

# Optimisation Alignment.

<http://www.stats.ox.ac.uk/~hein/lectures.htm>

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

V-LSPADKTNVKAANGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADAL  
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAF

TNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR  
SDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

**It often matches functional region with functional region**

**Determines homology at residue/nucleotide level.**

**Similarity/Distance between molecules can be evaluated**

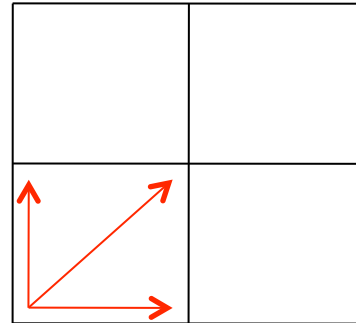
**Molecular Evolution studies.**

**Homology/Non-homology depends on it.**

# Number of alignments, $T(n,m)$

Alignments columns are equivalent to step (0,1), (1,0) and (1,1) in a  $[0,n][0,m]$  matrix.

Thus alignment by alignment search for best alignment is not realistic.



$T(n,m)$  is the number of alignments of  $s1[1,n]$  and  $s2[1,m]$  then

$$T(n,m) = T(n-1,m) + T(n,m-1) + T(n-1,m-1)$$

$$T(0,0) = 1 \quad T(n,m) > 3^{\min(n,m)}$$

If  $\begin{matrix} n- \\ -n \end{matrix}$  is equivalent to  $\begin{matrix} -n \\ n- \end{matrix}$

then alignments are equivalent to choosing two subsets of  $s1$  and  $s2$  that has to be matched, thus

$$T(n,m) = \sum_{i=1}^{\min(n,m)} \binom{n}{i} \binom{m}{i}$$

|   |  |   |   |    |     |     |     |
|---|--|---|---|----|-----|-----|-----|
|   |  | 1 | 9 | 41 | 129 | 321 | 681 |
| T |  | 1 | 7 | 25 | 63  | 129 | 231 |
| G |  | 1 | 5 | 13 | 25  | 41  | 61  |
| T |  | 1 | 3 | 5  | 7   | 9   | 11  |
| T |  | 1 | 1 | 1  | 1   | 1   | 1   |
|   |  | C | T | A  | G   | G   |     |

# Parsimony Alignment of two strings.

Sequences: s1=CTAGG s2=TTGT. 5, indels (g) 10.

Basic operations:

transitions 2 (C-T & A-G), transversions 5, indels (g) 10.

---

Cost Additivity

$$\begin{array}{ccccccc} & \text{CTAG} & & & \text{CTA} & & \text{G} \\ & & = & & & + & \\ & \text{TT-G} & & & \text{TT-} & & \text{G} \end{array}$$


---

$$\begin{array}{l} \{CTAG, TTG\}_{AL} = \text{Min} \left[ \begin{array}{l} \{CTA, TT\}_{AL} + GG \quad (A) \\ \{CTA, TTG\}_{AL} + G- \quad (B) \\ \{CTAG, TT\}_{AL} + -G \quad (C) \end{array} \right] \end{array}$$

12
12
4
10
32
0
10
10

---

$$D_{i,j} = \min\{D_{i-1,j-1} + d(s1[i],s2[j]), D_{i,j-1} + g, D_{i-1,j} + g\}$$

Initial condition:  $D_{0,0}=0$ . ( $D_{i,j} := D(s1[1:i], s2[1:j])$ )

|   |    |    |    |    |    |   |  |    |
|---|----|----|----|----|----|---|--|----|
|   |    |    |    |    |    |   |  |    |
| F | 40 | 32 | 22 | 14 | 9  |   |  | 17 |
| G | 30 | 22 | 12 | 4  | 12 |   |  | 22 |
| F | 20 | 12 | 2  | 12 | 22 |   |  | 32 |
| F | 10 | 2  | 10 | 20 | 30 |   |  | 40 |
| F | 0  | 10 | 20 | 30 | 40 |   |  | 50 |
|   |    | C  | T  | A  | G  | G |  |    |

Alignment:

CTAGG

i v

TT-GT

Cost 17

# Complexity of Accelerations of pairwise algorithm.

**Dynamical Programming:  $(n+1)(m+1)3=O(nm)$**

**Backtracking:  $O(n+m)$**

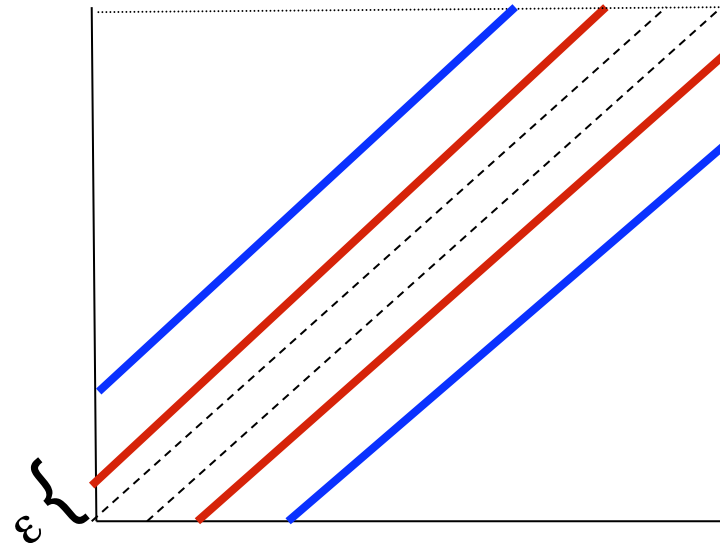
**Recursion without memory:  $T(n,m) > 3^{\min(n,m)}$**

Exact acceleration (Ukkonen, Myers).

Assume all events cost 1.

If  $d_\varepsilon(s_1, s_2) < 2\varepsilon + |l_1 - l_2|$ , then

$d(s_1, s_2) = d_\varepsilon(s_1, s_2)$



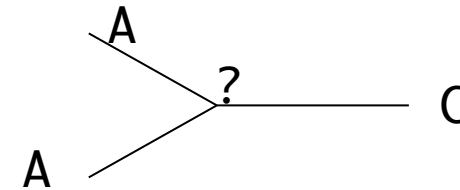
Heuristic acceleration: Smaller band & larger acceleration, but no guarantee of optimum.

# Alignment of three sequences.

s1=ATCG    s2=ATGCC    s3=CTCC

Alignment:    **AT-CG**  
                   **ATGCC**  
                   **CT-CC**

**A**  
**A**  
**C**



Consensus sequence: ATCC

Configurations in an alignment column:

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| - | - | n | n | n | - | n | - |
| - | n | - | n | - | n | n | - |
| n | - | - | - | n | n | n | - |

Recursion:     $D_{i,j,k} = \min\{D_{i-i',j-j',k-k'} + d(i,i',j,j',k,k')\}$

Initial condition:     $D_{0,0,0} = 0.$

Running time:  $l_1 * l_2 * l_3 * (2^3 - 1)$  Memory requirement:  $l_1 * l_2 * l_3$

New phenomena: ancestral/consensus sequence.

# Parsimony Alignment of four sequences

s1=ATCG    s2=ATGCC    s3=CTCC    s4=ACGCG

Alignment:    AT-CG                    G  
                   ATGCC                    C  
                   CT-CC                    C  
                   ACGCG                    G



Configurations in alignment columns:

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | n | - | - | - | n | n | n | - | n | n | n | n | - |
| - | - | n | - | n | n | - | n | - | - | n | - | n | n | n | - |
| - | n | - | - | n | - | n | - | n | - | n | n | - | n | n | - |
| n | - | - | - | - | n | n | - | - | n | n | n | n | - | n | - |

Recursion:             $D_i = \min\{D_{i-\Delta} + d(i, \Delta)\} \quad \Delta \in [\{0, 1\}^4 \setminus \{0\}^4]$

Initial condition:     $D_0 = 0$ .    Memory :  $l_1 * l_2 * l_3 * l_4$

Computation time:     $l_1 * l_2 * l_3 * l_4 * 2^4$     Memory :  $l_1 * l_2 * l_3 * l_4$

New Phenomena:    Cost and alignment is phylogeny dependent

# Alignment of many sequences.

s1=ATCG, s2=ATGCC, . . . . ., sn=ACGCG



Configurations in an alignment column:  $2^n - 1$   
 $\in$

Recursion:  $D_i = \min\{D_{i-\Delta} + d(i, \Delta)\} \Delta \in [ \{0, 1\}^n \setminus \{0\}^n ]$

Initial condition:  $D_{0,0,\dots,0} = 0.$

Computation time:  $l^n * (2^n - 1) * n$       Memory requirement:  $l^n$   
 (l: sequence length, n: number of sequences)

# Progressive Alignment

(Feng-Doolittle 1987 J.Mol.Evol.)

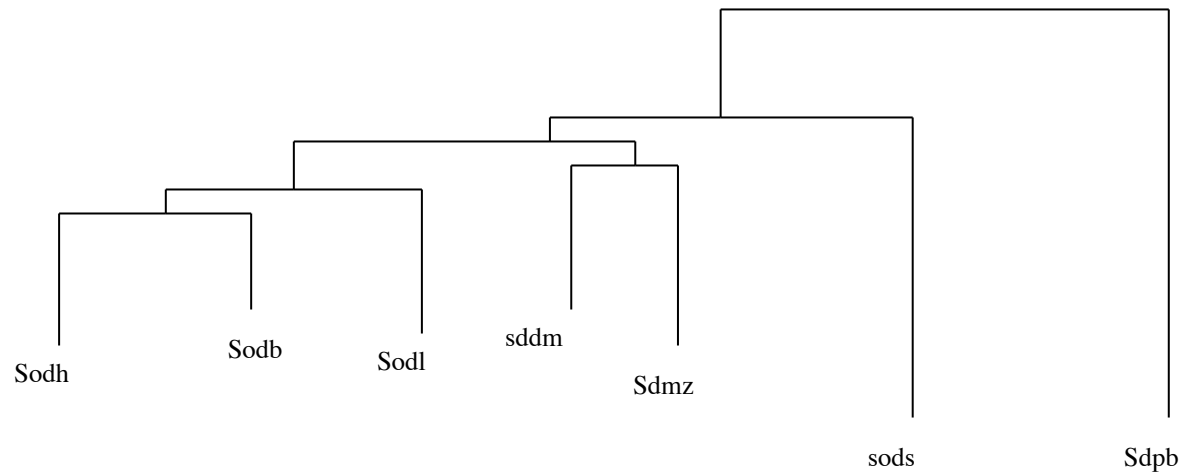
Can align alignments and given a tree make a multiple alignment.

```

      *
alkmny-trwq
akkmdyftrwq
kkkmemftrwq

      *
acdeqrt
acdehrt
  
```

$$[ P(n,q) + P(n,h) + P(d,q) + P(d,h) + P(e,q) + P(e,h) ] / 6$$



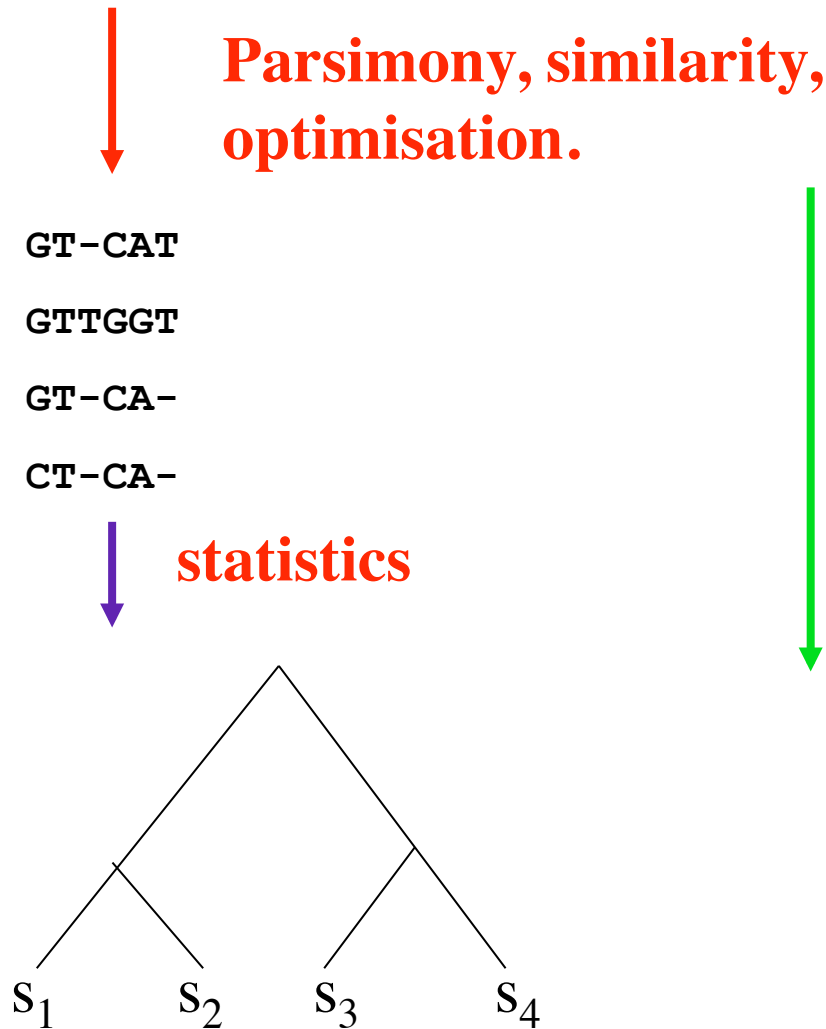
```

      *           *           *** * *           *           * * *
Sodh  atkavcvlkgdgpqvqgsinfeqkesdgpvkvwgsikglte-ghlgfhvhqfg----ndtagct  sagphfnp lsrk
Sodb  atkavcvlkgdgpqvqgtinfeak-gdtvkvwgsikglte--ghlgfhvhqfg----ndtagct  sagphfnp lsrk
Sodl  atkavcvlkgdgpqvqgsinfeqkesdgpvkvwgsikglte-ghlgfhvhqfg----ndtagct  sagphfnp lsrk
Sddm  atkavcvlkgdgpqvq -infeak-gdtvkvwgsikglte--ghlgfhvhqfg----ndtagct  sagphfnp lsrk
Sdmz  atkavcvlkgdgpqvq- infeqkesdgpvkvwgsikglte-ghlgfhvhqfg----ndtagct  sagphfnp Lsrk
Sods  vatkavcvlkgdgpqvq- infeak-gdtvkvwgsikgltepnglhgfhhqfg----ndtagct  sagphfnp lsrk
Sdpb  datkavcvlkgdgpqvq--infeqkesdgpv---wgsikgltghlgfhvhqfgscasndtagctvlggssagphfnpehtnk
  
```

# Approaches to Sequence Analysis

Data {GTCAT, GTTGGT, GTCA, CTCA}

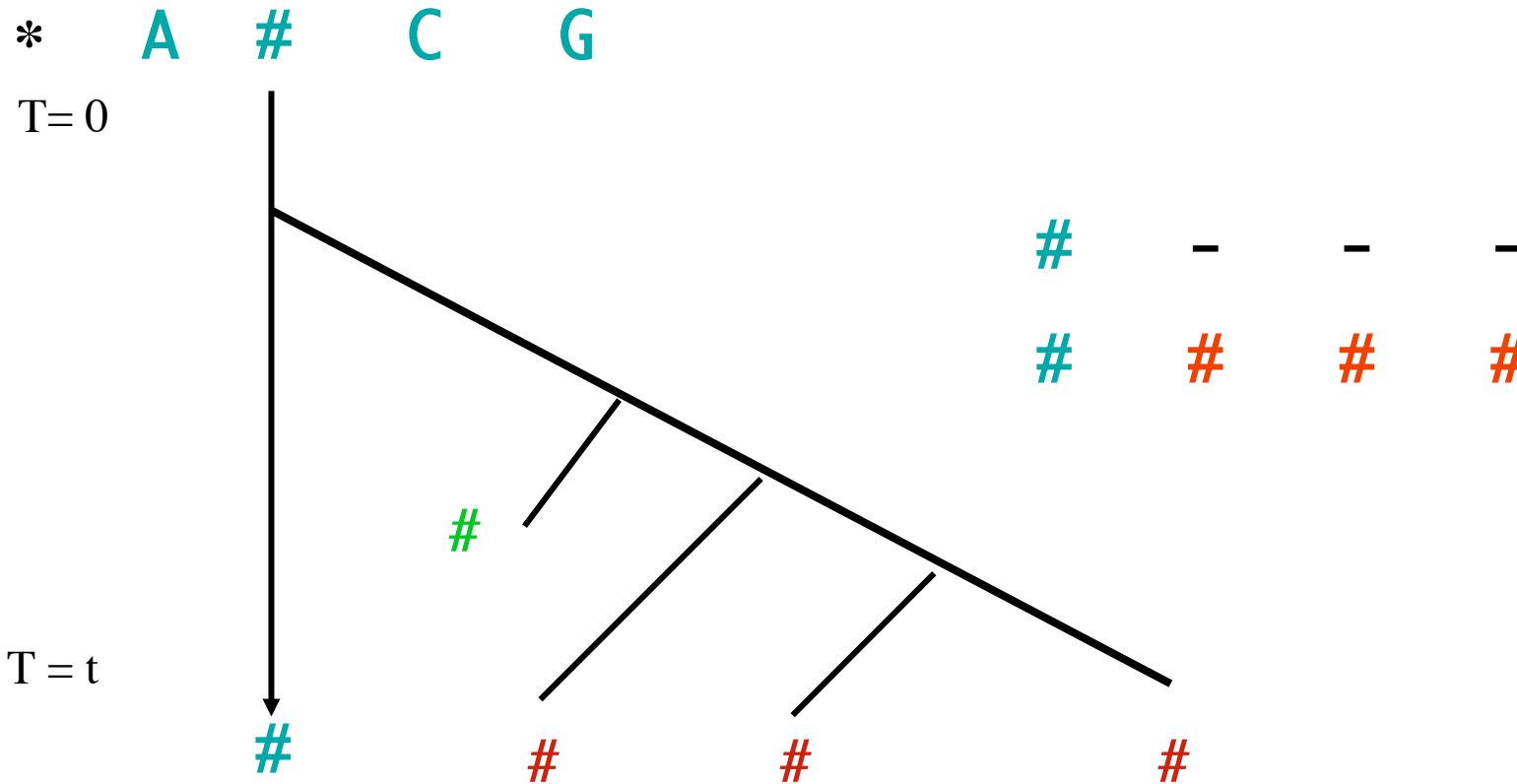
Actual Practice: 2 phase analysis.



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

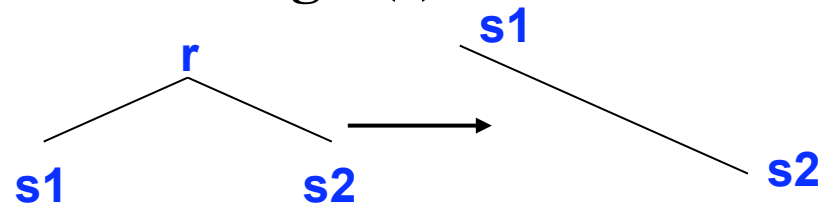


$\lambda$  (birth rate)  $<$   $\mu$  (death rate)

1.  $P(s) = (1 - \lambda/\mu)(\lambda/\mu)^l$

$\pi_A^{\#A} \dots \pi_T^{\#T}$   $l = \text{length}(s)$

2. Time reversible:  $s1 \rightleftharpoons s2$



# $\lambda$ & $\mu$ 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

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

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


---

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

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


---

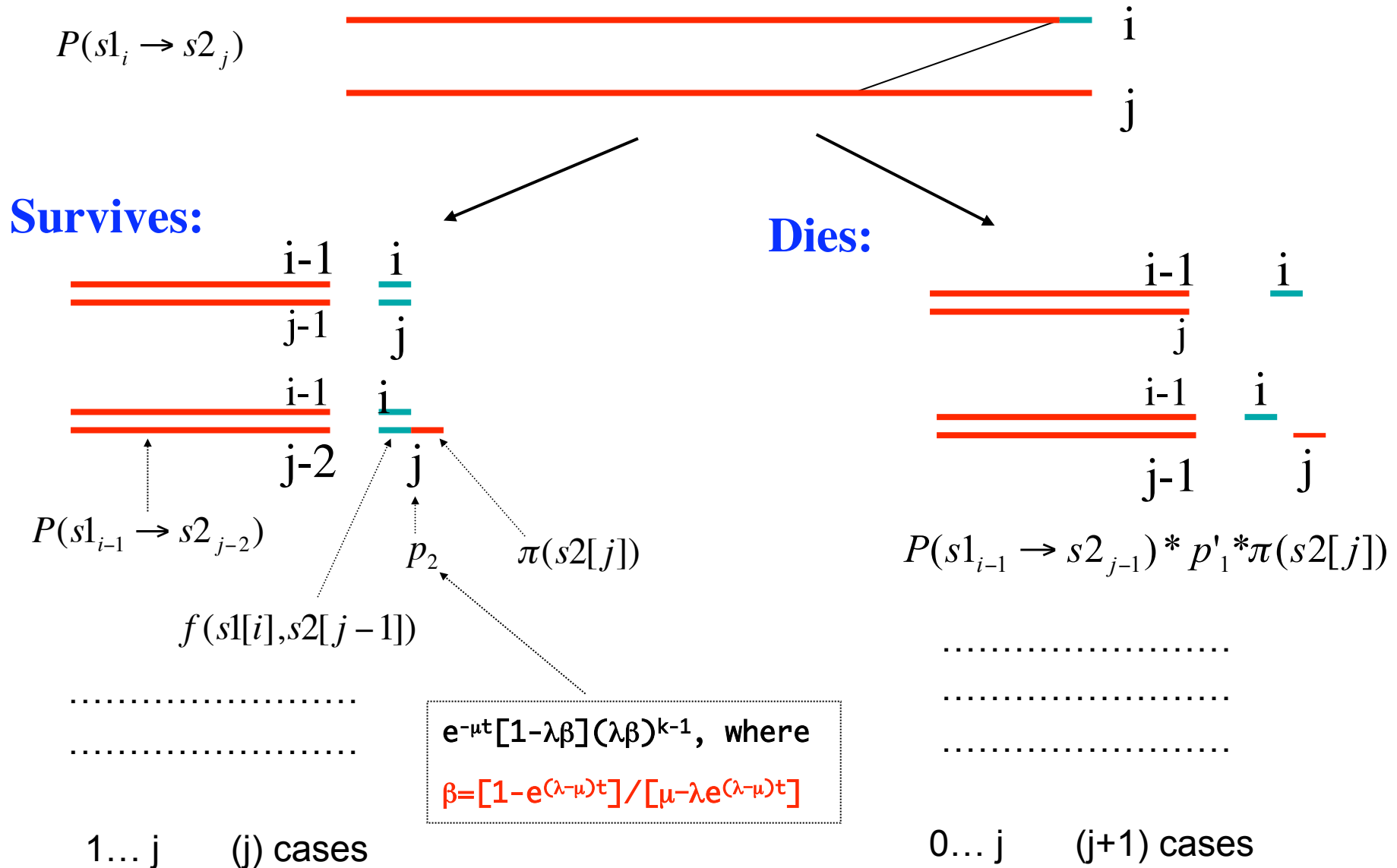
$$\begin{array}{cccccc} * & - & - & - & \dots & - \\ * & \# & \# & \# & \dots & \# \end{array}$$

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


---

Initial Conditions:  $p_k(0) = p''_k(0) = p'_k(0) = 0 \quad k > 1$   
 $p_1(0) = p''_0(0) = 1. \quad p'_0(0) = 0$

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

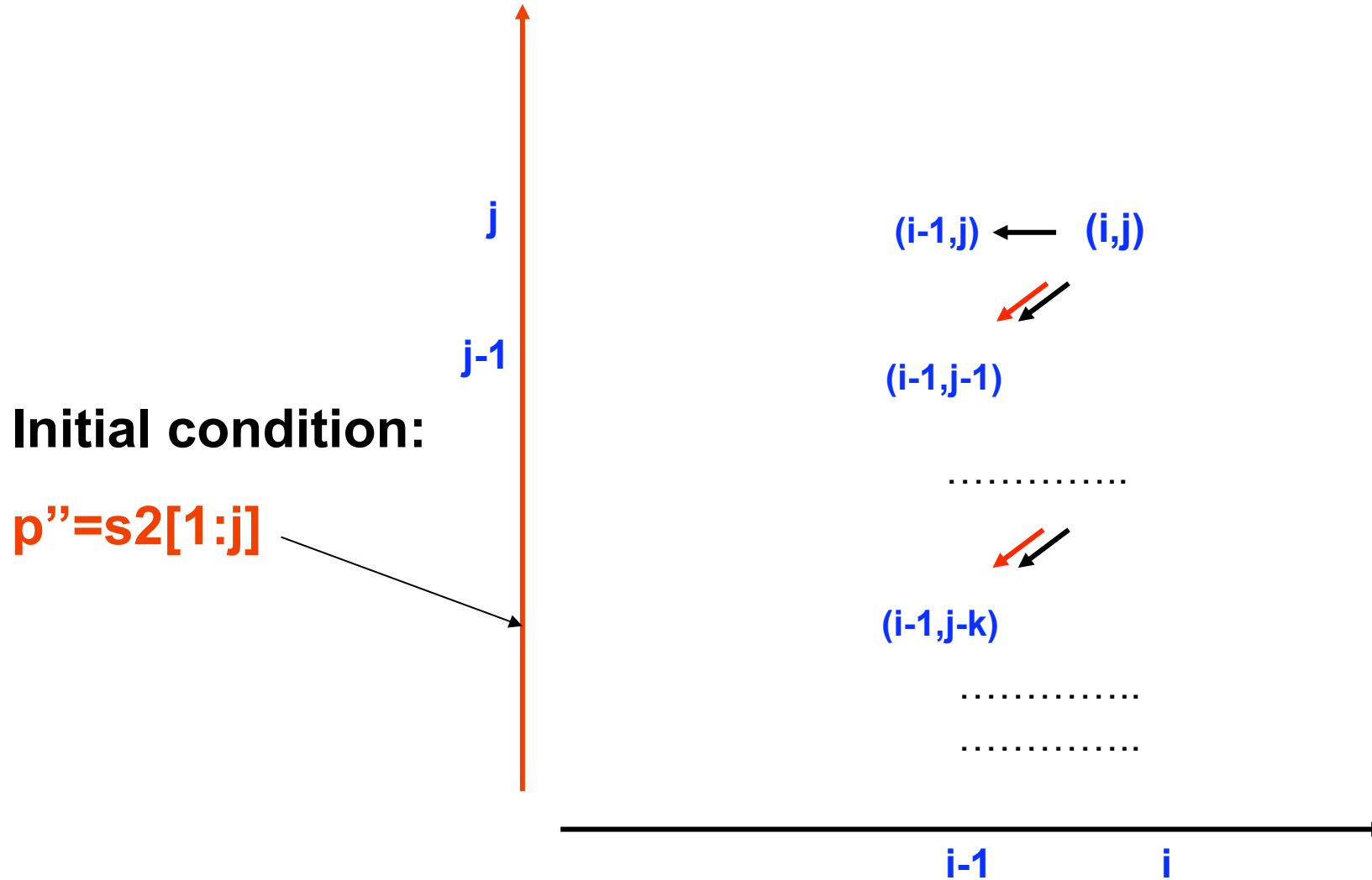


# Basic Pairwise Recursion ( $O(\text{length}^3)$ )

survive



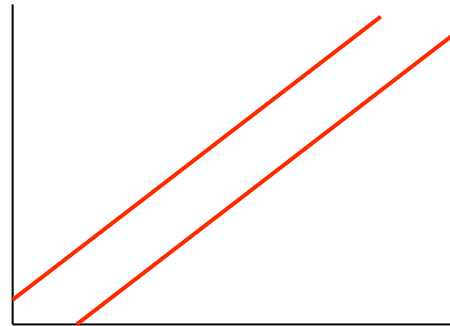
death



# Acceleration of Pairwise Algorithm

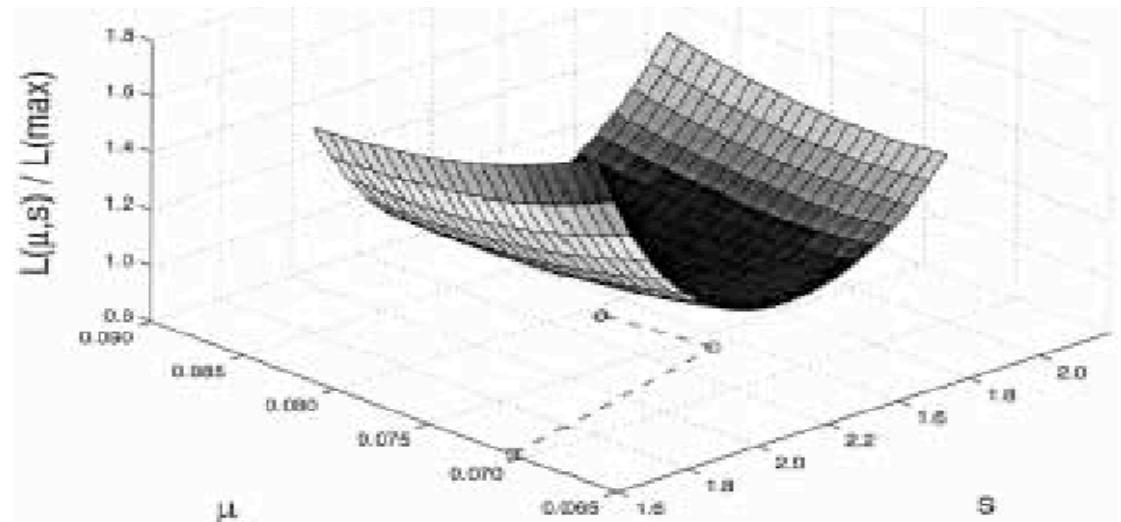
(From Hein,Wiuf,Knudsen,Moeller & Wiebling 2000)

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 ~ $10^6$

# $\alpha$ -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:

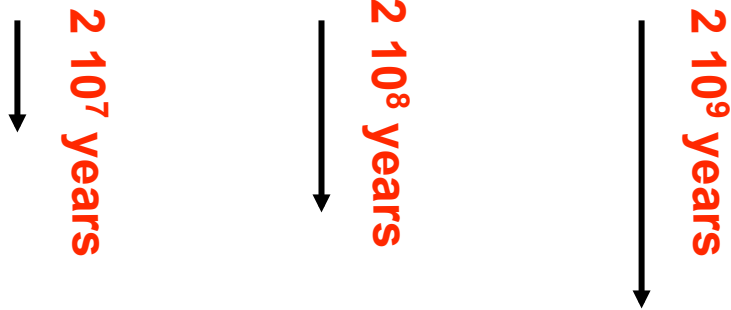
V-LSPADKTNVKAANGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT  
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR  
DGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLCVLAH HFGKEFTPPVQAAYQKVVAGVANALAHKYH

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

# The invasion of the immortal link

VLSPADNAL.....DLHAHKR 141 AA long

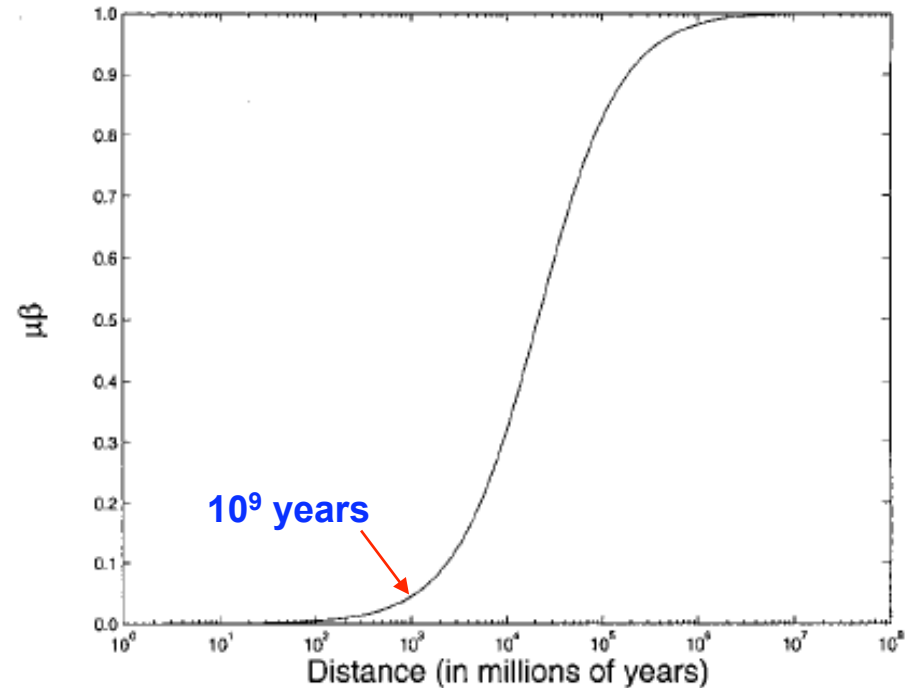
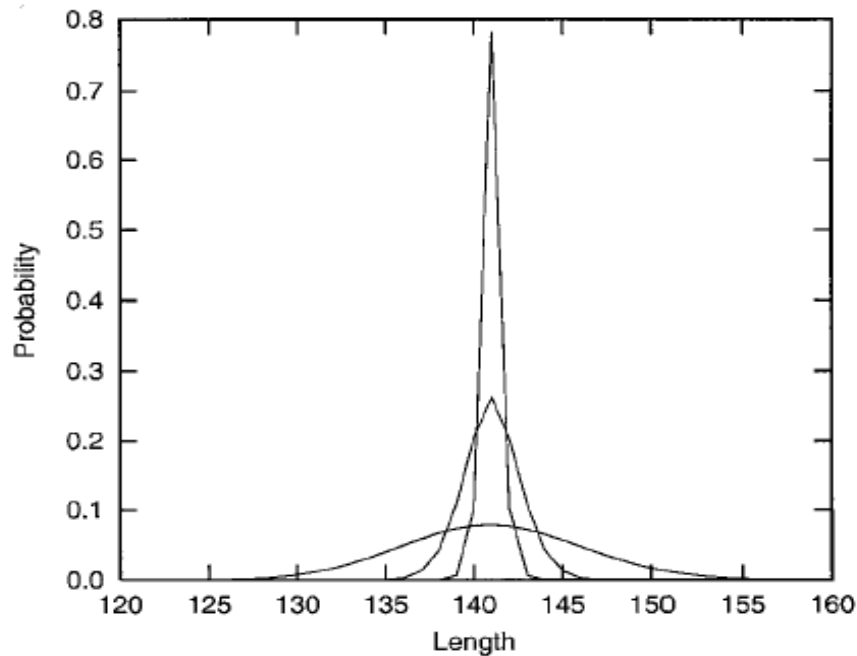


????????????????????????????? k AA long

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

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

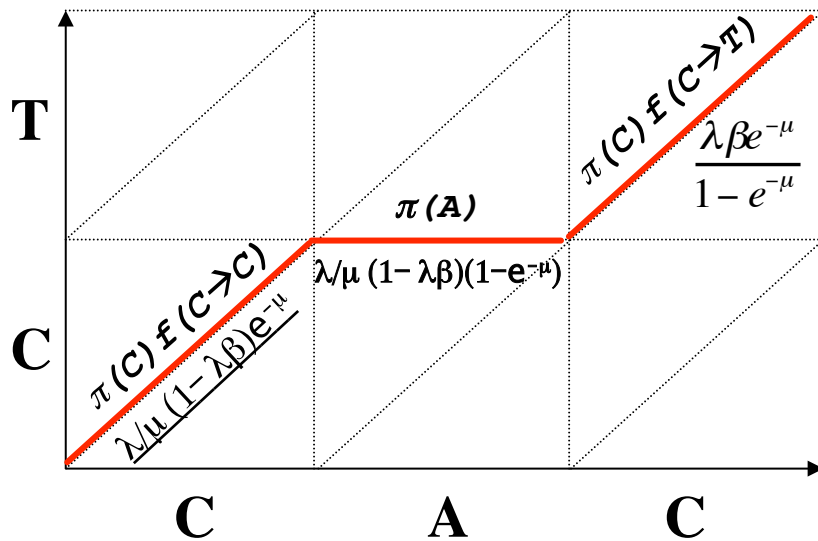
\*##### ... ###



# Statistical Alignment via Hidden Markov Models

Steel and Hein, 2001 + Holmes and Bruno, 2001

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



**HMM formulation allows:**

*Finding most probable alignment*

*Probability of sequence pair*

*Probability of specific edge*

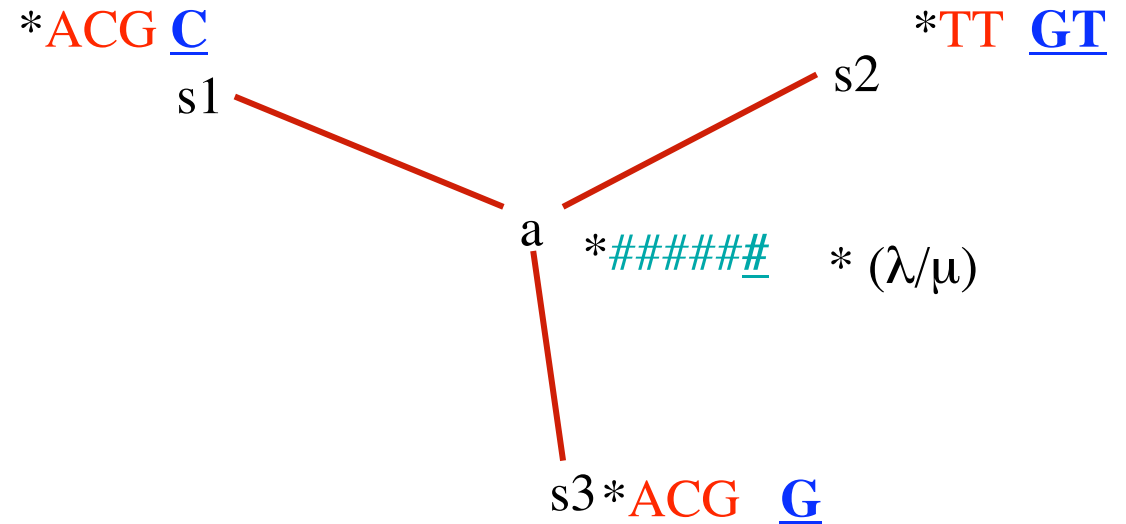
# Why multiple statistical alignment is non-trivial.

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

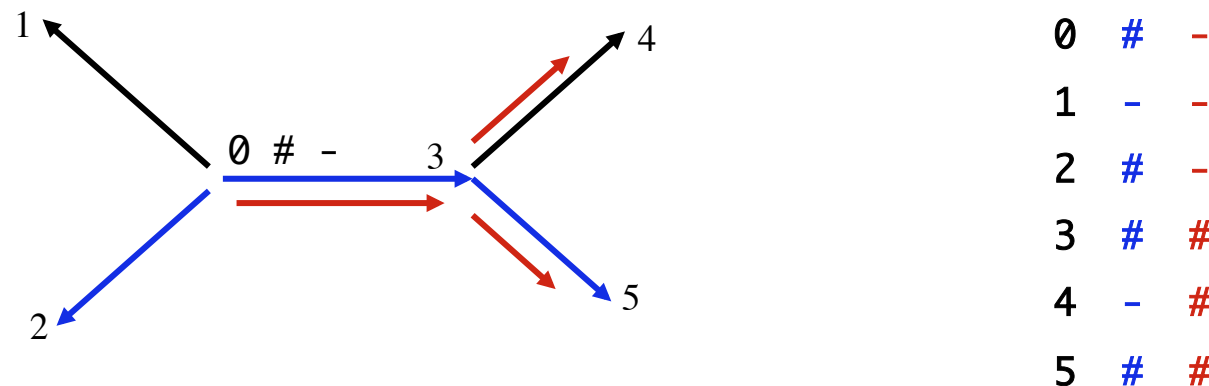
## Optimisation Alignment



## Statistical Alignment



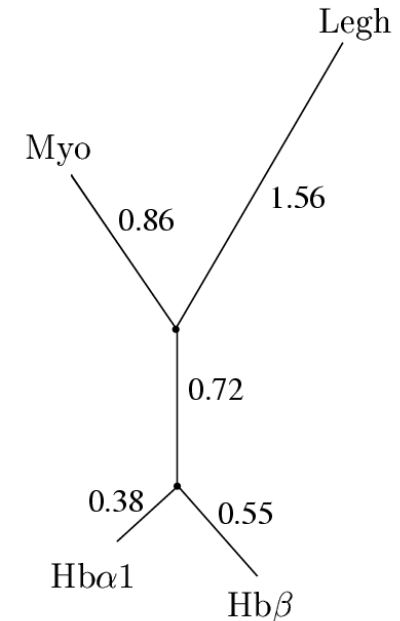
- An HMM generating alignment according to TKF91:



# 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--LSPADKTNVKAAWGKVG AHAGEYGAEALERMFLSFPTTKTYFPHF--DLS-H-----GSAQVKGHGKKVAD-AL-TNA-  
 Hbb: MV-HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESF-GDLSTPDAVM-GNPKVKAHGKKVLG-AF-SDG-  
 Myo: MG--LSDGEWQLVLNVWVKVEADIPGHGQEV LIRLFKGH PETLEKFDKFK-HLKSEDE-MKASEDLKKHGATVLT-AL-GGI-  
 Legh: MGA-FSEKQESLVKSSWEAFKQNPVPHHSAVFYTLILEKAPAAQNMFS-F---LSNGVD-P-NNPKLKAHA EKVFKMTVDSAVQ

VAHVDDMPNALSALS DLHAHKL RVD PVNFK-LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL-TS-K---YR-  
 LAHLDNLKGT FATLSELHCDKLHVDPENFR-LLGNVLCVLAHFGKEFTPPVQAA YQKV VAGVANAL-AH-K---YH-  
 LKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECI IQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG  
 LRAKGEVVLADPTLGSVHVQKGVLDP-HFL-VVKEALLKTFKEAVGDKWNDELGNAWEVAYDELA AAI-KK-A-MGSA-

Gerton Lunter, Istvan Miklos, Alexei Drummond, Yun Song

# Metropolis-Hastings Statistical Alignment

Lunter, Drummond, Miklos, Jensen & Hein, 2005

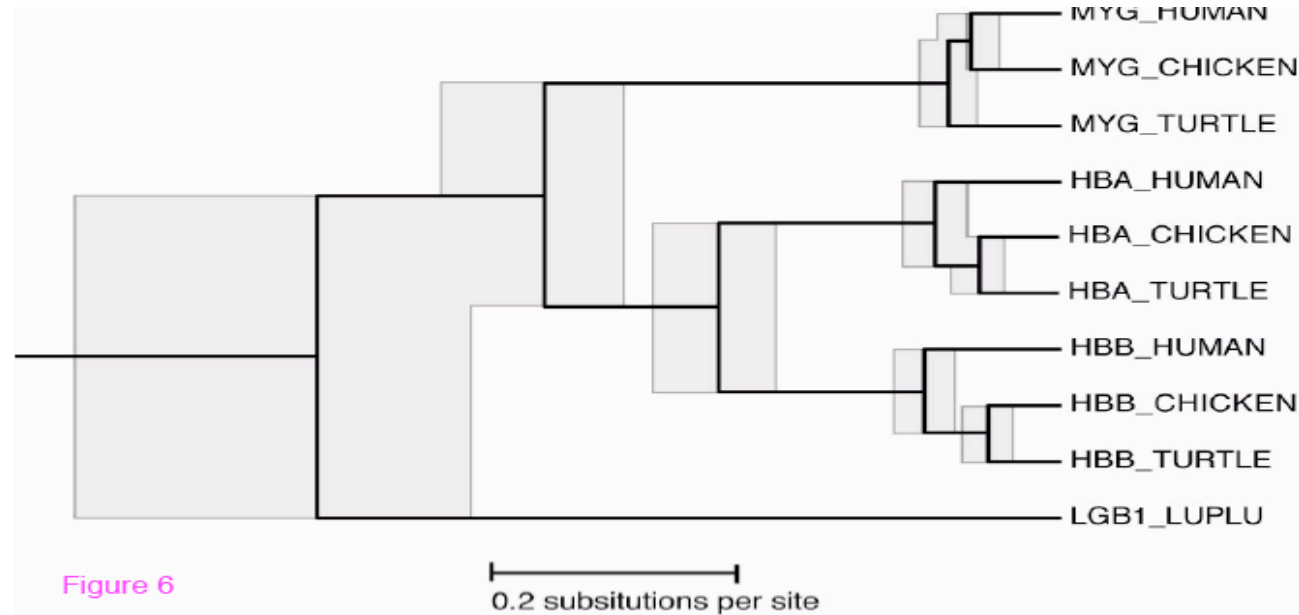
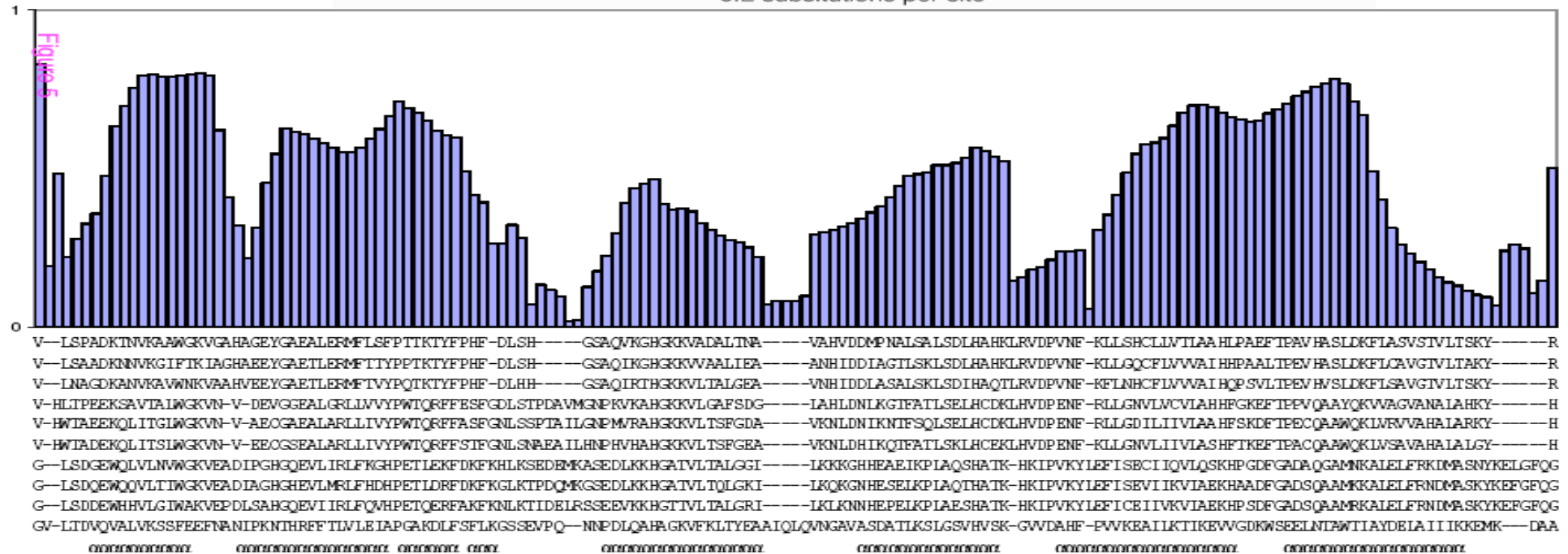
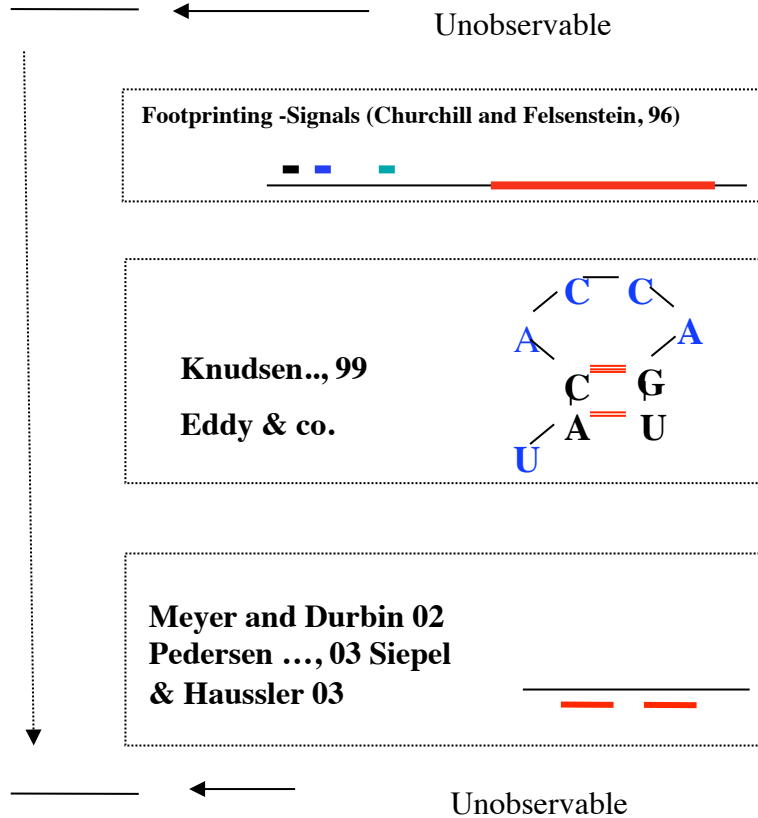


Figure 6

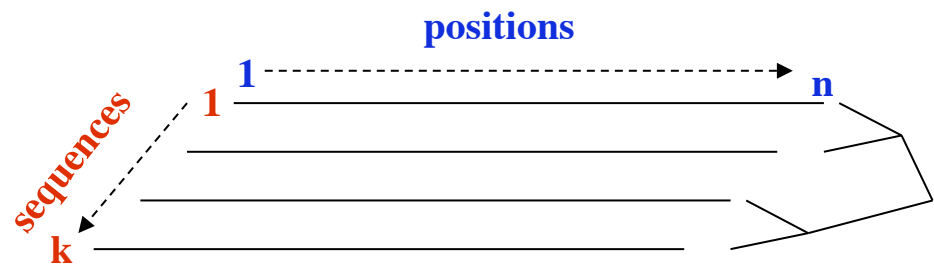
0.2 substitutions per site



# The Basics of Evolutionary Annotation



*Many aligned sequences  
related by a known phylogeny*



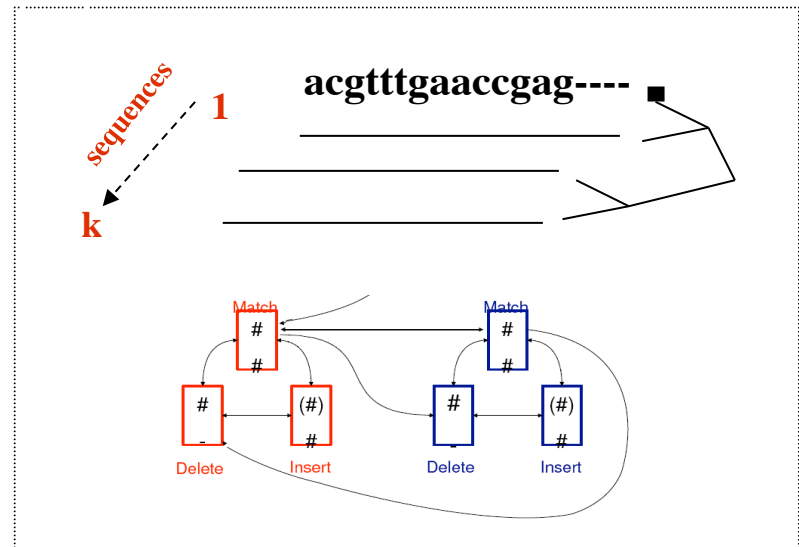
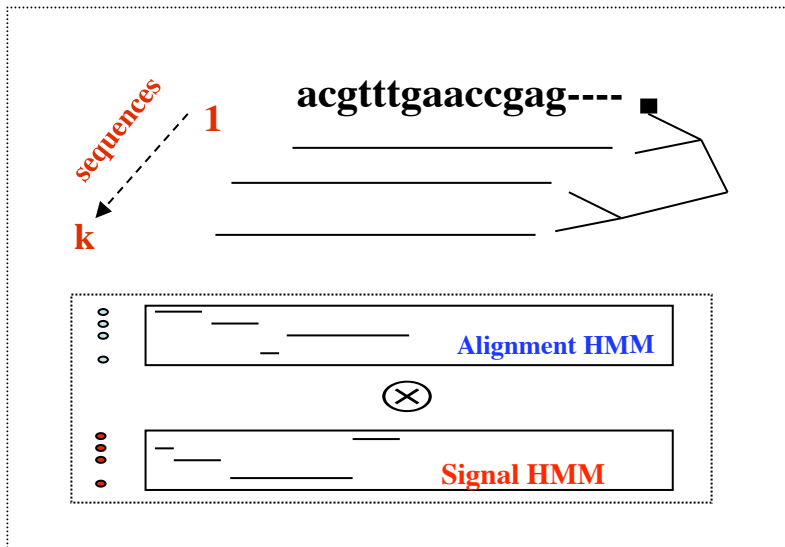
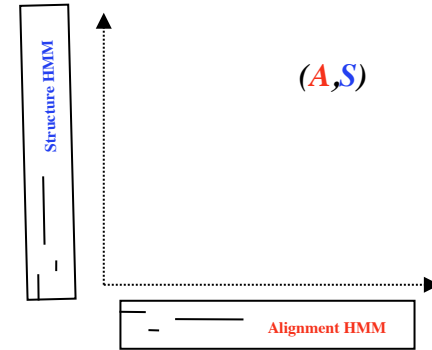
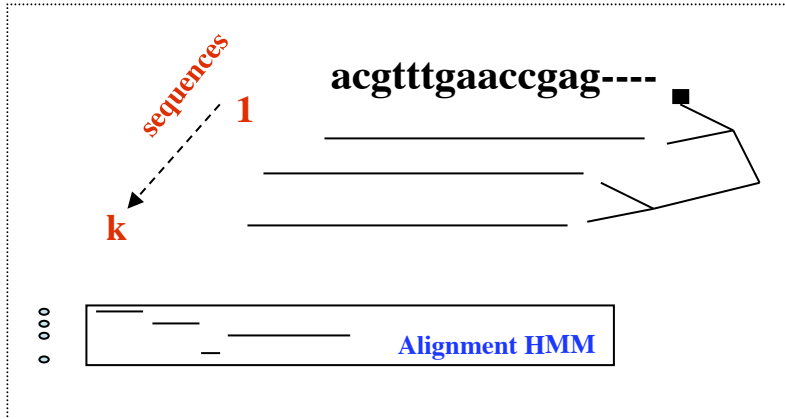
*HMM*



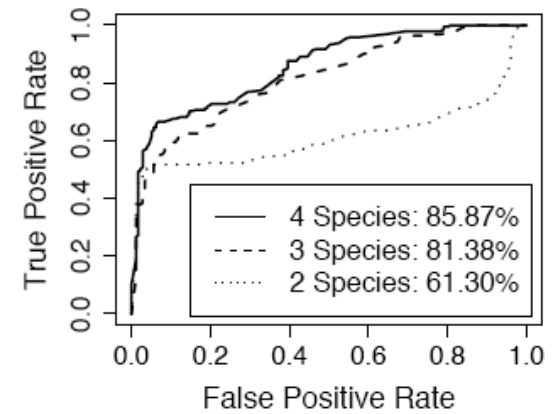
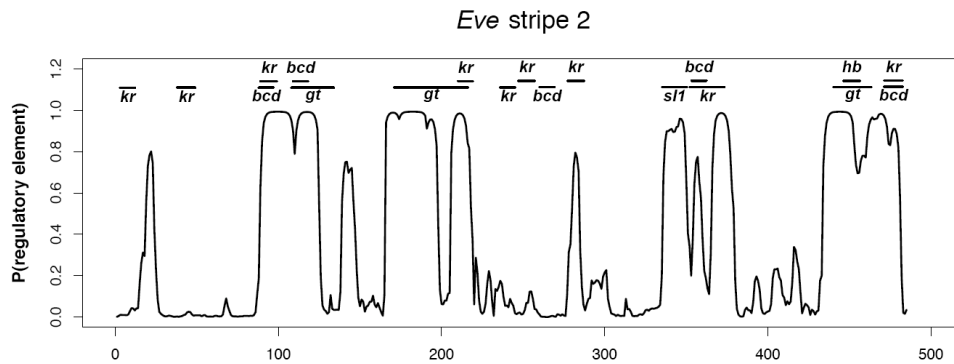
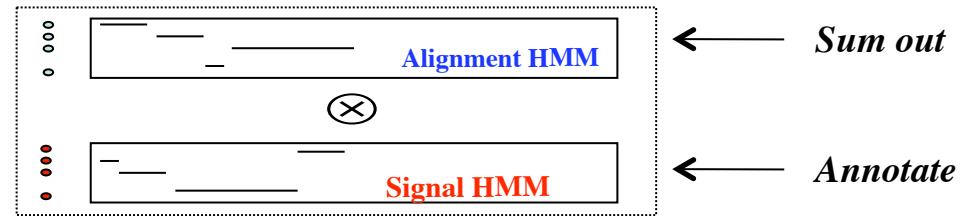
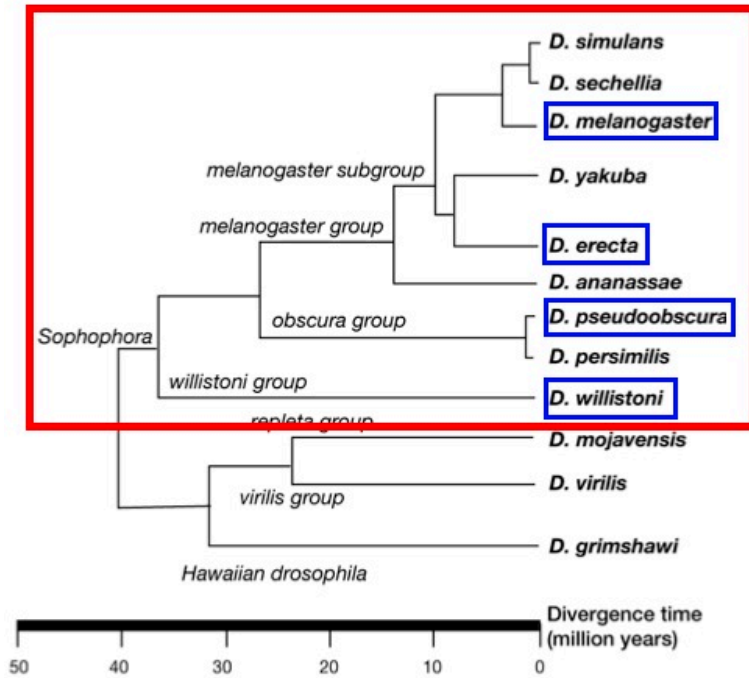
$$P(\text{Sequence}|\text{Structure})P(\text{Structure}) =$$

$$P(\text{Structure}|\text{Sequence})P(\text{Sequence})$$

# Statistical Alignment and Footprinting.



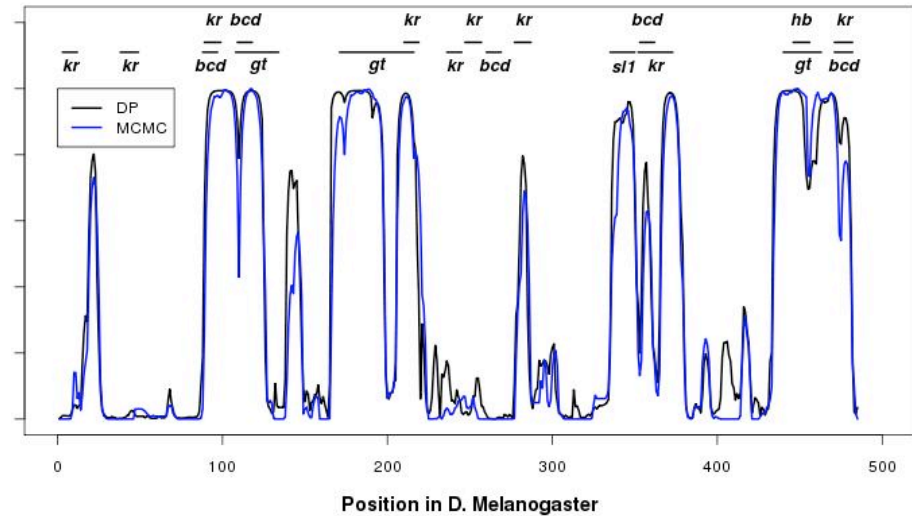
# SAPF - Statistical Alignment and Phylogenetic Footprinting



# BigFoot

- *Dynamical programming is too slow for more than 4-6 sequences*
- *MCMC integration is used instead – works until 10-15 sequences*
- *For more sequences other methods are needed.*

Eve stripe 2



Eve stripe 2

