Approaches to Sequence Analysis

Data {GTCAT,GTTGGT,GTCA,CTCA}

Parsimony, similarity, optimisation.

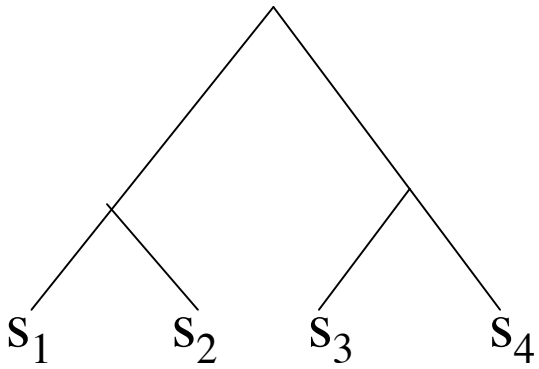
GT-CAT

GTTGGT

GT-CA-

CT-CA-

statistics

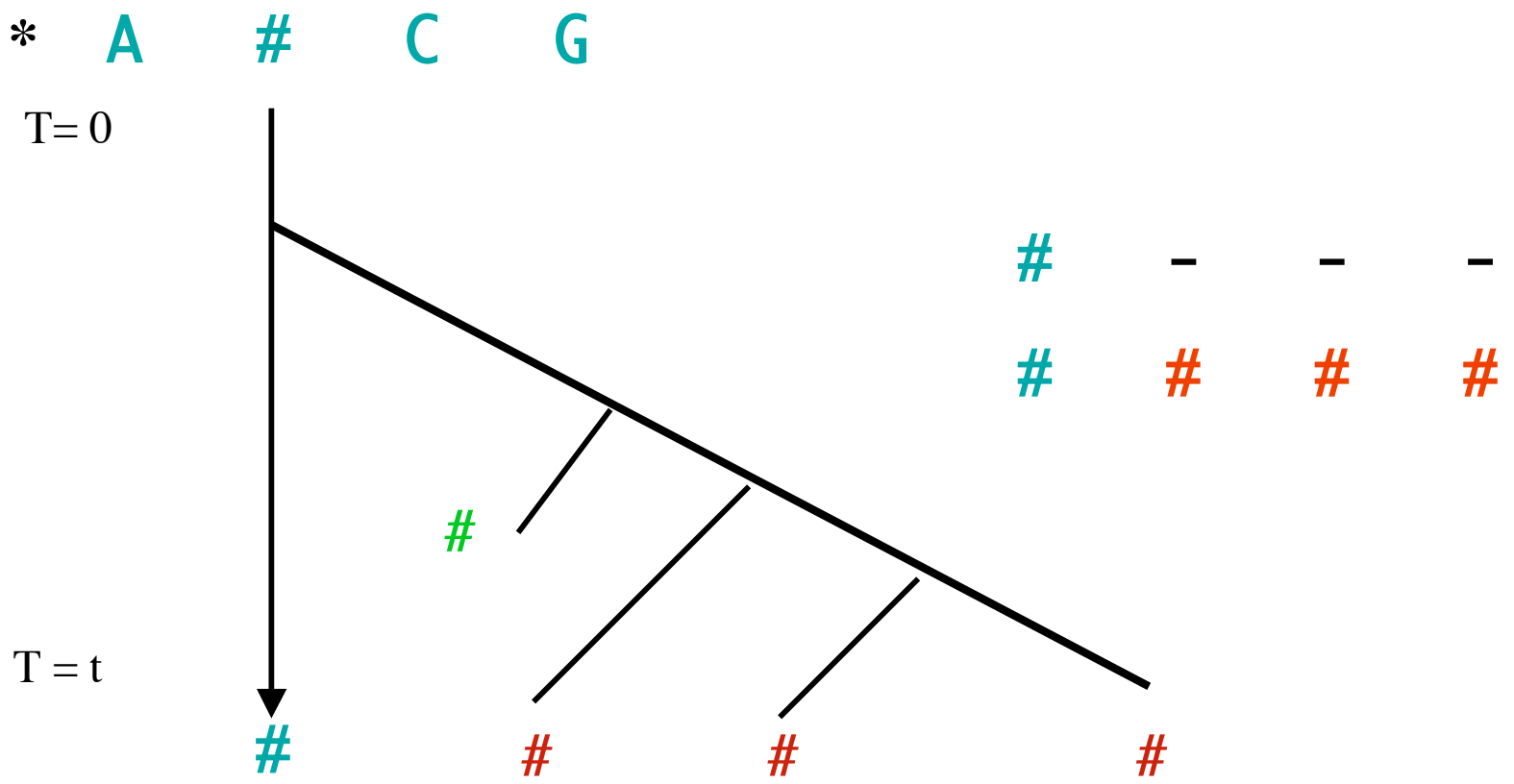


Actual Practice: 2 phase analysis.

Ideal Practice: 1 phase analysis.

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

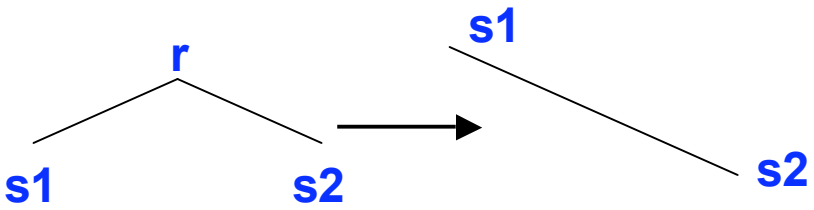
Thorne-Kishino-Felsenstein (1991) Process



λ (birth rate) < μ (death rate)

1. $P(s) = (1-\lambda/\mu)(\lambda/\mu)^{l} \quad \pi_A^{\#A} * .. * \pi_T^{\#T}$ **l =length(s)**

2. Time reversible: s1 ⇄ s2



λ & μ into Alignment Blocks

A. Amino Acids Ignored:

```
# - - -        # - - - -        * - - - -
# # # #        - # # # #        * # # # #
      k              k                k
```

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

$p_k(t)$ $p'_k(t)$ $p''_k(t)$

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

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

B. Amino Acids Considered:

```
T - - -
R Q S W
```

$$P_t(T\text{-->}R)*\pi_Q*..*\pi_W*p_4(t)$$

4

```
T - - - -
- R Q S W
```

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

Diff. Equations for p-functions

- - ... -
...

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

- - - ... -
- # # # ... #

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

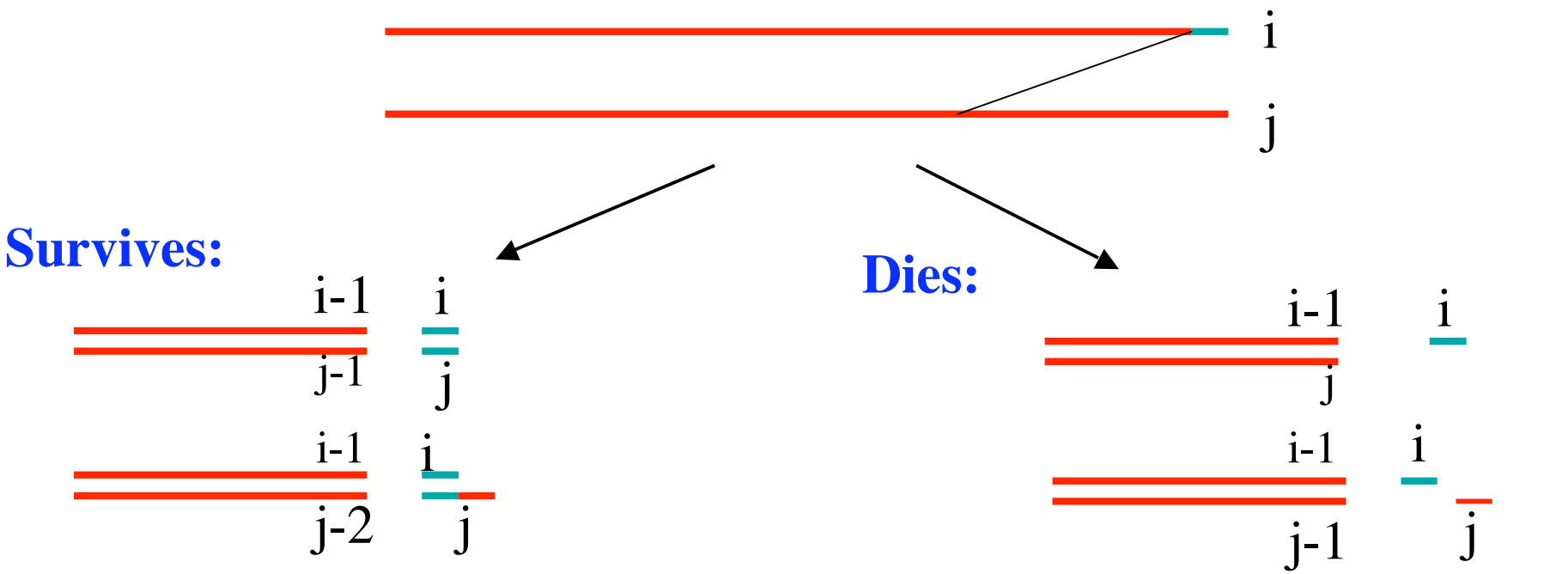
* - - - ... -
* # # # ... #

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

Initial Conditions: $p_k(0) = p_k''(0) = p'_k(0) = 0 \quad k>1$

$p_0(0) = p_0''(0) = 1. \qquad p'_0(0) = 0$

Basic Pairwise Recursion (O(length^3))



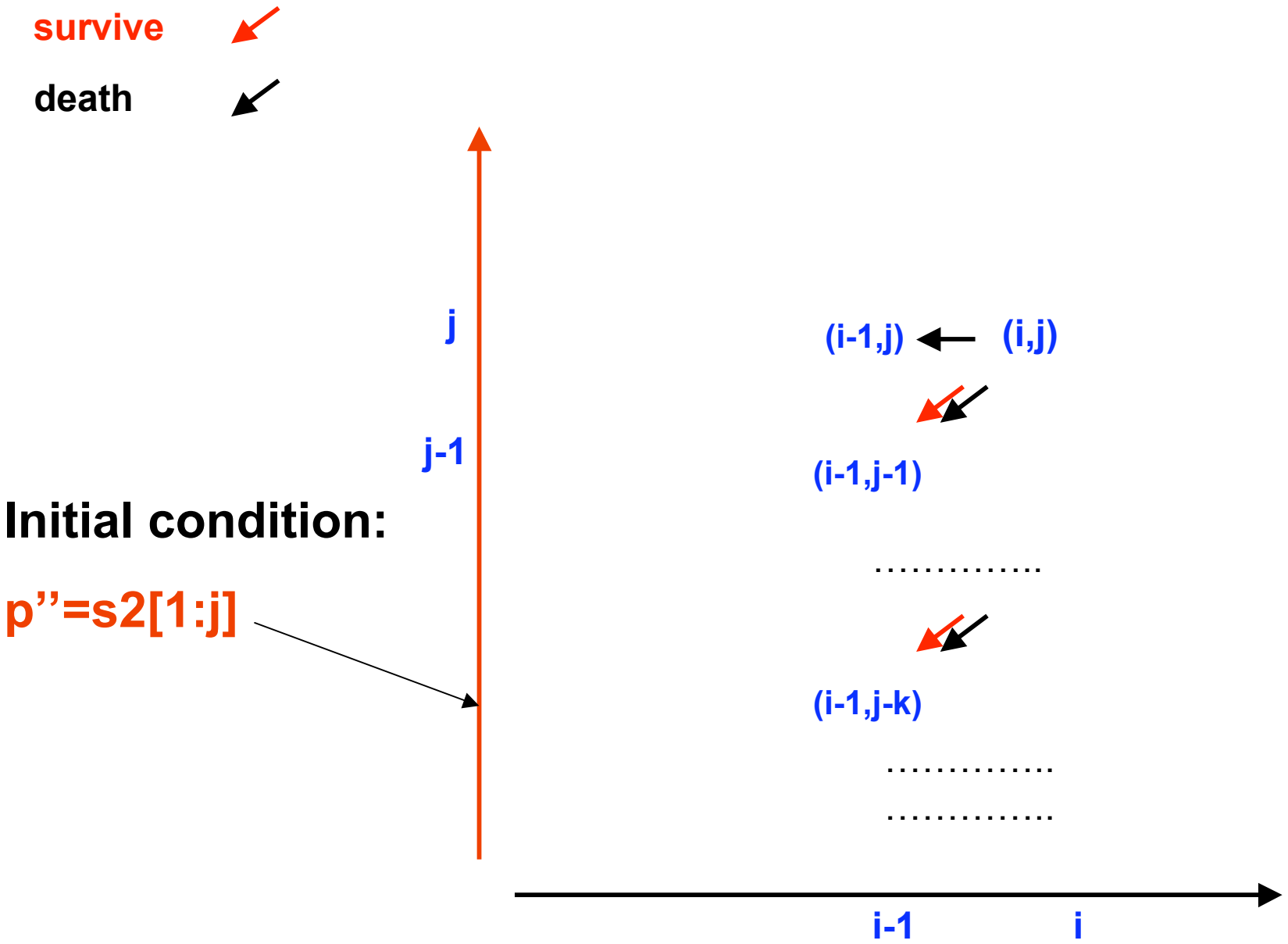
$$P(s1_{i-1} \rightarrow s2_{j-2}) * p_2 * f(s1[i], s2[j-1])\pi(s2[j])$$

$$P(s1_{i-1} \rightarrow s2_{j-1}) * p'_1 * \pi(s2[j])$$

1… j (j) cases

0… j (j+1) cases

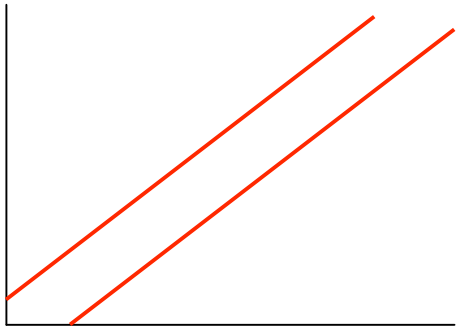
Basic Pairwise Recursion (O(length^3))



Accelleration of Pairwise Algorithm

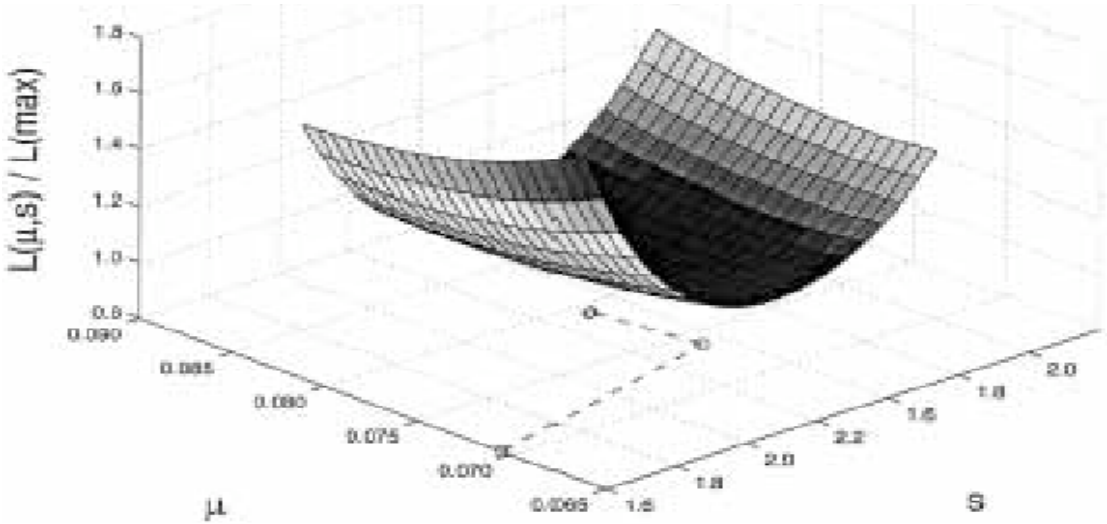
(From Hein,Wiuf,Knudsen,Moeller & Wiebling 2000)

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 ~10^6

α-globin (141) and β-globin (146)

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

430.108 : -log(α-globin)
327.320 : -log(α-globin --> β-globin)
730.428 : -log(α-globin, β-globin) = -log(l(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-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
DGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Ratio l(maxalign)/l(sumalign) = 0.00565064

The invasion of the immortal link

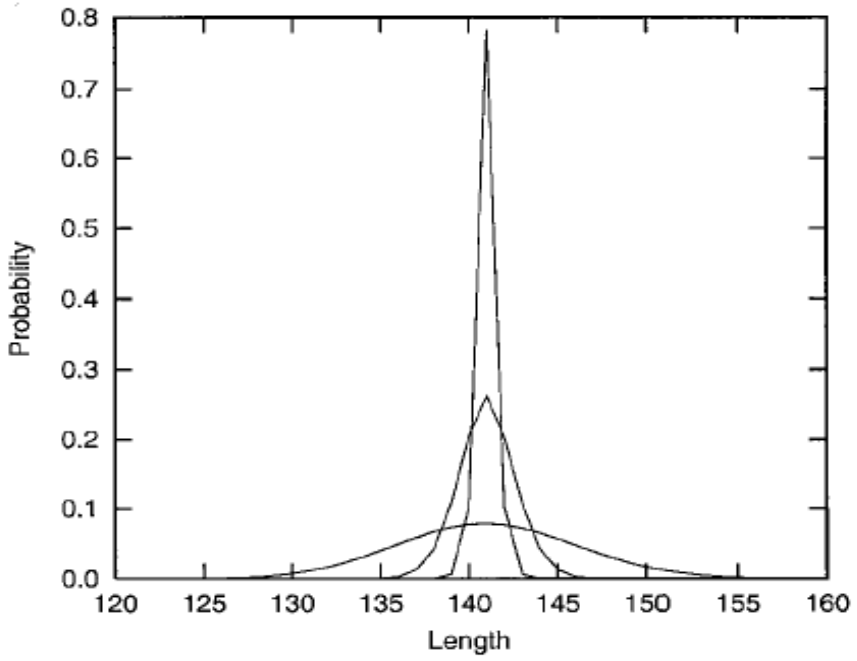
VLSPADNAL.....DLHAHKR 141 AA long

$2\ 10^7$ years

$2\ 10^8$ years

$2\ 10^9$ years

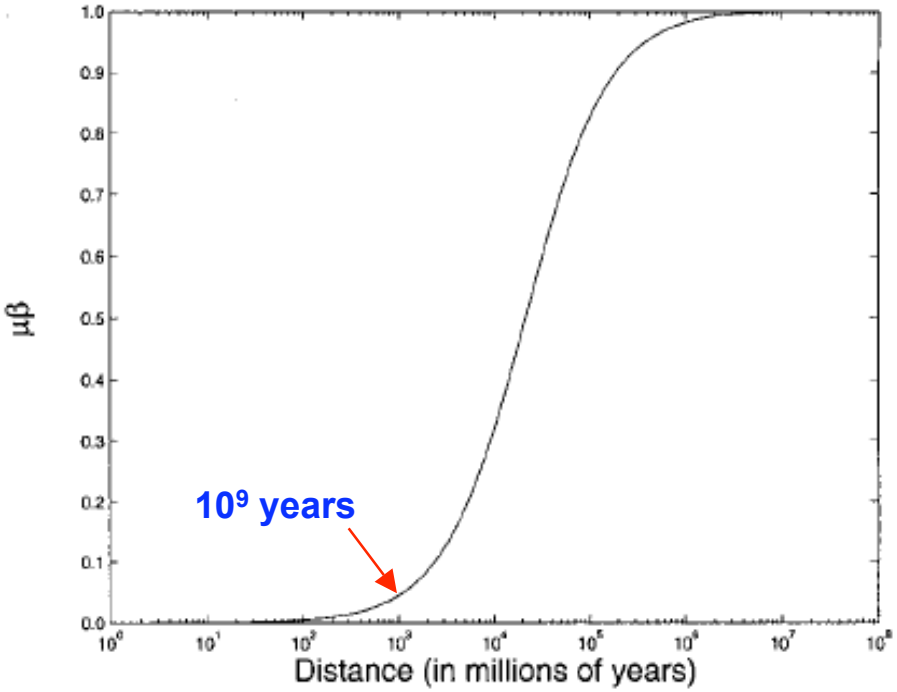
?????????????????? k AA long



*########## …. ### 141 AA long

*########### …. ###

*########### …. ###

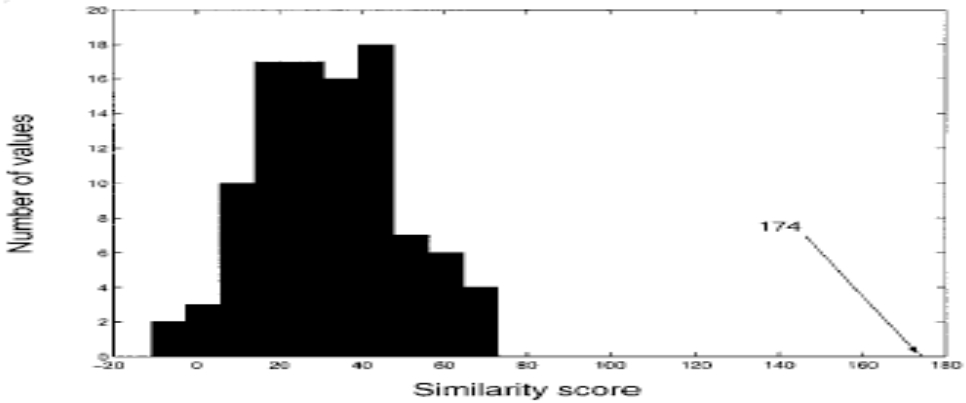


Homology test.

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

```
Real    s1  = ATWYFCAK-AC
        s2  = ETWYKCALLAD
              *** **   *
```

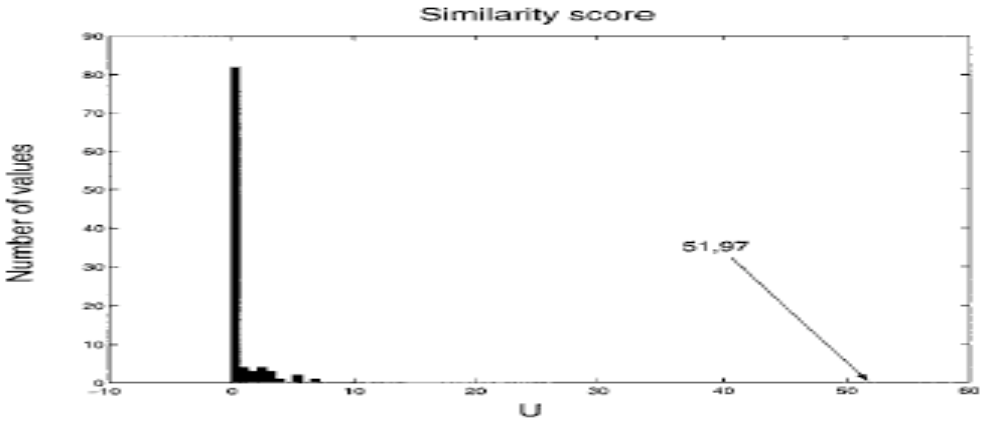
$W_{i,j} = -\ln(\pi_i * P^{2.5}{}_{i,j} / (\pi_i * \pi_j))$



D(s1,s2) is evaluated in D(s1,s2*)

α-, myoglobin homology tests

```
Random s1  = ATWYFC-AKAC
       s2* = LTAYKADCWLE
             *
```

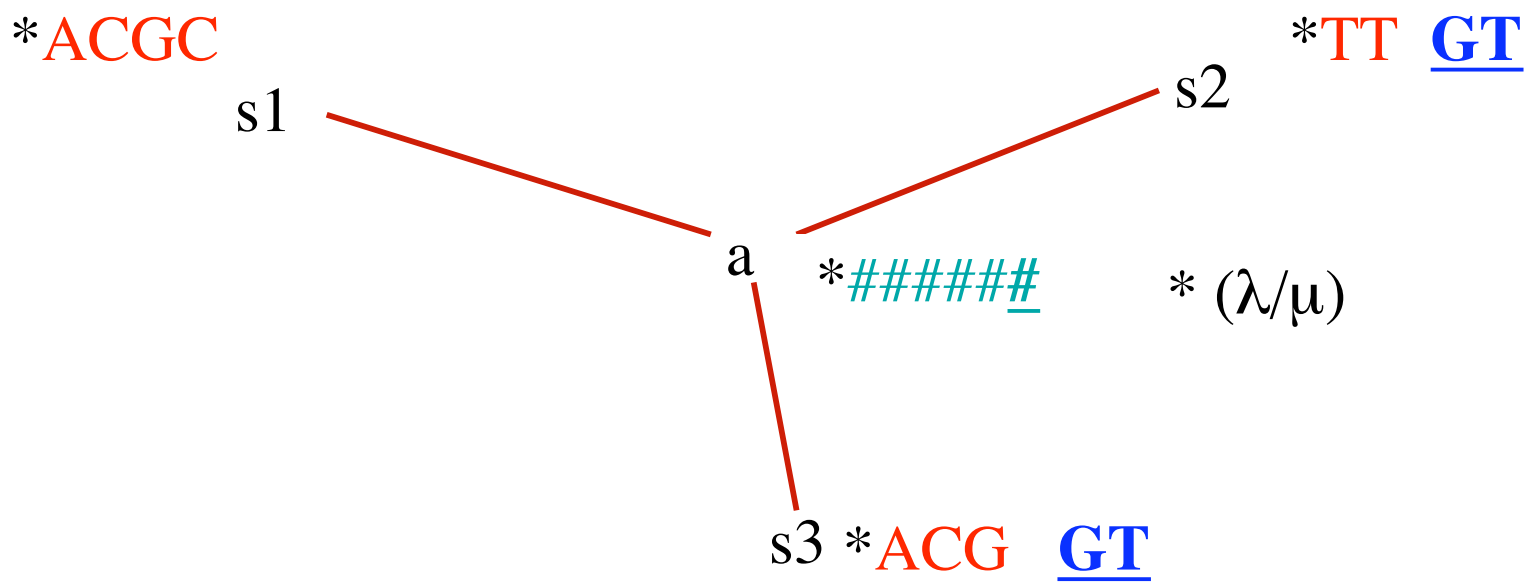


1. Test the competing hypothesis that 2 sequences are 2.5 events apart versus infinitely far apart.

2. It only handles substitutions "correctly". The rationale for indel costs are more arbitrary.

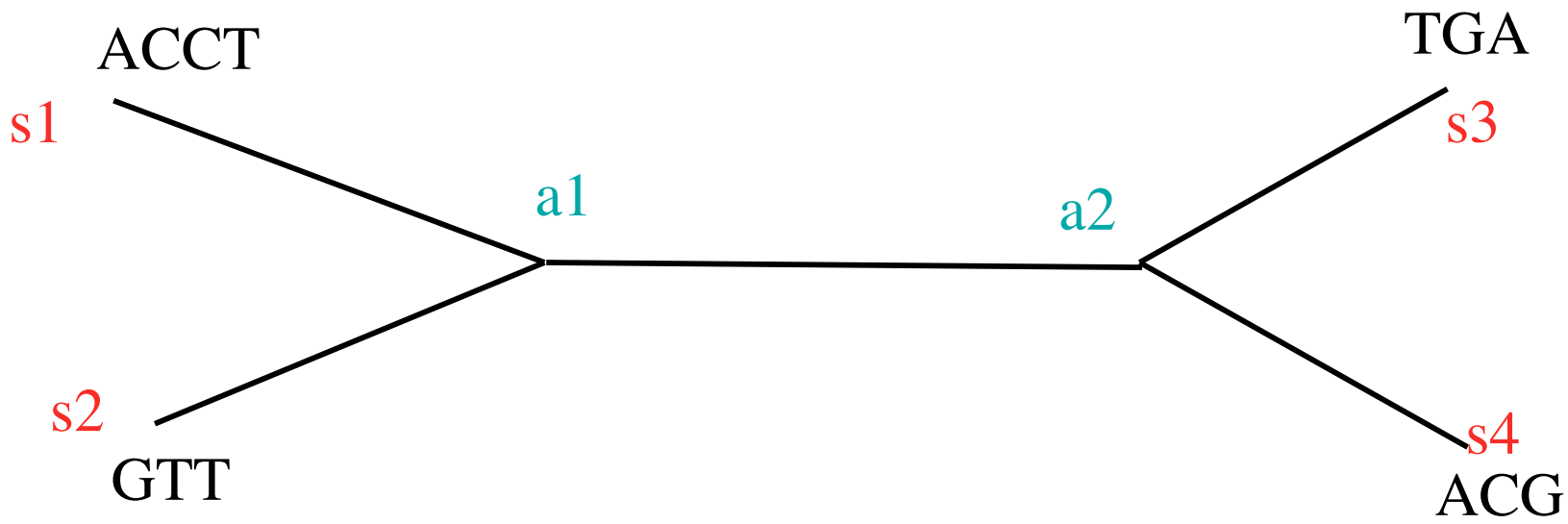
Algorithm for alignment on star tree (O(length^6))

(Steel & Hein, 2001)



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

Binary Tree Problem

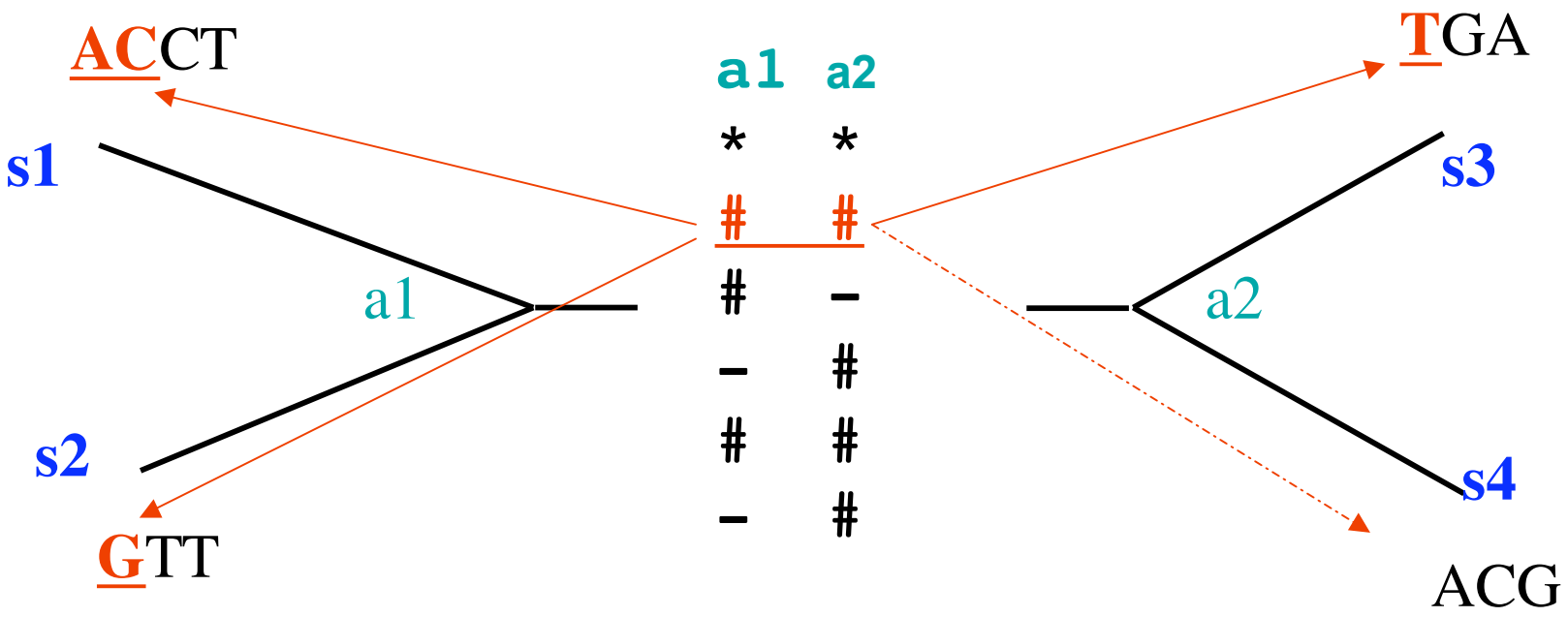


Binary Tree Problem

The problem would be simpler if:

i. **The ancestral sequences & their alignment was known.**

ii. **The alignment of ancestral alignment columns to leaf sequences was known.**



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

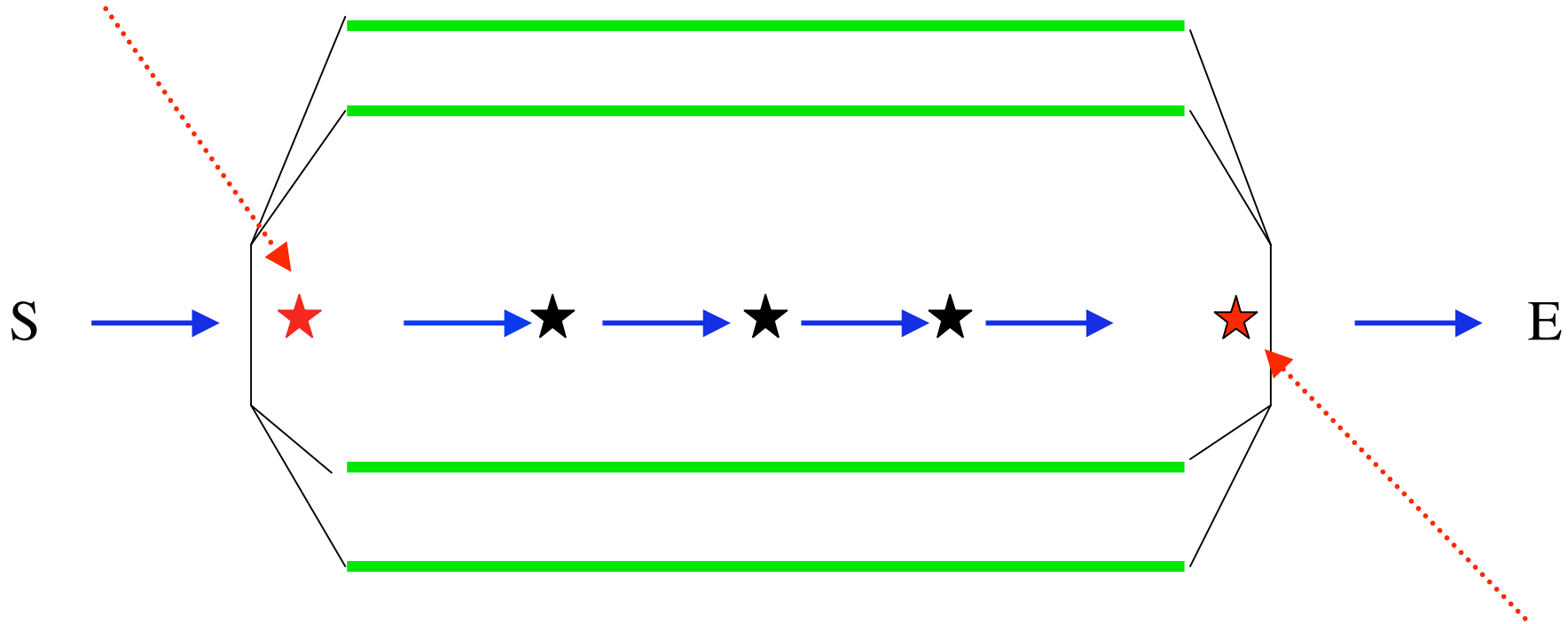
Generating Ancestral Alignments.

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

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 ($S[k+1:l]$) starting in given MC starts in α.

$P_\alpha(S_k) =$

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

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

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

Maximum likelihood phylogeny and alignment

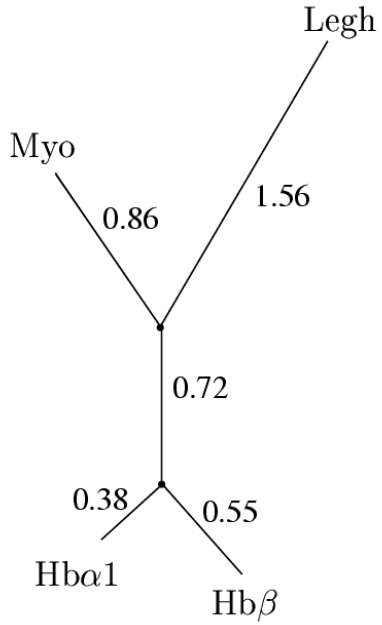
Human alpha hemoglobin;
Human beta hemoglobin;
Human myoglobin
Bean leghemoglobin

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

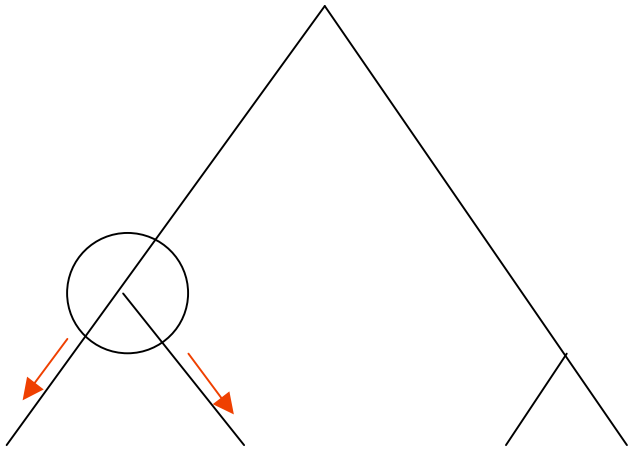
```
Hba1: MV--LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF--DLS-H-----GSAQVKGHGKKVAD-AL-TNA-
Hbb:  MV-HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESF-GDLSTPDAVM-GNPKVKAHGKKVLG-AF-SDG-
Myo:  MG--LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK-HLKSEDE-MKASEDLKKHGATVLT-AL-GGI-
Legh: MGA-FSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFS-F---LSNGVD-P-NNPKLKAHAEKVFKMTVDSAVQ

      VAHVDDMPNALSALSDLHAHKLRVDPVNFK-LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL-TS-K---YR-
      LAHLDNLKGTFATLSELHCDKLHVDPENFR-LLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL-AH-K---YH-
      LKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
      LRAKGEVVLADPTLGSVHVQKGVLDP-HFL-VVKEALLKTFKEAVGDKWNDELGNAWEVAYDELAAAI-KK-A-MGSA-
```

Gibbs Samplers for Statistical Alignment

Holmes & Bruno (2001):

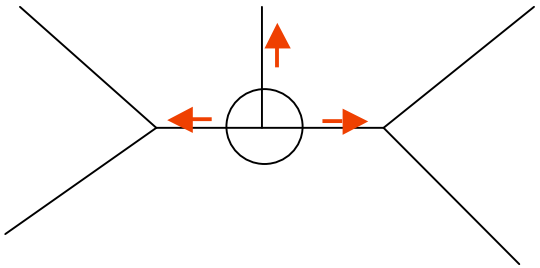
Sampling Ancestors to pairs.



Jensen & Hein (in press):

Sampling nodes adjacent to triples

Slower basic operation, faster mixing



Metropolis-Hastings Statistical Alignment.

Lunter, Drummond, Miklos, Jensen & Hein, 2005

The alignment moves:

We choose a random window in the current alignment

Then delete all gaps so we get back subsequences

Stochastically realign this part

The phylogeny moves:

As in Drummond et al. 2002

Metropolis-Hastings Statistical Alignment

Lunter, Drummond, Miklos, Jensen & Hein, 2005

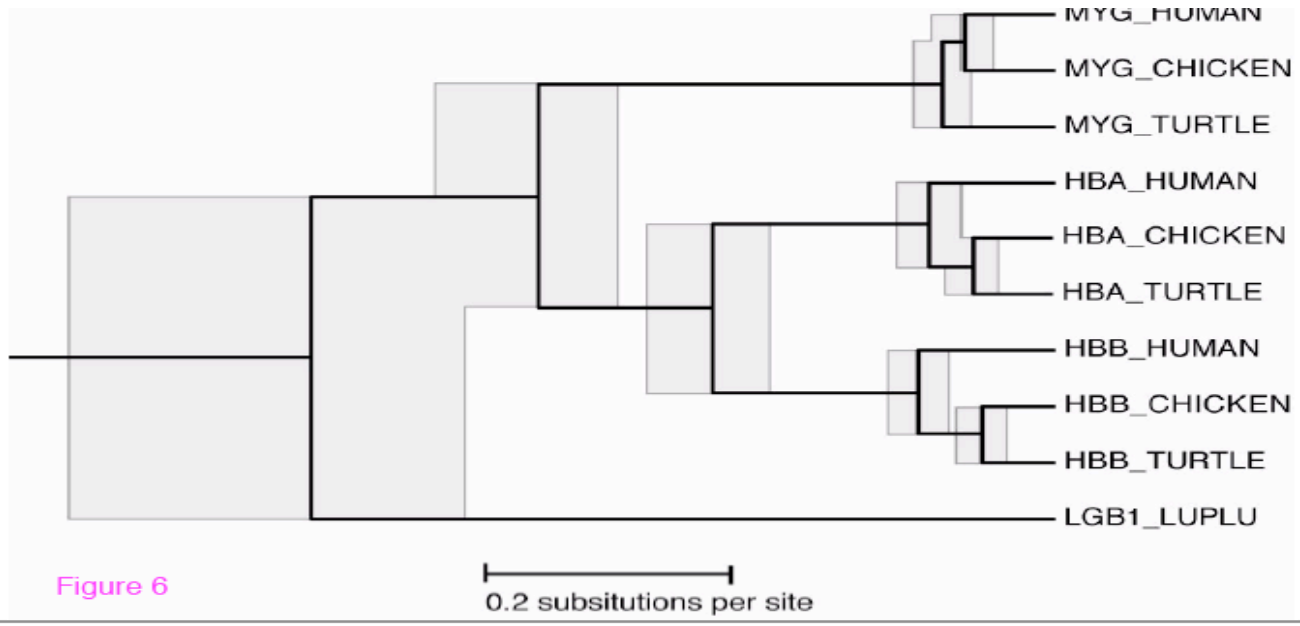
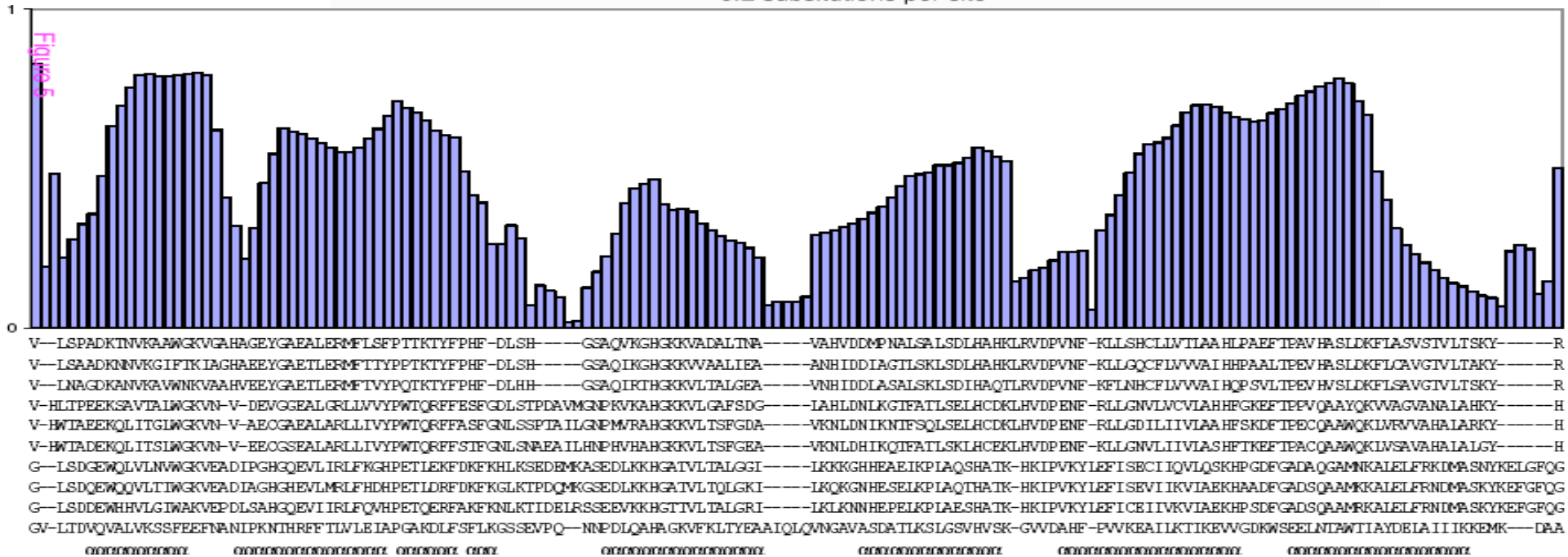


Figure 6



References Statistical Alignment

•Fleissner R, Metzler D, von Haeseler A. Simultaneous statistical multiple alignment and phylogeny reconstruction.Syst Biol. 2005 Aug;54(4):548-61.

•Hein,J., C.Wiuf, B.Knudsen, Møller, M., and G.Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. (J. Molecular Biology 302.265-279)

•Hein,J.J. (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a binary tree. (Pac.Symp.Biocompu. 2001 p179-190 (eds RB Altman et al.)

•Steel, M. & J.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. (Letters in Applied Mathematics)

•Hein JJ, J.L.Jensen, C.Pedersen (2002) Algorithms for Multiple Statistical Alignment. (PNAS) 2003 Dec 9;100(25):14960-5.

• **Holmes, I.** (2003) Using Guide Trees to Construct Multiple-Sequence Evolutionary HMMs. *Bioinformatics*, special issue for ISMB2003, 19:147i–157i.

• Jensen, J.L. & **Hein, J.** (2004) A Gibbs sampler for statistical multiple alignment. *Statistica Sinica*, in press.

• **Miklós, I., Lunter, G.A. & Holmes, I.** (2004) A 'long indel' model for evolutionary sequence alignment. *Mol. Biol. Evol.* 21(3):529–540.

• **Lunter, G.A.**, Miklós, I., Drummond, A.J., Jensen, J.L. & **Hein, J.** (2005) Bayesian Coestimation of Phylogeny and Sequence Alignment. *BMC Bioinformatics*, 6:83

• **Lunter, G.A., Miklós, I., Drummond, A.,** Jensen, J.L. **& Hein, J.** (2003) Bayesian phylogenetic inference under a statistical indel model. ps pdf *Lecture Notes in Bioinformatics, Proceedings of WABI'03*, 2812:228–244.

• **Lunter, G.A., Miklós, I., Song, Y.S. & Hein, J** (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.*, 10(6):869–88Miklos, Lunter & Holmes (2002) (submitted ISMB)

•Miklos, I & Toroczkai Z. (2001) An improved model for statistical alignment, in WABI2001, Lecture Notes in Computer Science, (O. Gascuel & BME Moret, eds) 2149:1-10. Springer, Berlin

•Metzler D. "Statistical alignment based on fragment insertion and deletion models." Bioinformatics. 2003 Mar 1;19(4):490-9.

•Miklos, I (2002) An improved algorithm for statistical alignment of sequences related by a star tree. Bul. Math. Biol. 64:771-779.

•Miklos, I: Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution Disc. Appl. Math. accepted.

•Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution.J Mol Evol. 1992 Jan;34(1):3-16.

•Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences.J Mol Evol. 1991 Aug;33(2):114-24. Erratum in: J Mol Evol 1992 Jan;34(1):91.

•Thorne JL, Churchill GA. Estimation and reliability of molecular sequence alignments.Biometrics. 1995 Mar;51(1):100-13.