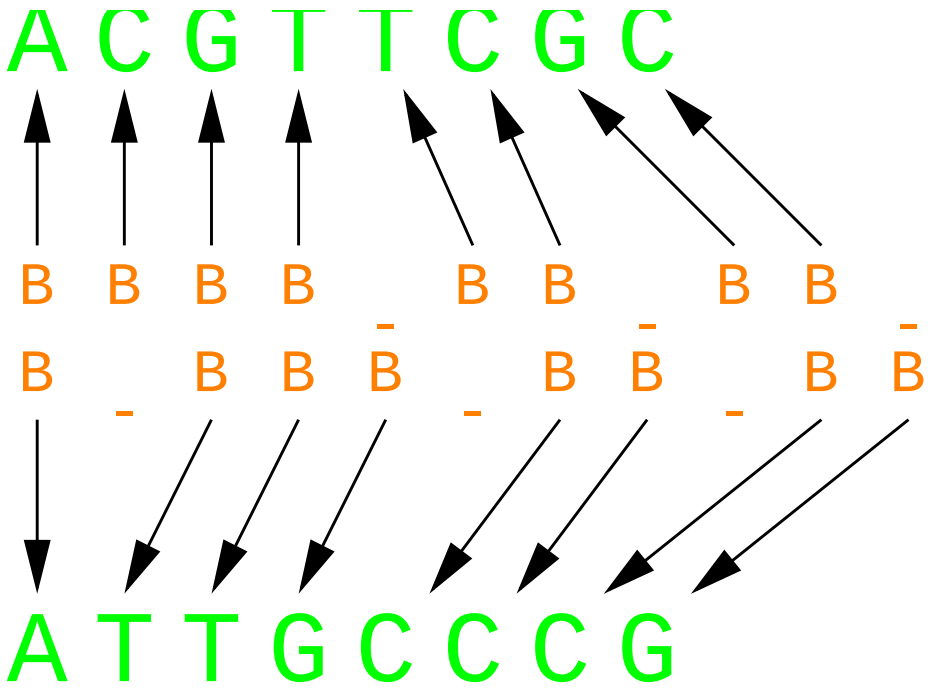



is generated by a Markov chain:

| from \ to | $\begin{smallmatrix}\mathrm{B}\\ \mathrm{B}\end{smallmatrix}$ | $\begin{smallmatrix}\mathrm{B}\\ -\end{smallmatrix}$ | $\overline{\mathrm{B}}$ |
|---|---|---|---|
| $\begin{smallmatrix}\mathrm{B}\\ \mathrm{B}\end{smallmatrix}$ | $(1-\lambda\beta)\frac{\lambda}{\mu}e^{-\mu}$ | $(1-\lambda\beta)\frac{\lambda}{\mu}(1-e^{-\mu})$ | $\lambda\beta$ |
| $\begin{smallmatrix}\mathrm{B}\\ -\end{smallmatrix}$ | $\lambda\beta\frac{e^{-\mu}}{1-e^{-\mu}}$ | $\lambda\beta$ | $\frac{1-e^{-\mu}-\mu\beta}{1-e^{-\mu}}$ |
| $\overline{\mathrm{B}}$ | $(1-\lambda\beta)\frac{\lambda}{\mu}e^{-\mu}$ | $(1-\lambda\beta)\frac{\lambda}{\mu}(1-e^{-\mu})$ | $\lambda\beta$ |

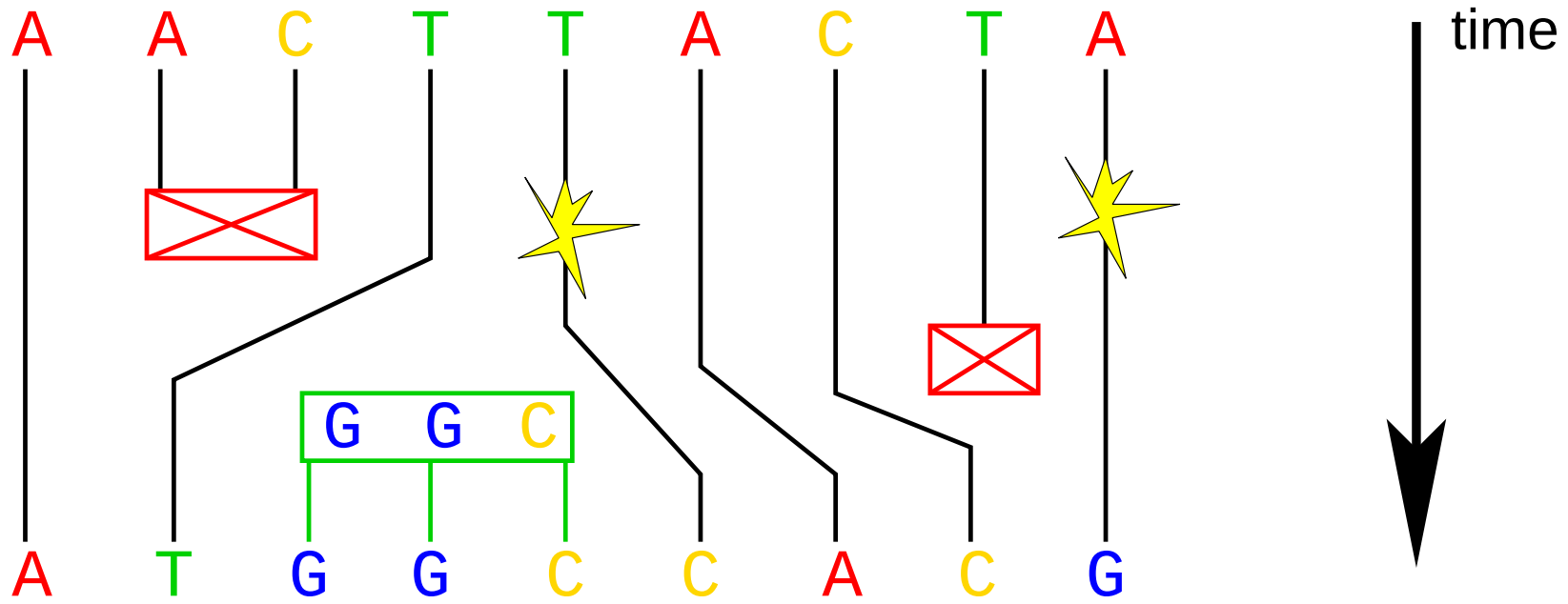
transition probabilies im (model: TKF'91), $\beta = \frac{1-e^{\lambda-\mu}}{\mu-\lambda e^{\lambda-\mu}}$

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




*pair Hidden Markov Model (pair HMM)*

# InDels are usually longer than 1 position



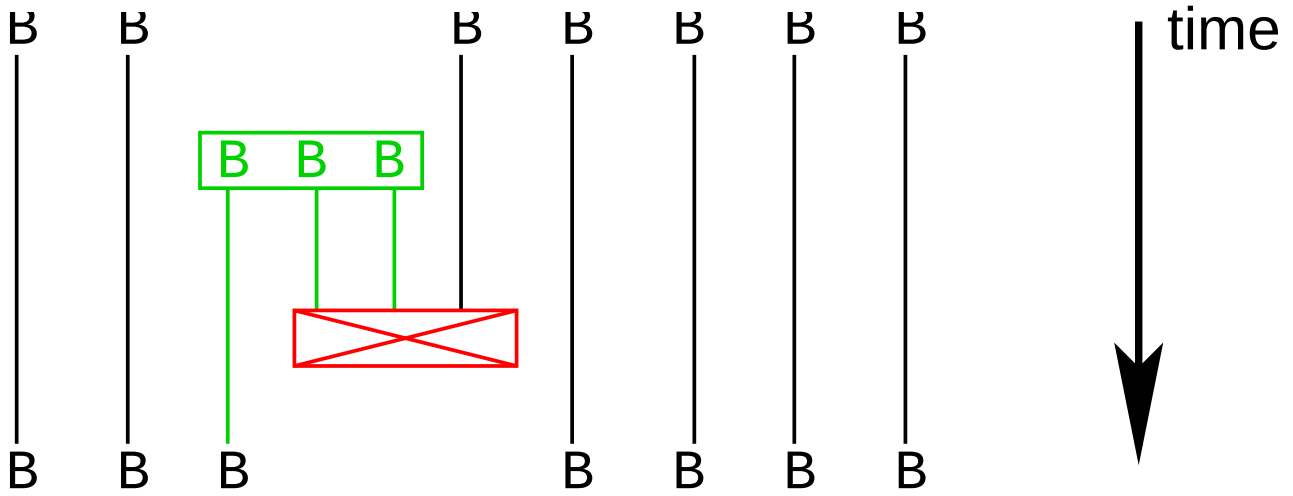
J.L. Thorne, H. Kishino, J. Felsenstein (1992) Inching towards reality: an improved likelihood model for sequence evolution. *J. Mol. Evol.*, **34**, 3-16.
D. Metzler (2003) Statistical alignment based on fragment insertion and deletion models, *Bioinformatics* 19:490-499.

FID Model (also a pairHMM):

- instead of single nucleotides, fragments are inserted an 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)

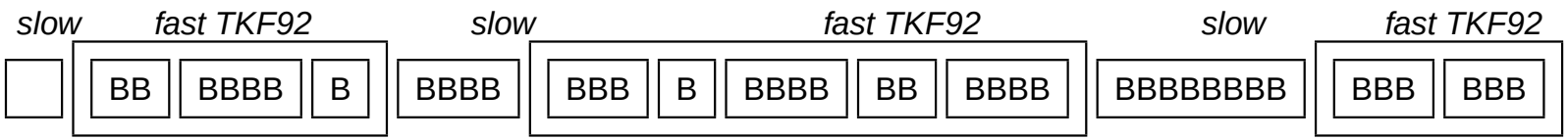
How good are FID-based methods when GID/“Long Indel Model” is true?

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

Maybe generate mixed-geometric gap-length with two types of fragments

# InDel Model for detecting conserved regions

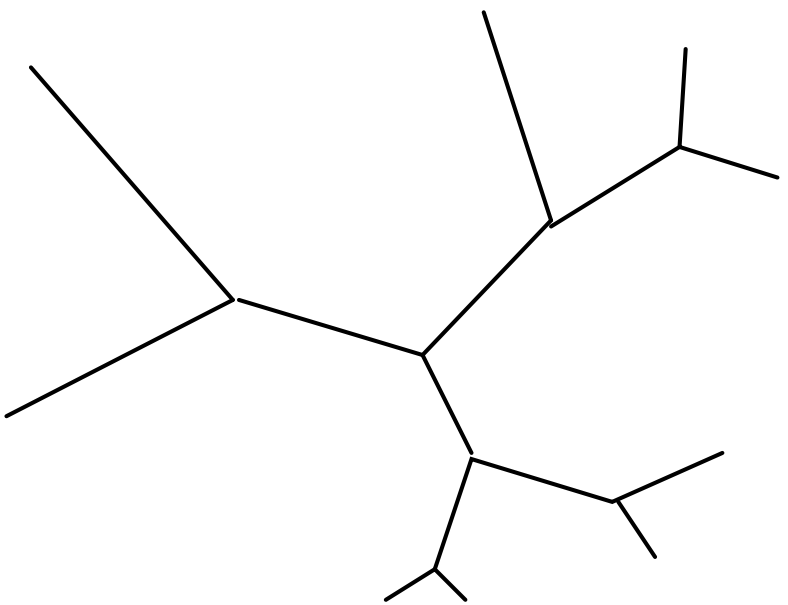
A. Arribas-Gil, D. Metzler, J.-L. Plouhinec (2007)



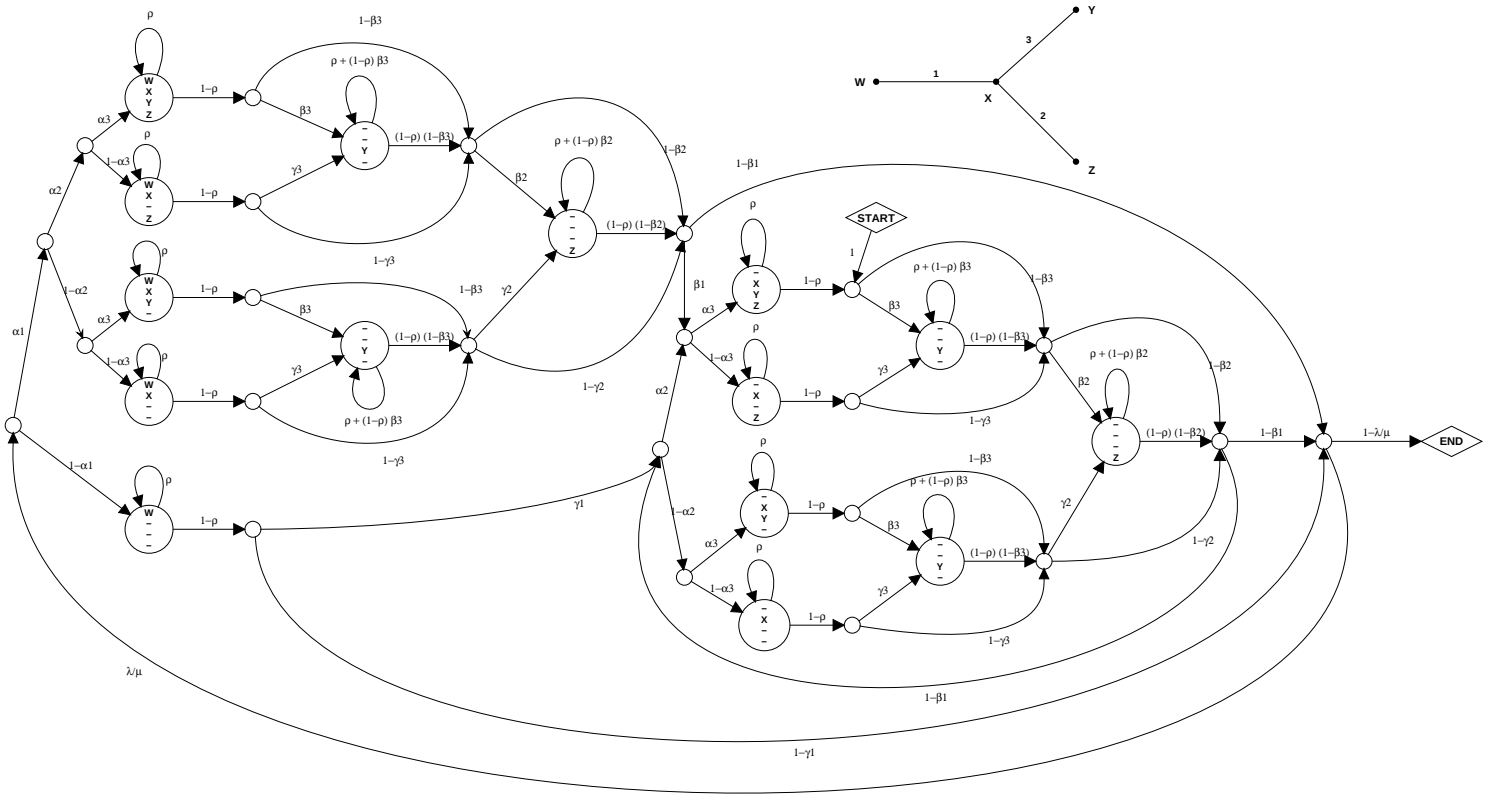
# Outline

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

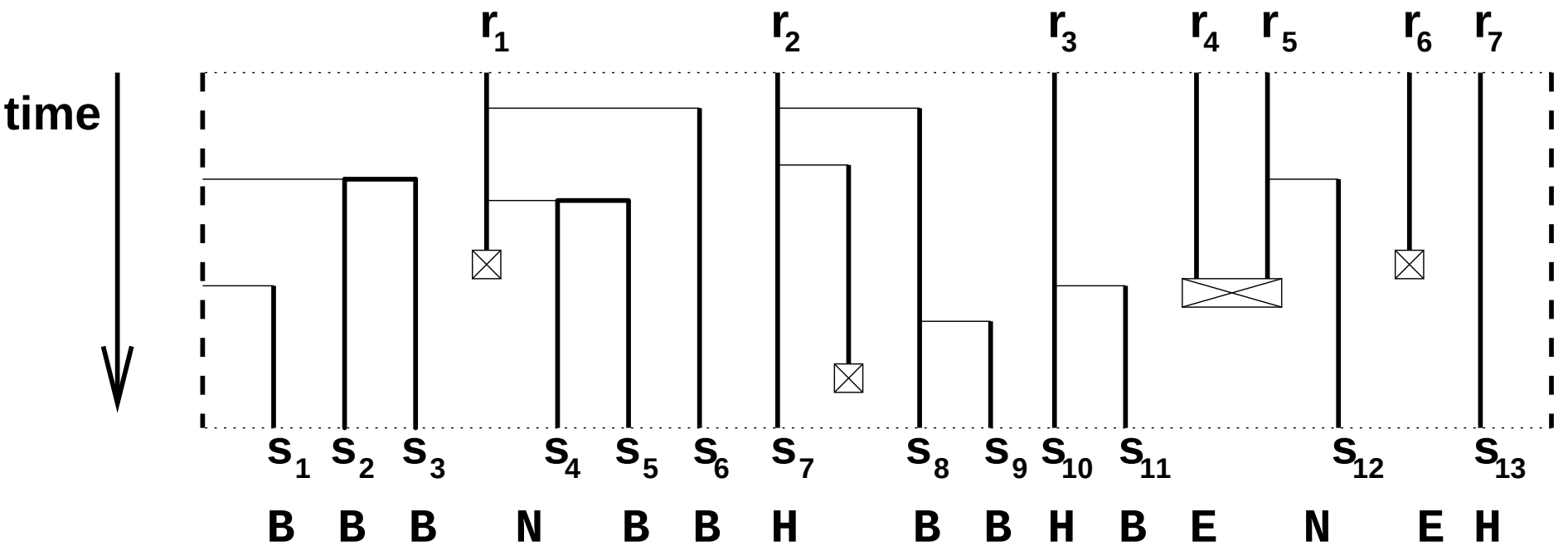
R. Fleißner, D. Metzler, A. von Haeseler (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology* 54(4):548-61.

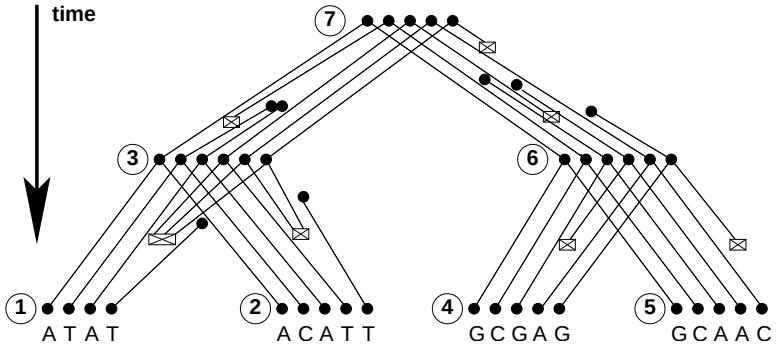


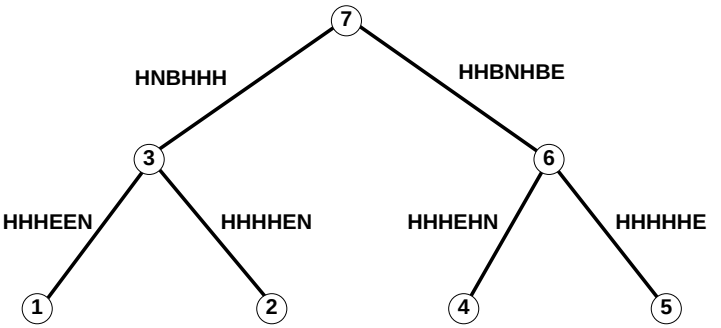
# multiple HMM for sampling a sequence given its neighbours

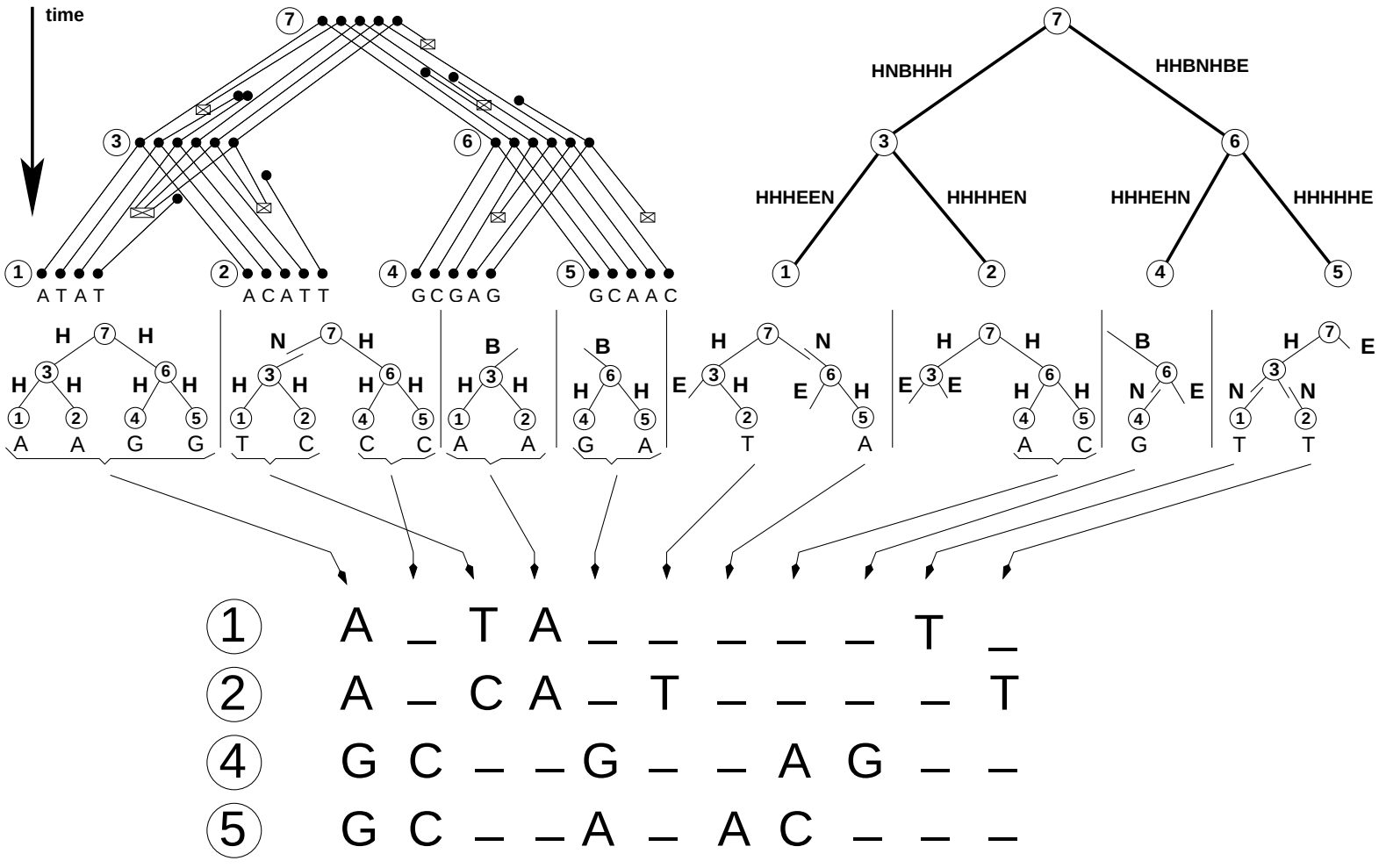


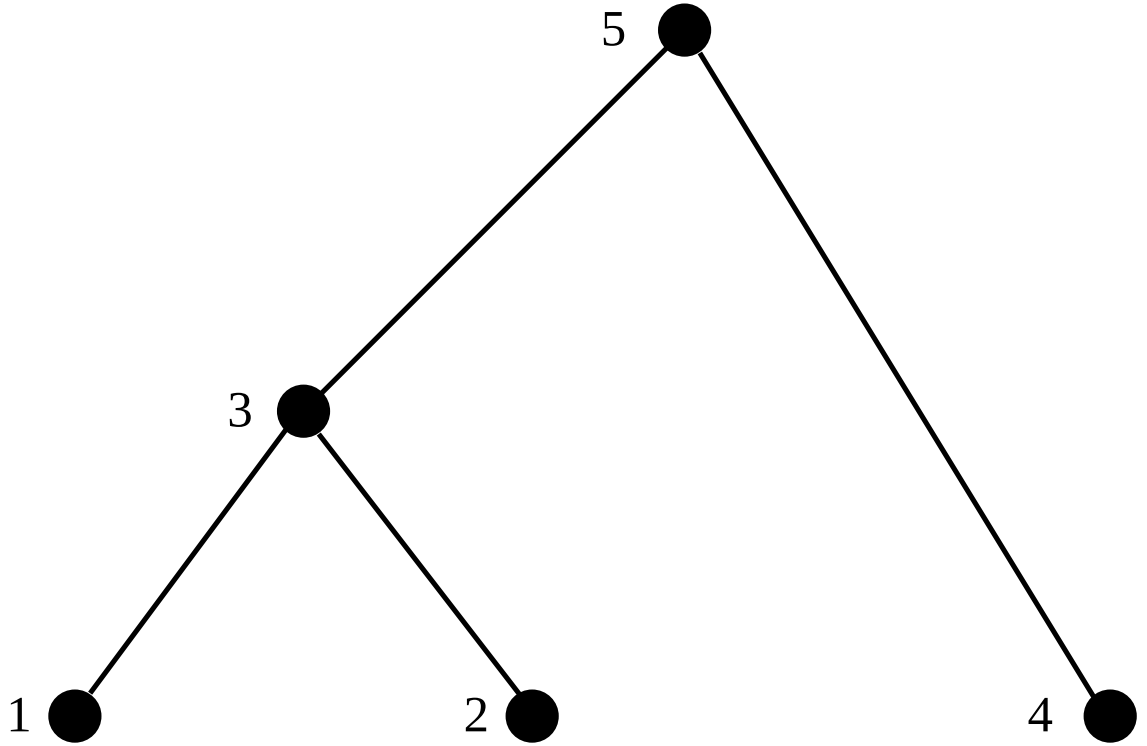
G.A. Lunter, I. Miklós, Y.S. Song, J. Hein (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.* 10(6):869-889.

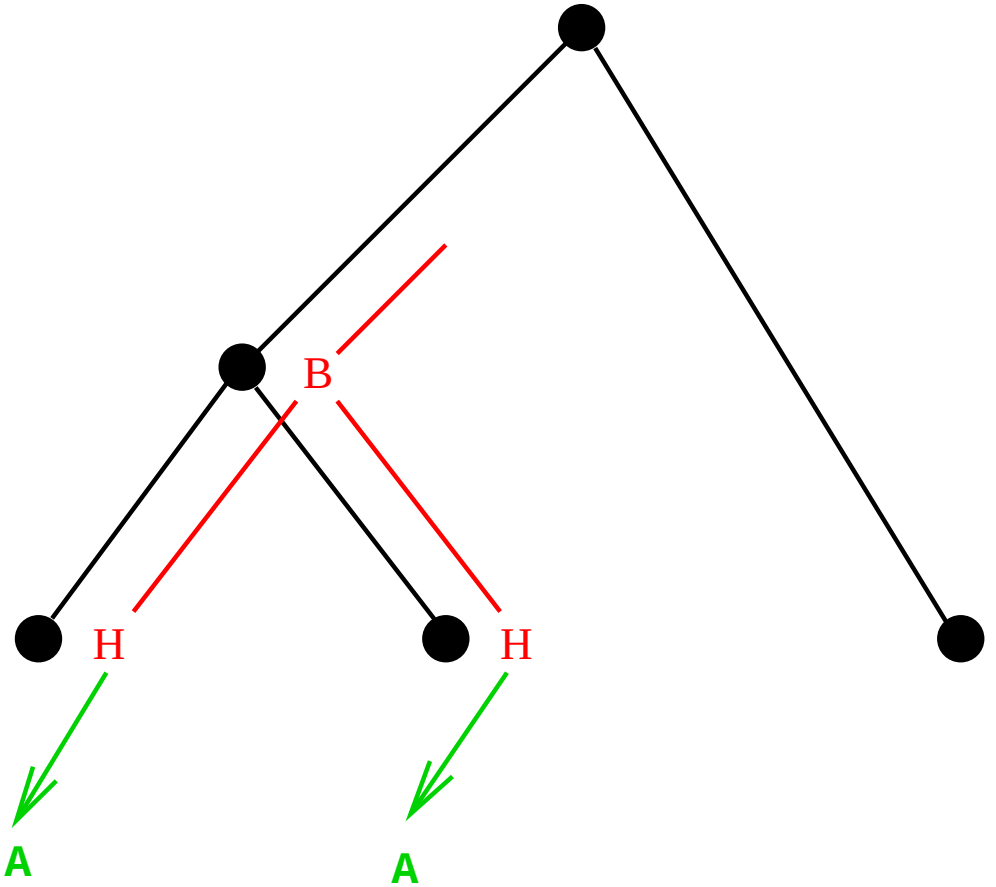



time
7
3
6
1
2
4
5
A T A T
A C A T T
G C G A G
G C A A C

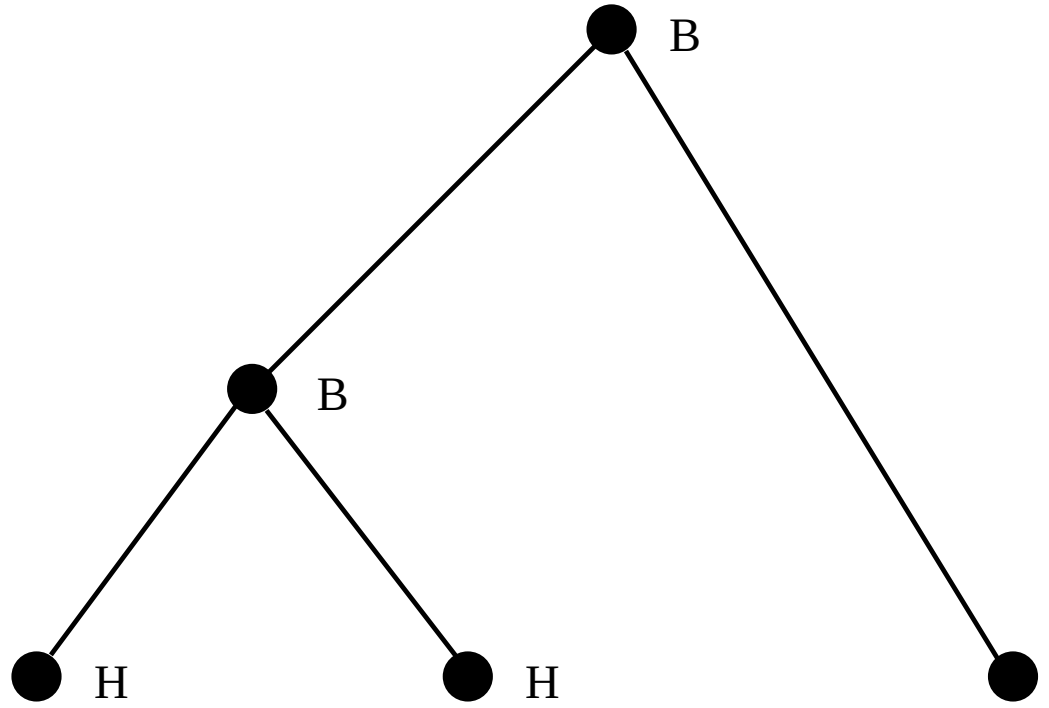

7
HNBHHH
HHBNHBE
3
6
HHHEEN
HHHHEN
HHHEHN
HHHHHE
1
2
4
5


time
A T A T
A C A T T
G C G A G
G C A A C
HNBHHH
HHBNHBE
HHHEEN
HHHHEN
HHHEHN
HHHHHE
1 A _ T A _ _ _ _ _ T _
2 A _ C A _ T _ _ _ _ T
4 G C _ _ G _ _ A G _ _
5 G C _ _ A _ A C _ _ _


5
3
1
2
4



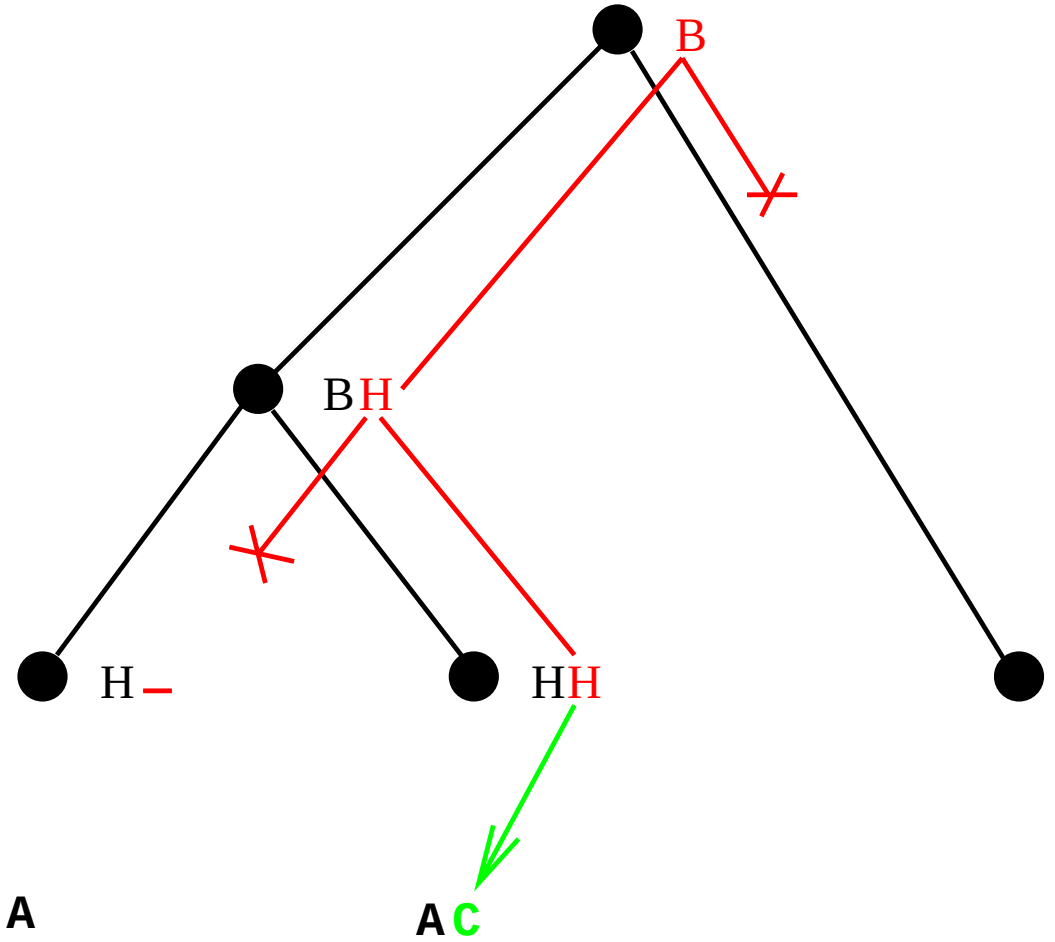

A
A
–



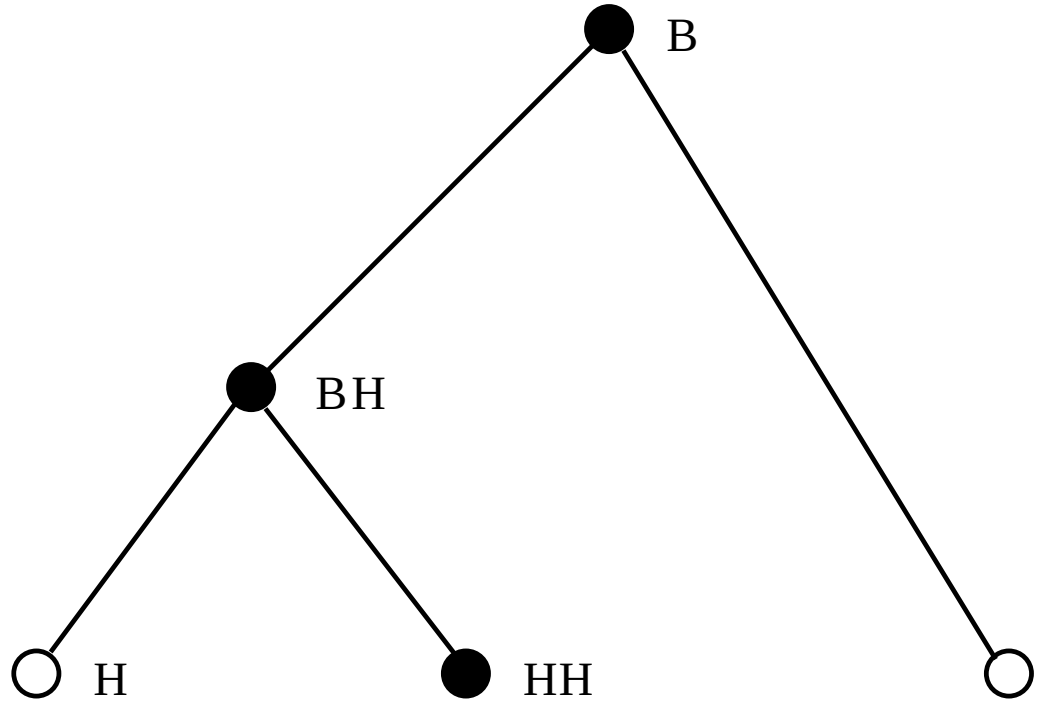

**A**

**A**

**A**
**A**
**–**



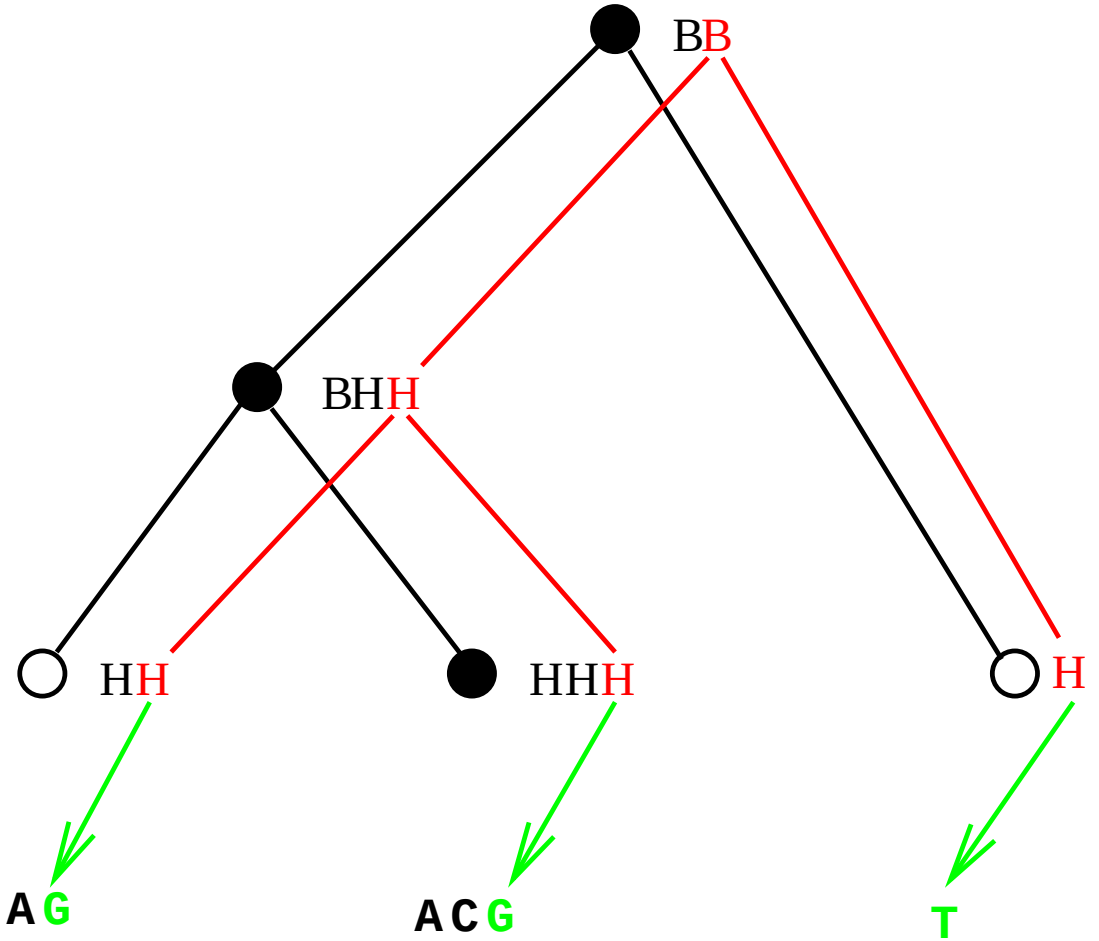

A –
A C
– –



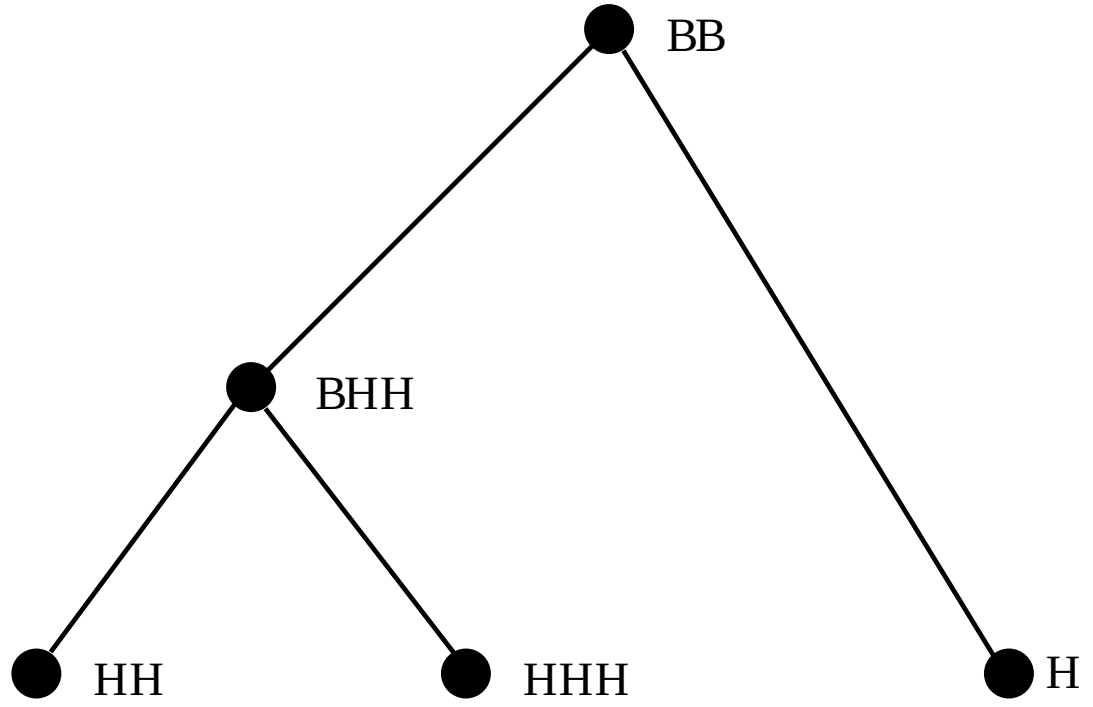

**A**

**AC**

**A –**
**A C**
**– –**

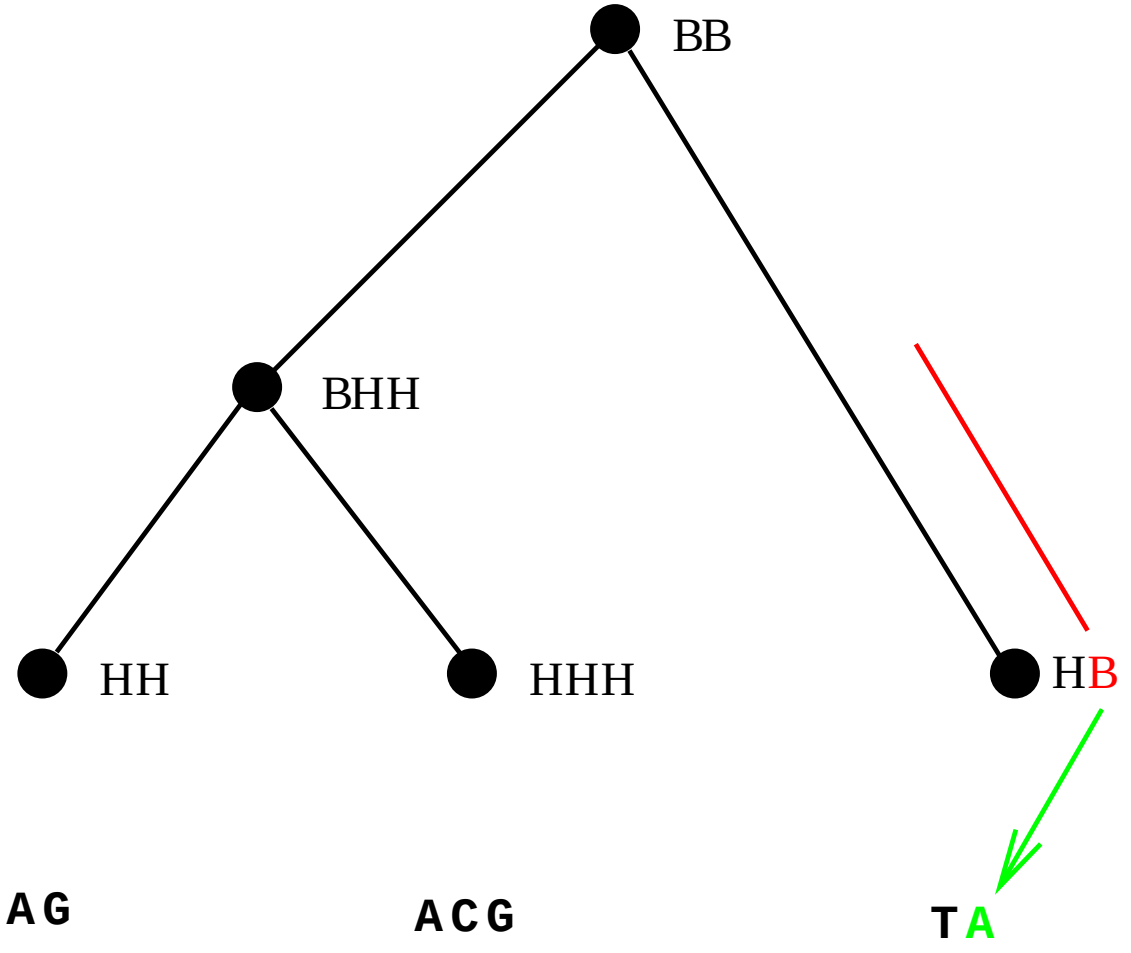



A – G
A C G
– – T

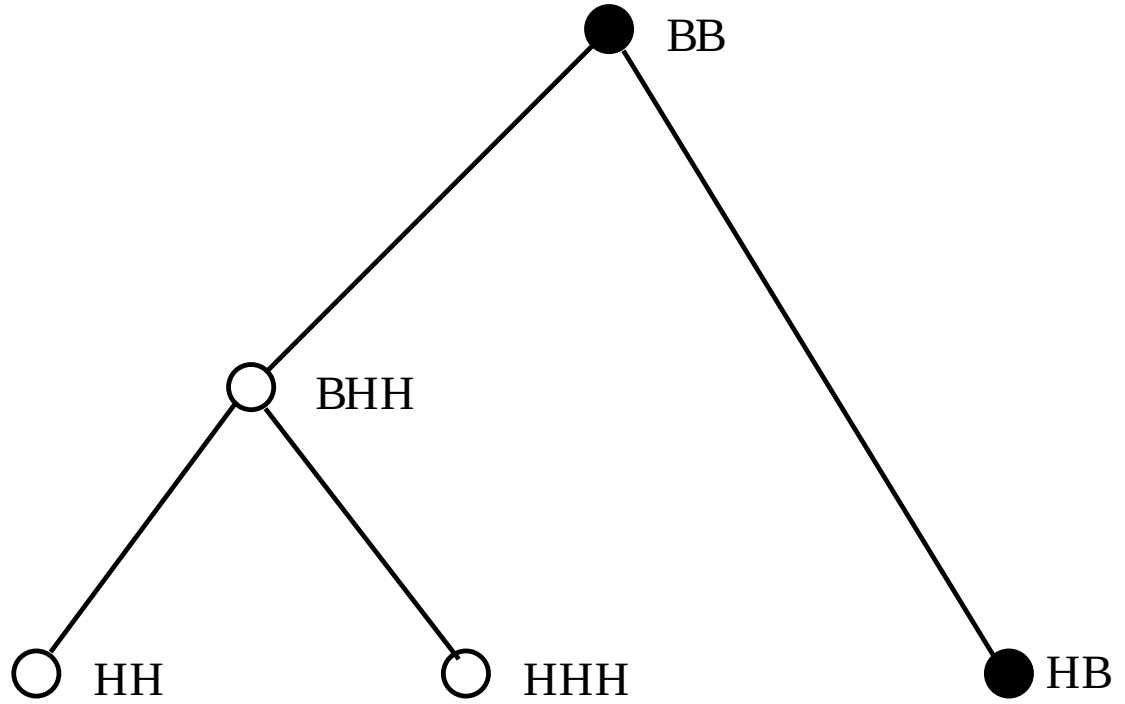



**AG** **ACG** **T**

**A – G**
**A C G**
**– – T**

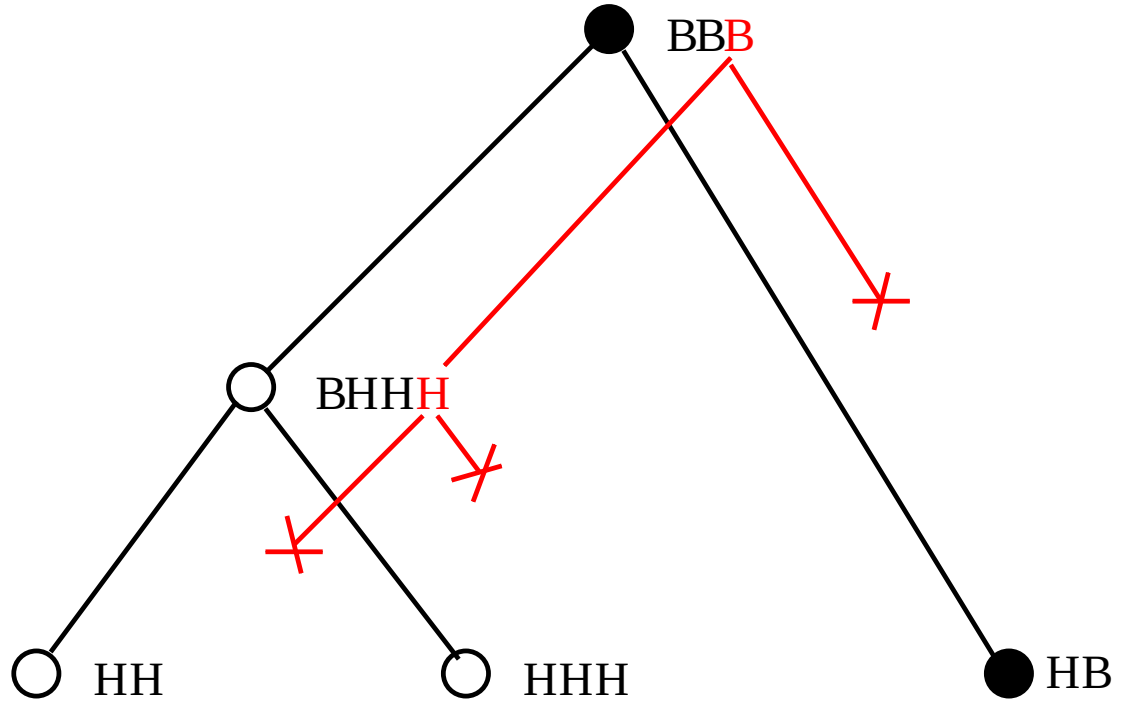



A – G A
A C G –
– – T –



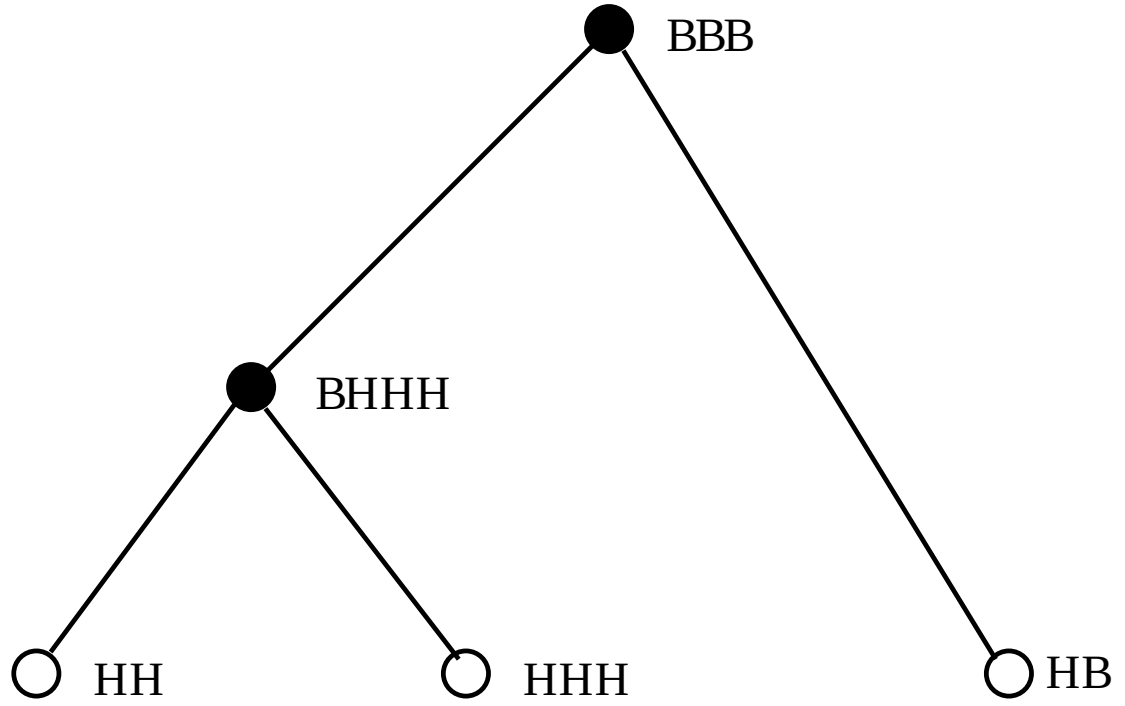

**AG** **ACG** **TA**

**A – G A**
**A C G –**
**– – T –**



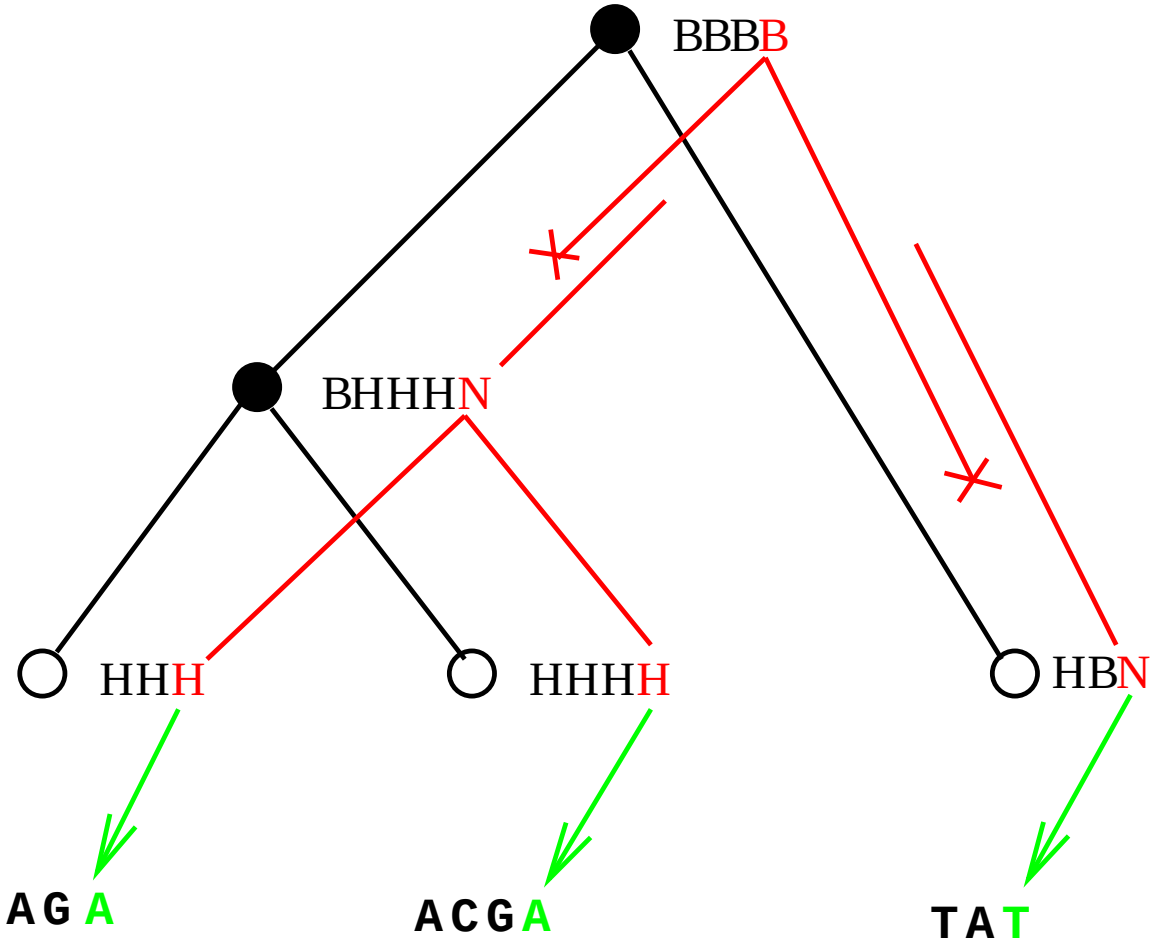

**AG** **ACG** **TA**

**A – G A**
**A C G –**
**– – T –**

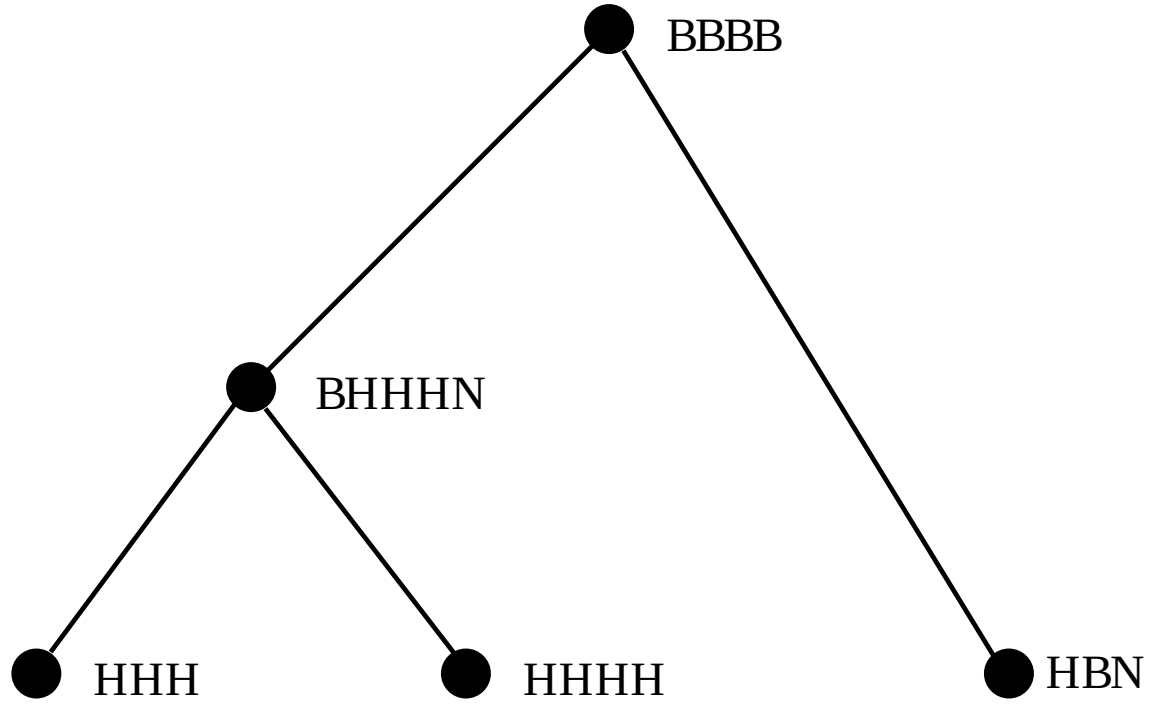



**AG** **ACG** **TA**

**A – G A**
**A C G –**
**– – T –**

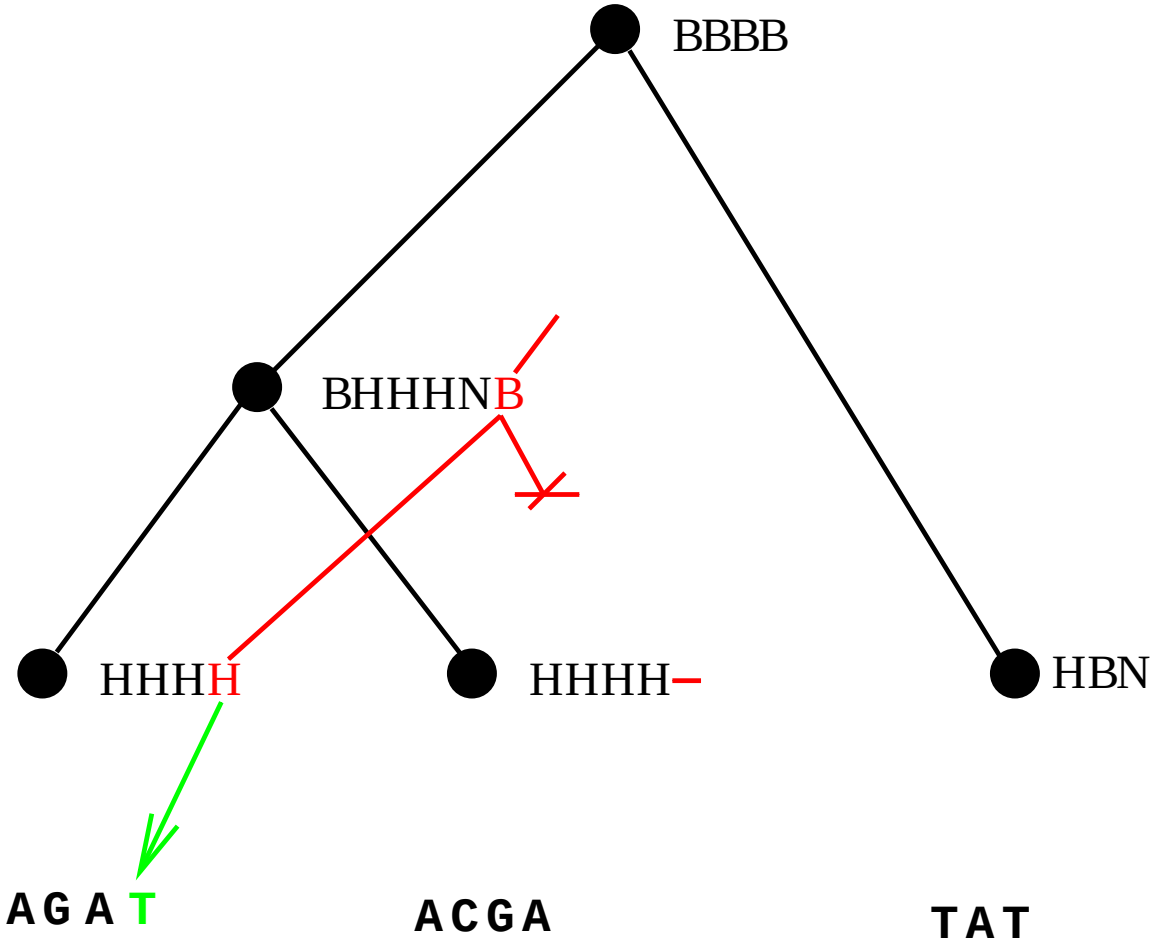



A – G A A –
A C G – A –
– – T – – T

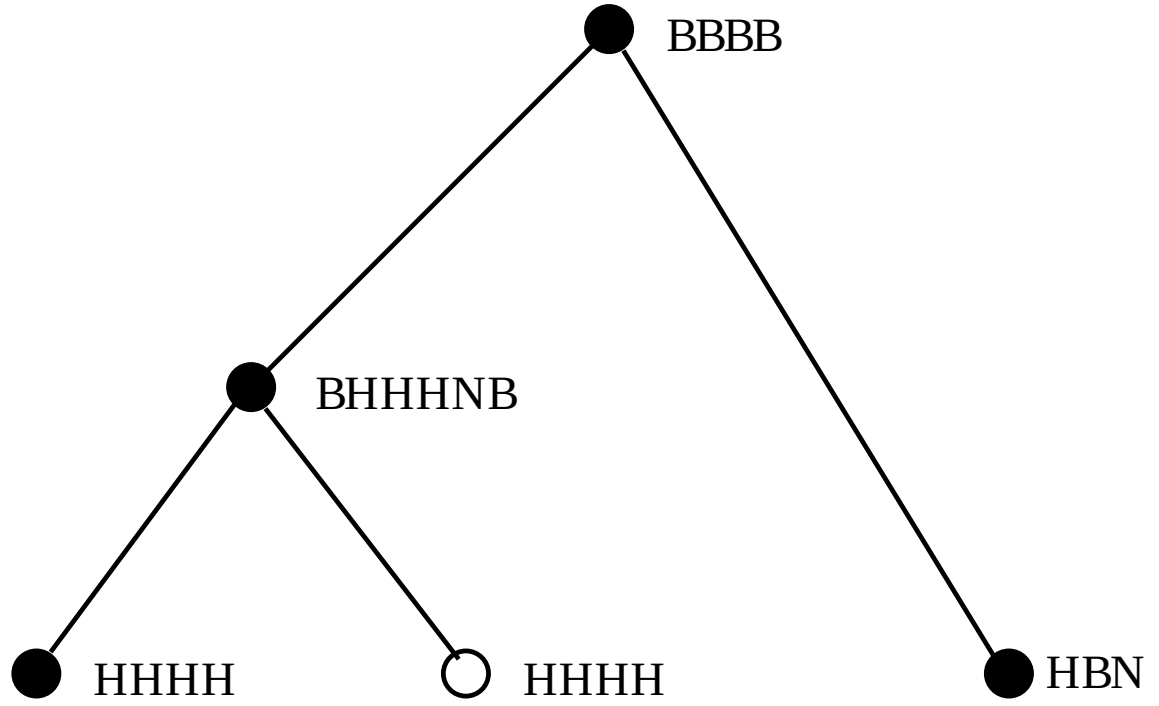



**AGA** **ACGA** **TAT**

**A – G A A –**
**A C G – A –**
**– – T – – T**




A – G A A _ T
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).

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

where

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

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

$P_{\mathcal{S}}(k)$ : 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 $e$ emit

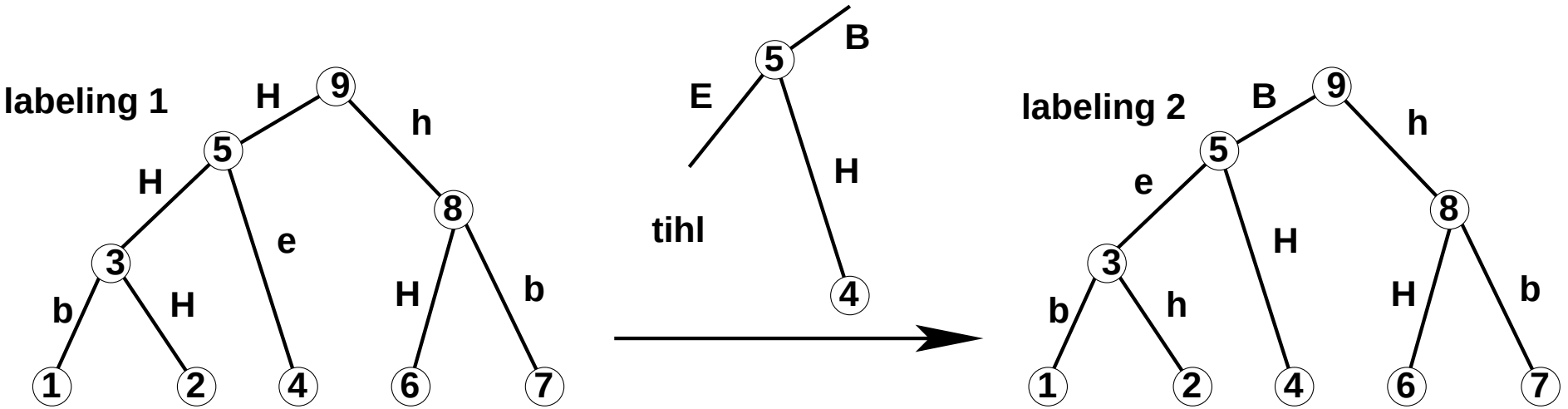
TKF91: states of hidden Markov chain are the Sets Of Active Nodes (soans).

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)

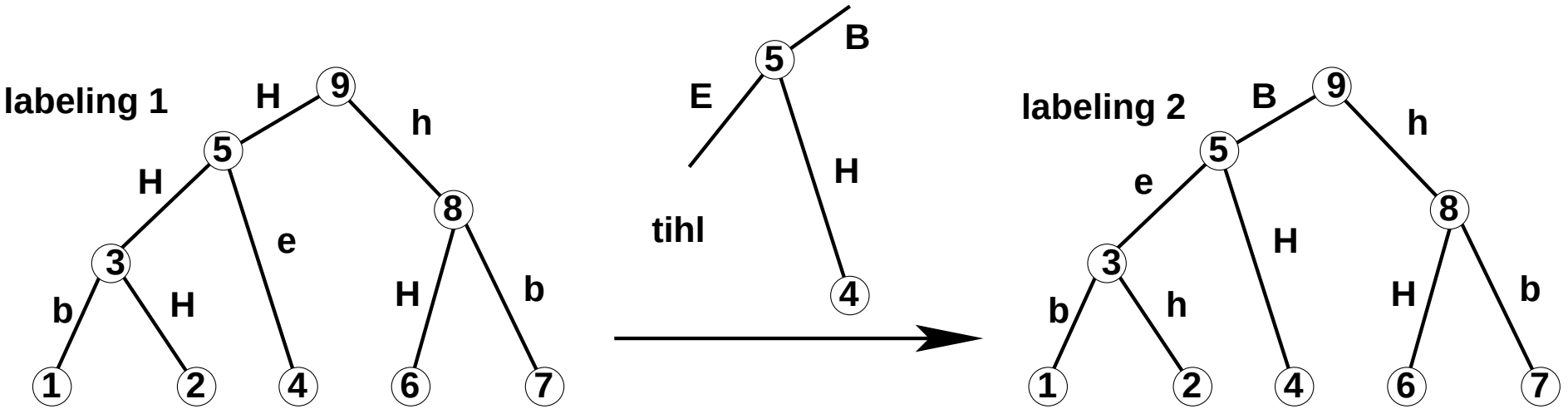
- D. Metzler, R. Fleißner, A. Wakolbinger, A. von Haeseler (2005) Stochastic insertion-deletion processes and statistical sequence alignment.
- D. Metzler, R. Fleißner (2007) Sequence Evolution Models for Simultaneous Alignment and Phylogeny Reconstruction.

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




tihl = tree indexed heirs line

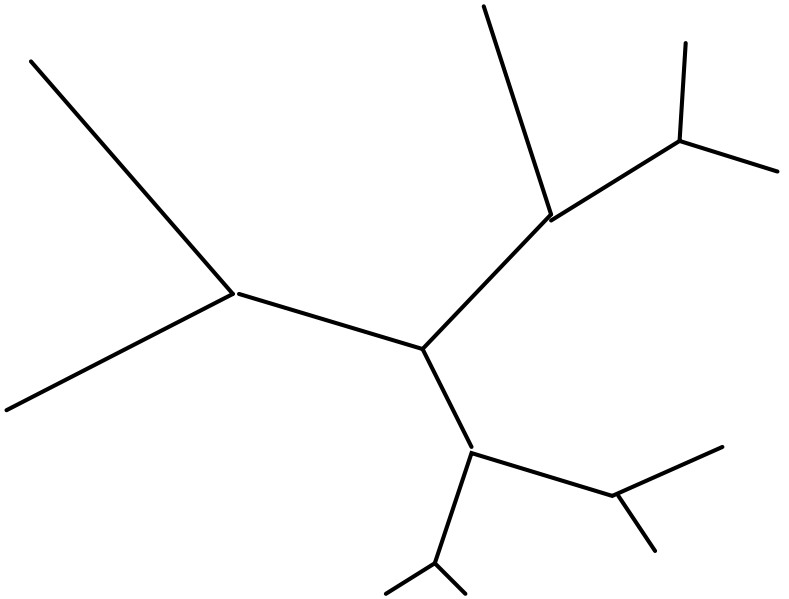
state space: edge-labellings with $\{B, H, e, b, h\}$.



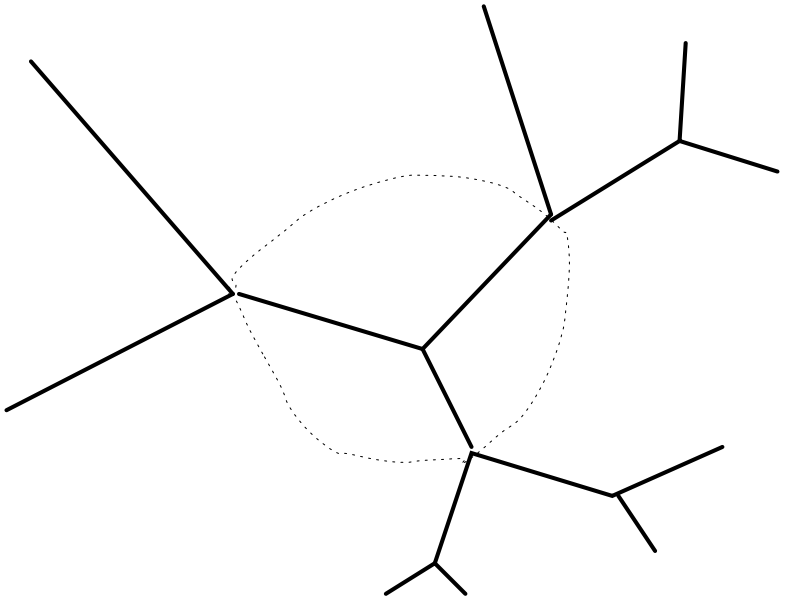

tihl = tree indexed heirs line

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

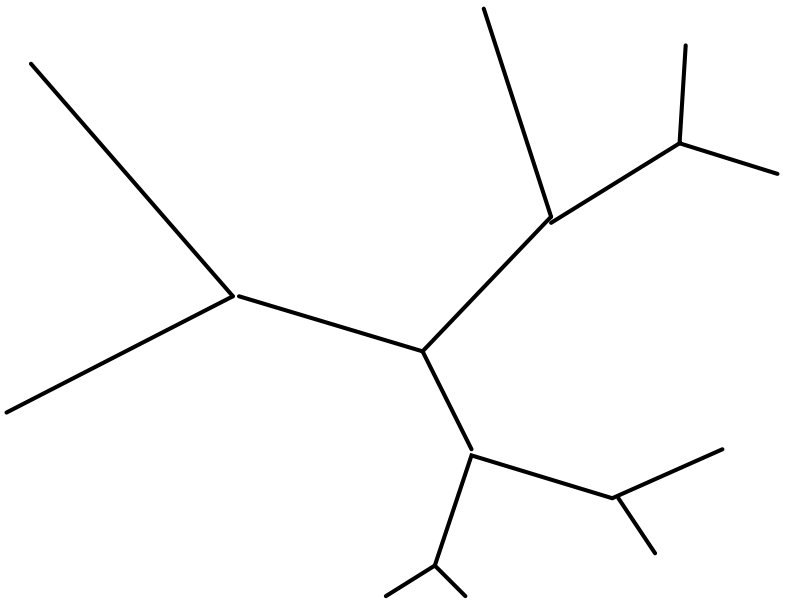


- re-sample alignments of 3-star subtrees (like J.L. Jensen and J. Hein, 2005, do for TKF91)
- do this only for limited parts of the sequences
- Can non-emitting tihls be ignored?
- assing sequences to internal nodes or use nucleotide (or AA) distributions?

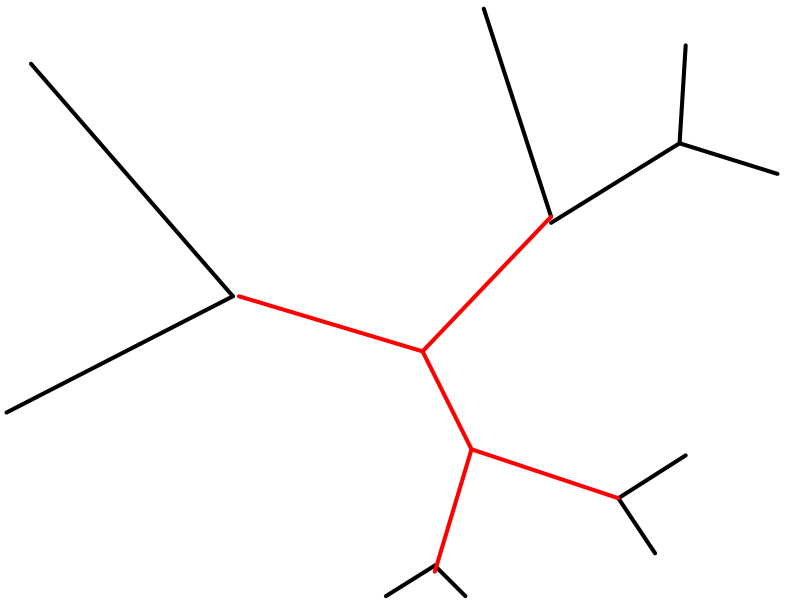


- re-sample alignments of 3-star subtrees (like J.L. Jensen and J. Hein, 2005, do for TKF91)
- do this only for limited parts of the sequences
- Can non-emitting tihls be ignored?
- assing sequences to internal nodes or use nucleotide (or AA) distributions?

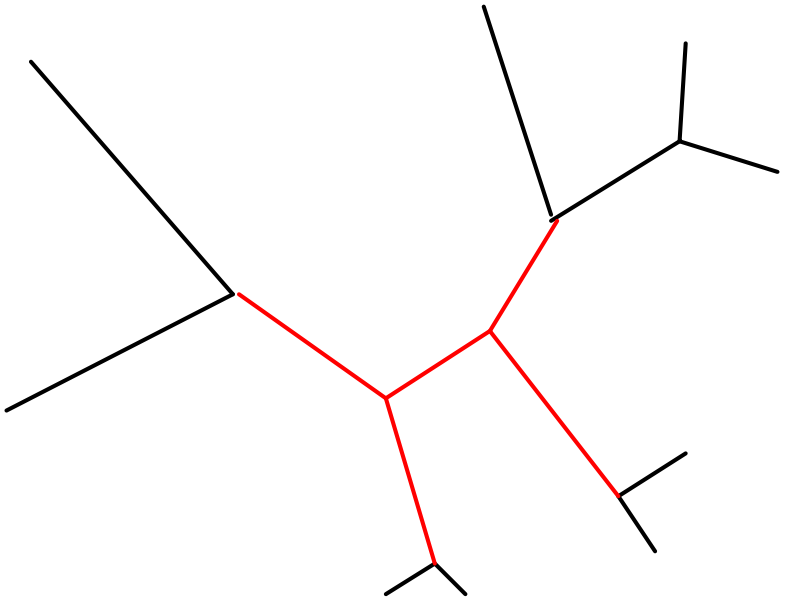
When changing the tree topology...



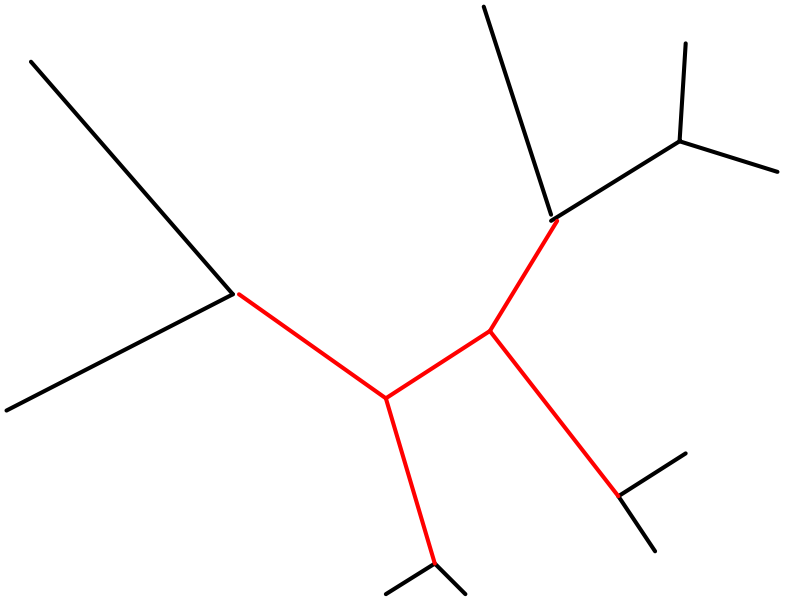
When changing the tree topology...



When changing the tree topology...



When changing the tree topology...



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

# Conclusions

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

# Conclusions

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

# Conclusions

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