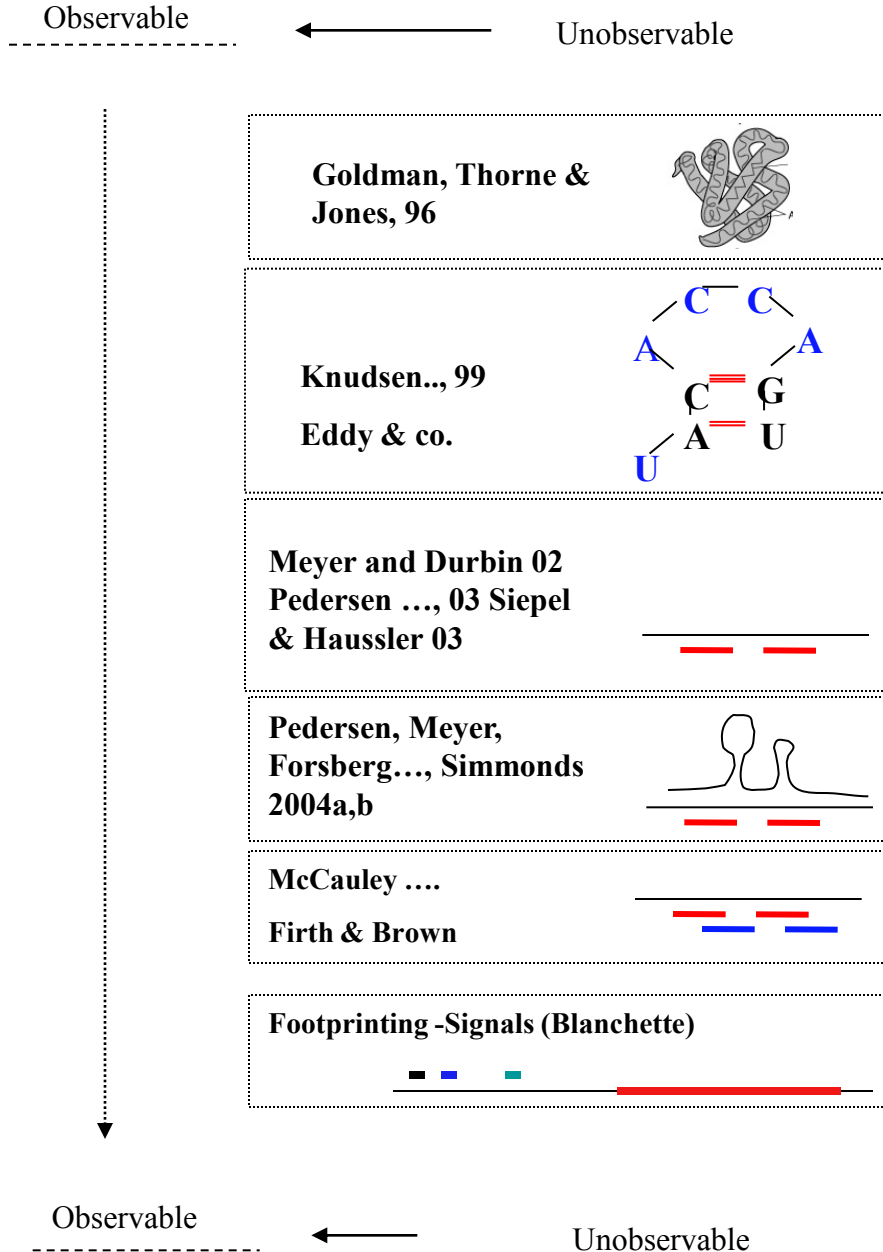


Co-Modelling and Conditional Modelling



AGGTATATAATGCG..... $P_{\text{coding}}\{\text{ATG} \rightarrow \text{GTG}\}$ or
AGCCATTTAGTGCG..... $P_{\text{non-coding}}\{\text{ATG} \rightarrow \text{GTG}\}$



• Conditional Modelling

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

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

Needs:

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

ii. $P(\text{Structure})$

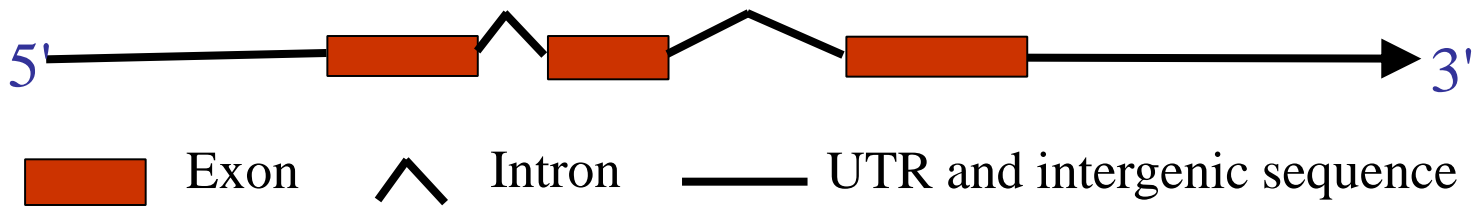
Ab Initio Gene prediction

Ab initio gene prediction: prediction of the location of genes (and the amino acid sequence it encodes) given a raw DNA sequence.

... tttttgcagtagtcccgggcccctctgttggggcctccccttctctccagggtggagtcgaggaggcggggctcggggcctccttatctctagagccggcctggctctctggcgccg
ggggccccttagtccgggctttttgccatggggctctctgttcccctctgtcgtgctgtgtttttttggcgccgcctaccgggagttgggagcgcgctgggacgcgccgactaagcggggcgc
aaagccccaagggtagccctctcgcgccctccgggacctcagtgcccttctgggtgcgcatgagcccggagttcgtggctgtgcagccggggaagtcaagtgcagctcaattgcagcaaca
gctgtccccagccgcagaattccagcctccgcctccgctgcgccgaaggcaagacgctcagaggccgggttggtgtcttaccagctgctcgcagctgagggcctggagctcccctcgcgc
actgcctcgtgacctgcgagggaaaaacacgctgggcatctccaggttctactgctacagtgagggatggggtctcccccggctgggggtgaggggagggggctggaagaggtggggg
aagggtagttgacagtcgctctatagggagcggggggcctctggggatccccttgggtgagccctccagggcgtgattttggagcctccggctttaaagggcagggaaat
acactttgcgctgccacgtgacgcaggtgttcccgggtgggctctgggtgggtgacctgagggcatggaagccgggtcatctattccgaaagcctggagcgccttaccggcctggatctgg
ccaacgtgaccttgacctacgagtttgctgctggaccccgcgacttctggcagcccgtgatctgccacgcgcgcctcaatctcgcagggcctgggtgggtccgcaacagctcggcaccatta
cactgatgctcgggtgagggaccccctgtaacctggggactaggaggaagggggcagagagagttatgaccccagaggggcgcacagaccaagcgtgagctccacgcgggtcgacagacct
cccctgtgttccggttccctaattctcgccttctgctcccagcttggagccccgcgccacagctttggcctccgggtccatcgctgccccttgtagggatcctcctcactgtggggcgtcgtg
acctatgcaagtgccctagctatgaagtcccaggcgtaaagggggatgttctatgcccggctgagcgagaaaaagaggaatatgaaacaatctgggggaaatggccatacatggtg....

Input data

Output:

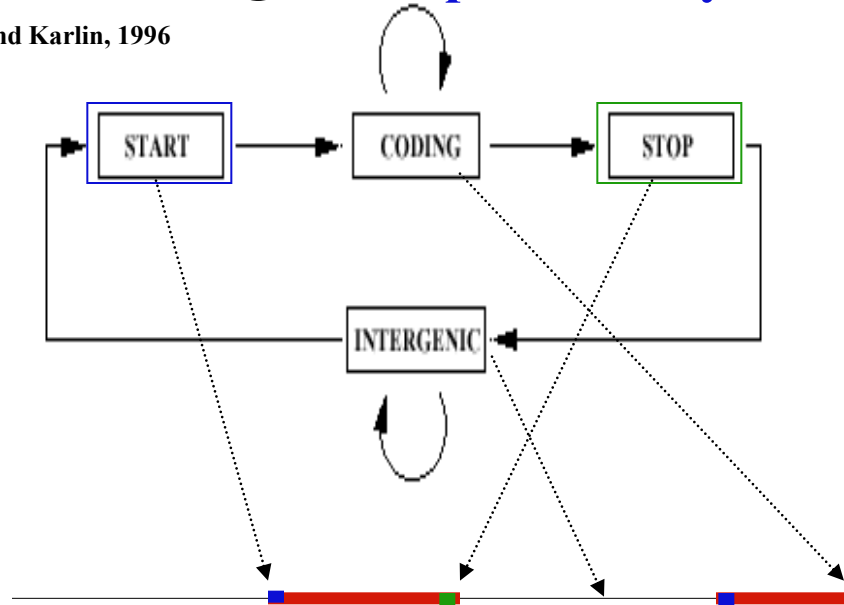


5'... tttttgcagtagtcccgggcccctctgttggggcctccccttctctccagggtggagtcgaggaggcggggctcggggcctccttatctctagagccggcctggctctctggcgccggggccccttagtccgggctttttgccATGGGGTCTCTGTTCCCTCTGTCGCTGCTGTTTTTTTTTGGCGGCCGCCTACCCGGGAGTTGGGAGCGCGCTGGGACGCCGACTAAGCGGGCGCAAAGCCCAAGGGTAGCCCTCTCGGCCCTCCGGGACCTCAGTGCCCTTCTGGGTGCGCATGAGCCCAGGAGTTCGTGGCTGTGCAGCCGGGGAAGTCAGTGCAGCTCAATTGCAGCAACAGCTGTCCCCAGCCGCAGAATTCCAGCCTCCGCACCCCGCTGCGGCAAGGCAAGACGCTCAGAGGGCCGGGTTGGGTGTCTTACCAGCTGCTCGACGTGAGGGCCTGGAGCTCCCTCGCACTGCCTCGTGACCTGCGCAGGAAAAACACGCTGGGCCACCTCCAGGATCACCGCTACAgtgaggggacaggggctcgggtcccggctgggggtgaggggagggggctggaagaggtggggaa
gggtagttgacagtcgctctatagggagcggccgagacctcactcagaggetccccttgccttagAACCGCCCCACAGCGTGATTTTGGAGCCTCCGGTCTTAAAGGGCAGGAAATACACTTTGCGCTGCCACGTGACGCAGGTGTTCCCGGTGGGCTACTTGGTGGTGACCCTGAGGCATGGAAGCCGGGTCATCTATTCCGAAAGCCTGGAGCGCTTACCAGGCTGGATCTGGCCAACGTGACCTTGACCTACGAGTTTGTGCTGGACCCCGCACTTCTGGCAGCCCCTGATCTGCCACGCgcgcctcaatctcgcagggcctgggtgggtccgcaacagctcggcaccatta
TGGCAACGTGACCTTGACCTACGAGTTTGTGCTGGACCCCGCACTTCTGGCAGCCCCTGATCTGCCACGCgcgcctcaatctcgcagggcctgggtgggtccgcaacagctcggcaccatta
CAGCTCGGCACCCATTACACTGATGCTCGgtgagggacacctgtaacctggggactaggaggaagggggcagagagagttatgaccccagaggggcgcacagaccaagcgtgagctccacgcgggtcgacagacctcccctgtgt
ccgttccctaattctcgccttctgctcccagCTTGGAGCCCCCGCCCCACAGCTTTGGCCTCCGGTTCATCGCTGCCCTTGTAGGGATCCTCCTCACTGTGGGCG
CTGCGTACCTATGCAAGTGCCCTAGCTATGAAGTCCCAGGCGTAAagggggatgttctatgcccggctgagcgagaaaaagaggaatatgaaacaatctgggggaaatggccatacatggtg.... 3'

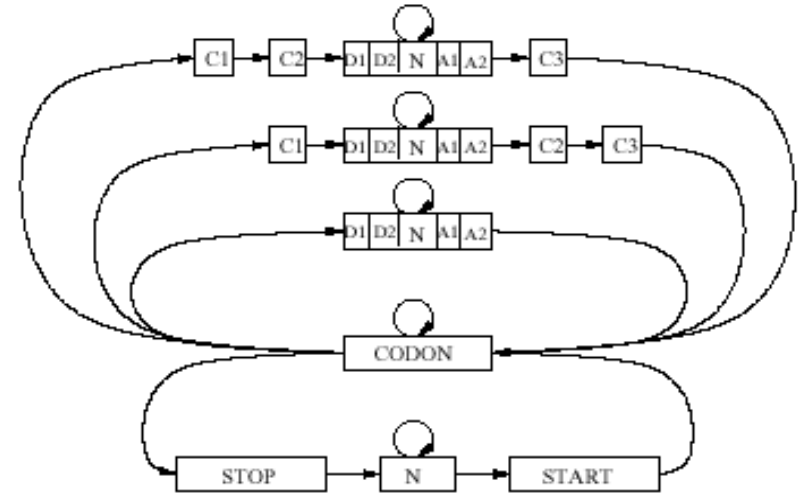
HMM Examples

Gene Finding: Simple Prokaryotic

Burge and Karlin, 1996

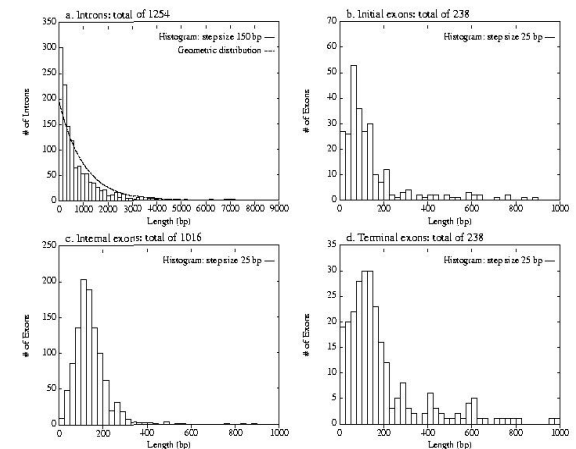


Simple Eukaryotic



• Intron length > 50 bp required for splicing

• Length distribution is not geometric



Legend: Intron, exon length data from 238 multi-exon genes of GENSCAN learning set (Appendix A1).

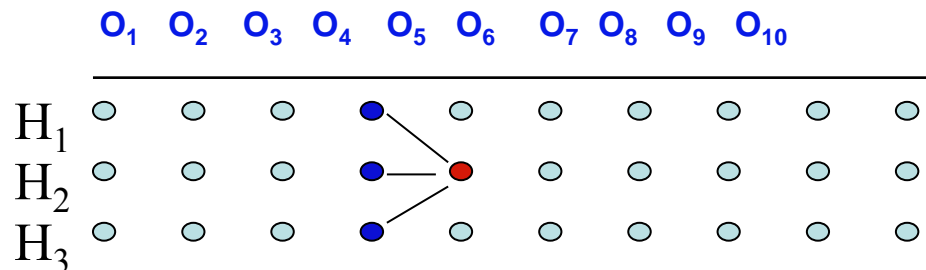
What is the probability of the data?

The probability of the observed is $P(\vec{O}) = \sum_{\vec{H}} P(\vec{O}|\vec{H})P(\vec{H})$, which could be hard to calculate. However, these calculations can be considerably accelerated. Let $P_{O_k=i}^{H_k=j}$ the probability of the observations (O_1, \dots, O_k) conditional on $H_k=j$. Following recursion will be obeyed:

$$i. P_{O_k=i}^{H_k=j} = P(O_k = i | H_k = j) \sum_{H_{k-1}=r} P_{O_{k-1}}^{H_{k-1}=r} p_{r,i}$$

$$ii. P_{O_1=i}^{H_1=j} = P(O_1 = i | H_1 = j) \pi_j \quad (\text{initial condition})$$

$$iii. P(O) = \sum_{H_n=j} P_{O_n=i}^{H_n=j}$$



$$P_{O_5=i}^{H_5=2} = P(O_5 = i | H_5 = 2) \sum_{H_4=j} P_{O_4}^{H_4=j} p_{j,i}$$

What is the most probable "hidden" configuration?

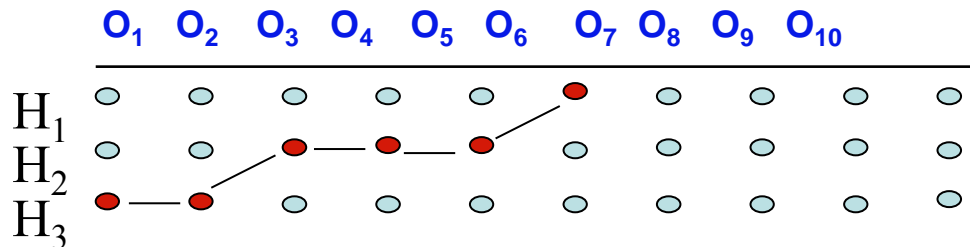
Let H^* be the sequences of hidden states in the most probably hidden path ie $\text{ArgMax}_H [P\{O|H\}]$. Let H_k^j be the probability of the most probable path up to k ending in hidden state j .

Again recursions can be found:

$$i. H_1^j = \pi_j e(O_1, 1) \quad ii. H_k^j = \max_i \{H_{k-1}^i p_{i,j}\} e(O_k, j)$$

The actual sequence of hidden states H_k^* can be found recursively by

$$iii. H_{k-1}^* = \{i \mid H_{k-1}^i p_{i,j} e(O_k, j) = H_k^{H_k^*}\}$$



$$H_6^1 = \max_j \{H_6^j * p_{j,1} * e(O_6, 1)\}$$

$$H_5^* = \{i \mid H_5^i * p_{i,1} * e(O_6, 1) = H_6^1\}$$

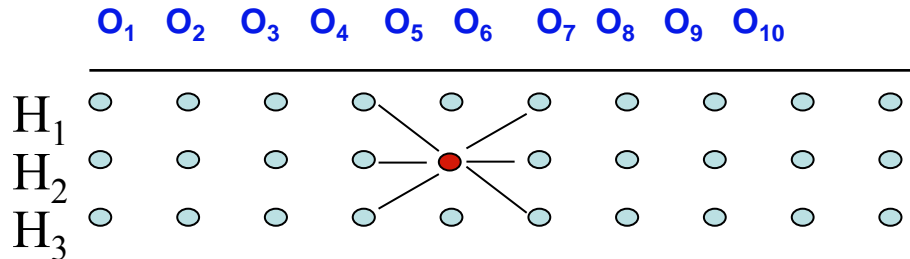
What is the probability of specific "hidden" state?

Let Q_k^j be the probability of the observations from $k+1$ to n given $H_k=j$. These will also obey recursions:

$$Q_k^j = \sum_{H_{k+1}=i} P(O_{k+1} | H_{k+1} = i) p_{j,i} Q_{k+1}^i$$

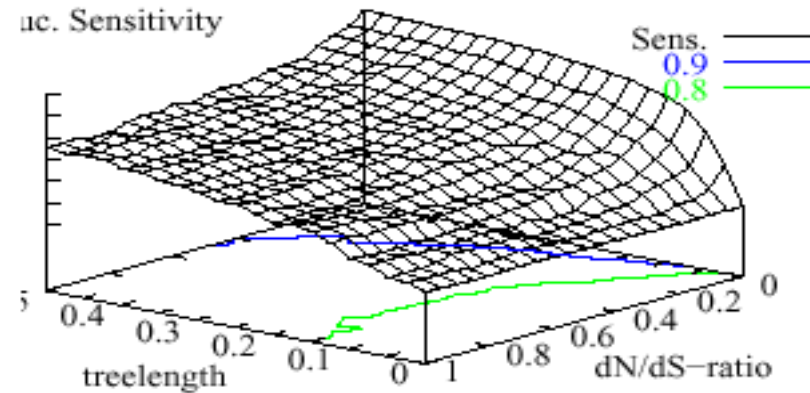
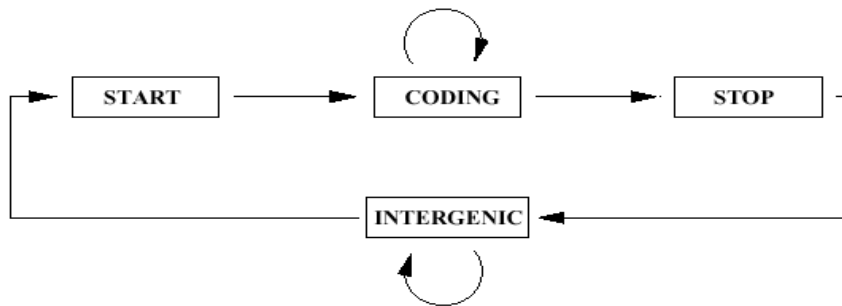
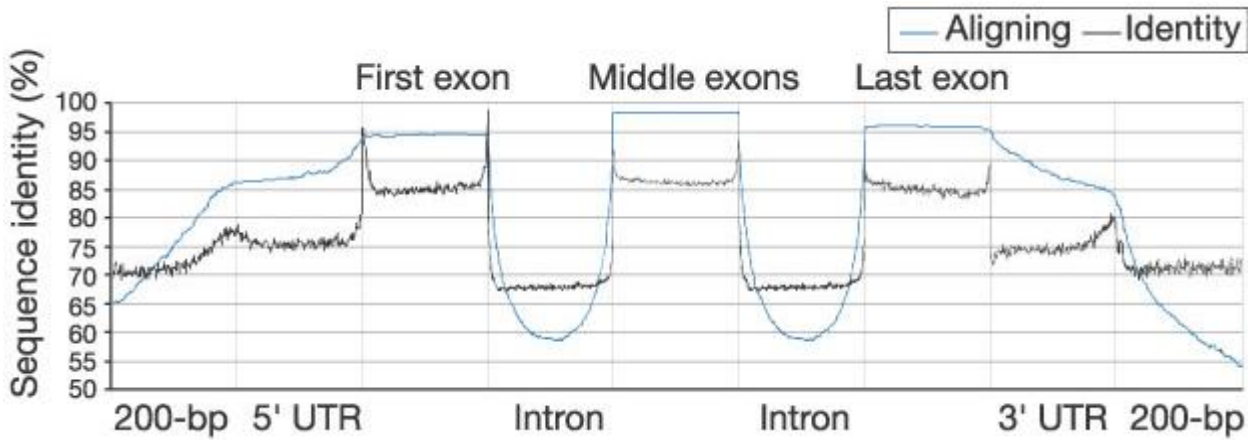
The probability of the observations and a specific hidden state can be found as: $P\{O, H_k = j\} = P_k^j Q_k^j$

And of a specific hidden state can be found as: $P\{H_k = j\} = P_k^j Q_k^j / P(O)$



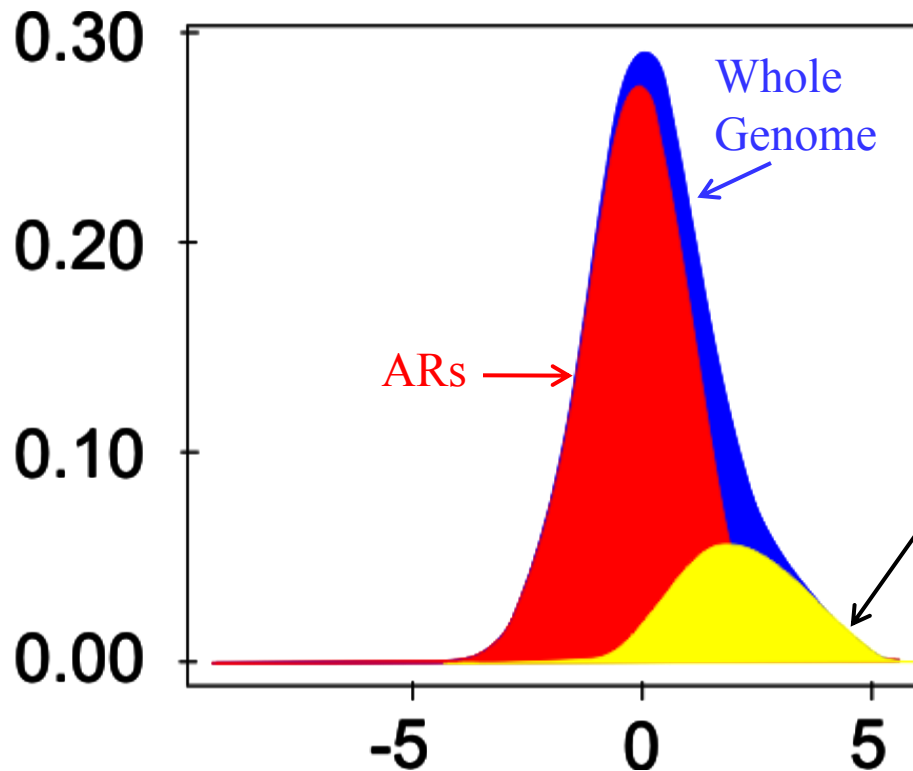
$$P\{H_5 = 2\} = P_5^2 Q_5^2 / P(O)$$

Comparative Gene Annotation



AGGTATATA**ATGCG**..... $P_{\text{coding}}\{\text{ATG} \rightarrow \text{GTG}\}$ or
 AGCCATTTA**GTGCG**..... $P_{\text{non-coding}}\{\text{ATG} \rightarrow \text{GTG}\}$

~5% of the Human genome is under conservation
(Chiaromonte *et al.*)



(Whole Genome – ARs)

Table 1. Estimates of the Share of the Human Genome under Selection for Different Window Sizes (W) and Required Number of Aligned Bases (T)

| W | T | $p_1 = (1 - p_0)$ | Coverage | α_{sel} (%) |
|-----|-----|-------------------|-----------------|--------------------|
| 30 | 20 | 0.15 | 846472K (30.4%) | 4.51 |
| | 25 | 0.17 | 743308K (26.7%) | 4.50 |
| | 30 | 0.23 | 439501K (15.8%) | 3.65 |
| 50 | 40 | 0.19 | 756051K (27.1%) | 5.19 |
| | 45 | 0.22 | 623286K (22.4%) | 4.90 |
| | 50 | 0.31 | 292506K (10.5%) | 3.31 |
| 100 | 80 | 0.23 | 739836K (26.6%) | 6.15 |
| | 90 | 0.29 | 550530K (19.8%) | 5.8 |
| | 100 | 0.52 | 122437K (4.4%) | 2.29 |
| 200 | 160 | 0.31 | 708701K (25.4%) | 7.92 |
| | 180 | 0.40 | 467954K (16.8%) | 6.68 |
| | 200 | 0.81 | 328668K (1.2%) | 0.96 |

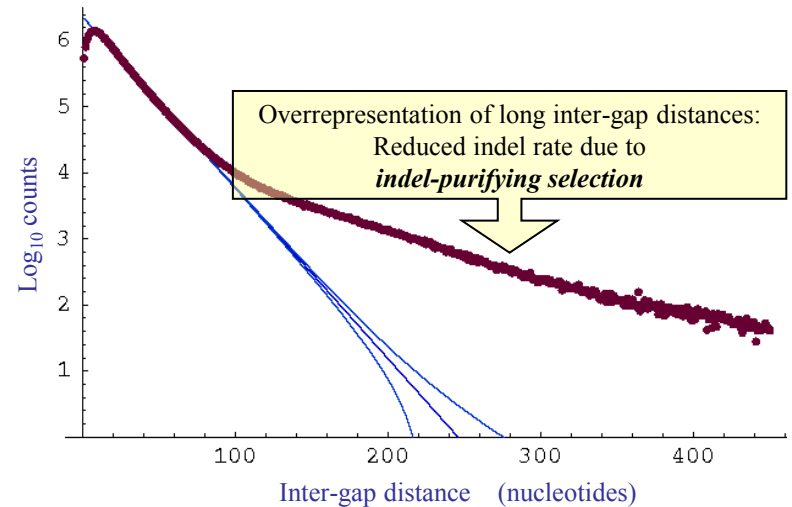
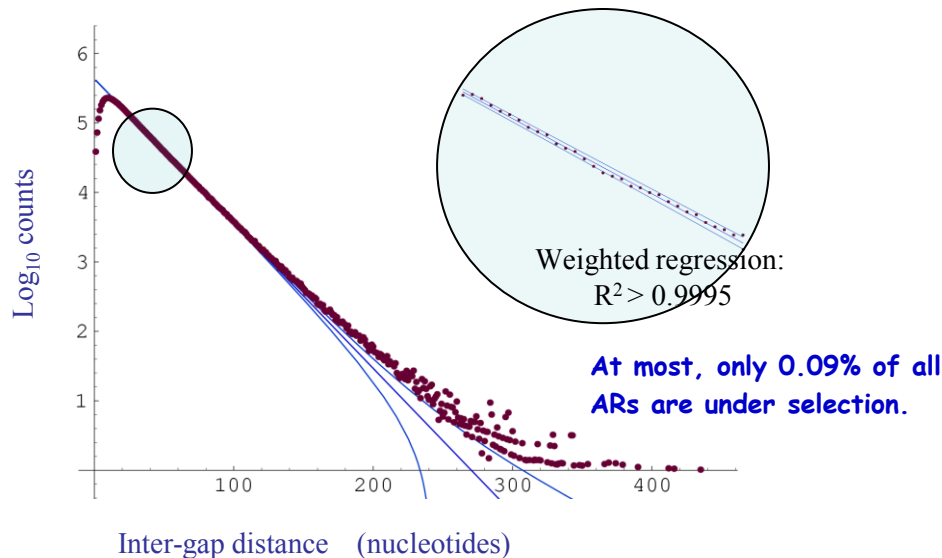
Due to this work, people often say
5% of the genome is constrained

Percentage of Genome under Purifying Selection

CGACATTAA--ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----CAGACGCA



Consider lengths of *inter-gap segments*! Do they follow a geometric distribution?

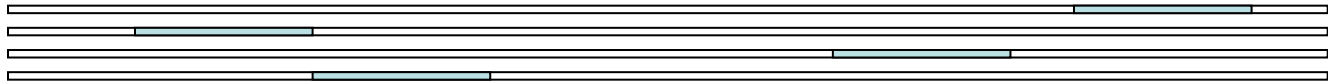


Finding Regulatory Signals in Genomes

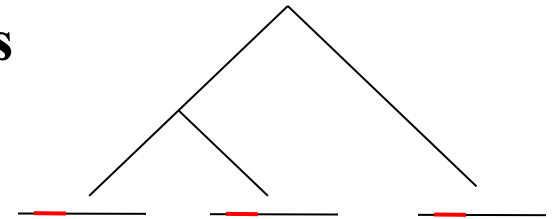
Searching for known signal in 1 sequence



Searching for unknown signal common to set of unrelated sequences

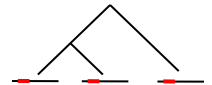


Searching for conserved segments in homologous

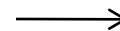
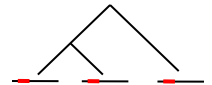


Challenges

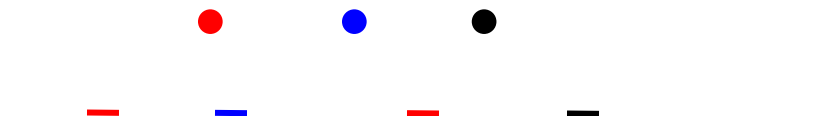
Combining homologous and non-homologous analysis



Merging Annotations



Predicting signal-regulatory protein relationships



Weight Matrices & Sequence Logos

Set of signal sequences:

$f_{b,i}$ b's in position i , $s(b)$ pseudo count.

$$\text{corrected probability} : p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \text{ nucleo}} s(b')}$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|
| 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

B R M C W A W H R W G G B M

| A | -1.93 | .79 | .79 | -1.93 | .45 | 1.50 | .79 | .45 | 1.07 | .79 | .0 | -1.93 | -1.93 | .79 |
|---|-------|-------|-------|-------|-------|-------|-------|-----|-------|-------|-------|-------|-------|-------|
| C | .45 | -1.93 | .79 | 1.68 | -1.93 | -1.93 | -1.93 | .45 | -1.93 | -1.93 | -1.93 | -1.93 | .0 | .79 |
| G | .0 | .45 | -1.93 | -1.93 | -1.93 | -1.93 | -1.93 | .66 | -1.93 | 1.3 | 1.68 | 1.07 | -1.93 | |
| T | .15 | .66 | -1.93 | -1.93 | 1.07 | .66 | .79 | .0 | .79 | -1.93 | -1.93 | -1.93 | .66 | -1.93 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

T T G C A T A A G T A G T C
 .45 -.66 .79 1.66 .45 -.66 .79 .45 -.66 .79 .0 1.68 -.66 .79



Position Frequency Matrix - PFM

Consensus sequence:

Position Weight Matrix - PWM

$$PWM : W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

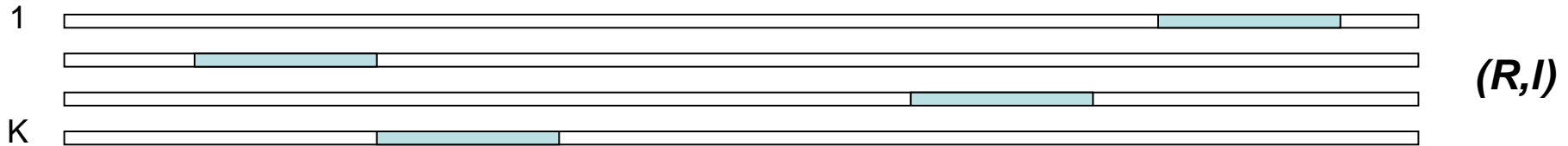
Score for New Sequence $S = \sum_{i=1}^w W_{b,i}$

Sequence Logo & Information content

$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i}$$

Motifs in Biological Sequences

- 1990 Lawrence & Reilly “An Expectation Maximisation (EM) Algorithm for the identification and Characterization of Common Sites in Unaligned Biopolymer Sequences Proteins 7.41-51.
 1992 Cardon and Stormo Expectation Maximisation Algorithm for Identifying Protein-binding sites with variable lengths from Unaligned DNA Fragments L.Mol.Biol. 223.159-170
 1993 Lawrence... Liu “Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment” Science 262, 208-214.



$\Theta = (\theta_{1,A}, \dots, \theta_{w,T})$ probability of different bases in the window

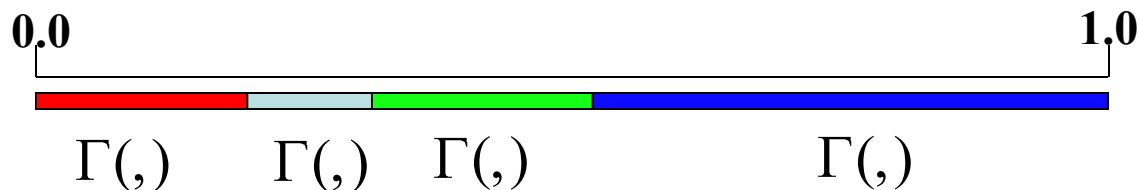
$A = (a_1, \dots, a_K)$ – positions of the windows

$\theta_0 = (\theta_A, \dots, \theta_T)$ – background frequencies of nucleotides.

$$p(R | \theta_0, \Theta, A) = \theta_0^{h(R_{\{A\}^c})} \prod_{j=1}^w \theta_j^{h(R_{A+j-1})} = \theta_0^{h(R)} \prod_{j=1}^w \left(\frac{\theta_j}{\theta_0} \right)^{h(R_{A+j-1})}$$

Priors A has uniform prior

Θ_j has Dirichlet($N_0\alpha$) prior – α base frequency in genome. N_0 is pseudocounts



Natural Extensions to Basic Model I

Multiple Pattern Occurances in the same sequences:

Liu, J. "The collapsed Gibbs sampler with applications to a gene regulation problem," *Journal of the American Statistical Association* **89** 958-966.

Prior: any position i has a small probability p to start a binding site:

$$A = (a_1, \dots, a_k) \quad P(A) \approx p_0^k (1 - p_0)^{N-k} \quad (\text{with nonoverlap ping constraints})$$



Composite Patterns:

BioOptimizer: the Bayesian Scoring Function Approach to Motif Discovery *Bioinformatics*



Natural Extensions to Basic Model II

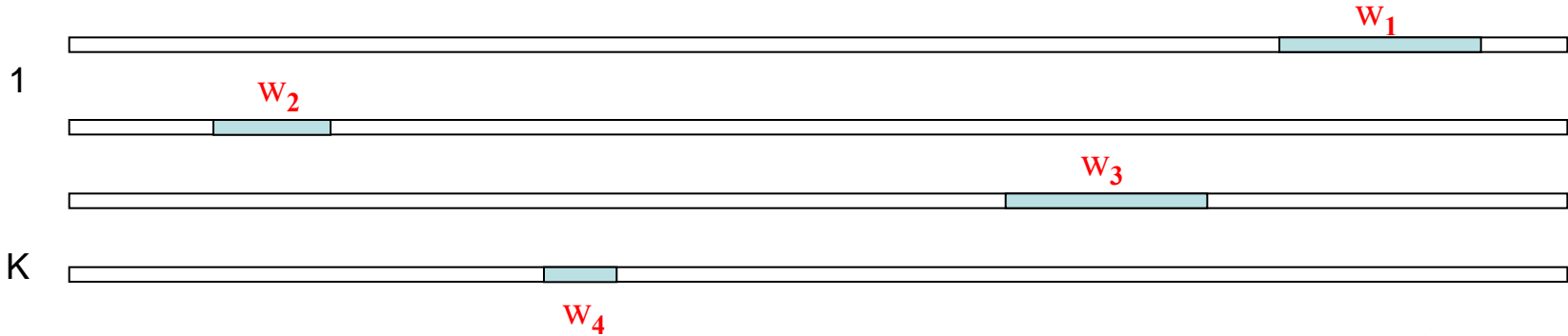
Correlated in Nucleotide Occurrence in Motif:

Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 6, 909-916.



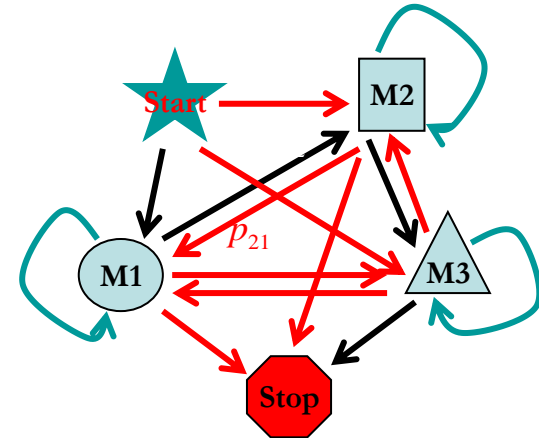
Insertion-Deletion

BALSA: Bayesian algorithm for local sequence alignment *Nucl. Acids Res.*, 30 1268-77.



Regulatory Modules:

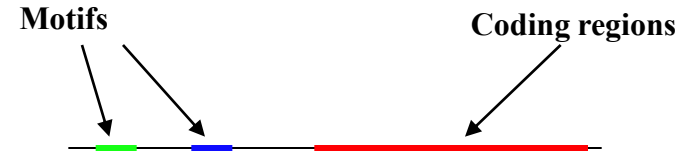
De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Nat'l Acad Sci USA*, 102, 7079-84



Combining Signals and other Data

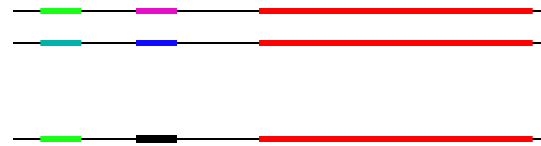
Expression and Motif Regression:

Integrating Motif Discovery and Expression Analysis Proc.Natl.Acad.Sci. 100.3339-44



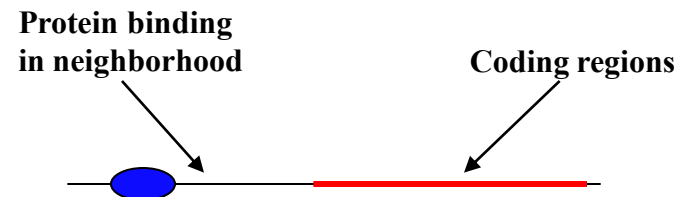
1. Rank genes by $E = \log_2(\text{expression fold change})$
2. Find “many” (hundreds) candidate motifs
3. For each motif pattern m , compute the vector S_m of matching scores for genes with the pattern

4. Regress E on S_m
$$Y_g = \alpha + \beta_m S_{mg} + \epsilon_g$$



ChIP-on-chip - 1-2 kb information on protein/DNA interaction:

An Algorithm for Finding Protein-DNA Interaction Sites with Applications to Chromatin Immunoprecipitation Microarray Experiments *Nature Biotechnology*, 20, 835-39



Phylogenetic Footprinting (homologous detection)

Term originated in 1988 in Tagle et al. **Blanchette et al.:** For unaligned sequences related by phylogenetic tree, find all segments of length **k** with a history costing less than **d**. Motif loss an option.

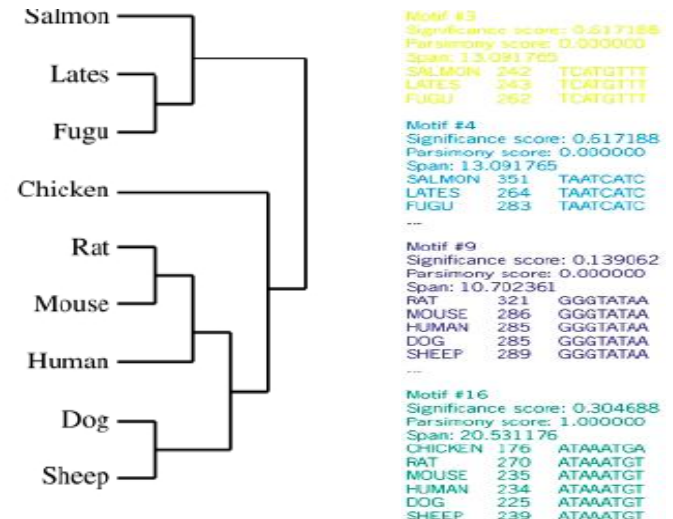
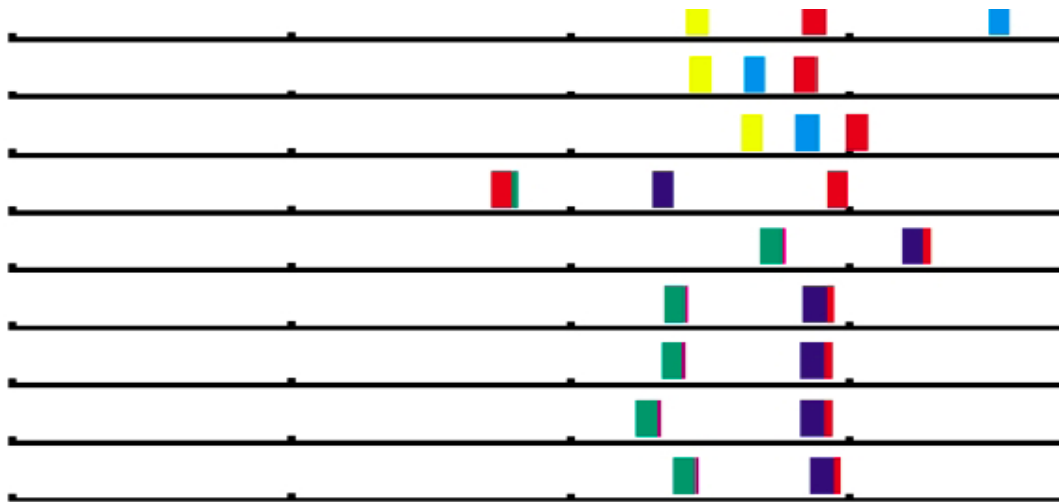
$$D_i^{begin} = \min\{ D_{i,\Delta}^{begin} + d(i,\Delta) \}$$

$$D_i^{signal,1} = \min\{ D_{i,\Delta}^{begin} + d(i,\Delta) \}$$

$$D_i^{signal,j} = \min\{ D_{i,\Delta}^{signal,j-1} + d(i,\Delta) \}$$

...

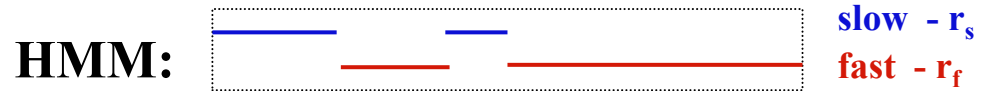
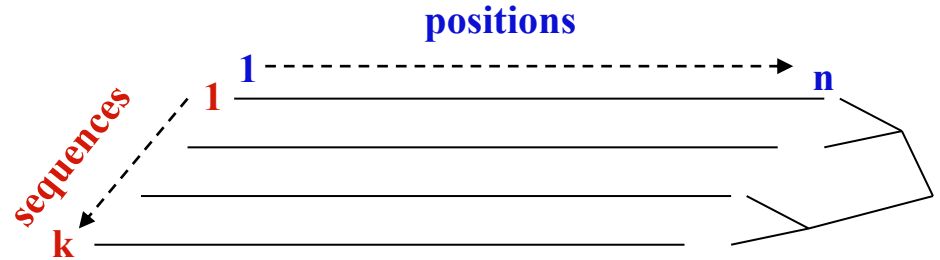
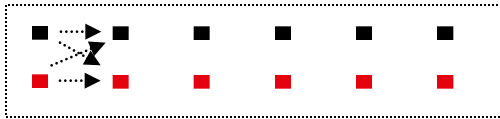
$$D_i^{end} = \min\{ D_{i,\Delta}^{end} + d(i,\Delta) \}$$



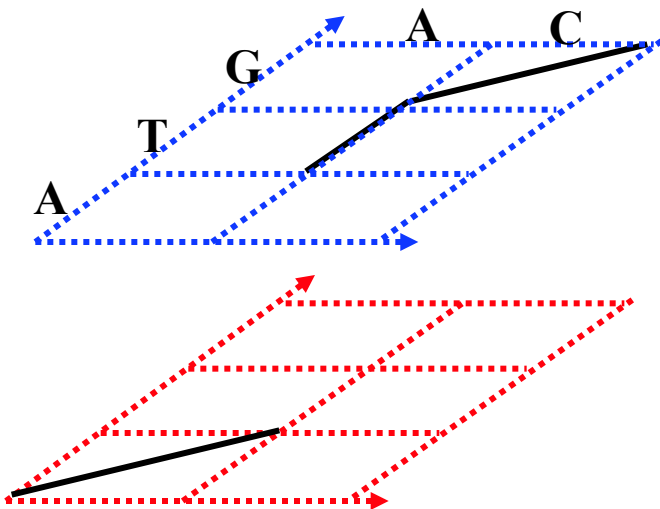
The Basics of Footprinting

- Many aligned sequences related by a known phylogeny:

HMM:



- Two un-aligned sequences:



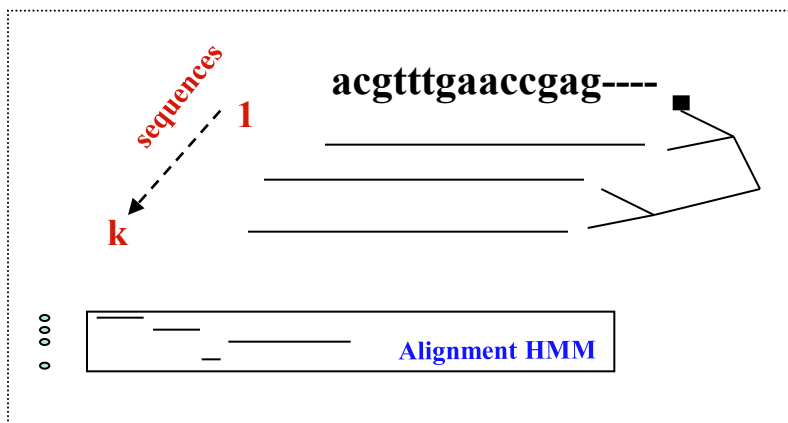
ATG

A-C

Statistical Alignment and Footprinting.

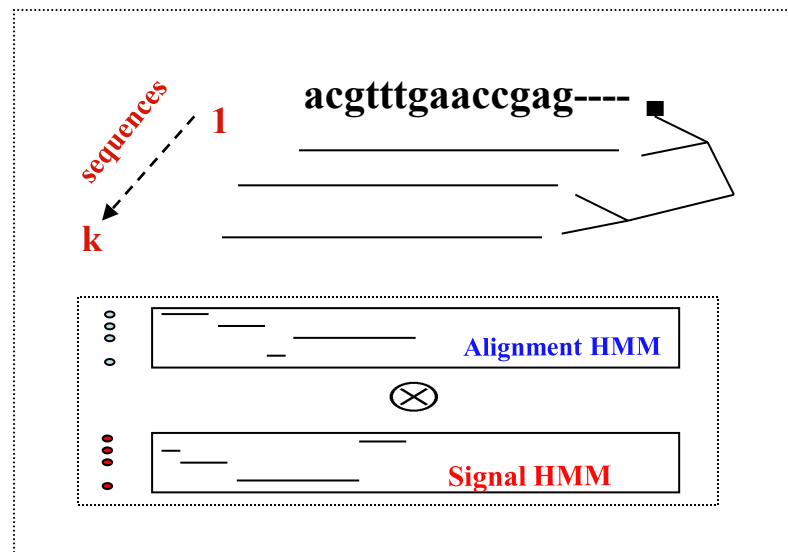
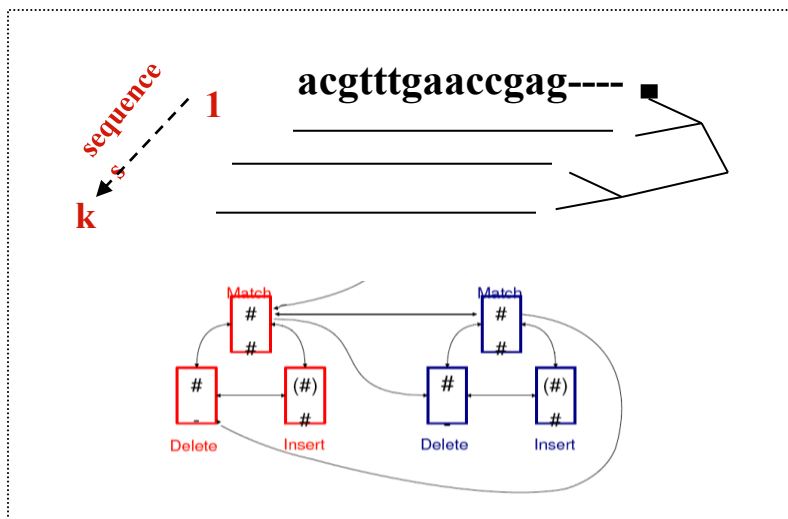
• Many un-aligned sequences related by a known phylogeny:

- Conceptually simple, computationally hard
- Dependent on a single alignment/no measure of uncertainty

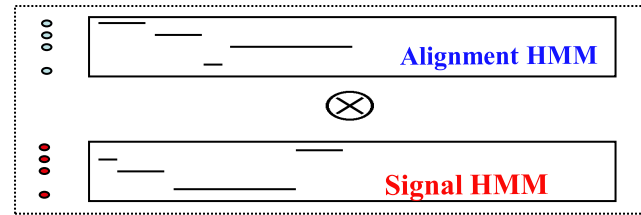
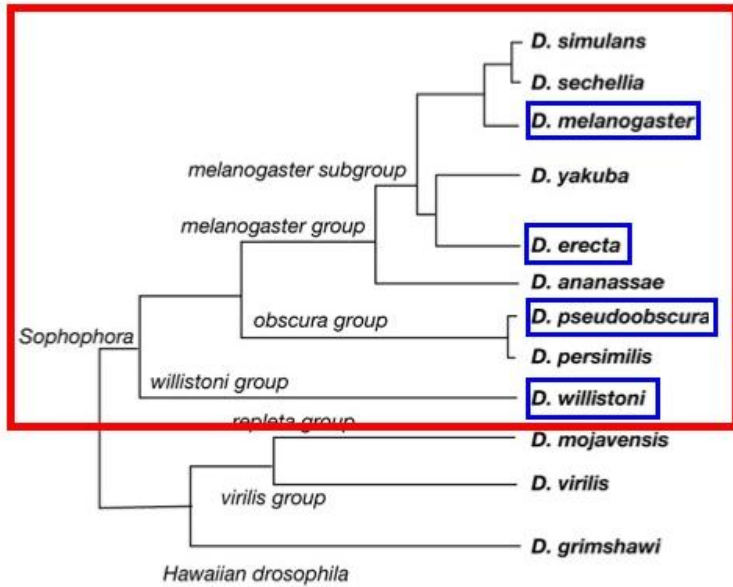


Solution:

Cartesian Product of HMMs

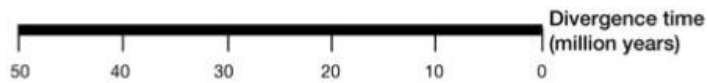


SAPF - Statistical Alignment and Phylogenetic Footprinting

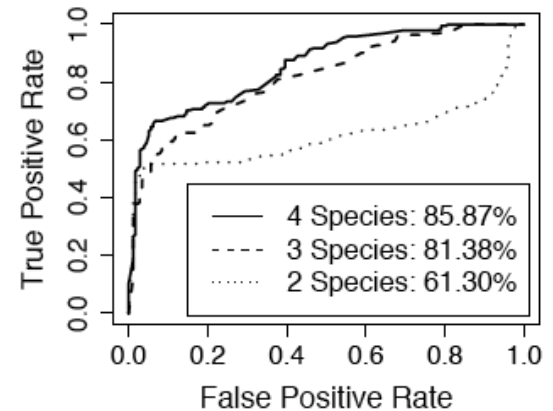
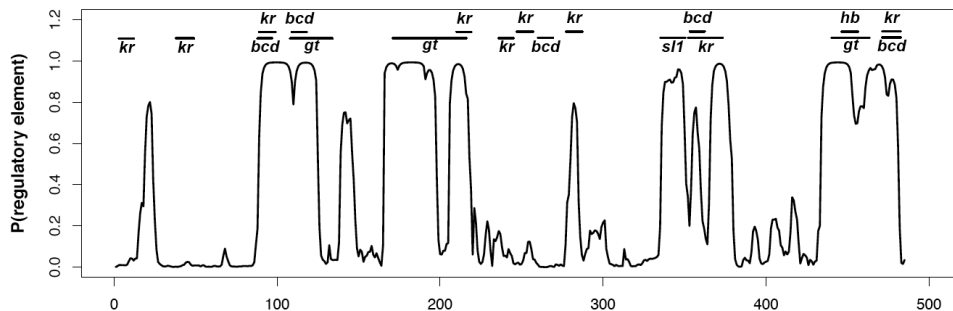


← Sum out

← Annotate



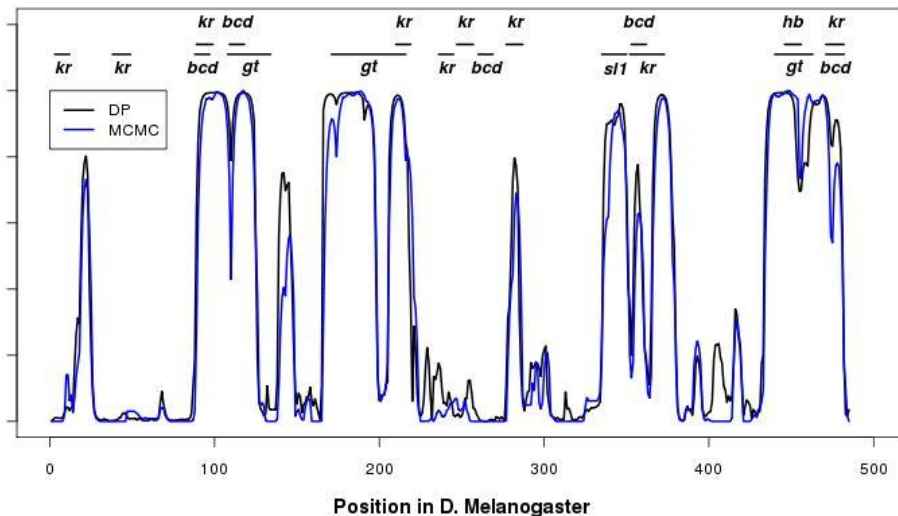
Eve stripe 2



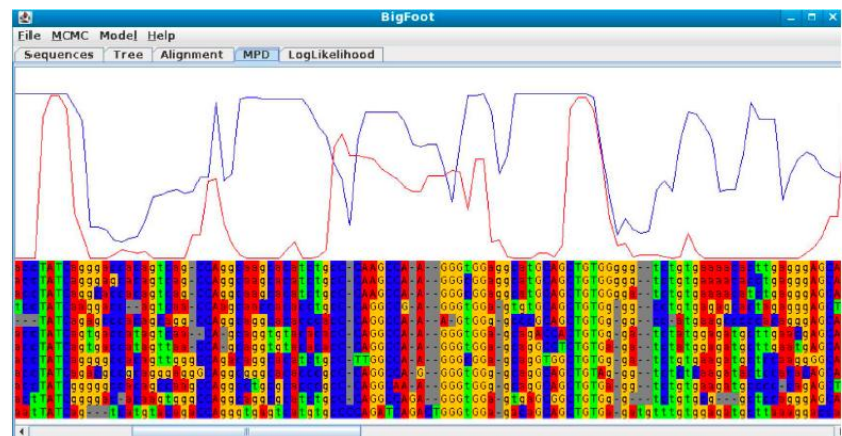
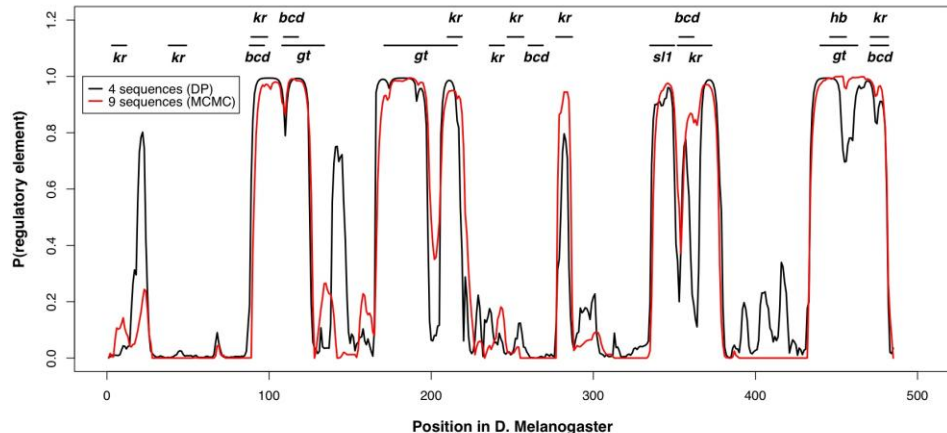
BigFoot

- *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



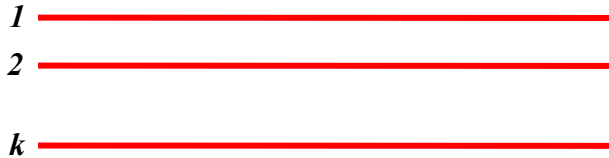
Eve stripe 2



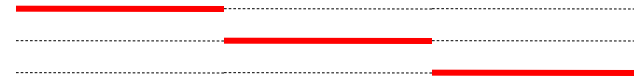
FSA - Fast Statistical Alignment

Pachter, Holmes & Co

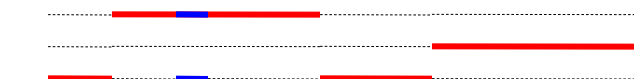
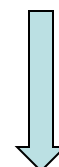
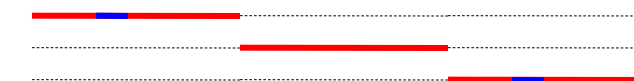
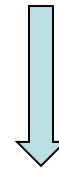
Data – k genomes/sequences:



Iterative addition of homology statements to shrinking alignment:



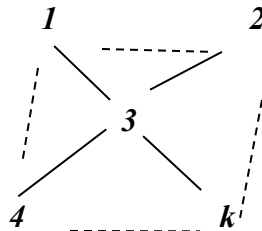
Add most certain homology statement from pairwise alignment compatible with present multiple alignment



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
- 12 2,k 1,4 4,k

Rate of Molecular Evolution versus estimated Selective Deceleration

Neutral Process

| | A | C | G | T |
|---|-----------|-----------|-----------|-----------|
| A | - | $q_{A,C}$ | $q_{A,G}$ | $q_{A,T}$ |
| C | $q_{C,A}$ | - | $q_{C,G}$ | $q_{C,T}$ |
| G | $q_{G,A}$ | $q_{G,C}$ | - | $q_{G,T}$ |
| T | $q_{T,A}$ | $q_{T,C}$ | $q_{T,G}$ | - |

How much selection?

Selection => deceleration

Selected Process

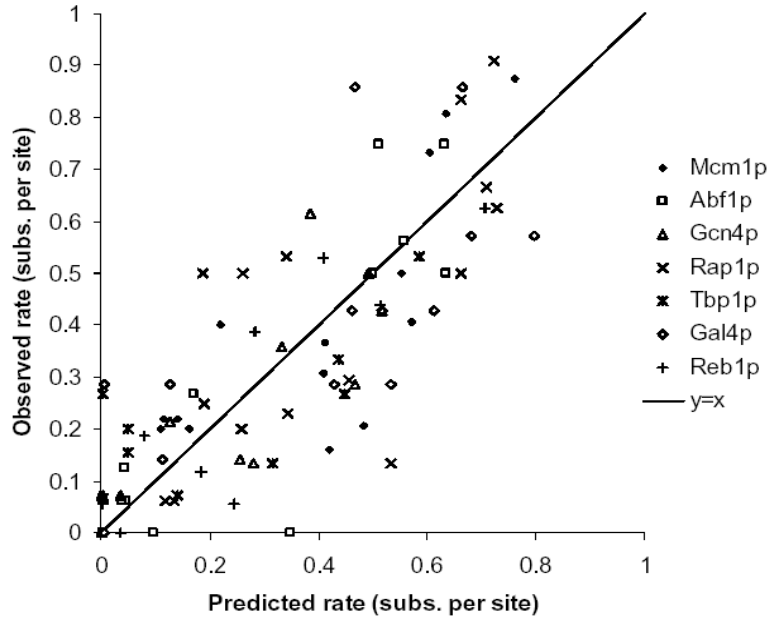
| | A | C | G | T |
|---|------------|------------|------------|------------|
| A | - | $q'_{A,C}$ | $q'_{A,G}$ | $q'_{A,T}$ |
| C | $q'_{C,A}$ | - | $q'_{C,G}$ | $q'_{C,T}$ |
| G | $q'_{G,A}$ | $q'_{G,C}$ | - | $q'_{G,T}$ |
| T | $q'_{T,A}$ | $q'_{T,C}$ | $q'_{T,G}$ | - |

Neutral Equilibrium

$$(\pi_A, \pi_C, \pi_G, \pi_T)$$

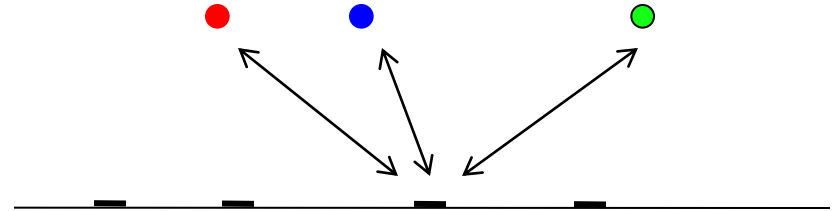
Observed Equilibrium

$$(\pi'_A, \pi'_C, \pi'_G, \pi'_T)$$



Signal Factor Prediction

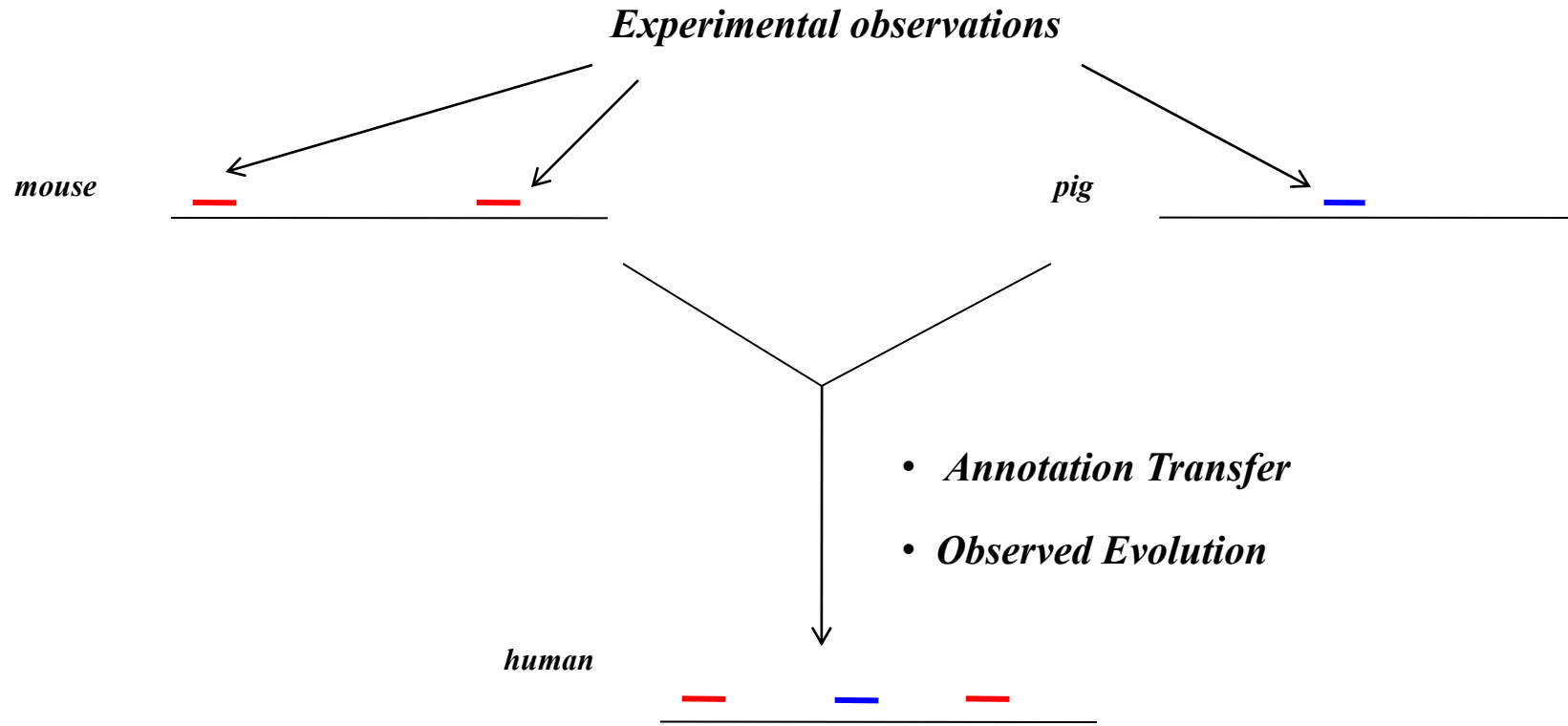
- *Given set of homologous sequences and set of transcription factors (TFs), find signals and which TFs they bind to.*



- *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.*

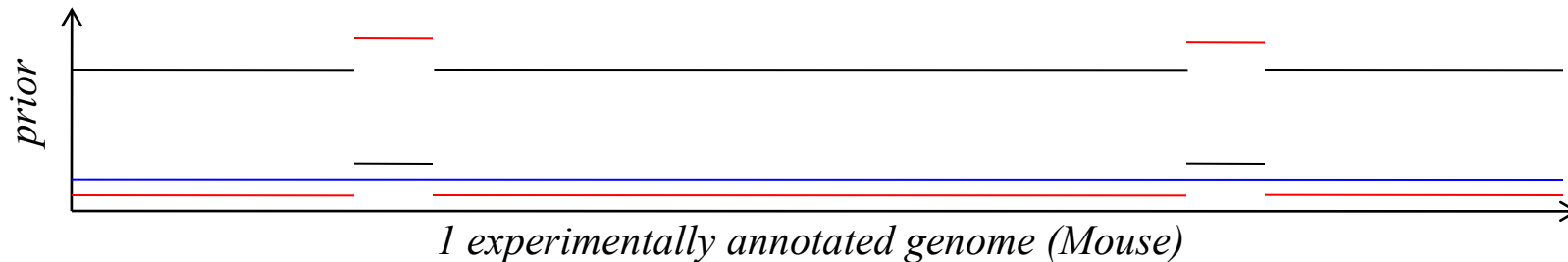
Knowledge Transfer and Combining Annotations



Must be solvable by Bayesian Priors

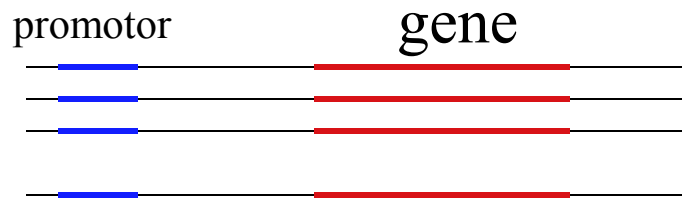
Each position p_i probability of being j 'th position in k 'th TFBS

If no experiment, low probability for being in TFBS

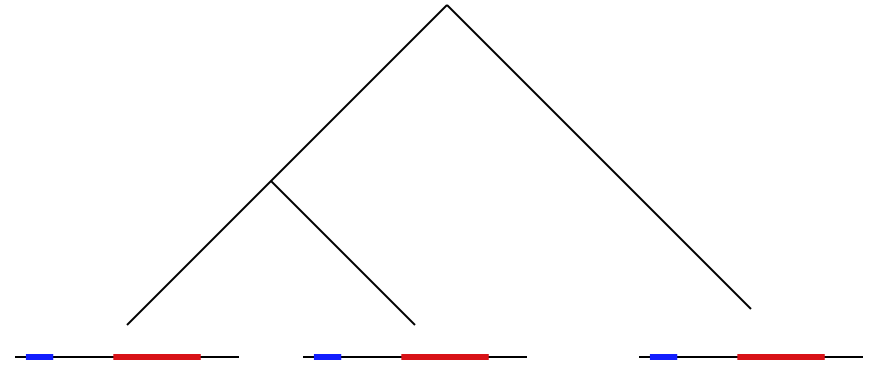


(Homologous + Non-homologous) detection

Unrelated genes - similar expression



Related genes - similar expression



Combine above approaches

