

Approaches to Sequence Analysis

Data {GTCAT, GTTGGT, GTCA, CTCA}



**Parsimony, similarity,
optimisation.**

GT-CAT

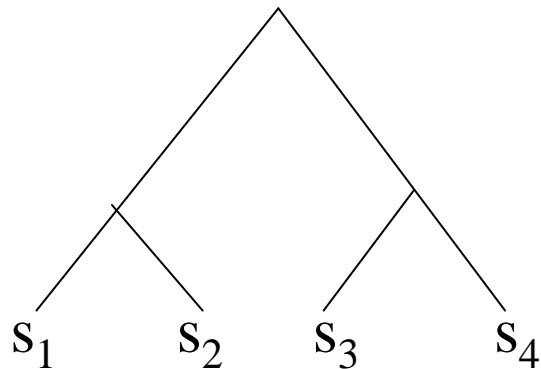
GTTGGT

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

Lunter et al.(2003) 381-387. 390-96

Actual Practice: 2 phase analysis.

λ & μ 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

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

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

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

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

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

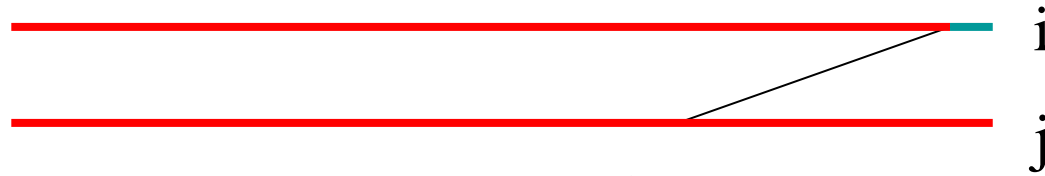
$$\Delta p''_k = \Delta t * [\lambda * k * p''_{k-1} + \mu * (k+1) * p''_{k+1} - [(k+1)\lambda + k\mu] * p''_k]$$

Initial Conditions:

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

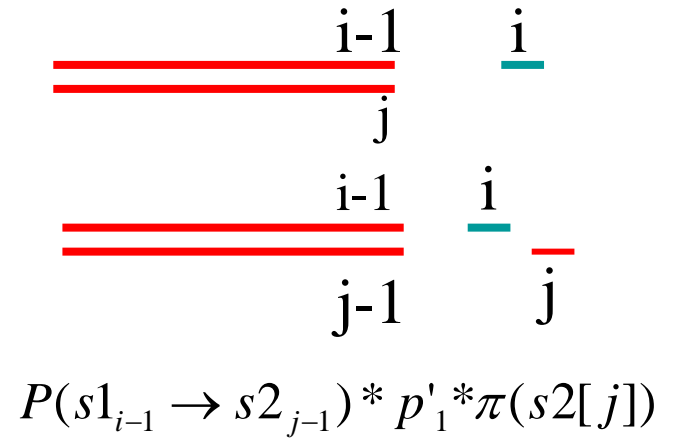
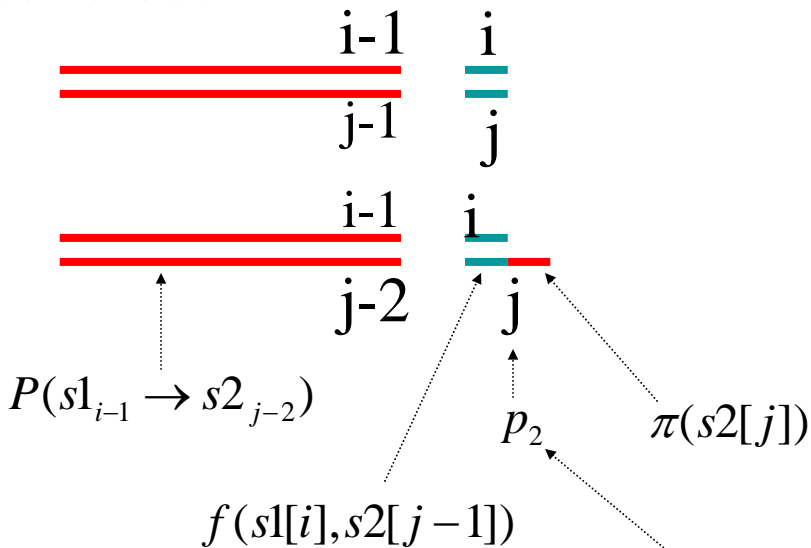
Basic Pairwise Recursion ($O(\text{length}^3)$)

$$P(s1_i \rightarrow s2_j)$$



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

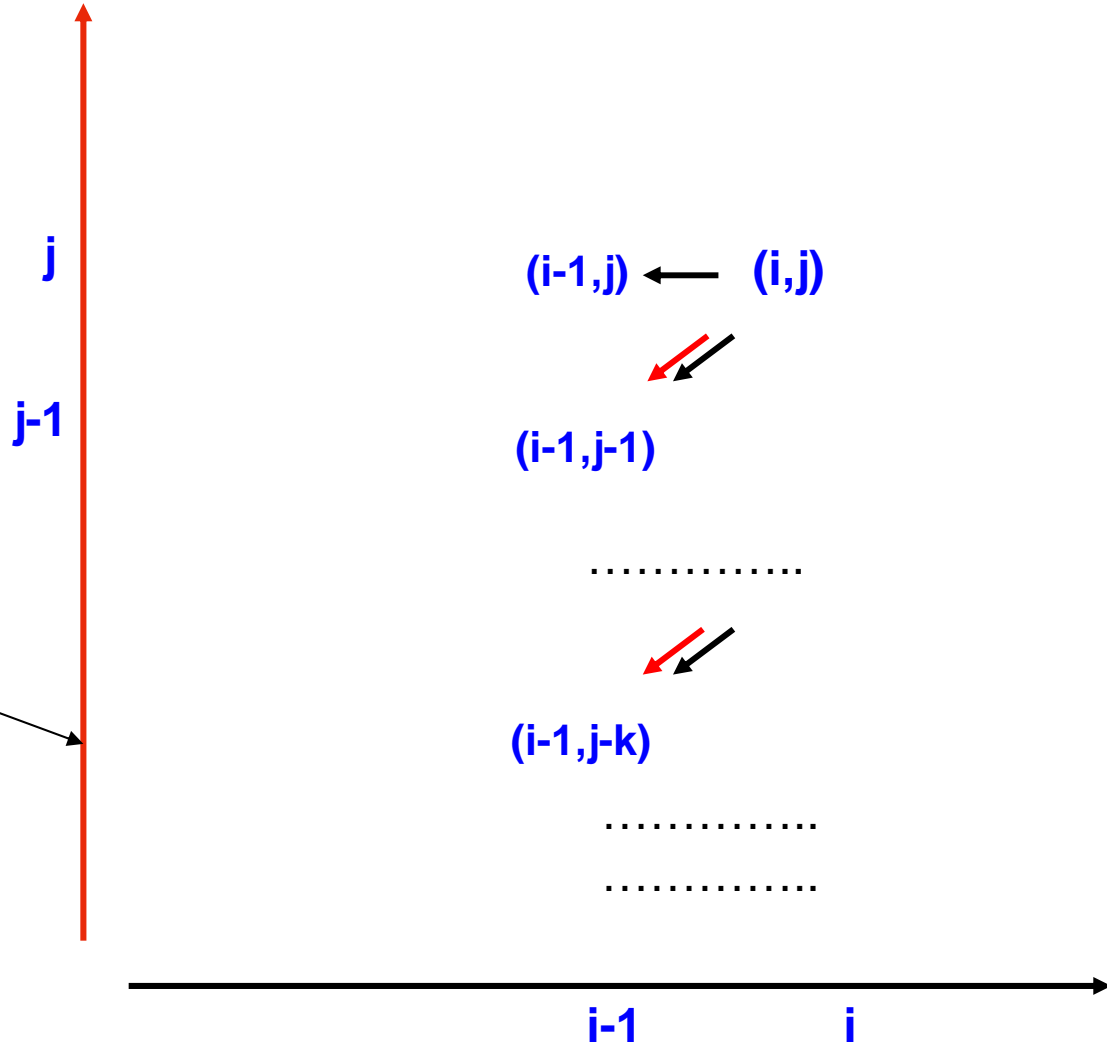
0... j (j+1) cases

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

survive

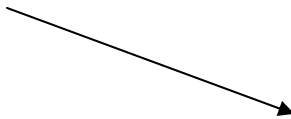


death



Initial condition:

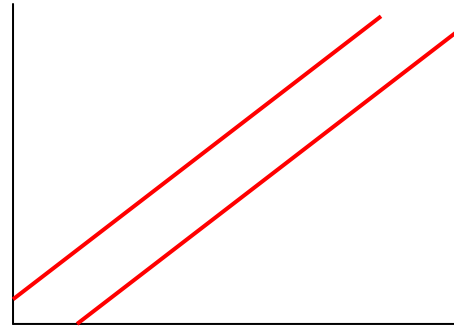
$p'' = s2[1:j]$



Acceleration of Pairwise Algorithm

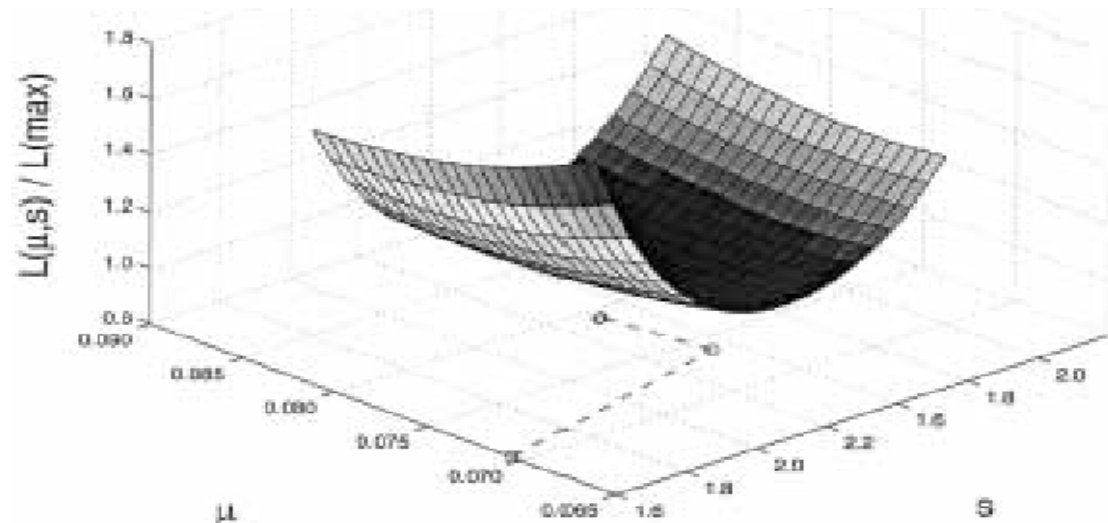
(From Hein,Wiuf,Knudsen,Moeller & Wiebling 2000)

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 ~10⁶

α -globin (141) and β -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 : $-\log(\alpha\text{-globin}, \beta\text{-globin}) = -\log(l(\text{sumalign}))$

λ^*t : 0.0371805 +/- 0.0135899
 μ^*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

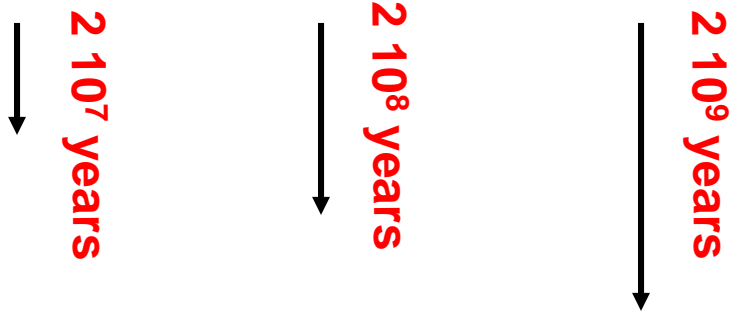
NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
DGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVLCVLAHFGKEFTPPVQAAYQKVVAGVANALAHKYH

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

The invasion of the immortal link

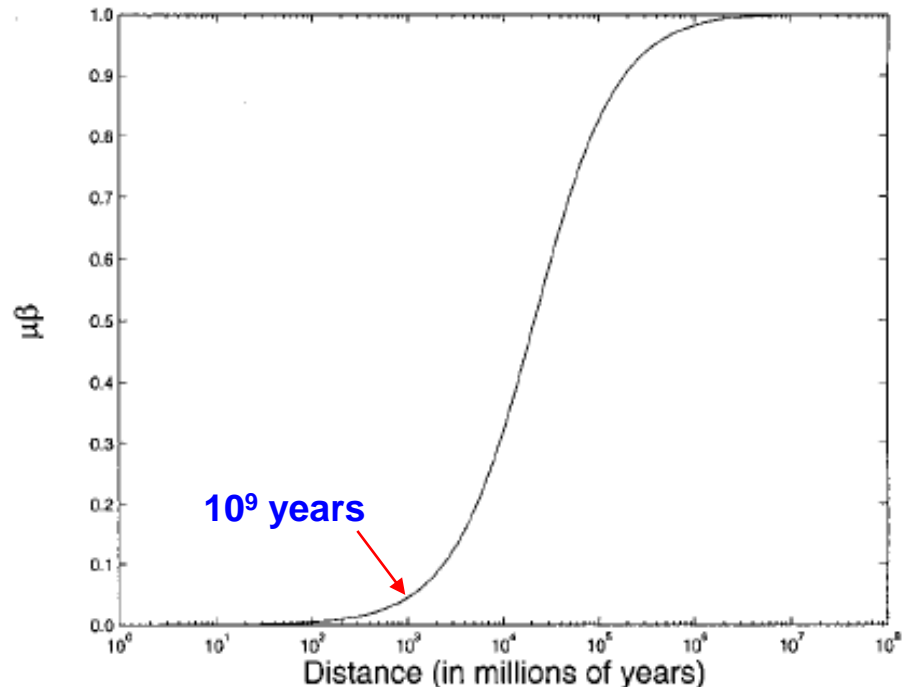
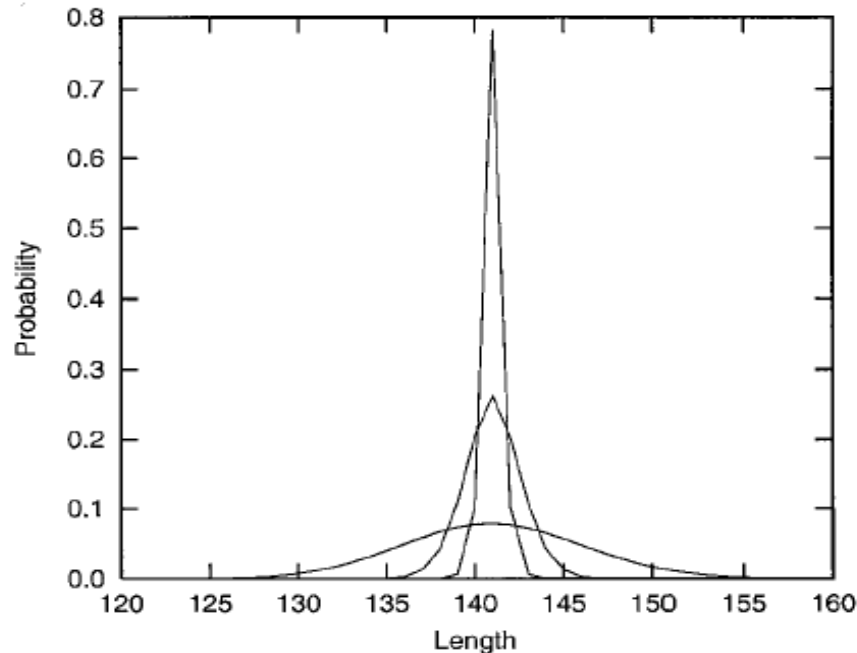
VLSPADNAL.....DLHAHKR 141 AA long

*##### ... ### 141 AA long



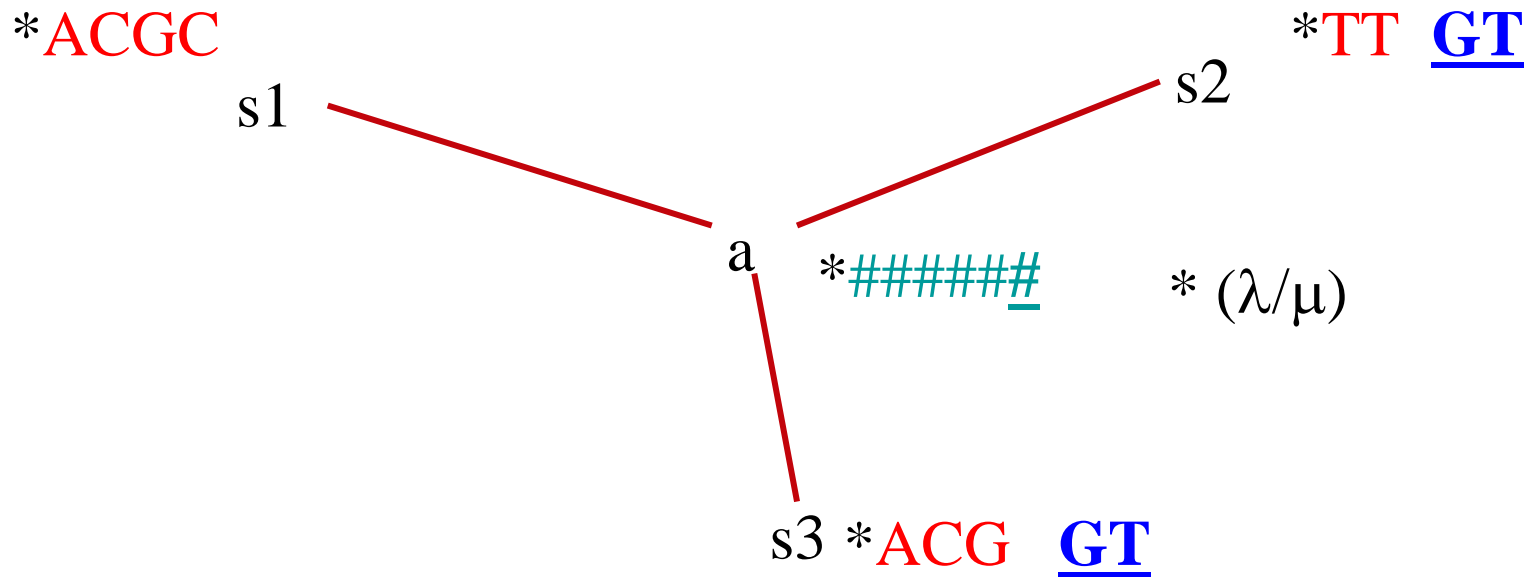
????????????????????? k AA long

*##### ... ###



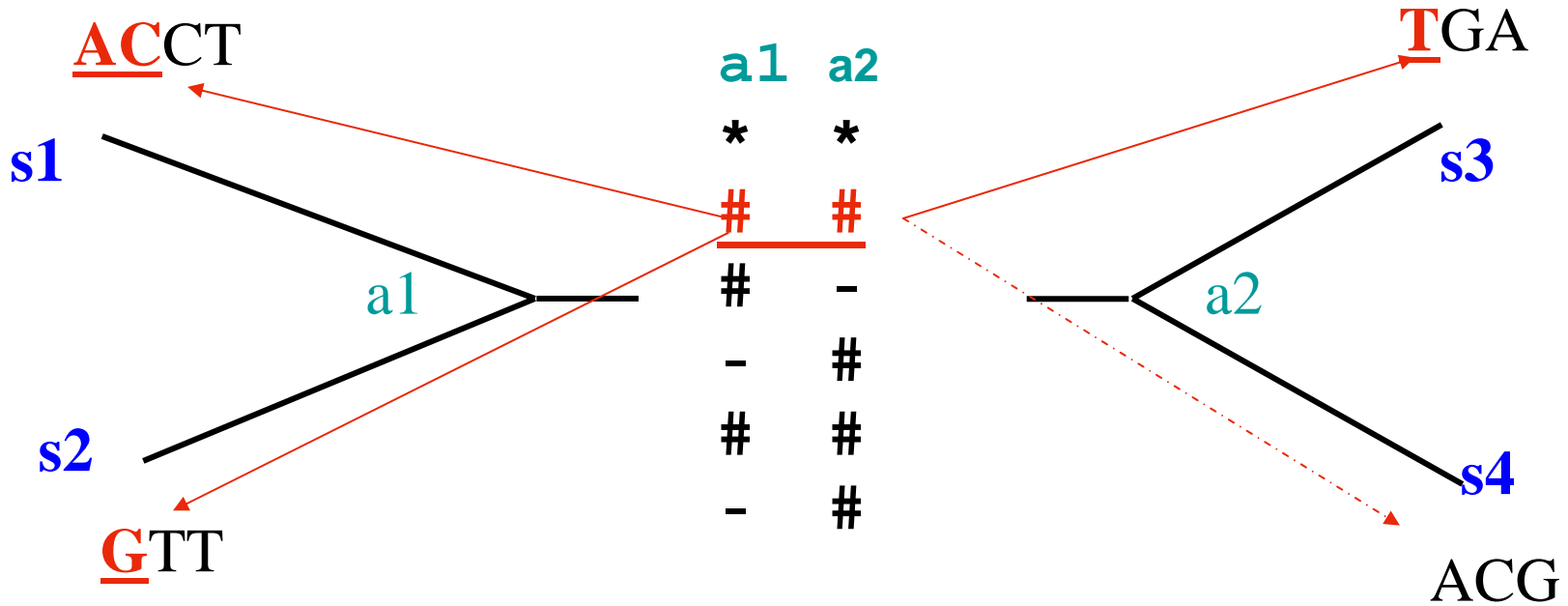
Algorithm for alignment on star tree ($O(\text{length}^6)$)

(Steel & Hein, 2001)



$$P(S) = \left(1 - \frac{\lambda}{\mu}\right) [P_*(S) + \frac{\lambda}{\mu} \sum P_{\#}(Tail) P(S - Tail)]$$

Binary Tree Problem



The problem would be simpler if:

- The ancestral sequences & their alignment was known.
- The alignment of ancestral alignment columns to leaf sequences was known

How to sum over all possible ancestral sequences and their alignments?:

A Markov chain generating ancestral alignments can solve the problem!!

Generating Ancestral Alignments

- # # E
- E

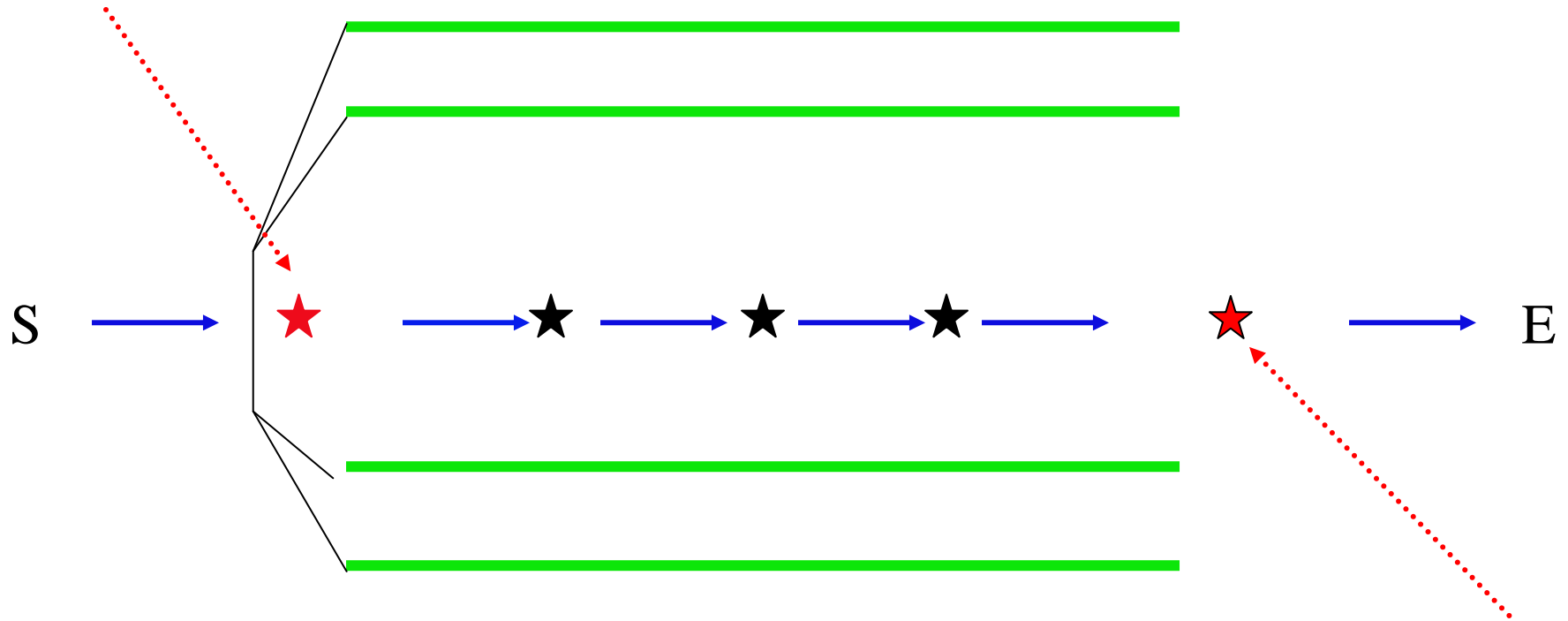
*	$\lambda\beta$	<u>$\lambda/\mu (1-\lambda\beta)e^{-\mu}$</u>	$\lambda/\mu (1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu) (1-\lambda\beta)$
*				
#	$\lambda\beta$	$\lambda/\mu (1-\lambda\beta)e^{-\mu}$	$\lambda/\mu (1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu) (1-\lambda\beta)$
#				
-	$\lambda\beta$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

a1 * - # E
a2 * # # E

$\lambda\beta$ $\lambda/\mu (1-\lambda\beta)e^{-\mu}$ $(1-\lambda/\mu) (1-\lambda\beta)$

The Basic Recursion

”Remove 1st step” - recursion:



”Remove last step” - recursion:

Last/First step removal are inequivalent, but have the same complexities.
First step algorithm is the simplest.

Sequence Recursion: First Step Removal

$P_\alpha(S_k)$: Epifixes ($\underline{S}[\underline{k}+1:\underline{l}]$) starting in given MC starts in α .

$$P_\alpha(S_k) =$$

$$\sum_{\varepsilon} \sum_{i \in S_\alpha} \sum_{H \in C_\alpha} P'({}^k S_i, H | \alpha) P(\alpha \rightarrow \varepsilon) P_\varepsilon(S_i)$$

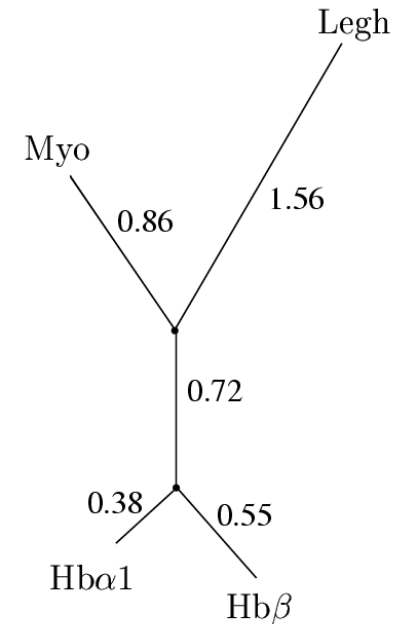
Where $P'({}^k S_i, H | \alpha) =$

$$F({}^k S_i, H) \left(\prod_{j: H(j)=0} p'_k(t_j) \pi_{s_j} [i(j) : k(j)] \right) \left(\prod_{j: H(j)=1} p_k(t_j) \pi_{s_j} [i(j)+1 : k(j)] \right)$$

Maximum likelihood phylogeny and alignment

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

Gerton Lunter
Istvan Miklos
Alexei Drummond
Yun Song



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

VAHVDDMPNALSALSDLHAHKL R VDPVNFK-LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL-TS-K---YR-
LAHLDNLKGT FATLSELHCDKLHVDPENFR-LLGNVLVCVLAHFGKEFTPPVQAA YQKV VAGVANAL-AH-K---YH-
LKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
LRAKGEVVLADPTLGSVHVQKGVLDP-HFL-VVKEALLKTFKEAVGDKWNDELGN AWEVAYDELA AAI-KK-A-MGSA-

Metropolis-Hastings Statistical Alignment

Lunter, Drummond, Miklos, Jensen & Hein, 2005

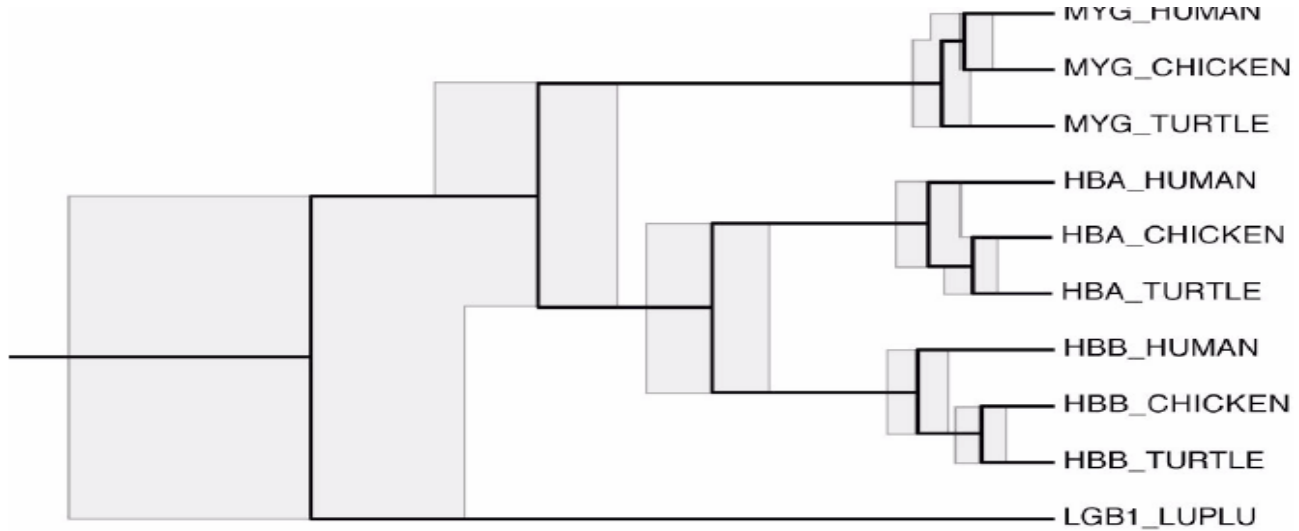


Figure 6

0.2 substitutions per site

