1) A motif can be generate from any collection of aligned DNA or protein sequences (T/F)? True – a motif model is simply made by counting up A/G/C/T for each position.

2) What is the maximum information content of the following degenerate motif, T[AG]A? As information content for a motif increases, the number of hits (above some cutoff) decreases (T/F)? 5 bits. 2 bits for all T and all A, A/G = 1 (if 50% each). True

3) Footprinting tells you where a protein binds DNA (T/F)? True by protecting DNA from cleavage


4) Gel shift can tell you whether a protein binds: a) probe DNA, b) competitor 'cold' DNA, c) both (c) both by shift of band with (a) and loss of band with (b)

5) Which motif has higher information content? Top one since there are more bits and lower entropy (randomness)

6) What is the probability of TGA with the following motif model? 0.2 x 0.2 x 0.1 = 0.004

$$PPM = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.2 & .5 & .1 \\ 0.3 & .1 & .1 \\ 0.3 & .2 & .1 \\ 0.2 & .2 & .7 \end{bmatrix}$$

7) Would you expect more or fewer matches to a high compared to a low information content motif in a genome? fewer

8) Both the binding energy model and log-likelihood ratio model of a motif can be derived from the position probability matrix [T/F]. True

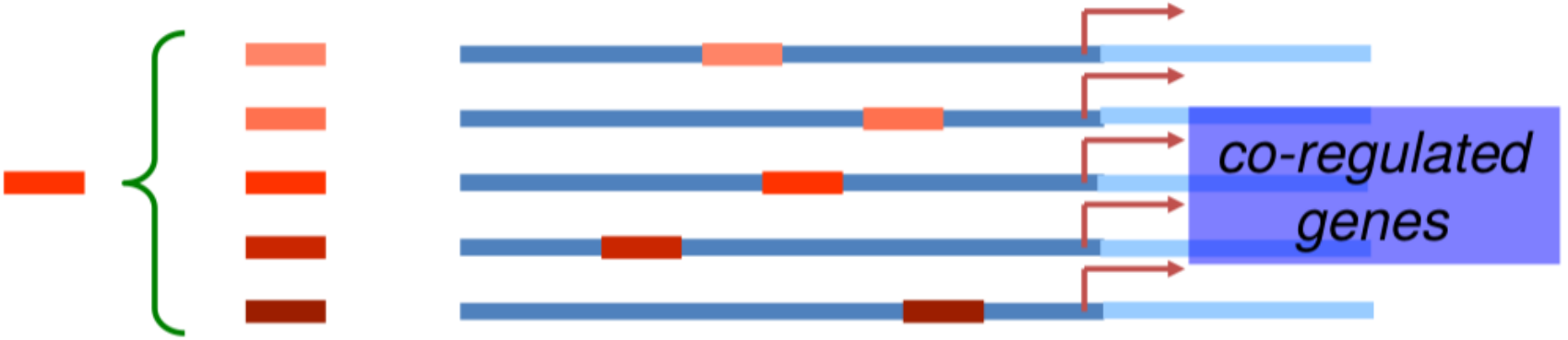
9) What is the advantage and disadvantgae of combinatorial motif search over greedy (consensus) search? Combinatorial guarantees the best (word) motif, but is slow.

# Today's objectives

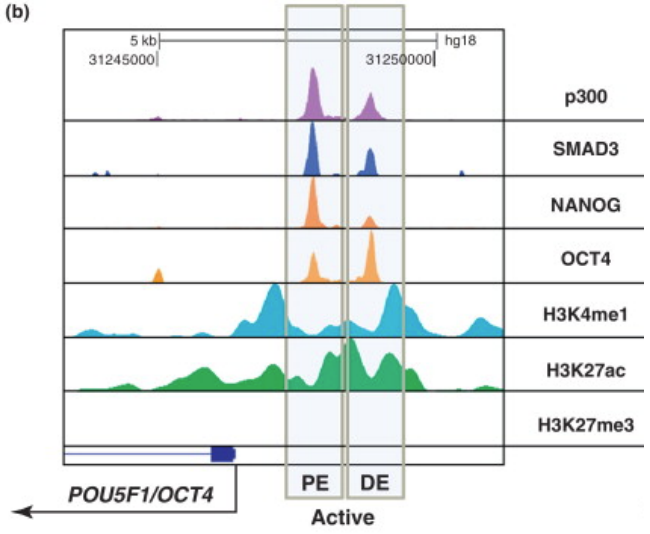
- Co-regulated genes: ChIP-seq, Clustering
- Expectation Maximization (EM)
- Gibbs sampling
- Phylogenetic footprinting
- Motif models of evolution

# Where do we expect to find motifs?

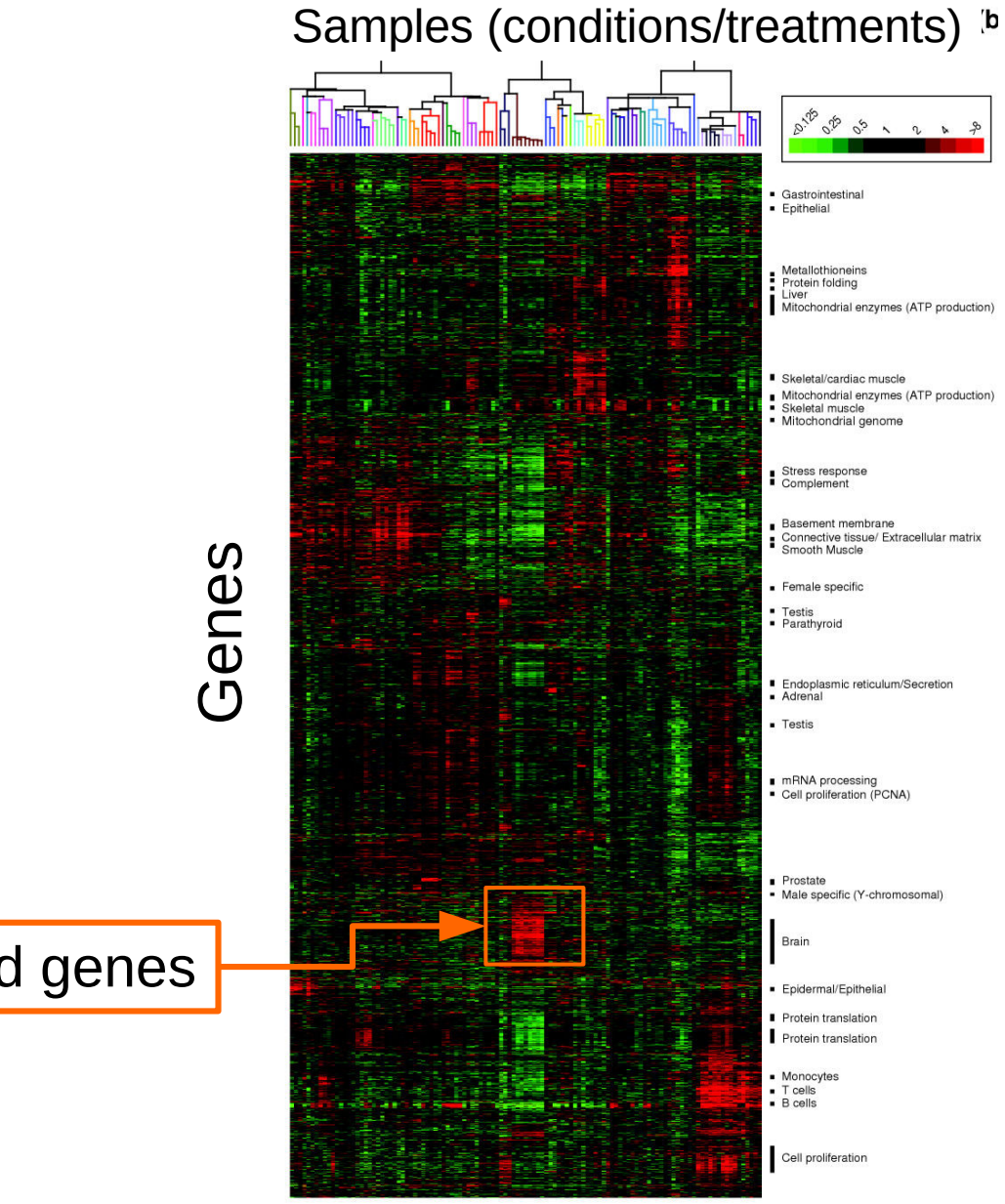
Find motif (binding sites) in a set of co-regulated genes (clustering)



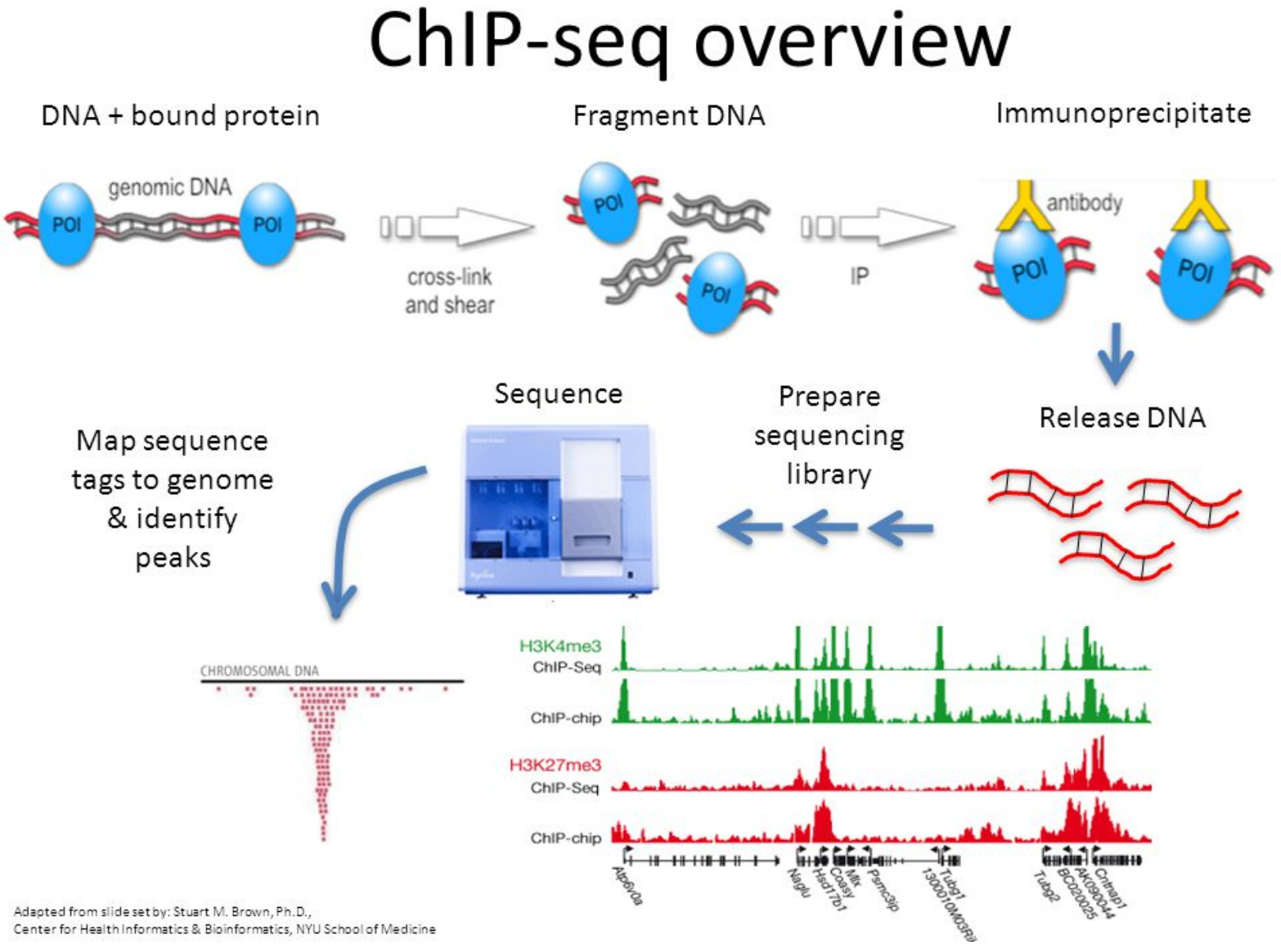

Find motif (binding sites) in a set of bound regions (e.g. ChIP-seq)



# Gene expression clustering (next week)



# Chromatin Immunoprecipitation (ChIP)-seq

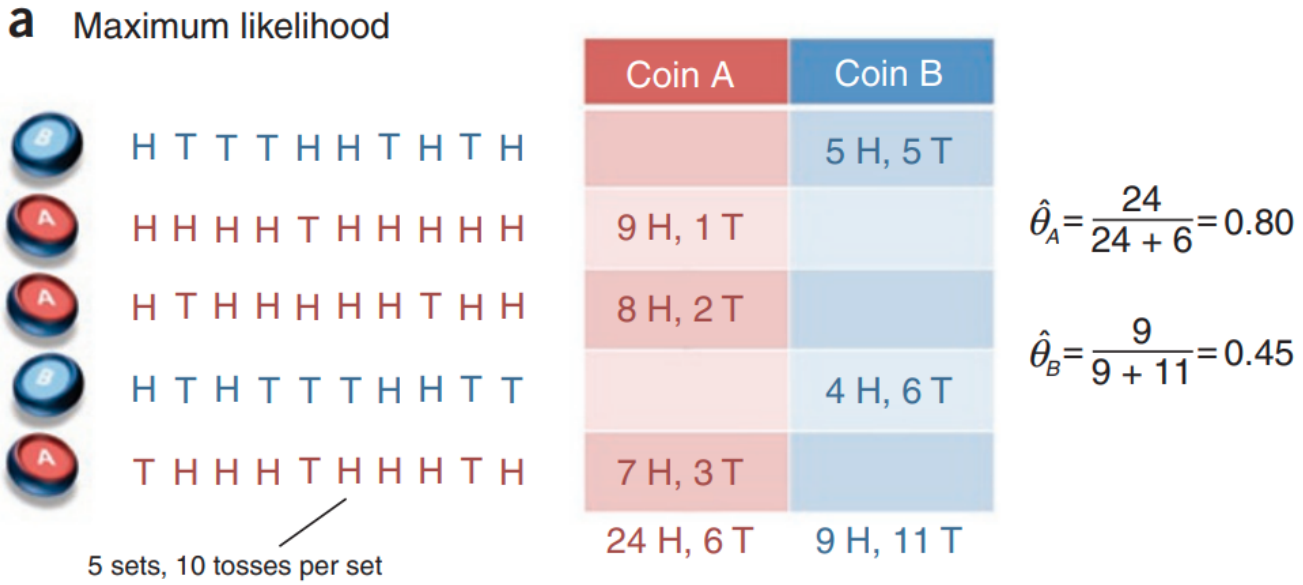



Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

# Maximum Likelihood

Given a pair of coins A and B of unknown biases, $\theta_A$ and $\theta_B$

What is the maximum likelihood estimate (MLE) of $\theta_A$ and $\theta_B$?



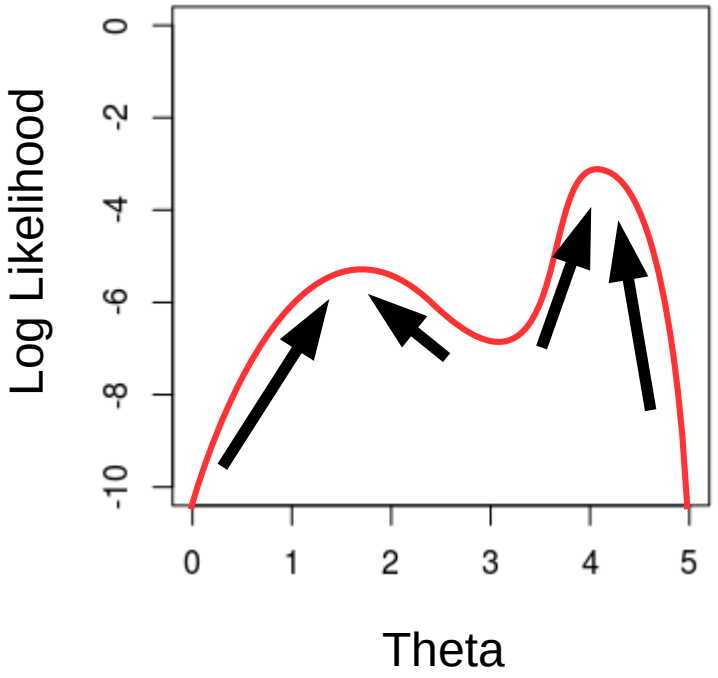

Coin used and toss outcomes are known

What if the coin used is unknown?

Do CB, Batzoglou S. What is the expectation maximization algorithm? Nature biotechnology. 2008 Aug;26(8):897-9.

# Expectation maximization

- used when missing information (parameters) makes ML too difficult
- iterative method of finding maximum likelihood
  - E-Step. Estimate the missing variables in the dataset.
  - M-Step. Maximize the parameters of the model in the presence of the data.
- can be limited to local optimum, no guarantee that the global maximum

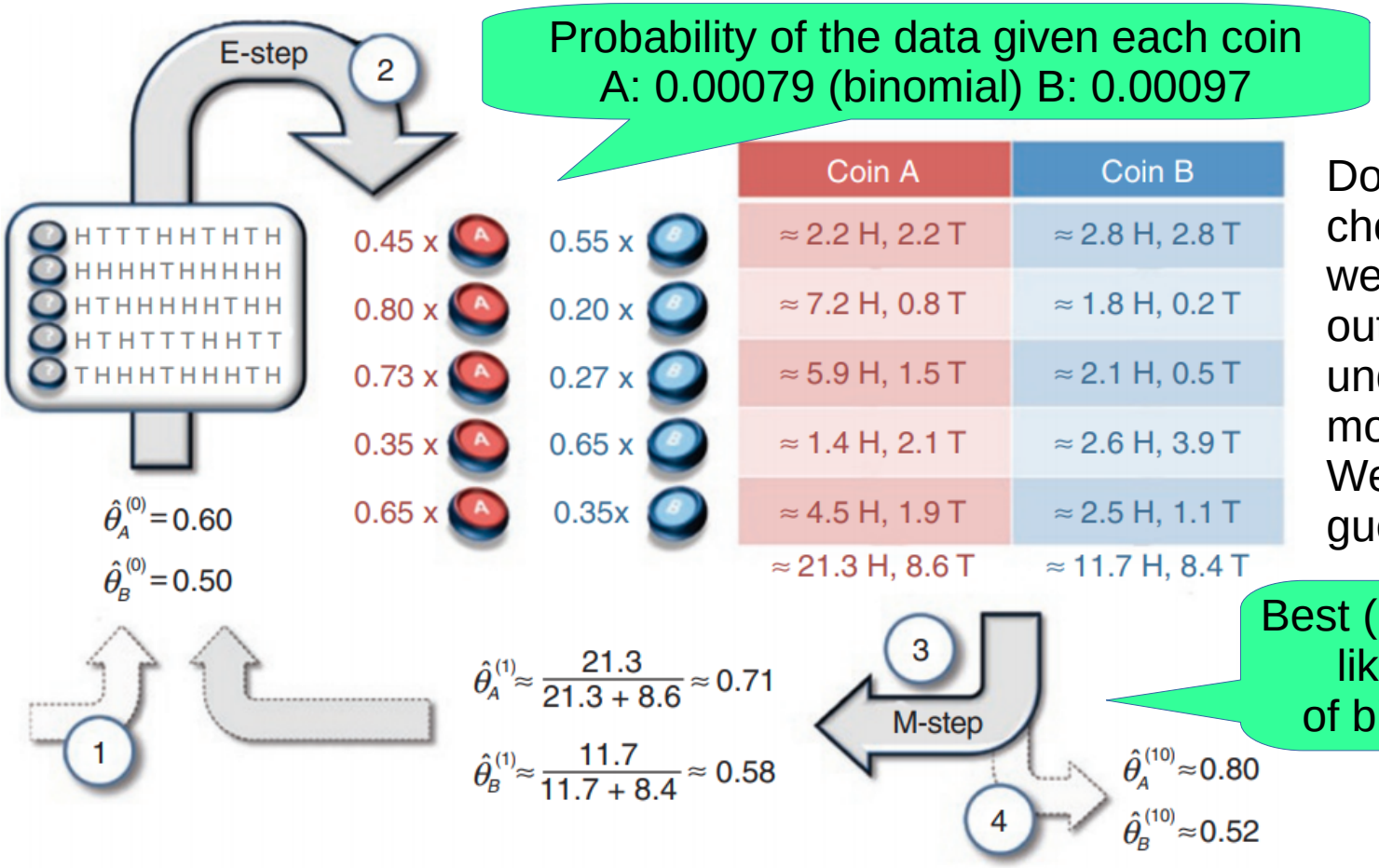


# Expectation Maximization

Given a pair of coins A and B of unknown biases, $\theta_A$ and $\theta_B$, and the coin used is not known:

Coin used is not known; toss outcomes are known

What is the EM estimate of $\theta_A$ and $\theta_B$?




Don't use best choice; use weighted outcomes under each model. Weights=best guess at coin

Do CB, Batzoglou S. What is the expectation maximization algorithm? Nature biotechnology. 2008 Aug;26(8):897-9.

# EM Algorithm

$$L(\theta; X, Z) = P(X, Z|\theta)$$ Likelihood

X = sequences
Z = positions (hidden)
θ = PWM

$$L(\theta; X) = P(X|\theta) = \int P(X, Z|\theta) dZ$$

Marginal likelihood, Z is marginalized, intractable

- Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $Z$ given $X$ under the current estimate of the parameters $\theta$:

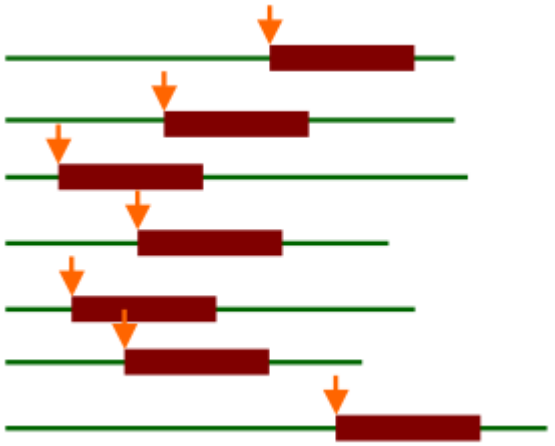
$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}[log(L(\theta; X, Z))]$$

- Maximization step (M step): Find the parameters that maximize this quantity:

$$\theta^{t+1} = argmax\ Q(\theta|\theta^t)$$

Chicken and the egg problem:
If we know Z, we can estimate θ
If we know θ, we can estimate Z

# Motif Finding by Expectation Maximization



Given a set of sequences find the best motif (p) within them

- positions in sequences are unknown (easy to find with p)
- position weight matrix (motif) is unknown (easy to calculate with positions)

$p =$

| | 1 | 2 | 3 |
|---|---|---|---|
| A | 0.1 | 0.5 | 0.2 |
| C | 0.4 | 0.2 | 0.1 |
| G | 0.3 | 0.1 | 0.6 |
| T | 0.2 | 0.2 | 0.1 |

Expectation Maximization

- assume fixed width W
- iteratively
  - update motif position using p
  - update p based on occurrences

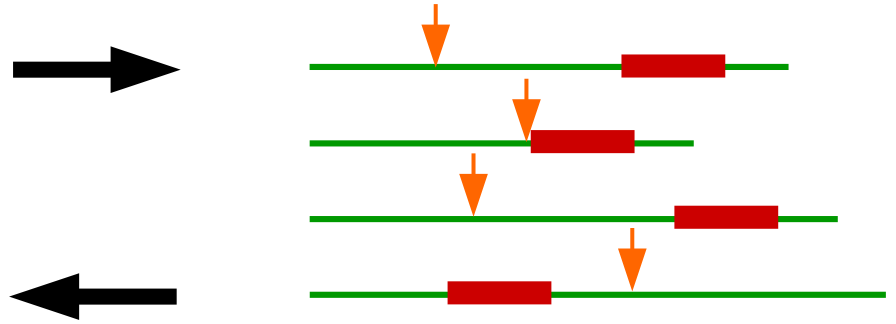
# EM Algorithm

X = sequences
Z = positions (hidden)
p = PWM

1. First, initialize the parameters p to some random values.
2. Compute the probability of each possible value of Z, given p.
3. Then, use the just-computed values of Z to compute a better estimate for the parameters p.
4. Iterate steps 2 and 3 until convergence.

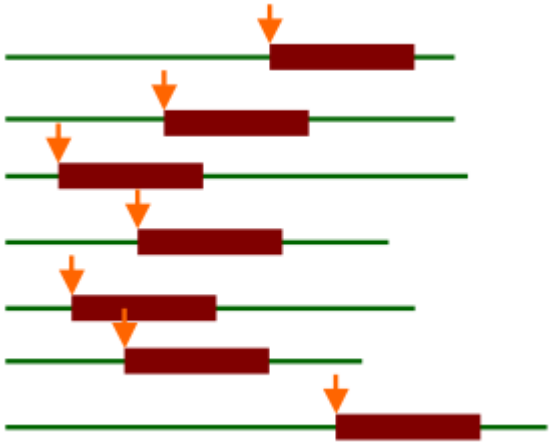
$p =$

| | 1 | 2 | 3 |
|---|---|---|---|
| A | 0.1 | 0.5 | 0.2 |
| C | 0.4 | 0.2 | 0.1 |
| G | 0.3 | 0.1 | 0.6 |
| T | 0.2 | 0.2 | 0.1 |



# Notation

- A motif is represented by a matrix of probabilities: $p_{ck}$ represents the probability of character c in column k
- $p_{c0}$ represents the probability of character c in the background



$$p_0 = \begin{matrix} A & 0.26 \\ C & 0.24 \\ G & 0.23 \\ T & 0.27 \end{matrix}$$

$$p = \begin{matrix} & 1 & 2 & 3 \\ A & 0.1 & 0.5 & 0.2 \\ C & 0.4 & 0.2 & 0.1 \\ G & 0.3 & 0.1 & 0.6 \\ T & 0.2 & 0.2 & 0.1 \end{matrix}$$

What is the probability of a motif at each position?

# Calculating Z

The element $Z_{ij}$ of the matrix Z represents the probability that the motif starts in position *j* in sequence *i*

$$\Pr(X_i \mid Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{after motif}}$$

$X_i$ is the *i*th sequence

$Z_{ij}$ is 1 if motif starts at position *j* in sequence *i*

$c_k$ is the character at position *k* in sequence *i*

# Example of $Z_{ij}$

$$X_i = \texttt{G C } \boxed{\texttt{T G T}} \texttt{ A G}$$

$$p = \begin{array}{cccccc} & & 0 & 1 & 2 & 3 \\ & \texttt{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ & \texttt{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ & \texttt{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ & \texttt{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$\Pr(X_i \mid Z_{i3} = 1, p) =$$

$$p_{\mathrm{G},0} \times p_{\mathrm{C},0} \times p_{\mathrm{T},1} \times p_{\mathrm{G},2} \times p_{\mathrm{T},3} \times p_{\mathrm{A},0} \times p_{\mathrm{G},0} =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

# Calculating $Z_{ij}$

$P(X_i|Z_{ij}=1, p)$ Probability of a sequence given motif (p) starting at position j in the sequence i

$P(Z_{ij}=1|X_i, p)$ Probability of motif at position j in sequence i, given sequence (X) and model (p)

$$P(Z_{ij}=1|X_i, p)=\frac{P(X_i|Z_{ij}=1, p)P(Z_{ij}=1)}{P(X_i)}$$

Bayes rule

$$P(X_i)=\sum_{k=1}^{L-W+1} P(X_i|Z_{ik}=1, p)P(Z_{ik}=1)$$

# Calculating $Z_{ij}$

- to estimate the starting positions in $Z$ at step $t$

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p^{(t)}) \Pr(Z_{ij} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, p^{(t)}) \Pr(Z_{ik} = 1)}$$

Z = probability matrix;
For each sequence (i), and position (j):
What is the probability of motif at a position (j) relative to all other positions (k)

E-step is done: probability of motif at each position

# Example

$X_i =$ G C T G T A G

$p =$

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| A | 0.25 | 0.1 | 0.5 | 0.2 |
| C | 0.25 | 0.4 | 0.2 | 0.1 |
| G | 0.25 | 0.3 | 0.1 | 0.6 |
| T | 0.25 | 0.2 | 0.2 | 0.1 |

$$Z_{i1} = \underline{0.3 \times 0.2 \times 0.1} \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

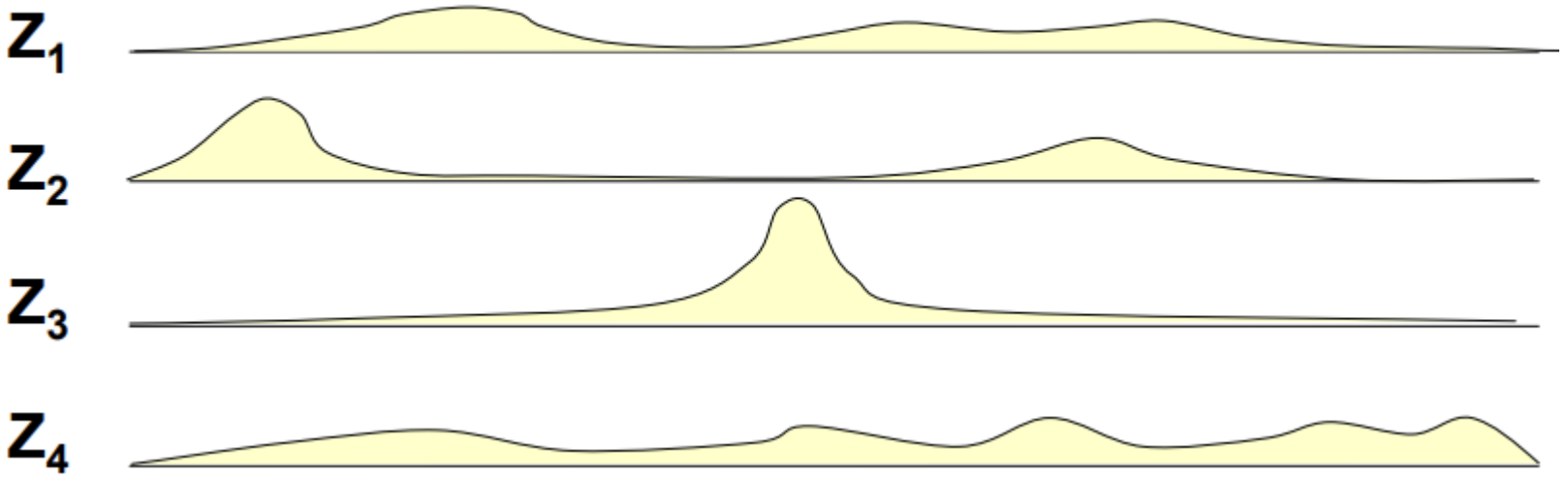
$$Z_{i2} = 0.25 \times \underline{0.4 \times 0.2 \times 0.6} \times 0.25 \times 0.25 \times 0.25$$

$\vdots$

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$

# $Z_{ij}$ over the sequences

**Some examples:**




M-step: estimating p (motif model)

$p_{c,k}$ = probability of character c in position k

$p_{c,0}$ = background nucleotide frequencies

$p_{c,1}$ = base c in first position (k=1)

$$p_{c,1} = \frac{n_{c,1}}{\sum_b n_{b,1}}$$

Pr (base c at first pos)

Pr (any base at first)

# The M-step: estimating p

- recall $p_{c,k}$ represents the probability of character $c$ in position $k$ ; values for position 0 represent the background

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

weighted counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1} = c\}} Z_{ij} & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \end{cases}$$

total # of c's in data set

# Example: estimating p

**A C A G C A**

$$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$$

**A G G C A G**

$$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$$

**T C A G T C**

$$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$$

A: $Z_{1,1}$ $Z_{1,3}$
G: $Z_{1,4}$
C: $Z_{1,2}$
T: NA

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \ldots + Z_{3,3} + Z_{3,4} + 4}$$

A in first position

[A,G,C,T] in first position

# Example: estimating p

**A C A G C A**

$$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$$

**A G G C A G**

$$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$$

**T C A G T C**

$$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$$

$$p_{G,2} = \frac{Z_{1,3} + Z_{2,1} + Z_{2,2} + Z_{3,3}}{Z_{1,1} + Z_{1,2} \ldots Z_{3,3} + Z_{3,4}}$$

# EM algorithm

- Typically converges quickly to a local optimum. There can be multiple optimum (motifs)
- Convergence by number of iterations or change in log-likelihood
- Sensitive to the starting conditions
- Multiple EM for Motif Elicitation (MEME)
  - try many start conditions
  - more than one motif
  - OOPS-one occurrence per sequence
  - ZOOPS-zero or one occurrence / seq
  - Two component mixture (zero or more)

# ZOOPS (zero or one occurrences)

λ prior probability of that a position in a sequence is the start of a motif

γ = (L-W+1)λ prior probability of a sequence containing a motif

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p^{(t)})\lambda^{(t)}}{\Pr(X_i \mid Q_i = 0, p^{(t)})(1-\gamma^{(t)}) + \sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, p^{(t)})\lambda^{(t)}}$$

$$Q_i = \sum_{i=1}^{L-W+1} Z_{i,j}$$

Q is random variable that is 0 to indicate sequence has no motif

$$\lambda^{(t+1)} = \frac{\gamma^{(t+1)}}{(L-W+1)} = \frac{1}{n(L-W+1)} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j}^{(t)}$$

# Gibbs Sampling

- A Markov Chain Monte Carlo (MCMC) algorithm to approximate the joint distribution of variables, or to approximate the marginal distribution of one of the variables. Basic version is Metropolis-Hastings.
- A stochastic version of EM algorithm (deterministic)
- less susceptible to local maxima

Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy to sample from

given: length parameter $W$, training set of sequences
  choose random positions for $a$
  do
    pick a sequence $X_i$
    estimate $p$ given current motif positions $a$ (update step)
      (using all sequences but $X_i$)
    sample a new motif position $a_i$ for $X_i$ (sampling step)
  until convergence
return: $p$, $a$

# Sampling new motif positions

- for each possible starting position, $a_i = j$ , compute a weight

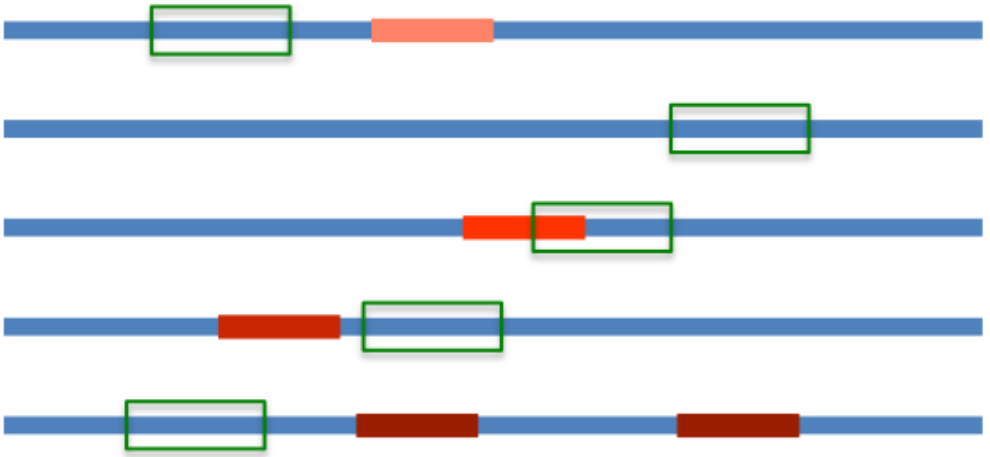
$$A_j = \frac{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k,0}}$$

- randomly select a new starting position $a_i$ according to these weights

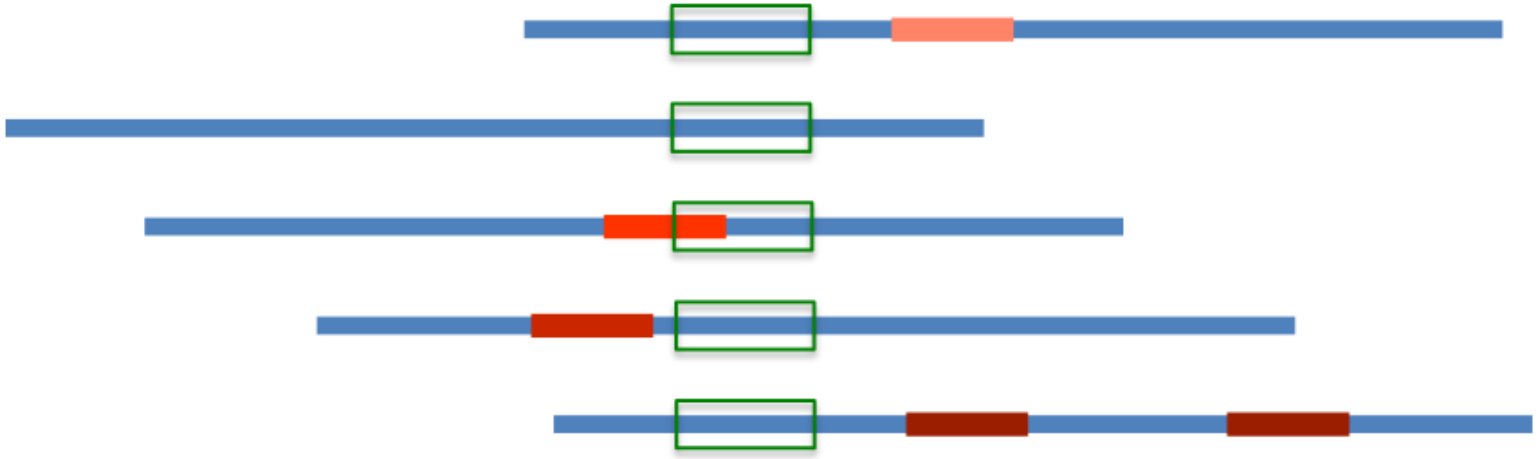
# Gibbs sampling

Initialization:

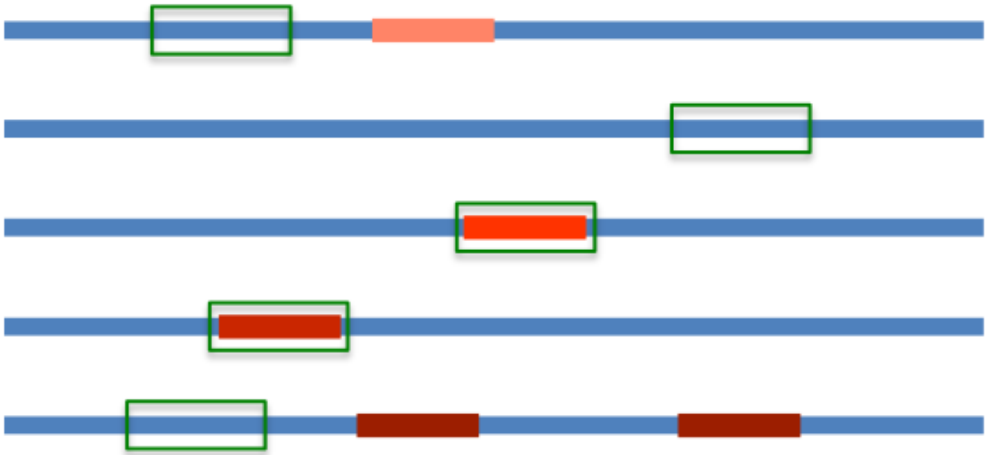
Random assignment of motif locations $a_1$-$a_k$



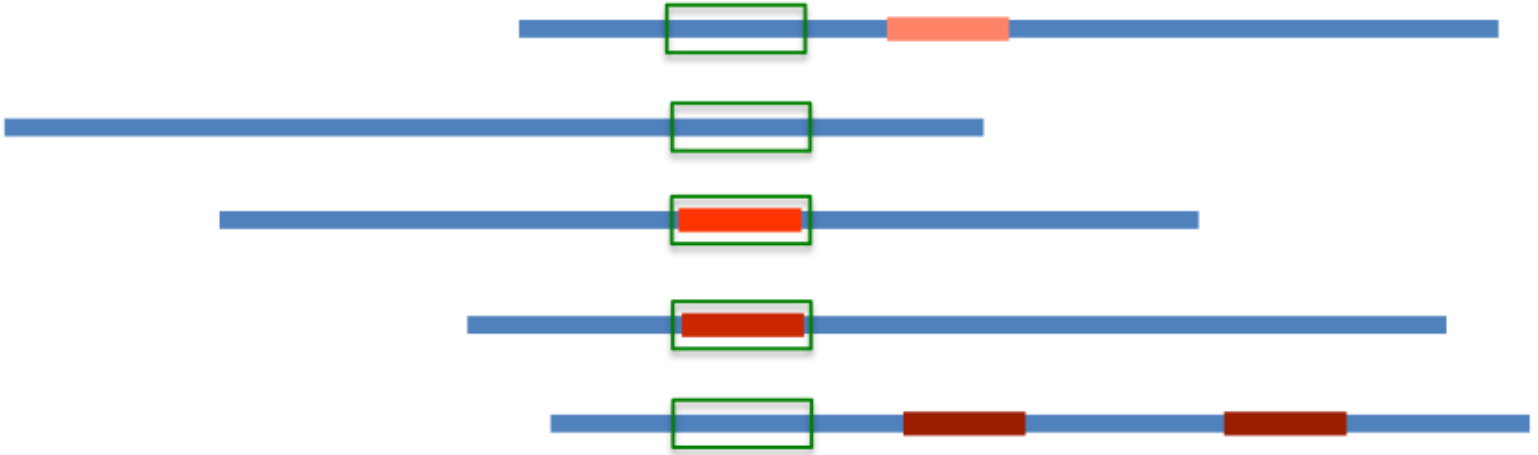
Construct initial matrix S from this alignment



How does it end?  Eventually you nucleate a few correct placements



The matrix has weak but sufficient scoring power



| | Logo | E-value | Sites | Width | More | Submit/Download |
|---|---|---|---|---|---|---|
| 1. |  | 5.7e-015 | 18 | 8 | | |
| 2. |  | 1.4e-009 | 18 | 11 | | |
| 3. |  | 2.0e+001 | 18 | 15 | | |

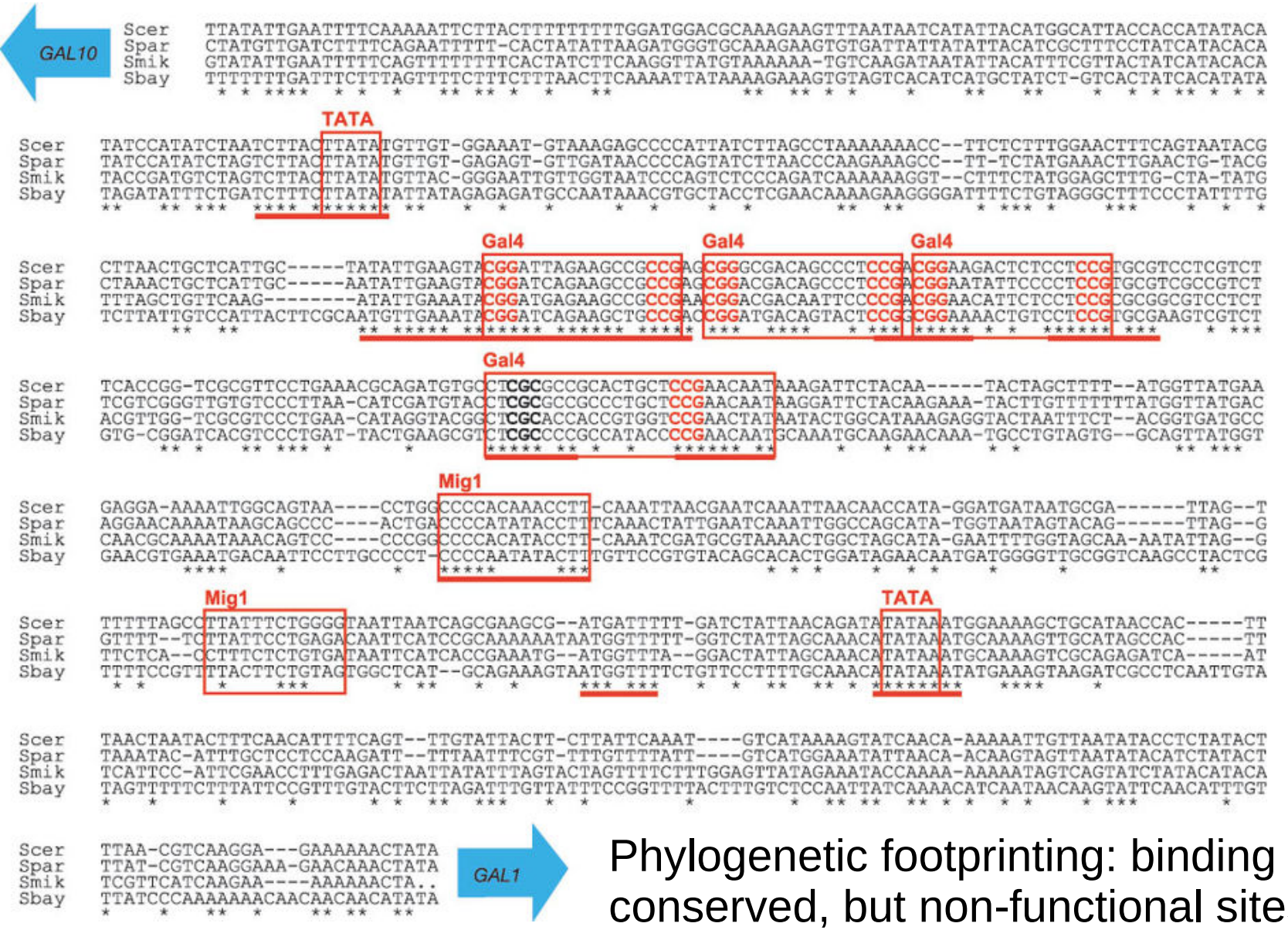
**Stopped because requested number of motifs (3) found.**

## Motif Locations

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences

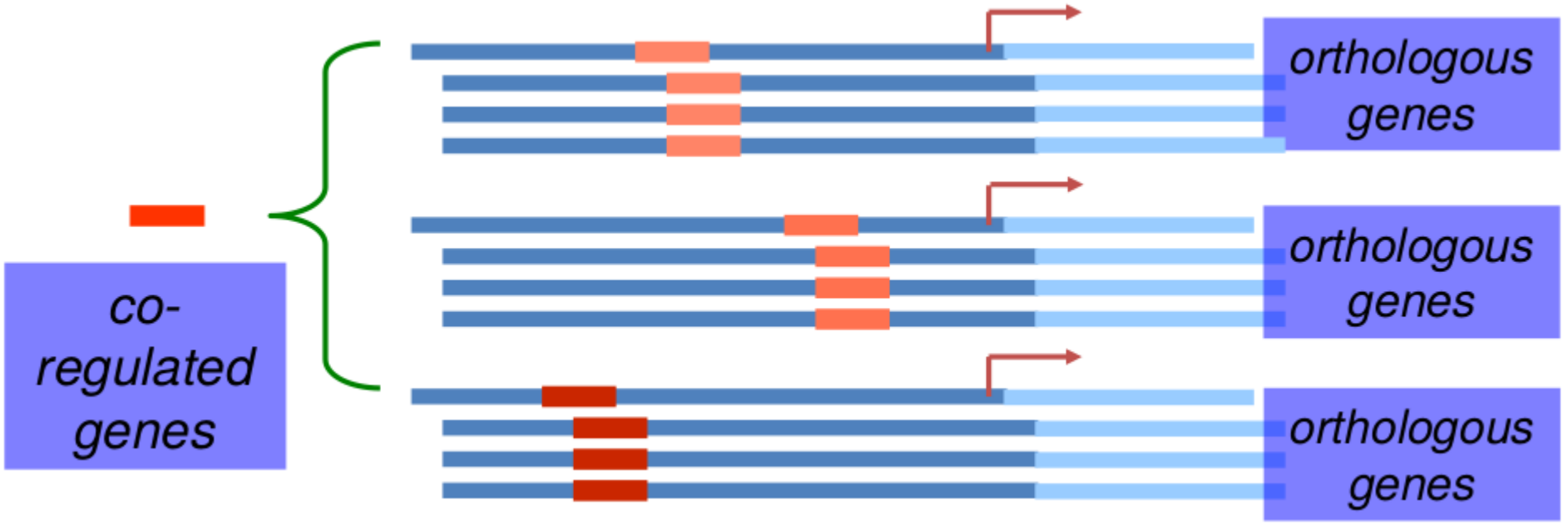
| | Name | *p*-value | Motif Location |
|---|---|---|---|
| 1. | SAM2 | 4.26e-7 | |
| 2. | MET30 | 7.66e-7 | |
| 3. | RAD59 | 4.52e-5 | |
| 4. | MET3 | 4.09e-5 | |
| 5. | MET28 | 5.70e-7 | |
| 6. | BNA3 | 4.06e-7 | |
| 7. | GSH1 | 5.96e-6 | |
| 8. | SER33 | 1.32e-4 | |
| 9. | MET6 | 2.06e-6 | |
| 10. | SUL2 | 1.23e-7 | |
| 11. | ADE3 | 4.17e-4 | |
| 12. | MET2 | 8.23e-7 | |
| 13. | MET1 | 2.58e-6 | |
| 14. | CYS4 | 2.90e-5 | |
| 15. | MET14 | 2.42e-4 | |
| 16. | MET22 | 1.27e-3 | |
| 17. | MET16 | 6.76e-4 | |
| 18. | MET8 | 1.38e-4 | |

# Phylogenetic Footprinting: Conserved Binding Sites




Phylogenetic footprinting: binding sites should be conserved, but non-functional sites will diverge

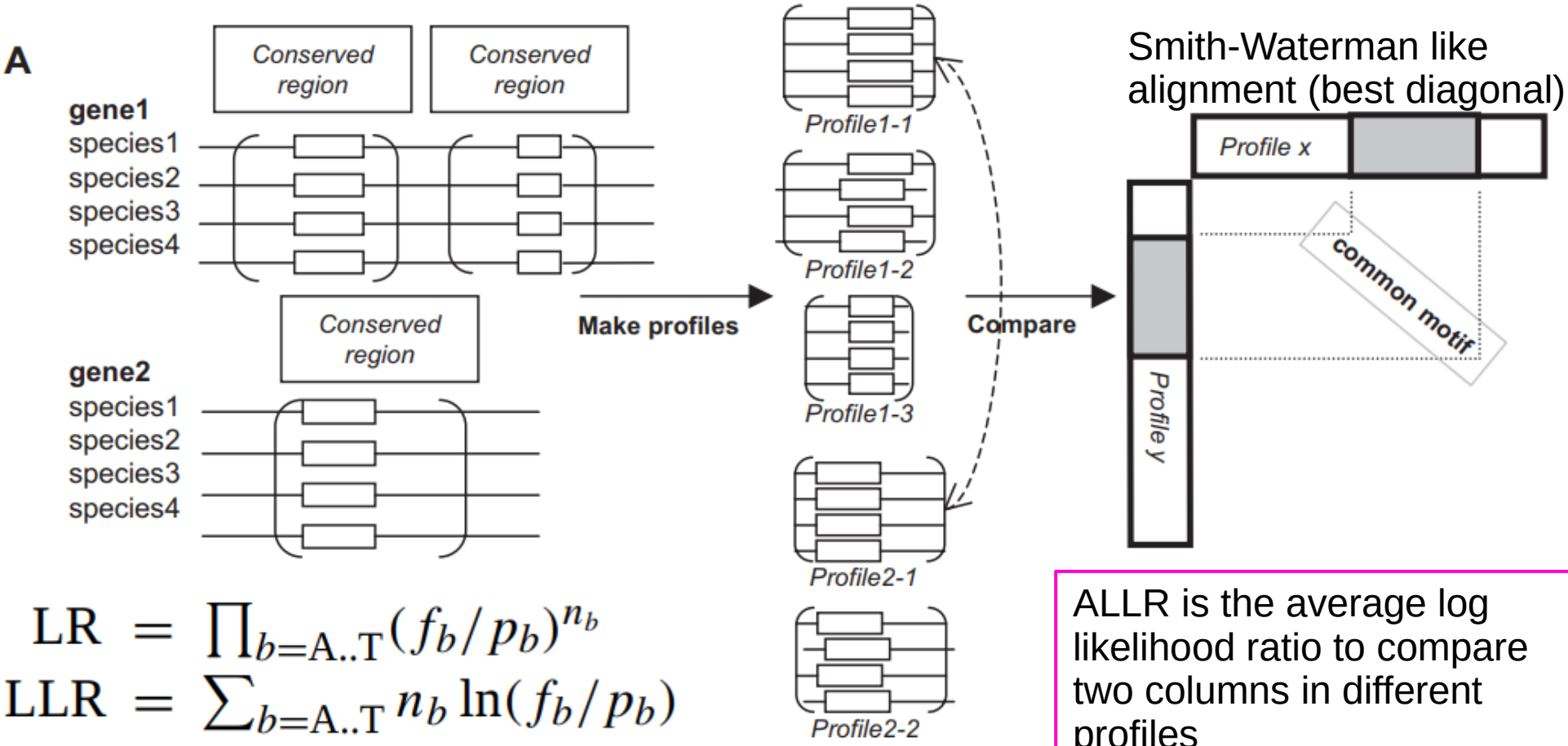
# Phylogenetic motif finding




Phylogenetic motif finding: increase the signal to noise - motif should be conserved across species, but random sequences should not.

Simple idea: motif finding in conserved regions
Implementations: Phylo-MEME, Phylo-Gibbs, PhyloCon

# Profiling algorithm: PhyloCon




$$\mathrm{LR} = \prod_{b=\mathrm{A..T}} (f_b / p_b)^{n_b}$$

$$\mathrm{LLR} = \sum_{b=\mathrm{A..T}} n_b \ln(f_b / p_b)$$

ALLR is the average log likelihood ratio to compare two columns in different profiles

$$\mathrm{ALLR} = \frac{\sum_{b=\mathrm{A..T}} n_{bj} \ln(f_{bi} / p_b) + \sum_{b=\mathrm{A..T}} n_{bi} \ln(f_{bj} / p_b)}{\sum_{b=\mathrm{A..T}} n_{bi} + n_{bj}}$$

i is for profile X and j is for profile X: how well do the counts from j fit profile X, and vice versa

# Models of Binding Site Evolution

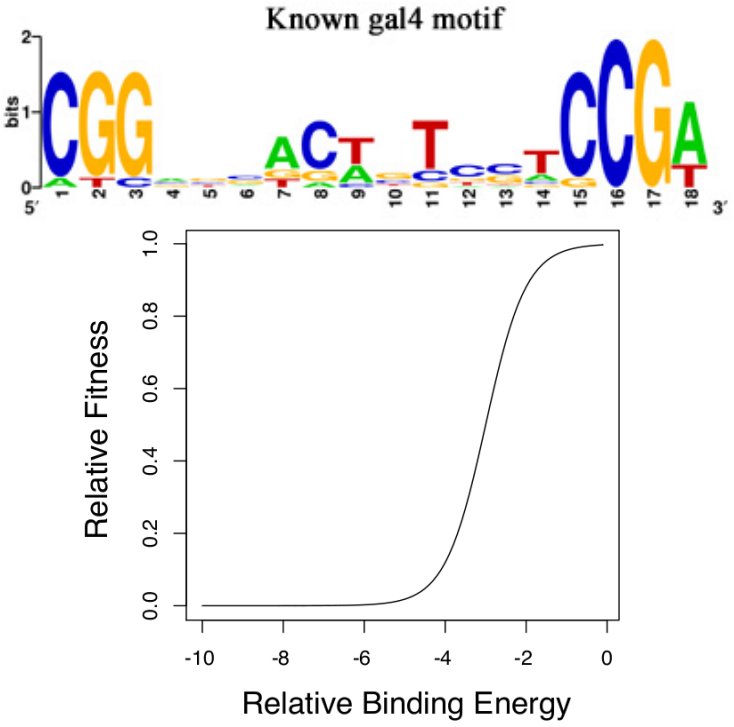
P(sequence | PWM)

AGAATCCGATCGATCGTCGCCGAAGTTA

P(alignment | PWM)

AGAATCCGAACGATCGTCGCCGAAGTTA
AGGAACCGGTCGATCGTCGCCGTAGTGA
GCTATCCGATTGATCGTCGCCGAGGGTA
AGAGACCGATCGATCGTCGCCGAACTTT

# Models of Binding Site Evolution




Kimura 1962 $f_{xy} \cong \frac{2s}{1-e^{-2Ns}}$ and $f_{yx} \cong \frac{-2s}{1-e^{2Ns}}$

Bulmer 1991 $\pi_{x_i}\mu_{xy_i}f_{xy_i} = \pi_{y_i}\mu_{yx_i}f_{yx_i}$

Moses et al. 2004 $f_{xy} = \frac{\ln\left(\frac{\pi_y \mu_{yx}}{\pi_x \mu_{xy}}\right)}{2N\left(1 - \frac{\pi_x \mu_{xy}}{\pi_y \mu_{yx}}\right)}$

$Q_{ij} = \mu_{ij} f_{ij}$

Equilibrium Frequencies

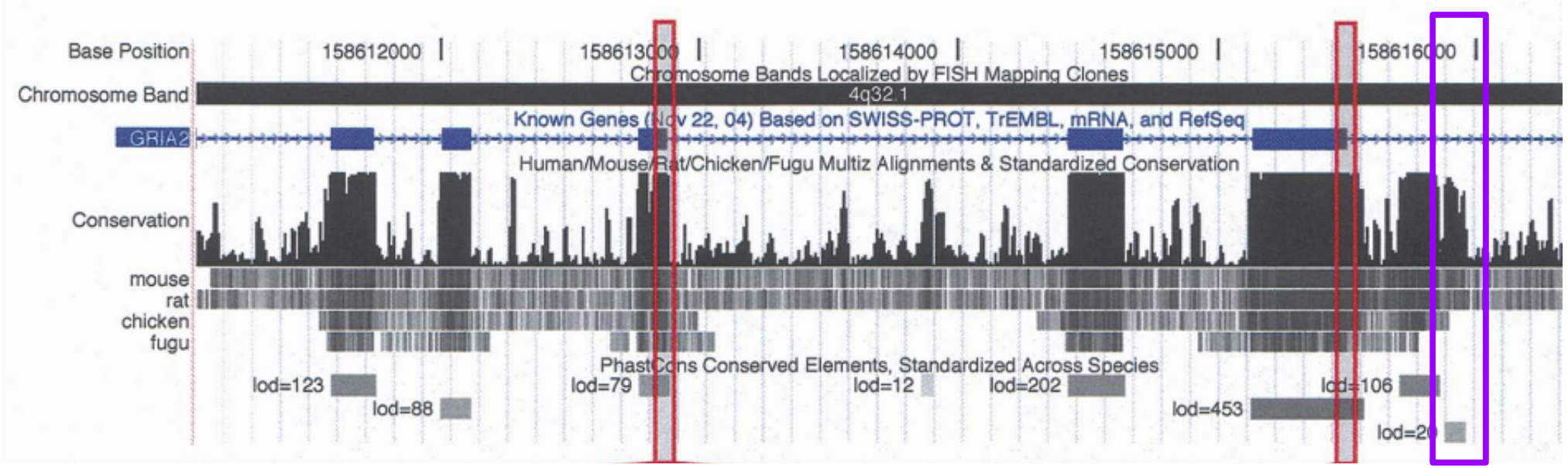
Mutation rates

PPM

$$Q = \begin{pmatrix} * & \pi_G \kappa & \pi_C & \pi_T \\ \pi_A \kappa & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \kappa \\ \pi_A & \pi_G & \pi_C \kappa & * \end{pmatrix}$$

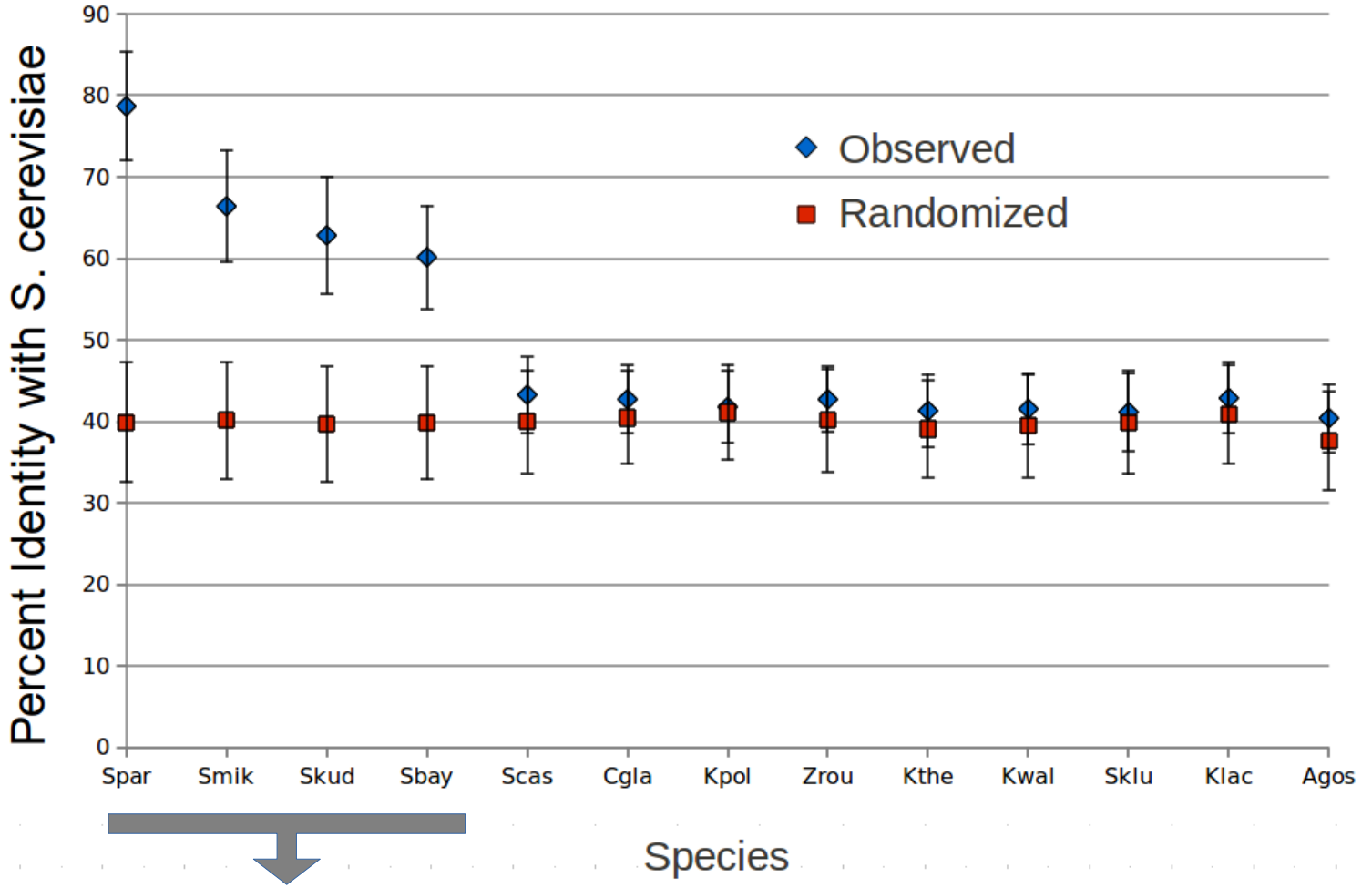
P(alignment | PWM, tree, branchlengths)

# Problem: regulatory sequence divergence w/o change in function




Fugu and Chicken: no alignment in regions that appear to be conserved in human-mouse-rate

# Promoter sequence divergence



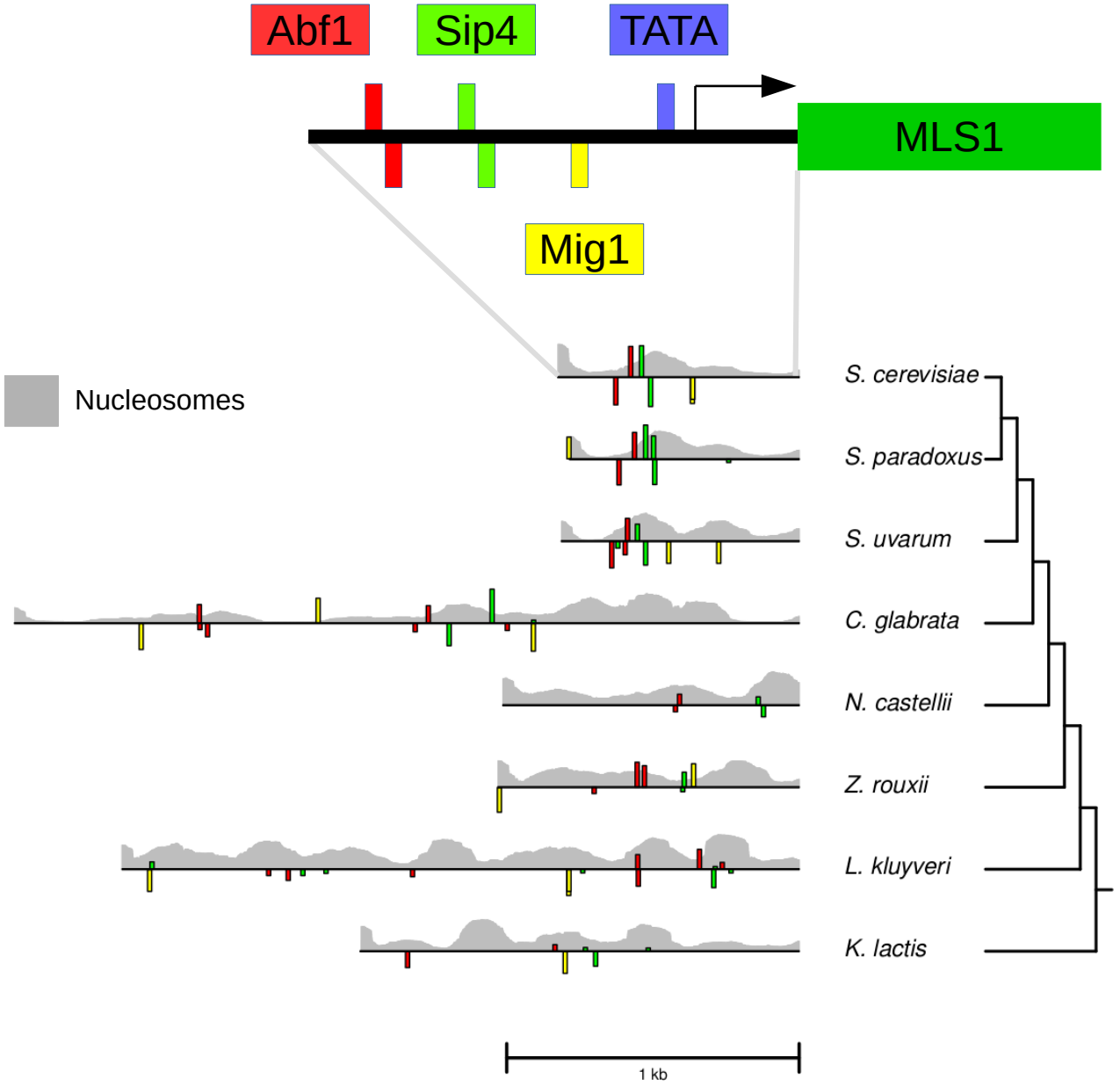

*sensu strictu*
Saccharomyces species

Closely related

Distantly related

Venkataram and Fay (2010)

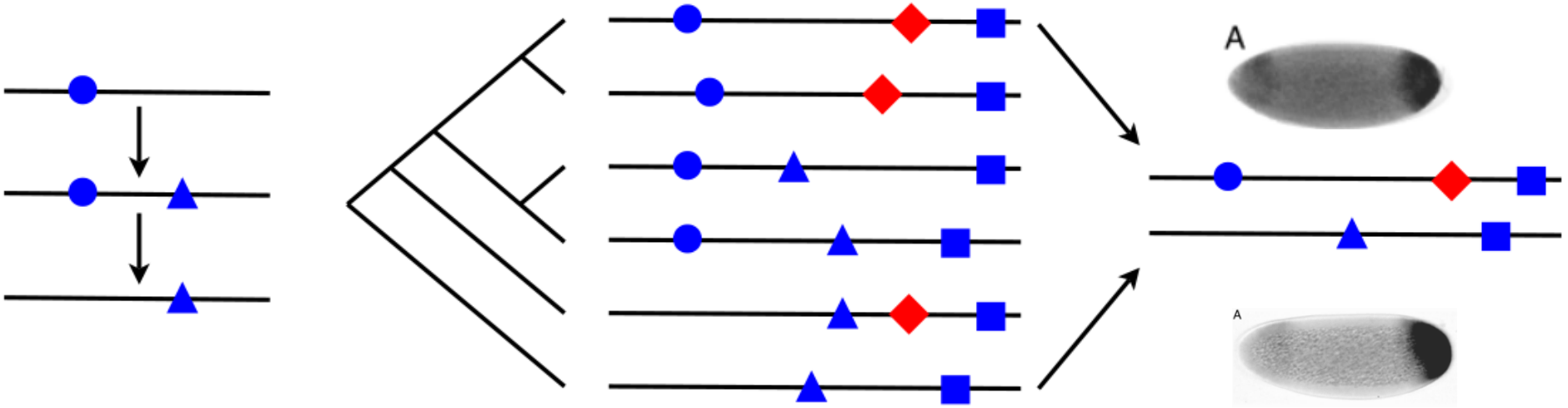
# Binding sites: present but change position & orientation




Bergen et al. (2016)

# Binding Site Turnover Model

Binding site turnover: explains divergence in sequence but conservation of function




Wratten et al. (2006)
M. domestica vs. D. melanogaster

Solution:

- many closely related species, providing power to find conserved motifs
- distantly related species without alignments

# Exercises

1. Calculate $Z_{i3}$ given $X_i$ = AATGCAT. What (position) *i* gives the maximum $Z_i$ value.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| *A* | .3 | 0.2 | 0.1 | 0.85 |
| *C* | .2 | 0.1 | 0.7 | 0.05 |
| *G* | .2 | 0.1 | 0.1 | 0.05 |
| *T* | .3 | 0.6 | 0.1 | 0.05 |

2. What is the probability of T in 3rd position of motif ($p_{T,3}$) in terms of $Z_{ij}$ given $X_i$ = AATGCAT

3. The EM algorithm may not find the best motif model for a collection of sequences. Give two reasons why.
4. Why is random sampling from probabilities for the motif position in Gibb's sampling better than taking the most likely position for the motif?
5. Phylogenetic footprinting: a) requires orthologous genes, b) requires conserved regulatory sequences. Answer T/F for each.
6. Which site is expected to have the slowest substitution rate in the binding site model given in question one.

# Exercises

7) In EM for a motif model. The E step calculates what probability given what?

8) In EM for a motif model. The M step calculates what probability given what?

9) Binding site turnover: explains divergence in sequence without divergence in function [T/F], is less common as divergence time increases [T/F], assumes gene expression/regulation changes between species [T/F], assumes chance gain and loss of redundant sites [T/F]