

Statistical Alignment and Footprinting

Rutgers – DIMACS 27.4.09

The Problem

- *Statistical Alignment - Annotation - Annotation & Statistical Alignment*

Statistical Alignment

- *The Model*
- *The Pairwise Algorithm – the HMM connection*
- *Multiple sequence alignment algorithms*

Annotation

- *The general problem*
- *protein secondary structure – protein genes – RNA structure - signals*

Annotation & Alignment

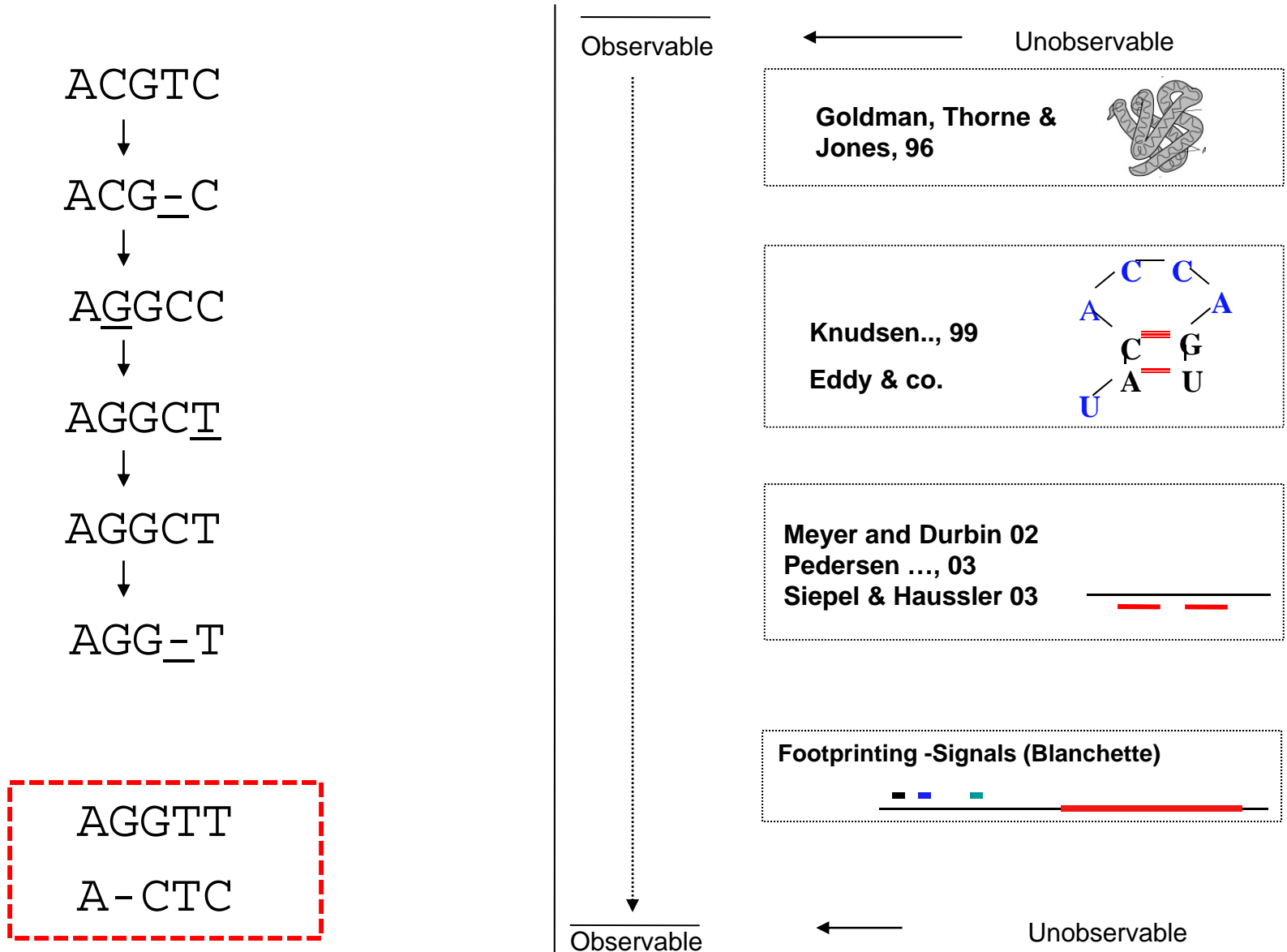
- *The general algorithm*
- ***Signals (footprinting)***
- *Protein Secondary Structure Prediction*

Ahead

- *Transcription Factor Prediction - Knowledge transfer - homologous/nonhomologous analysis*

Sequence Evolution and Annotation

Alignment and Footprinting



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

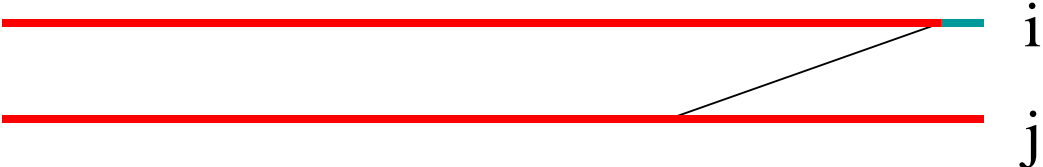
$$P_t(T \rightarrow R) * \pi_Q * \dots * \pi_W * p_4(t)$$

T - - - -
 - R Q S W
 4

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

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

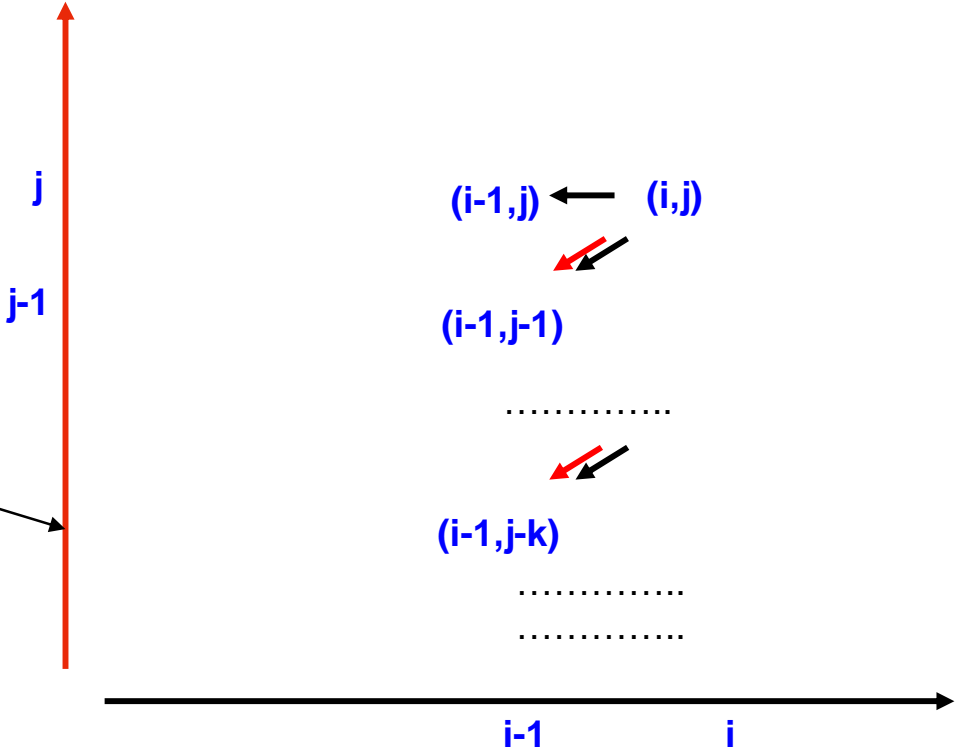
$$P(s1_i \rightarrow s2_j)$$



survive

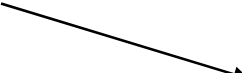


death



Initial condition:

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



α -globin (141) and β -globin (146)

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

α -globin

VLSPADKTNVKAAWGKVGGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHV
DDMPNALSALSSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

β -globin

VHLTPEEKSAVTALWGKVVNDEVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD
GLAHLNLSKGTFAFLSELHCDKLHVDPENFRLLGNVLCVLAHFFGKEFTPPVQAAYQKVVAGVANALAHKYH

λ^*t : 0.0371805 +/- 0.0135899

μ^*t : 0.0374396 +/- 0.0136846

s^*t : 0.91701 +/- 0.119556

430.108 : $-\log(\alpha\text{-globin})$

327.320 : $-\log(\alpha\text{-globin} \rightarrow \beta\text{-globin})$

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

Maximum contributing alignment:

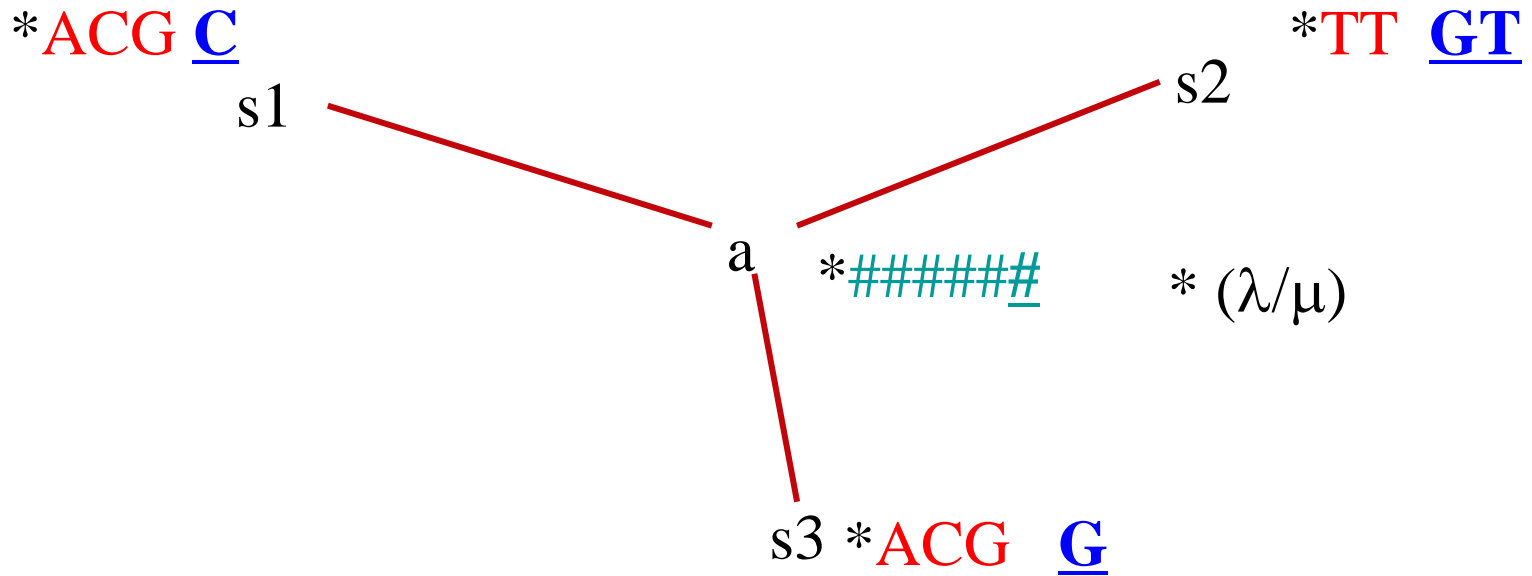
V-LSPADKTNVKAAWGKVGGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

NAVAHVDDMPNALSALSSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
DGLAHLNLSKGTFAFLSELHCDKLHVDPENFRLLGNVLCVLAHFFGKEFTPPVQAAYQKVVAGVANALAHKYH

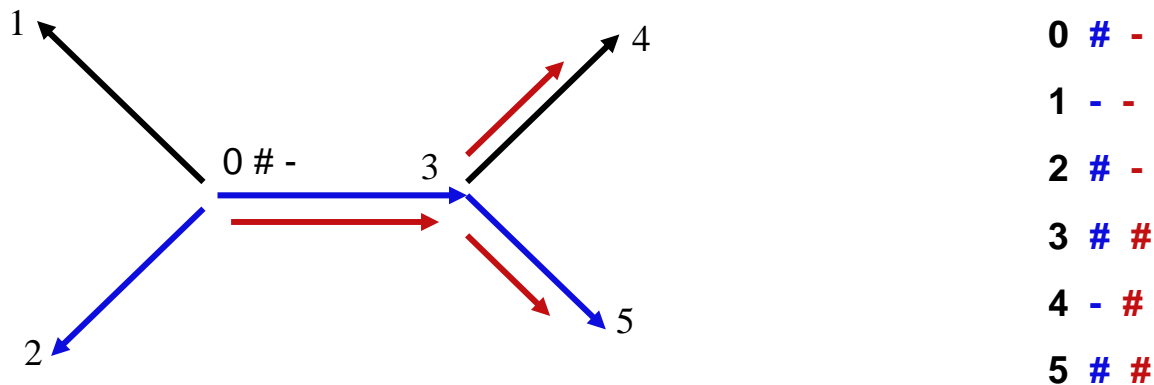
Ratio $l(\text{maxalign})/l(\text{sumalign}) = 0.00565064$

Why multiple statistical alignment is non-trivial.

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

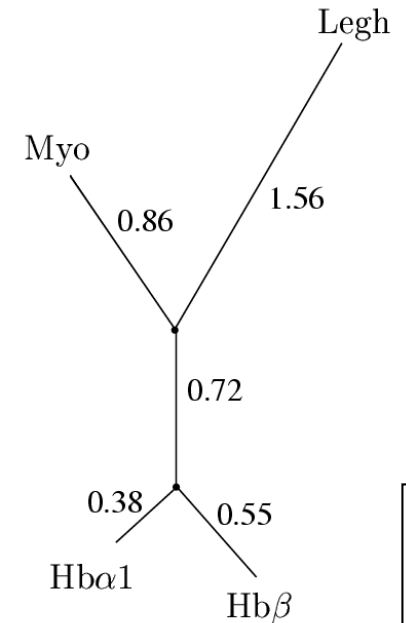


- An HMM generating alignment according to TKF91:



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--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-FSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFS-F---LSNGVD-P-NNPKLKAHA EKVF KMTVDSAVQ

VAHVDDMPNALSALS DLHAHKL RVD PVNFK-LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL-TS-K---YR-
LAHLDNLKGT FATLSELHCDKLHVDPENFR-LLGNVLCVLAH HF GKEFTPPVQAA 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

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

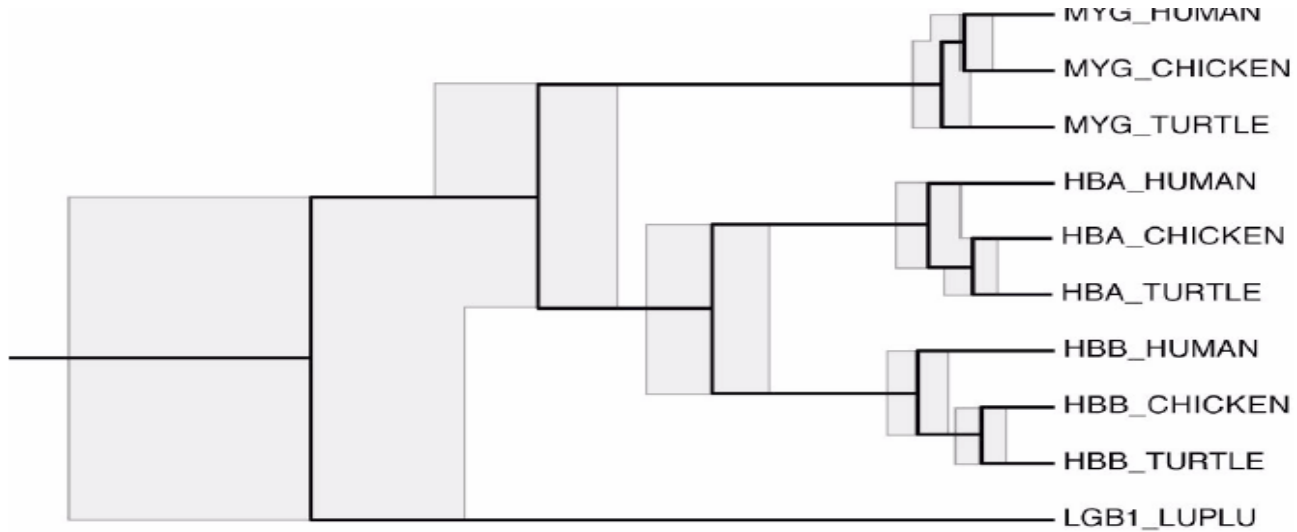
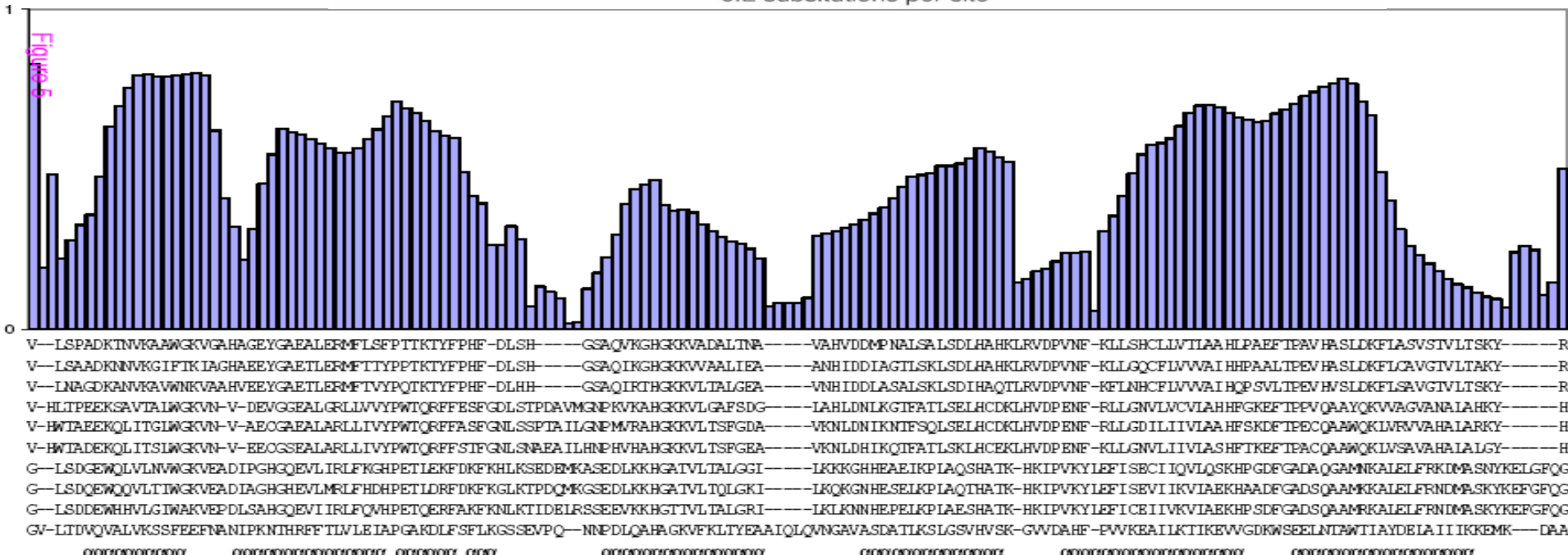


Figure 6

0.2 substitutions per site



How to proceed to many many sequences ??

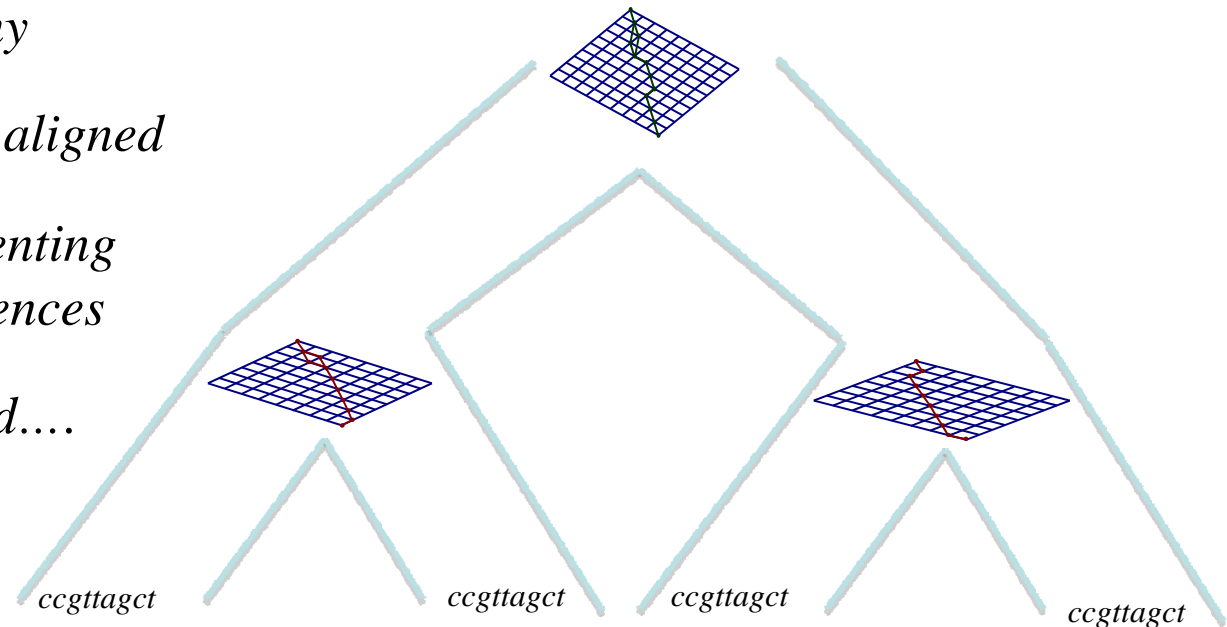
- *Dynamical Programming stops at 4-5 sequences*
- *MCMC stops at 10-13ish sequences*
- *Some approximations must be adopted*
 - *“Temporal Corner cutting”*
 - *Degenerate Genealogical Structures*

Many Sequences: Sequence Graphs (reticular alignment)

Istvan Miklos – Gerton Lunter – Miklos Csuros

Investigate a set of ancestral sequences/alignments that are computationally realistic

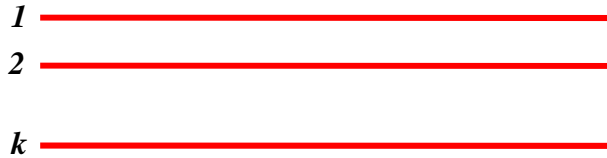
- *A set of homologous sequences are given*
- *With a known phylogeny*
- *Pairs of sequences are aligned*
- *Graphs defined representing alignment/ancestral sequences*
- *Pairs of graphs aligned....*



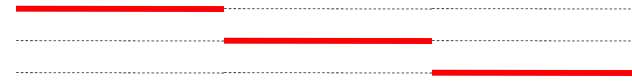
FSA - Fast Statistical Alignment

Pachter, Holmes & Co

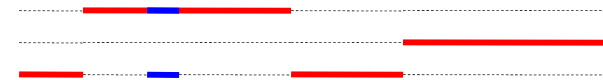
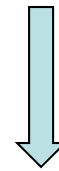
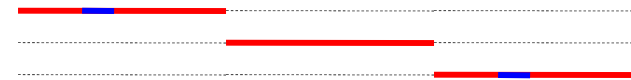
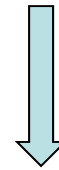
Data – k genomes/sequences:



Iterative addition of homology statements to shrinking alignment:

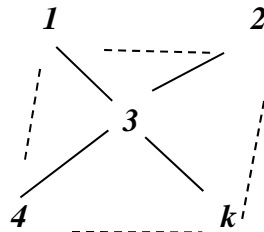


Add most certain homology statement from pairwise alignment compatible with present multiple alignment



i. Conflicting homology statements cannot be added
 ii. Some scoring on multiple sequence homology statements is used.

Spanning tree Additional edges

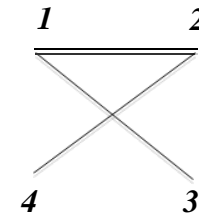
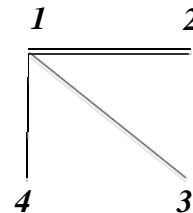
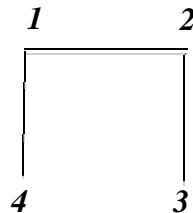
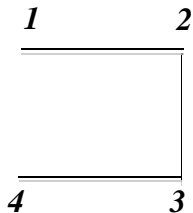
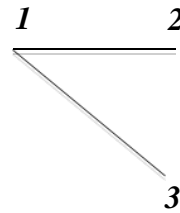
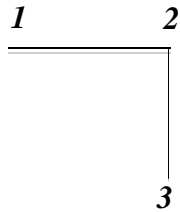
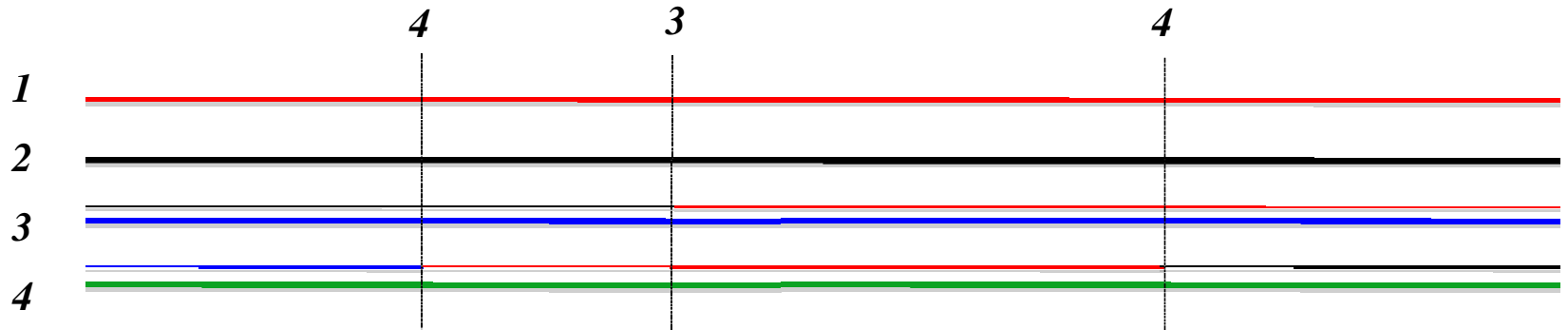


An edge – a pairwise alignment



- 1,3 2,3 3,4 3,k
- 1,2 2,k 1,4 4,k

Li-Stephens



Simplifications relative to the Ancestral Recombination Graph (ARG)

Local Trees are Spanning Trees – not phylogenies (Steiner Trees)

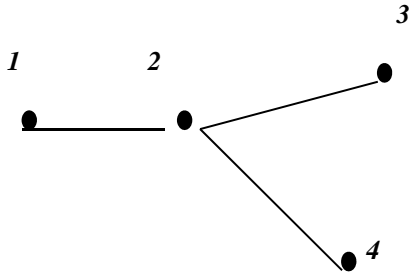
No non-ancestral bridges between ancestral material



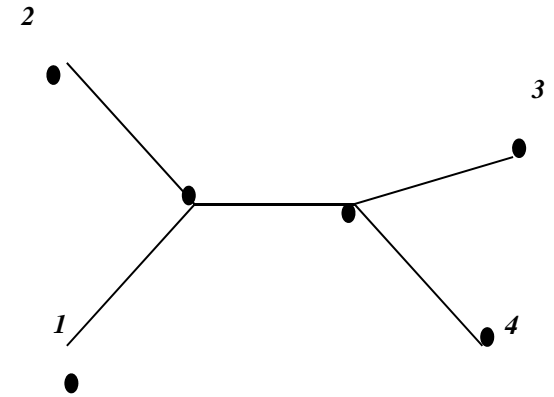
Are there intermediates between Spanning Trees and Steiner Trees?

Spannoids – k -restricted Steiner Trees

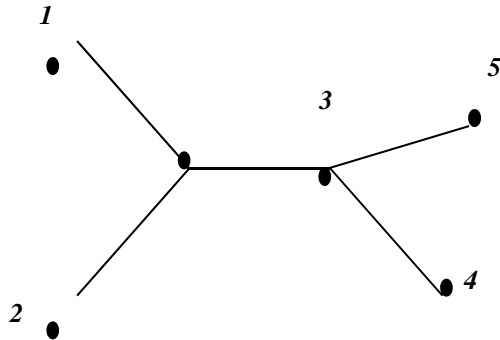
Baudis et al. (2000) Approximating Minimum Spanning Sets in Hypergraphs and Polymatroids



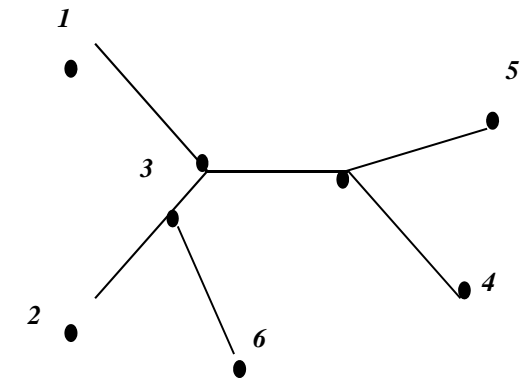
Spanning tree



Steiner tree



1-Spannoid



2-Spannoid

Advantage: Decomposes large trees into small trees

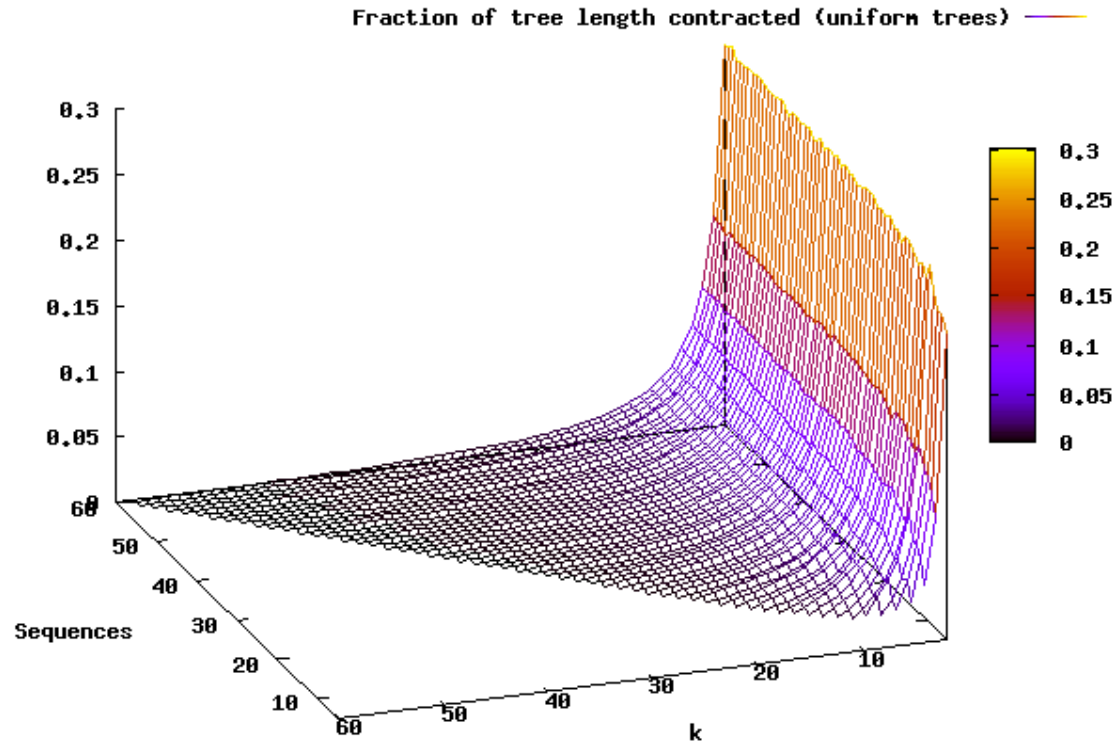
Questions: How to find optimal spannoid?

How well do they approximate?

Example – Contraction of Simulated Coalescent Trees

Simulation

- *Trees simulated from the coalescent*
- *Spannoid algorithm:*



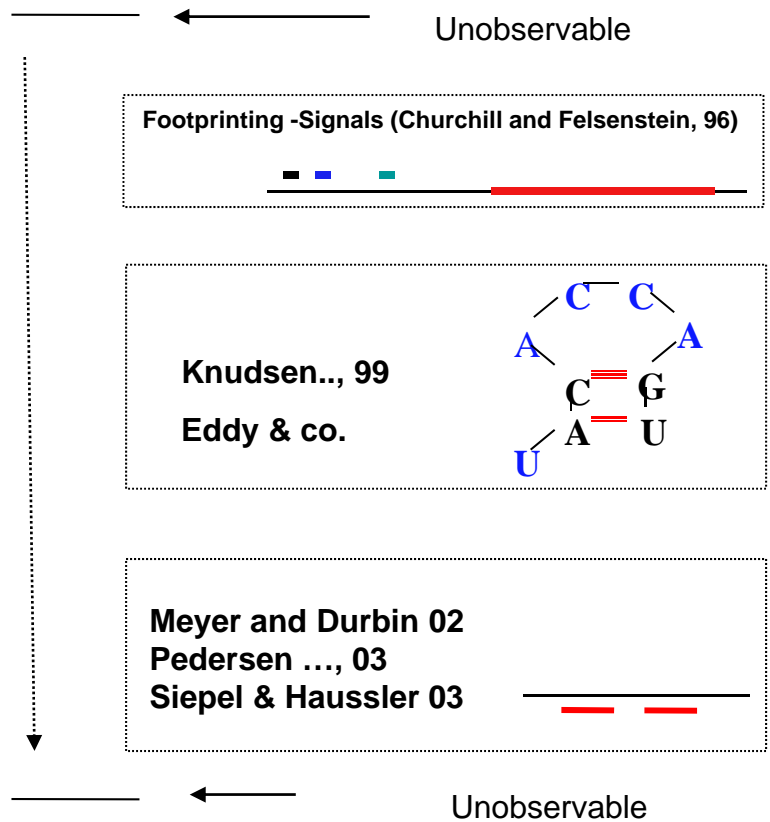
Conclusion

- *Approximation very good for $k > 5$*
- *Not very dependent on sequence number*

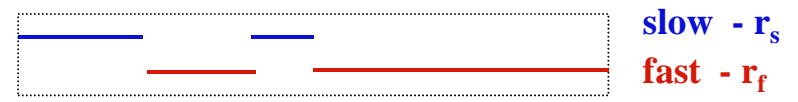
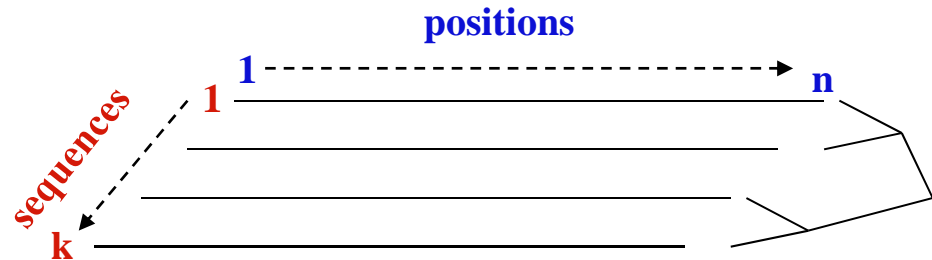
Annotation & Annotation with alignment

- *Annotation*
- *Annotation and alignment*
- *Footprinting*
 - *Three Programs*
 - *SAPF – dynamic programming up to 4 sequences*
 - *BigFoot– MCMCup to 13 sequences*
 - *GRAPEfoot – pairwise genome footprinting*

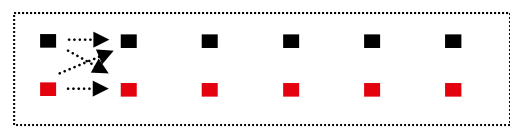
The Basics of Evolutionary Annotation



Many aligned sequences related by a known phylogeny



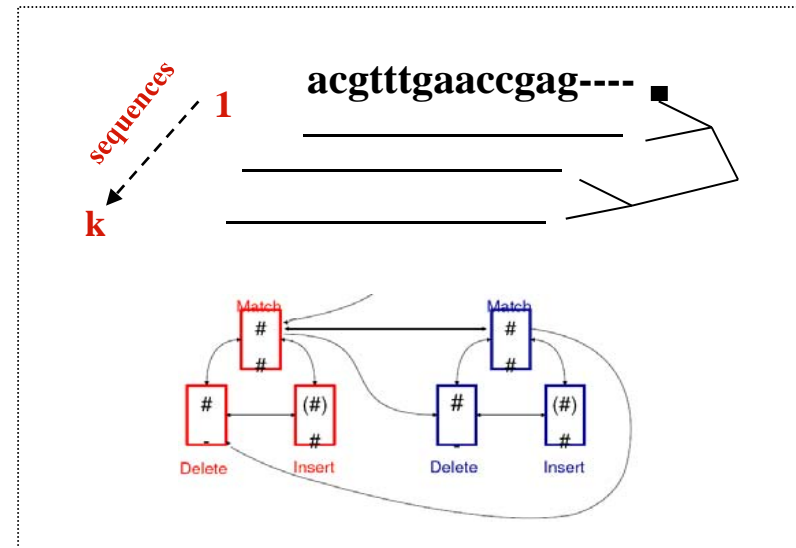
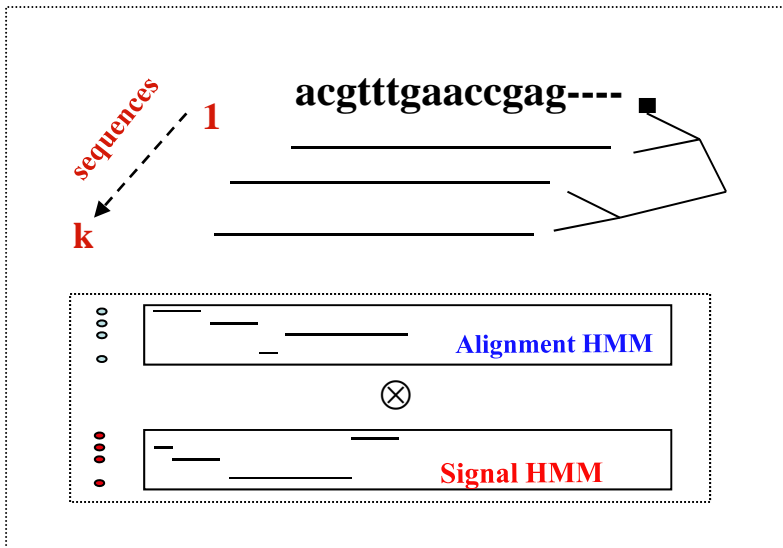
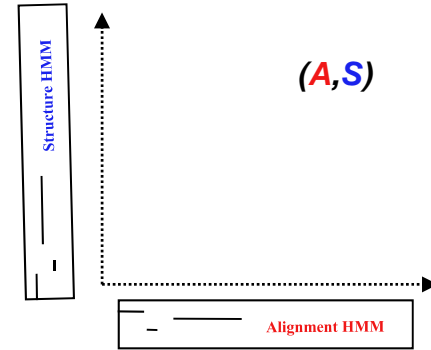
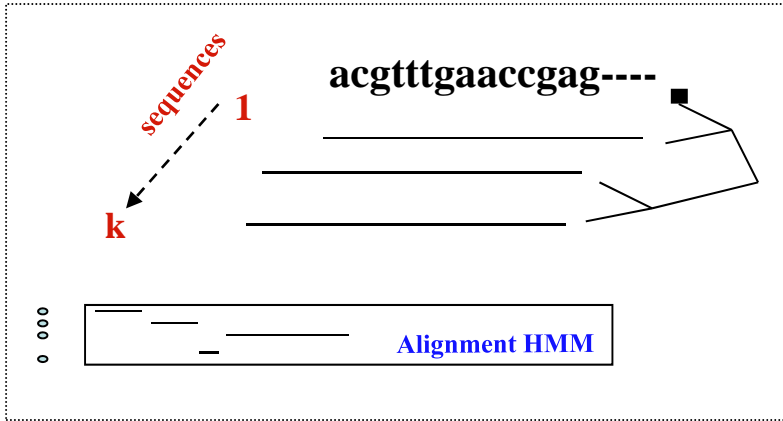
HMM



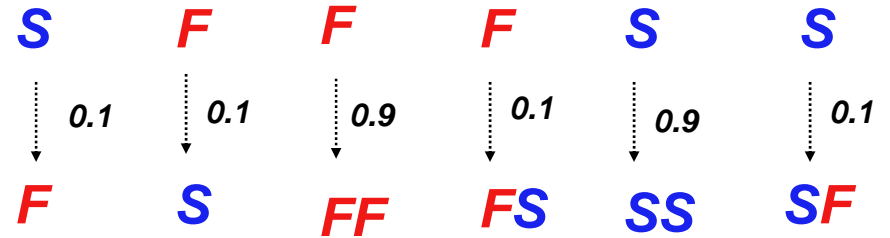
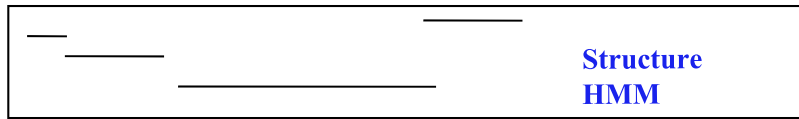
$$P(\text{Sequence}|\text{Structure})P(\text{Structure}) =$$

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

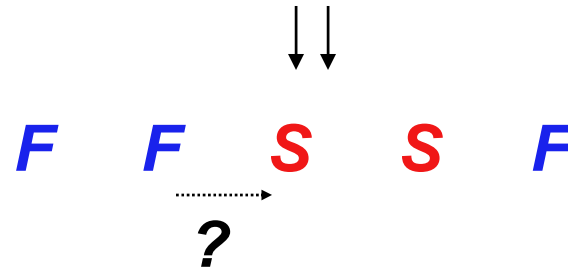
Statistical Alignment and Footprinting.



“Structure” does not stem from an evolutionary model



- The equilibrium annotation does not follow a Markov Chain:



- Each alignment in from the **Alignment HMM** is annotated by the **Structure HMM**.

- No ideal way of simulating:

using the **HMM at the alignment** will give other distributions on the leaves

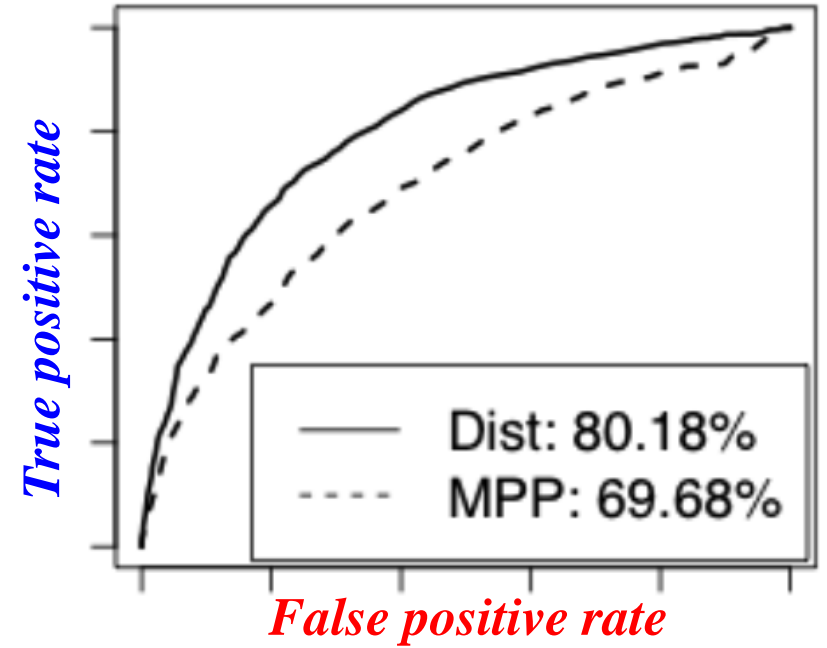
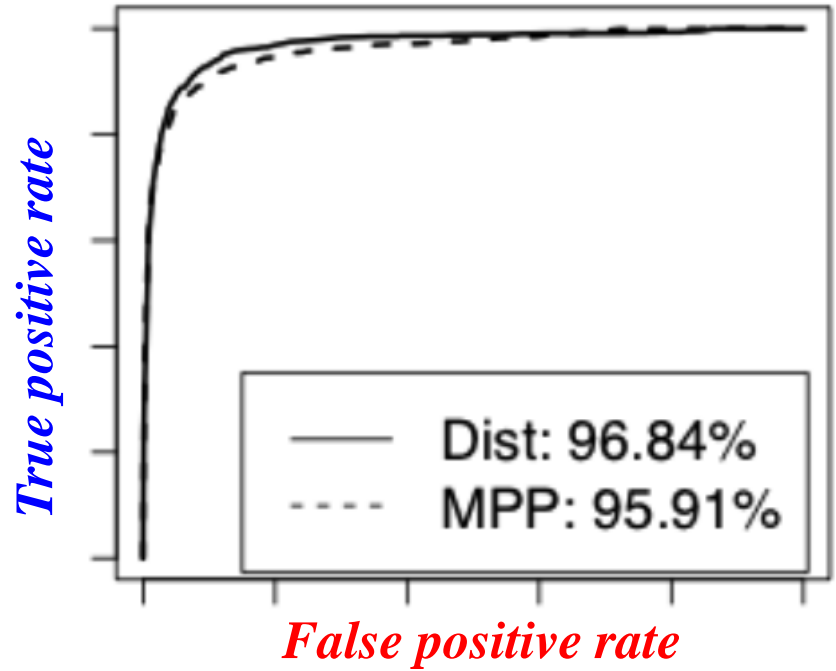
using the **HMM at the root** will give other distributions on the leaves

Summing Out is Better Satiya et al., 2008

Simulated data with parameter estimated from Eve Stripe 2.

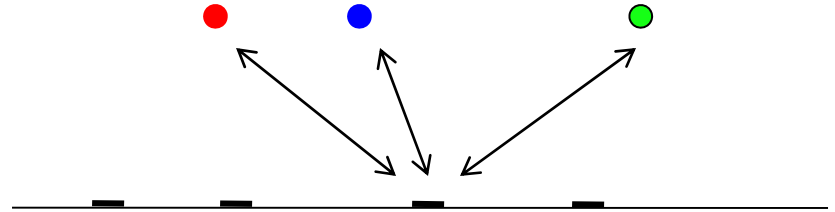
DIS – summing out alignments
MPP – fixing on 1 alignment

As above but with higher insertion-deletion rate.



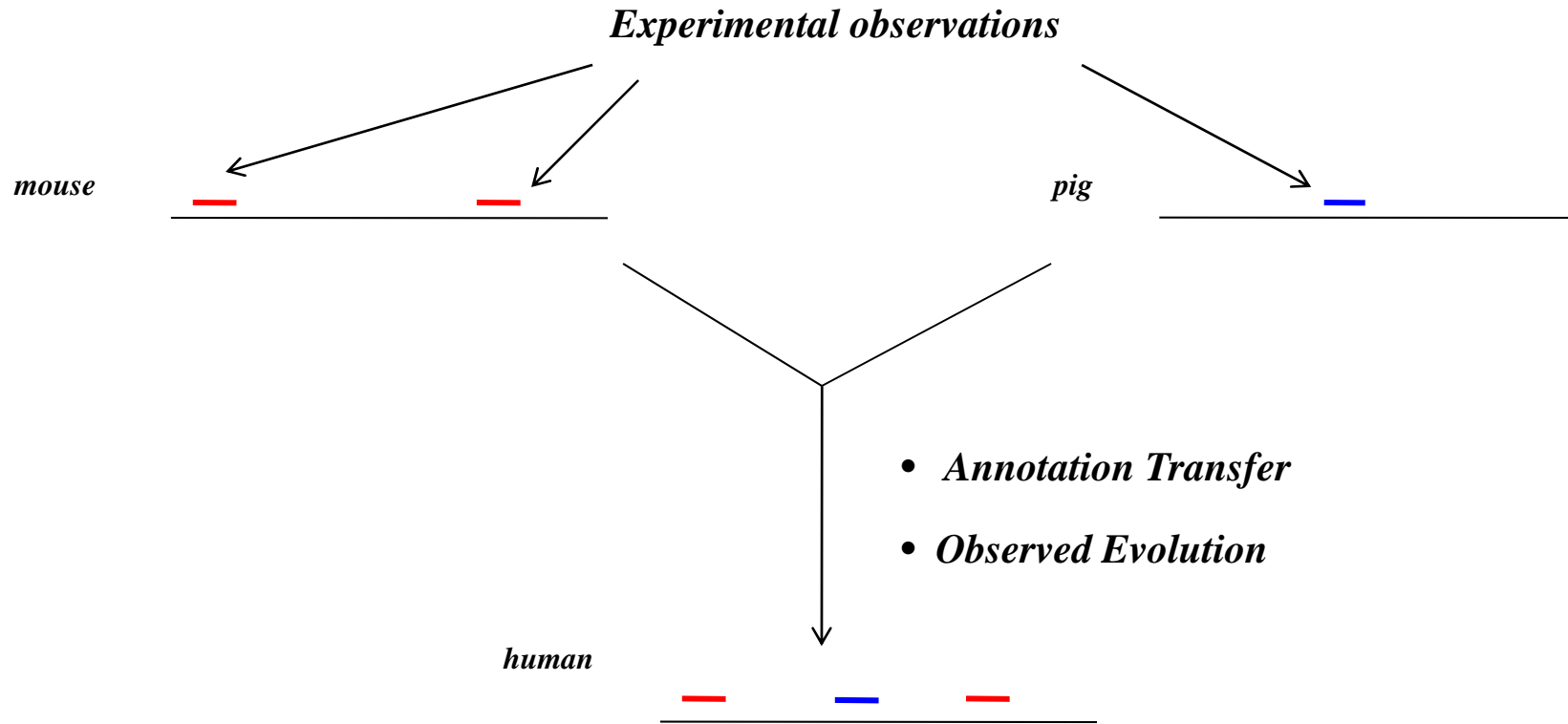
Signal Factor Prediction

- **Given set of homologous sequences and set of transcription factors (TFs), find signals and which TFs they bind to.**



- **Use PWM and Bruno-Halpern (BH) method to make TF specific evolutionary models**
- **Drawback BH only uses rates and equilibrium distribution**
- **Superior method: Infer TF Specific Position Specific evolutionary model**
- **Drawback: cannot be done without large scale data on TF-signal binding.**

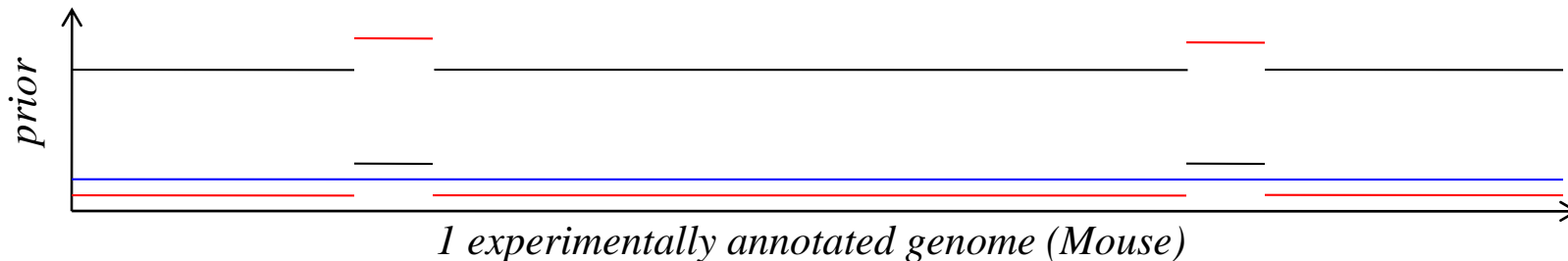
Knowledge Transfer and Combining Annotations



Must be solvable by Bayesian Priors

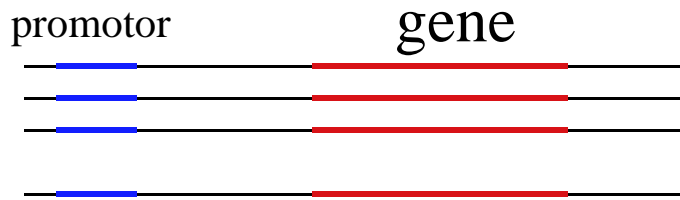
Each position p_i probability of being j 'th position in k 'th TFBS

If no experiment, low probability for being in TFBS

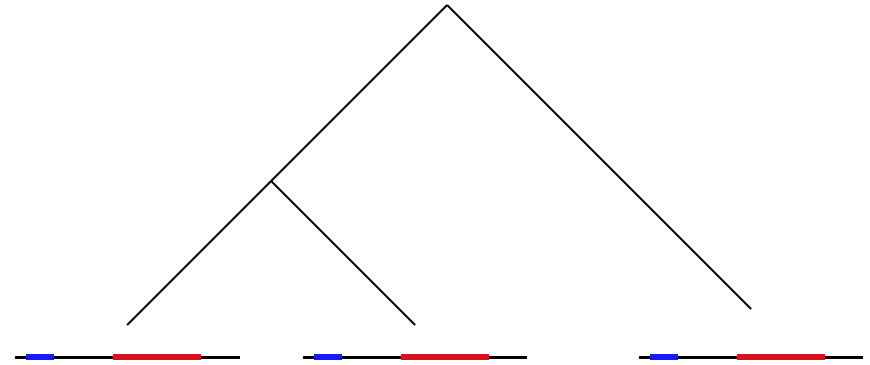


(Homologous + Non-homologous) detection

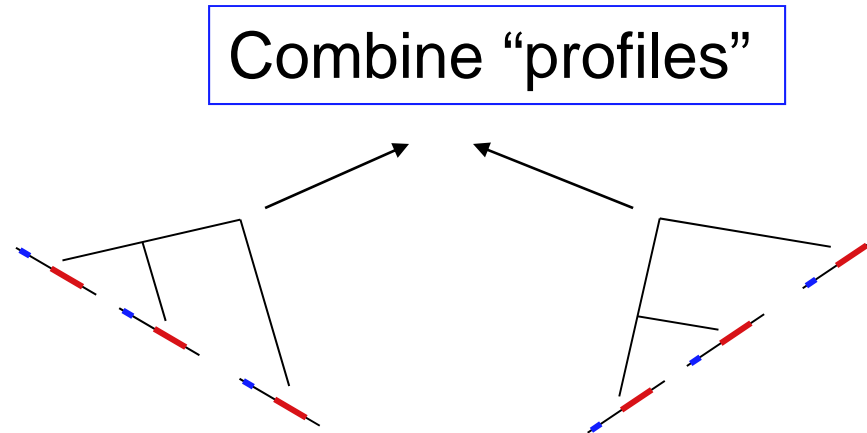
Unrelated genes - similar expression



Related genes - similar expression



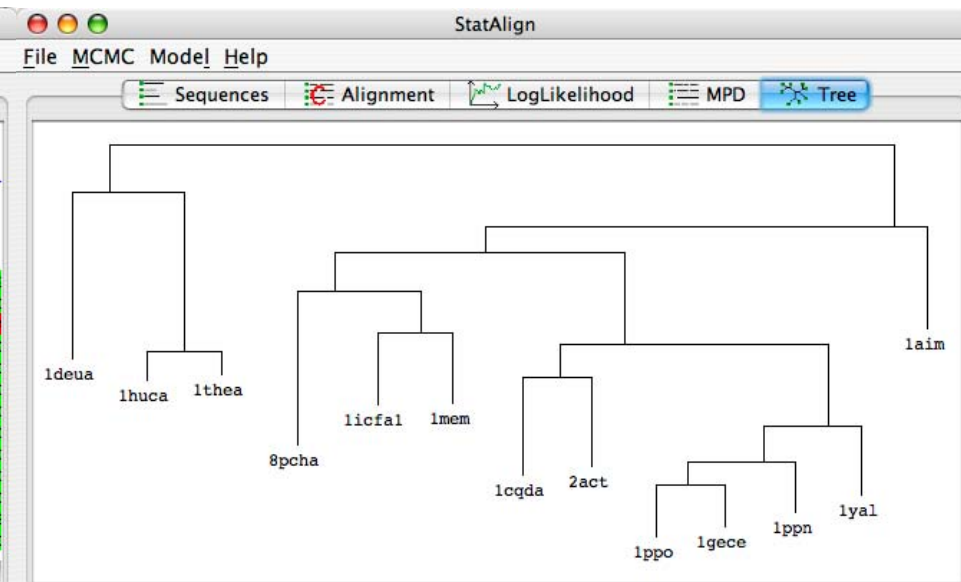
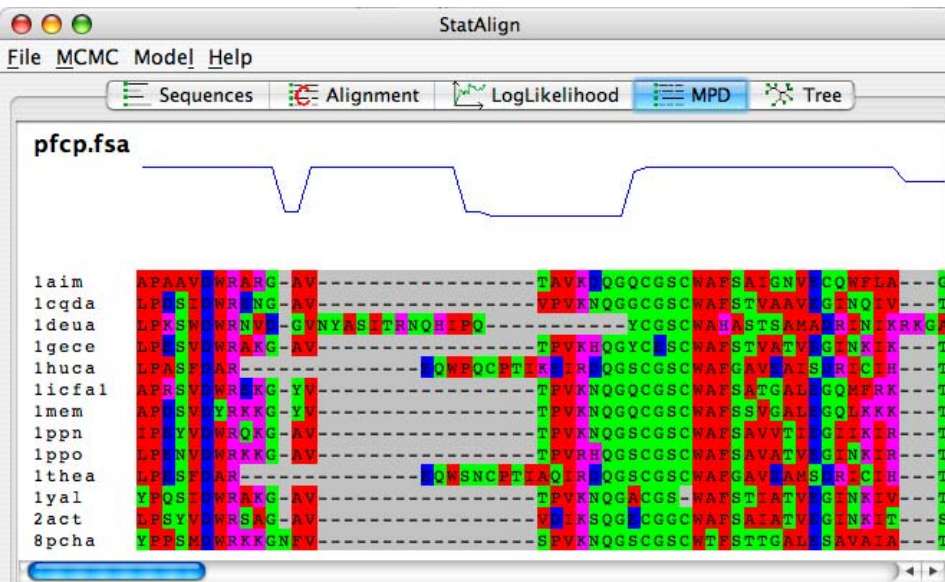
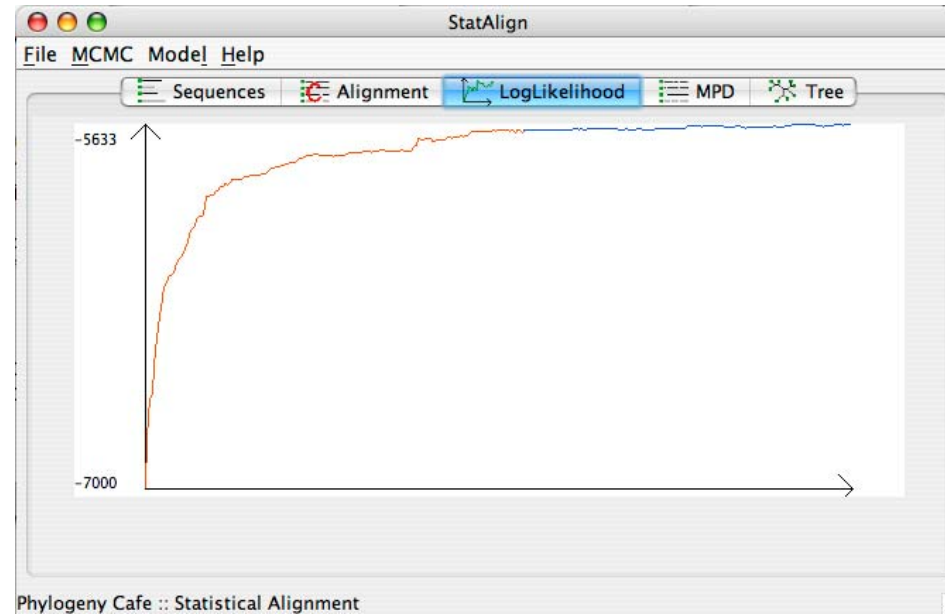
Combine above approaches



StatAlign software package

<http://phylogeny-café.elte.hu/StatAlign/stalign.tar.gz>

- Written in Java 1.5
- Platform-independent graphical interface
- Jar file is available, no need to instal
- Open source, extendable modules



Summary

The Problem

- *Statistical Alignment - Annotation - Annotation & Statistical Alignment*

Statistical Alignment

- *The Model*
- *The Pairwise Algorithm – the HMM connection*
- *the multiple sequence alignment algorithm*

Annotation

- *The general problem*
- *protein secondary structure – protein genes – RNA structure - signal*

Annotation & Alignment

- *The general algorithm*
- *Signals (footprinting)*
- *Protein Secondary Structure Prediction*

Ahead

- *Transcription Factor Prediction - Knowledge transfer - homologous/nonhomologous analysis*

Acknowledgements

Footprinting: *Rahul Satija*, Lior Pachter, Gerton Lunter

MCMC: Istvan Miklos, Jens Ledet Jensen, Alex Drummond,

Program: Adam Novak, Rune Lyngsø

Spannoids: Jesper Nielsen, Christian Storm

Earlier Statistical alignment Collaborators Mike Steel, Yun Song, Carsten Wiuf, Bjarne Knudsen, Gustav Wiebling, Christian Storm, Morten Møller,

Funding

BBSRC

MRC

Rhodes Foundation

Software

<http://phylogeny-café.elte.hu/StatAlign/statalign.tar.gz>

Next steps

<http://www.stats.ox.ac.uk/research/genome/projects>

Statistical Alignment and Footprinting

Statistical Alignment and Footprinting

Although bioinformatics perceived is a new discipline, certain parts have a long history and could be viewed as classical bioinformatics. For example, application of string comparison algorithms to sequence alignment has a history spanning the last three decades, beginning with the pioneering paper by Needleman and Wunch, 1970. They used dynamic programming to maximize a similarity score based on a cost of insertion-deletions and a score function on matched amino acids. The principle of choosing solutions by minimizing the amount of evolution is also called parsimony and has been widespread in phylogenetic analysis even if there is no alignment problem. This situation is likely to change significantly in the coming years. After a pioneering paper by Bishop and Thompson (1986) that introduced and approximated likelihood calculation, Thorne, Kishino and Felsenstein (1991) proposed a well defined time reversible Markov model for insertion and deletions (the TKF91-model), that allowed a proper statistical analysis for two sequences. Such an analysis can be used to provide maximum likelihood (pairwise) sequence alignments, or to estimate the evolutionary distance between two sequences. Steel et al. (2001) generalized this to any number of sequences related by a star tree. This was subsequently generalized further to any phylogeny and more practical methods based on MCMC has been developed. We have developed this into a generally available program package.

Traditional alignment-based phylogenetic footprinting approaches make predictions on the basis of a single assumed alignment. The predictions are therefore highly sensitive to alignment errors or regions of alignment uncertainty. Alternatively, statistical alignment methods provide a framework for performing phylogenetic analyses by examining a distribution of alignments. We developed a novel algorithm for predicting functional elements by combining statistical alignment and phylogenetic footprinting (SAPF). SAPF simultaneously performs both alignment and annotation by combining phylogenetic footprinting techniques with an hidden Markov model (HMM) transducer-based multiple alignment model, and can analyze sequence data from multiple sequences. We assessed SAPF's predictive performance on two simulated datasets and three well-annotated cis-regulatory modules from newly sequenced *Drosophila* genomes. The results demonstrate that removing the traditional dependence on a single alignment can significantly augment the predictive performance, especially when there is uncertainty in the alignment of functional regions. The transducer-based version of SAPF is currently able to analyze data from up to five sequences. We are currently developing an MCMC approach that we hope will be capable of analyzing data from 12-16 species, enabling the user to input sequence data from all 12 recently sequenced *Drosophila* genomes. We will present initial results from the MCMC version of SAPF and discuss some of the challenges and difficulties affecting the speed of convergence.