

# Approaches to Sequence Analysis

Data {GTCAT, GTTGGT, GTCA, CTCA}



**Parsimony, similarity,  
optimisation.**

GT-CAT

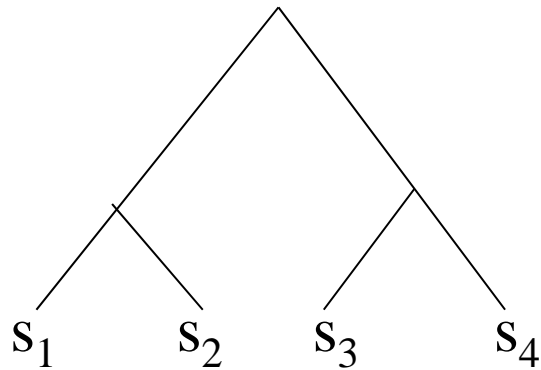
GTTGGT

GT-CA-

CT-CA-



**statistics**



**Ideal Practice: 1 phase analysis.**

1. TKF91 - The combined substitution/indel process.
2. Acceleration of Basic Algorithm
3. Many Sequence Algorithm
4. MCMC Approaches

**Actual Practice: 2 phase analysis.**



# $\lambda$ & $\mu$ into Alignment Blocks

## A. Amino Acids Ignored:

# - - -  
 # # # #  
 k

$$e^{-\mu t} [1 - \lambda\beta] (\lambda\beta)^{k-1}$$

$$p_k(t)$$

$$\beta = [1 - e^{-(\lambda-\mu)t}] / [\mu - \lambda e^{-(\lambda-\mu)t}]$$

# - - - -  
 - # # # #  
 k

$$[1 - \lambda\beta - \mu\beta] (\lambda\beta)^k$$

$$p'_k(t)$$

$$p'_0(t) = \mu\beta(t)$$

\* - - - -  
 \* # # # #  
 k

$$[1 - \lambda\beta] (\lambda\beta)^k$$

$$p''_k(t)$$

## B. Amino Acids Considered:

T - - -  
 R Q S W  
 4

$$P_t(T \rightarrow R) * \pi_Q * \dots * \pi_W * p_4(t)$$

T - - - -  
 - R Q S W  
 4

$$\pi_R * \pi_Q * \dots * \pi_W * p'_4(t)$$

# Differential Equations for p-functions

$$\begin{array}{cccccc} \# & - & - & \dots & - \\ \# & \# & \# & \dots & \# \end{array}$$

$$\Delta p_k = \Delta t * [\lambda * (k-1) p_{k-1} + \mu * k * p_{k+1} - (\lambda + \mu) * k * p_k]$$

$$\begin{array}{cccccc} \# & - & - & - & \dots & - \\ - & \# & \# & \# & \dots & \# \end{array}$$

$$\Delta p'_k = \Delta t * [\lambda * (k-1) p'_{k-1} + \mu * (k+1) * p'_{k+1} - (\lambda + \mu) * k * p'_k + \mu * p_{k+1}]$$

$$\begin{array}{cccccc} * & - & - & - & \dots & - \\ * & \# & \# & \# & \dots & \# \end{array}$$

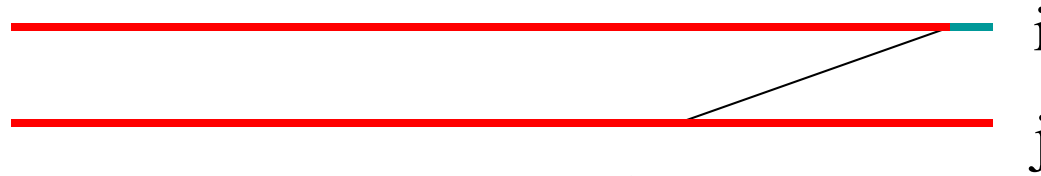
$$\Delta p''_k = \Delta t * [\lambda * k * p''_{k-1} + \mu * (k+1) * p''_{k+1} - [(k+1)\lambda + k\mu] * p''_k]$$

Initial Conditions:

$$\begin{array}{l} p_k(0) = p_k''(0) = p'_k(0) = 0 \quad k > 1 \\ p_1(0) = p_0''(0) = 1. \quad p'_0(0) = 0 \end{array}$$

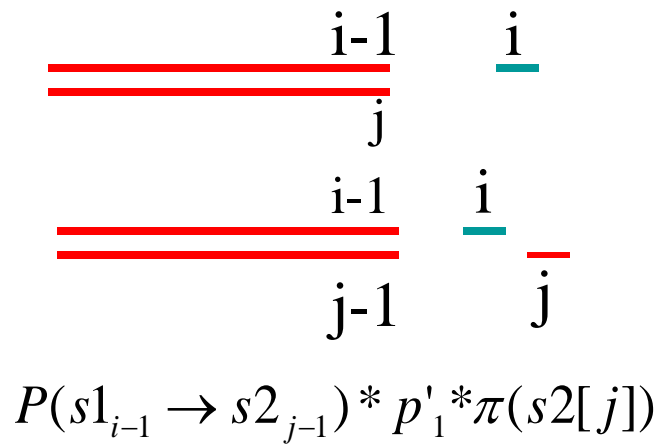
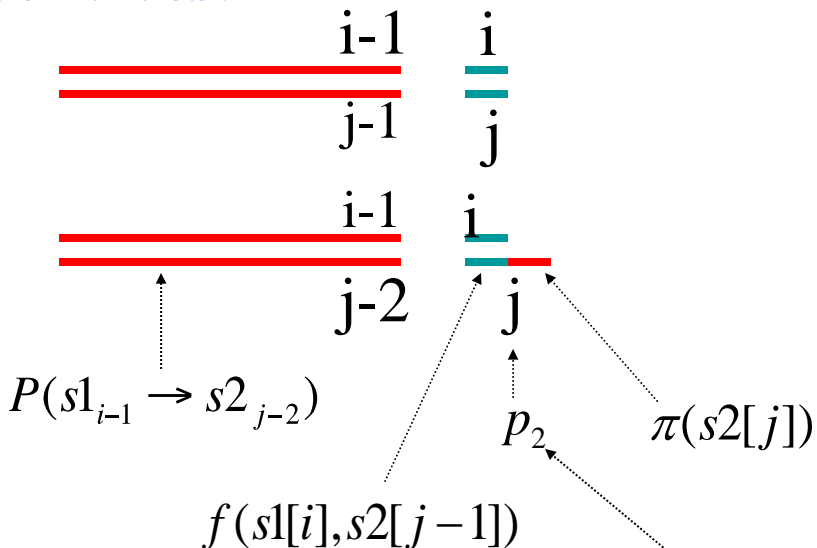
# Basic Pairwise Recursion ( $O(\text{length}^3)$ )

$$P(s1_i \rightarrow s2_j)$$



**Survives:**

**Dies:**



$$e^{-\mu t} [1 - \lambda \beta] (\lambda \beta)^{k-1}, \text{ where}$$

$$\beta = [1 - e^{-(\lambda - \mu)t}] / [\mu - \lambda e^{-(\lambda - \mu)t}]$$

1 ... j (j) cases

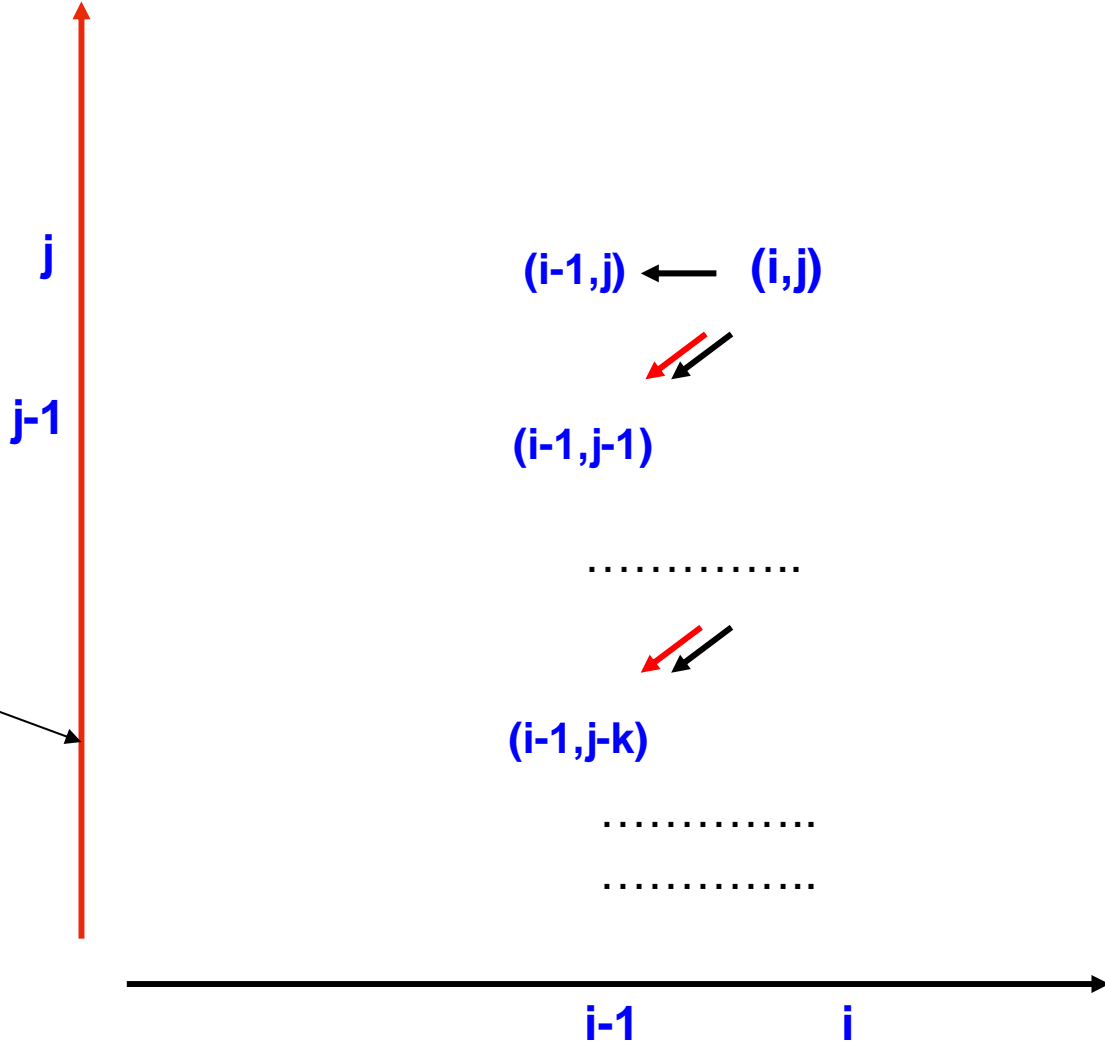
0 ... j (j+1) cases

# Basic Pairwise Recursion ( $O(\text{length}^3)$ )

survive

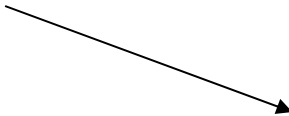


death



Initial condition:

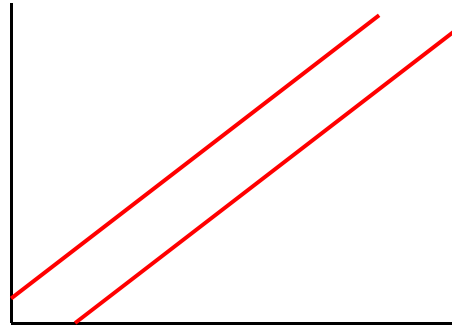
$p'' = s2[1:j]$



# Acceleration of Pairwise Algorithm

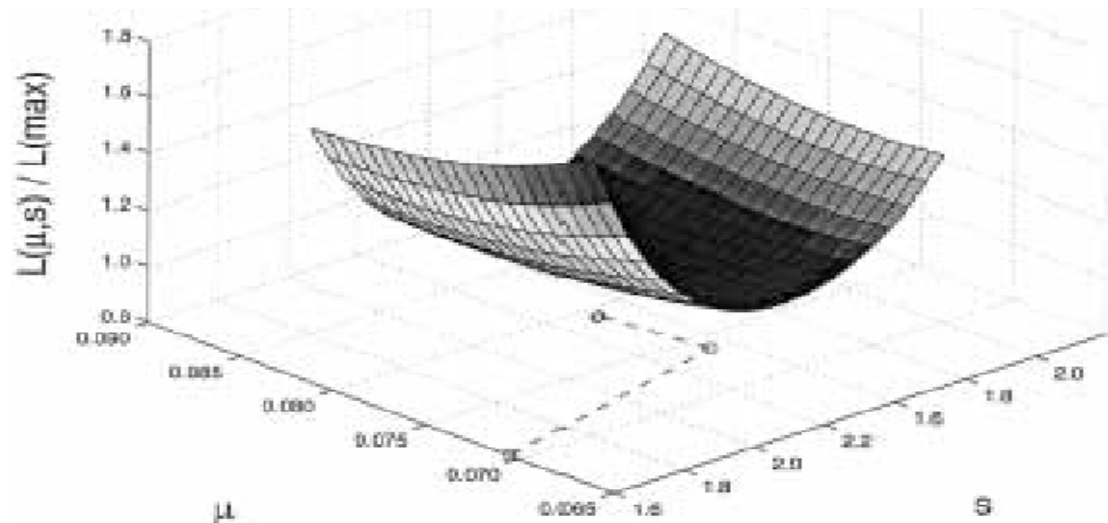
(From Hein,Wiuf,Knudsen,Moeller & Wiebling 2000)

Corner Cutting ~100-1000



Better Numerical Search ~10-100

Ex.: good start guess, 28 evaluations, 3 iterations



Simpler Recursion ~3-10

Faster Computers ~250

1991-->2000 ~10<sup>6</sup>

# $\alpha$ -globin (141) and $\beta$ -globin (146)

(From Hein,Wiuf,Knudsen,Moeller & Wiebling 2000)

430.108 :  $-\log(\alpha\text{-globin})$   
327.320 :  $-\log(\alpha\text{-globin} \rightarrow \beta\text{-globin})$   
747.428 :  $-\log(\alpha\text{-globin}, \beta\text{-globin}) = -\log(l(\text{sumalign}))$

$\lambda^*t$ : 0.0371805 +/- 0.0135899  
 $\mu^*t$ : 0.0374396 +/- 0.0136846  
 $s^*t$ : 0.91701 +/- 0.119556

E(Length)	E(Insertions,Deletions)	E(Substitutions)
143.499	5.37255	131.59

## Maximum contributing alignment:

V-LSPADKTNVKAAWGKVGHAHAGEYGAELERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT  
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

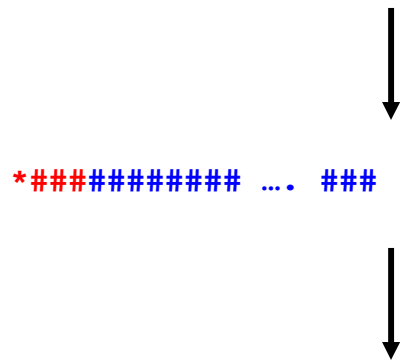
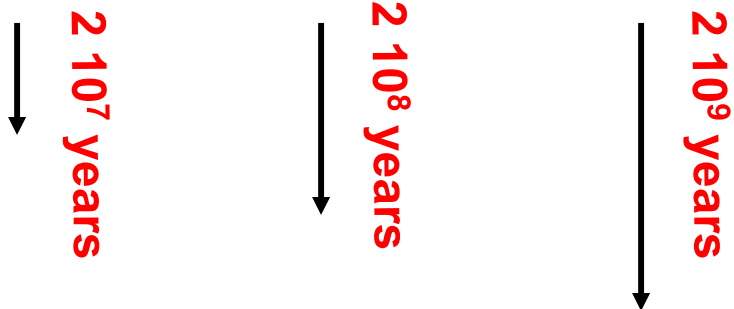
NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR  
DGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

Ratio  $l(\text{maxalign})/l(\text{sumalign}) = 0.00565064$

# The invasion of the immortal link

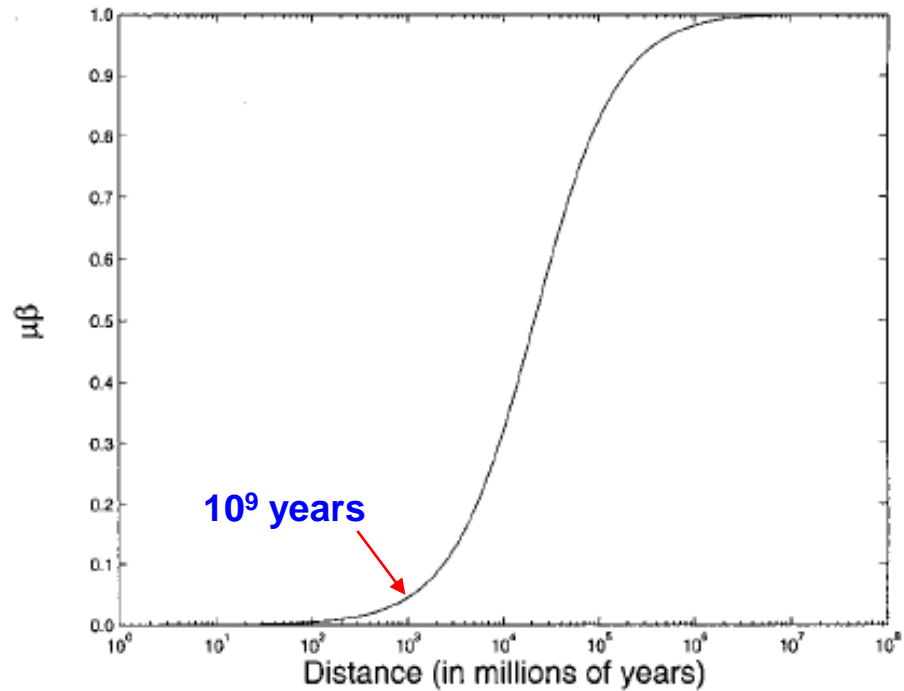
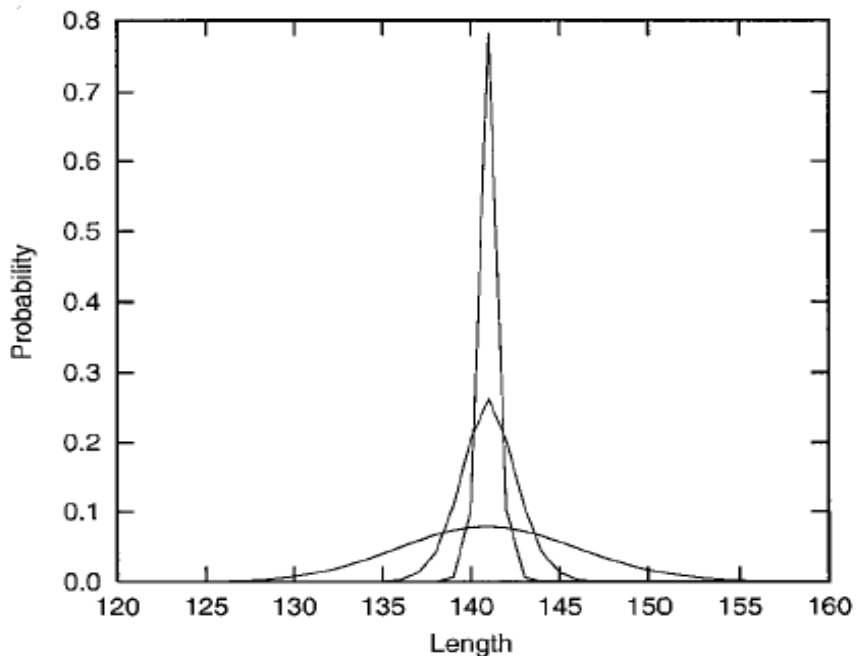
VLSPADNAL.....DLHAHKR      141 AA long

\*##### ... ###      141 AA long



????????????????????      k AA long

\*##### ... ###



# Markov Chains Generating the p-functions

## Ancestral Sequence Generator

	#	E				
*	$\lambda/\mu$	$1 - \lambda/\mu$	*	#	#	#
#	$\lambda/\mu$	$1 - \lambda/\mu$	#			

## p'' function generator

	- #	E					
*	$\lambda\beta$	$1 - \lambda\beta$	*	-	-	-	-
*			*	#	#	#	#
- #	$\lambda\beta$	$1 - \lambda\beta$					

## p'/p function generator

	- #	E
# #	$\lambda\beta$	$1 - \lambda\beta$
# -	$1 - \mu\beta$	$\mu\beta$
- #	$\lambda\beta$	$1 - \lambda\beta$

$$P \begin{pmatrix} \# \\ \# \end{pmatrix} = e^{-\mu}$$

#	-	-	-	-
#	#	#	#	#

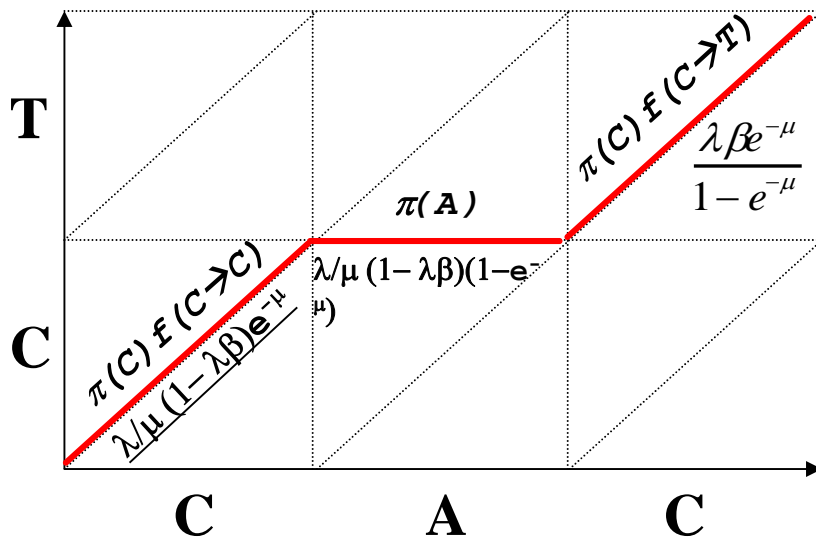
$$P \begin{pmatrix} \# \\ - \end{pmatrix} = 1 - e^{-\mu}$$

#	-	-	-	-
-	#	#	#	#

# Statistical Alignment via Hidden Markov Models

Steel and Hein, 2001 + Holmes and Bruno, 2001

	-	#	#	E
*	#	#	-	E
*	$\lambda\beta$	$\frac{\lambda\mu(1-\lambda\beta)e^{-\mu}}{1-e^{-\mu}}$	$\lambda\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda\beta)$
-	$\lambda\beta$	$\lambda\mu(1-\lambda\beta)e^{-\mu}$	$\lambda\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda\beta)$
#	$\lambda\beta$	$\lambda\mu(1-\lambda\beta)e^{-\mu}$	$\lambda\mu(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)(1-\lambda\beta)$
#	$\frac{1-\lambda\beta e^{-\mu}}{1-e^{-\mu}}$	$\frac{\lambda\beta e^{-\mu}}{1-e^{-\mu}}$	$\lambda\beta$	$\frac{(\mu-\lambda)\beta}{1-e^{-\mu}}$
-				



**HMM formulation allows:**

*Finding most probable alignment*

*Probability of sequence pair*

*Probability of specific edge*

# Why multiple statistical alignment is non-trivial.

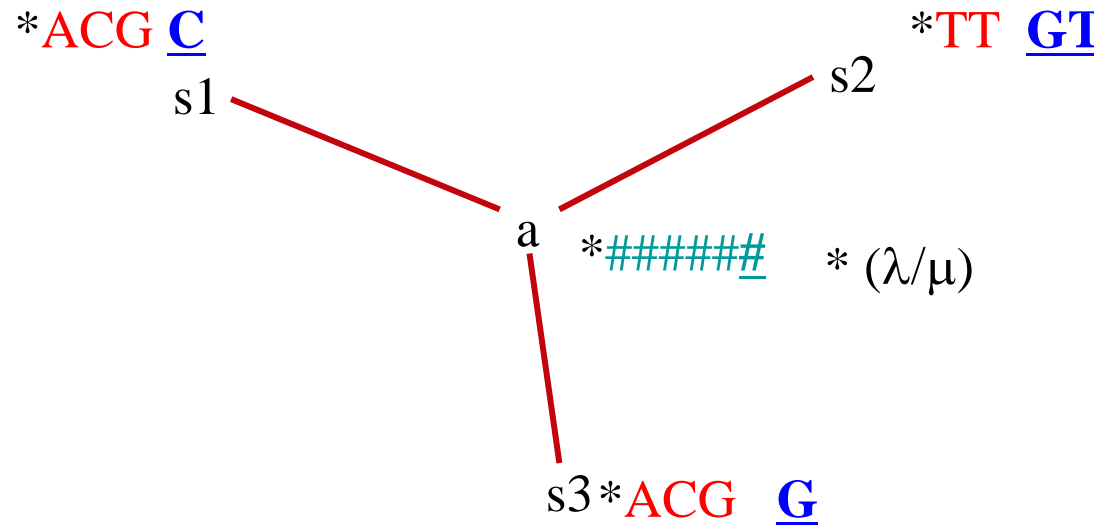
Steel & Hein, 2001, Hein, 2001, Holmes and Bruno, 2001

## Optimisation Alignment

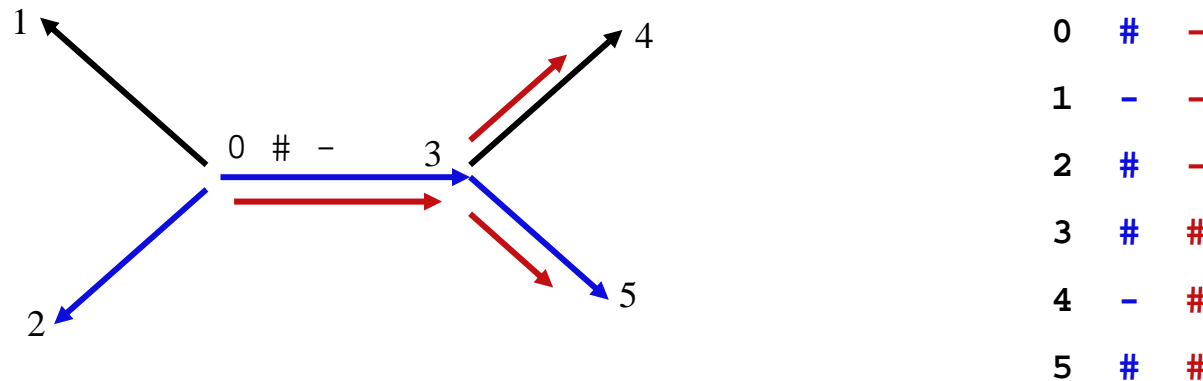
```

AT-C G
ATGC C
CT-C C
    
```

## Statistical Alignment

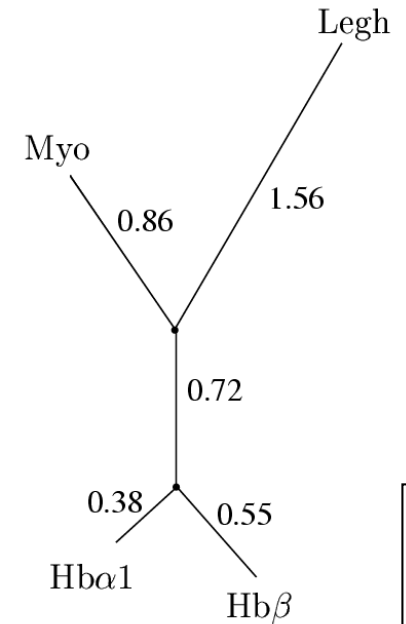


- An HMM generating alignment according to TKF91:



# Maximum likelihood phylogeny and alignment

Human alpha hemoglobin;  
 Human beta hemoglobin;  
 Human myoglobin  
 Bean leghemoglobin



Probability of data  $e^{-1560.138}$   
 Probability of data and alignment  $e^{-1593.223}$   
 Probability of alignment given data  $4.279 * 10^{-15} = e^{-33.085}$   
 Ratio of insertion-deletions to substitutions: 0.0334

Hba1: MV--LSPADKTNVKAAWGKVGAAHAGEYGAEALERMFLSFPTTKTYFPHF--DLS-H-----GSAQVKGHGKKVAD-AL-TNA-  
 Hbb: MV-HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESF-GDLSTPDAVM-GNPKVKAHGKKVLG-AF-SDG-  
 Myo: MG--LSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGH PETLEKFDKFK-HLKSEDE-MKASEDLKKHGATVLT-AL-GGI-  
 Legh: MGA-FSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFS-F---LSNGVD-P-NNPKLKAHA EKVF KMTVDSAVQ

VAHVDDMPNALSALS DLHAHKL RVD PVNFK-LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL-TS-K---YR-  
 LAHLDNLKGTFATLSELHCDKLHVDPENFR-LLGNVLVCVLAHFGKEFTPPVQAA YQKV VAGVANAL-AH-K---YH-  
 LKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG  
 LRAKGEVVLADPTLGSVHVQKGVLDP-HFL-VVKEALLKTFKEAVGDKWNDELGNAWEVAYDELA AAI-KK-A-MGSA-

# Metropolis-Hastings Statistical Alignment.

Lunter, Drummond, Miklos, Jensen & Hein, 2005

## The alignment moves:

*We choose a random window in the current alignment*

```
ALITL---GG
ALLTLTTLGG
---TLTSLGA
ALLGLTSLGA
```

```
QST--QCC-S
S-----CCS
---QST--QC
---QST--QC
```

```
TNQHVSTGN
GN-HVSTGK
TNQH-SCTLN
TNQHVSTLN
```

*Then delete all gaps so we get back subsequences*

```
ALITL---GG
ALLTLTTLGG
---TLTSLGA
ALLGLTSLGA
```

```
QSTQCCS
SCCS
QSTQC
QSTQC
```

```
TNQHVSTGN
GN-HVSTGK
TNQH-SCTLN
TNQHVSTLN
```

*Stochastically realign this part*

```
ALITL---GG
ALLTLTTLGG
---TLTSLGA
ALLGLTSLGA
```

```
QSTQCCS
-S--CCS
QSTQC--
QSTQC--
```

```
TNQHVSTGN
GN-HVSTGK
TNQH-SCTLN
TNQHVSTLN
```



## The phylogeny moves:

As in Drummond et al.  
2002

# Metropolis-Hastings Statistical Alignment

Lunter, Drummond, Miklos, Jensen & Hein, 2005

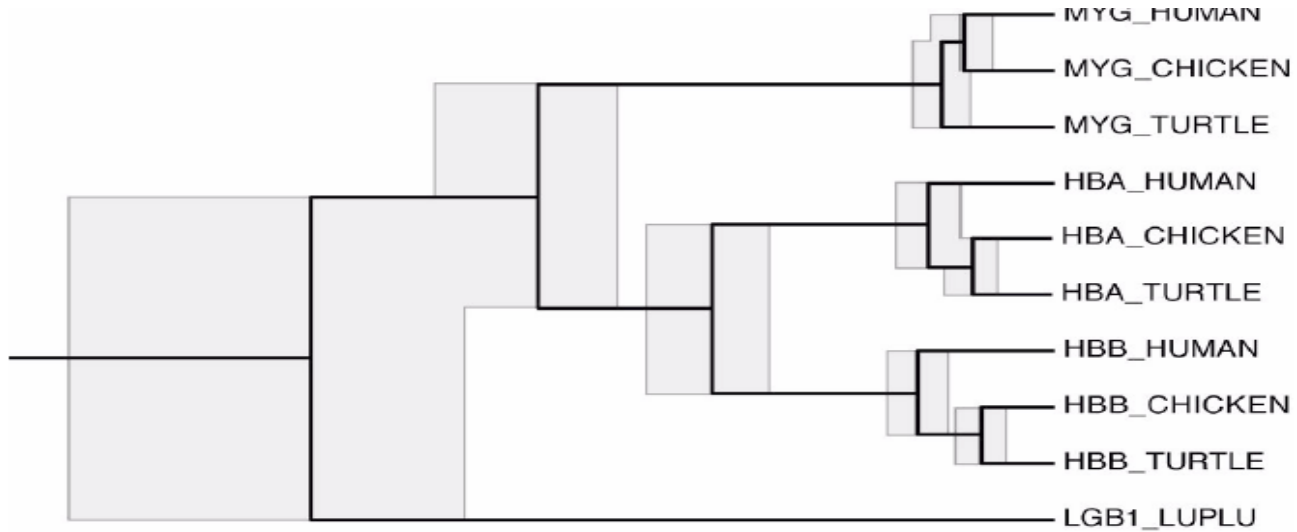


Figure 6

0.2 substitutions per site

