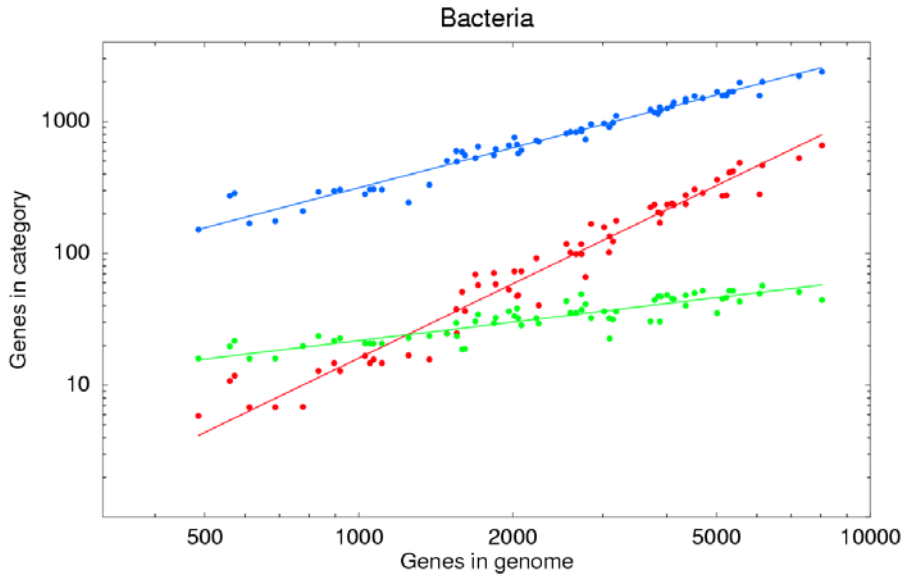
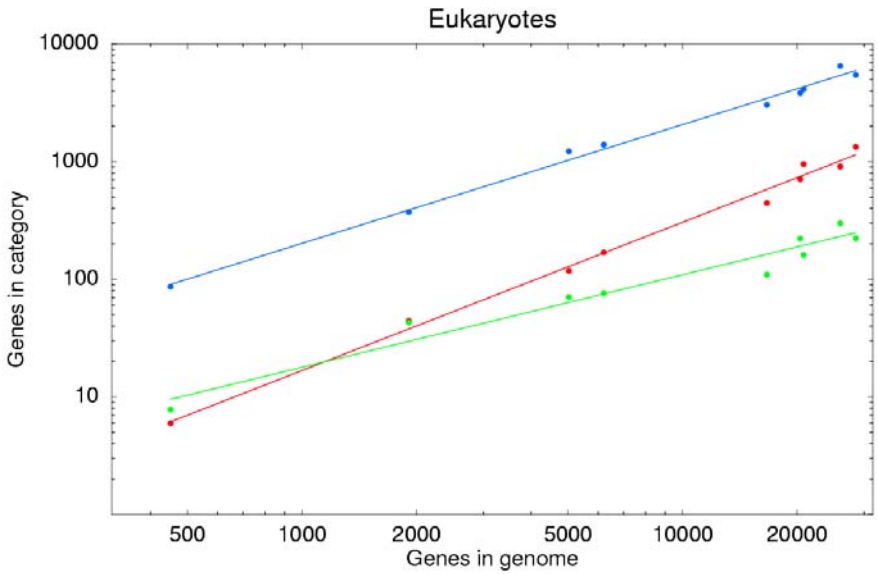
Numbers of genes in different functional categories grow as power-laws in genome size





Estimated Exponents		
Category	**Bacteria**	**Eukaryotes**
Transcription regulation	1.87+/-0.13	1.26+/-0.1
Metabolism	1.01+/-0.06	1.01+/-0.08
Cell cycle	0.47+/-0.08	0.79+/-0.16
Signal transduction	1.72+/-0.18	1.48+/-0.39
DNA repair	0.64+/-0.08	0.83+/-0.31
DNA replication	0.43+/-0.08	0.72+/-0.23
Protein biosynthesis	0.13+/-0.02	0.41+/-0.15
Protein degradation	0.97+/-0.09	0.90+/-0.11
Ion transport	1.42+/-0.28	1.43+/-0.20
Catabolism	0.88+/-0.07	0.92+/-0.08
Carbohydrate metabolism	1.01+/-0.11	1.36+/-0.36
Two-component systems	2.07+/-0.21	NA
Cell communication	1.81+/-0.19	1.58+/-0.34
Defense response	NA	3.35+/-1.41

From: E. van Nimwegen
Trends in Genetics **19** 479-484 (2003)

BIOZENTRUM

SIB

Probabilistic approaches to finding regulatory elements and motifs

- Finding sites for WMs.
- Discovery of regulatory modules.
- Finding motifs (inferring WMs).
- Site and motif finding in phylogenetically related sequences.
- Genome-wide site annotations: some interesting properties.

Erik van Nimwegen

Division of Bioinformatics
Biozentrum, Universität Basel,
Swiss Institute of Bioinformatics

From binding energies to weight matrices

1. We assume the binding energy of a sequence *s* is an additive function of the individual bases:

$$E(s) = \sum_{i=1}^{l} E_i(s_i)$$

2. The probability for the site to be bound is, roughly, a Fermi-function of energy *E*(*s*) and concentration of the transcription factor *c*

$$P_{\text{bound}}(s) = \frac{ce^{\beta E(s)}}{ce^{\beta E(s)} + K}$$

3. Assume that the only constraint on `functional binding sites' is that the need to have some characteristic average energy E. Then we get using *maximum entropy:*

$$P(s) = \frac{e^{\lambda E(s)}}{\sum_{s'} e^{\lambda E(s')}} = \prod_{i=1}^{l} \left[\frac{e^{\lambda E_i(s_i)}}{\sum_{\alpha} e^{\lambda E_i(\alpha)}} \right]$$

where the Lagrange multiplier λ is chosen such that $\sum_s E(s)P(s) = E$

From binding energies to weight matrices

$$P(s) = \frac{e^{\lambda E(s)}}{\sum_{s'} e^{\lambda E(s')}} = \prod_{i=1}^{l} \left[\frac{e^{\lambda E_i(s_i)}}{\sum_{\alpha} e^{\lambda E_i(\alpha)}} \right]$$

This can be rewritten in terms of a weight matrix (WM) *w*.

$$P(s) = \prod_{i=1}^{l} P_i(s_i) \equiv \prod_{i=1}^{l} w^i_{s_i} \qquad w^i_{\alpha} = \frac{e^{\lambda E_i(\alpha)}}{\sum_{\alpha'} e^{\lambda E_i(\alpha')}}$$

The probability that a binding site for the TF will have a sequence *s* given by:

$$P(s|w) = \prod_{i=1}^{l} w^i_{s_i}$$

Note: This is **not** the probability that sequence *s* is a binding site!

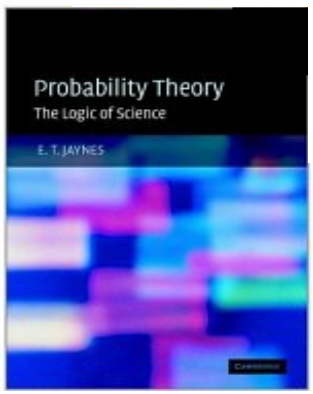
Finding sites for a known WM

Sidebar on the general Bayesian inference approach.

1. Given a dataset *D* enumerate all hypotheses *H* that could have accounted for the data.
2. Assign *prior probabilities* to each hypothesis *H.*
3. Define a likelihood model that gives the probability *P*(*D*|*H*) of obtaining the data *D* under each of the hypotheses *H*.
4. The posterior probability *P*(*H*|*D*) that hypothesis *H* produced the data is given by **Bayes' theorem**:

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_{\tilde{H}} P(D|\tilde{H})P(\tilde{H})}$$

The most useful book I ever read:

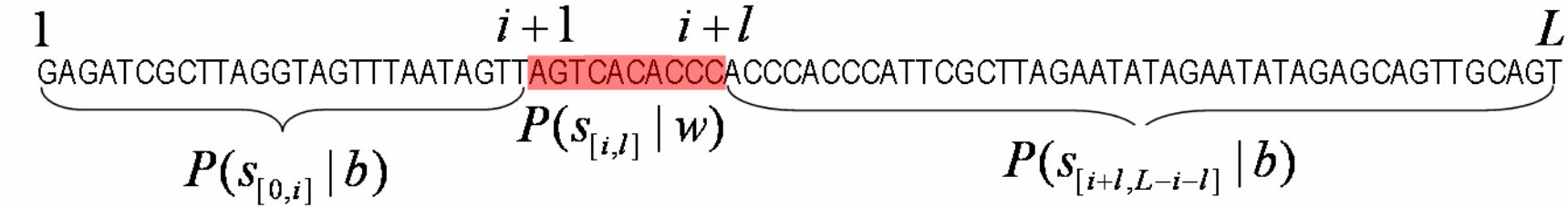


E.T. Jaynes: Probability Theory, the Logic of Science

Finding sites for a known WM

Simplest example: We are given a sequence *s* of length *L* that is known to contain a single site for the TF represented by *w*. **Task**: find where the site is.

- **Hypotheses:** Positions *i* at which the site could start. $0 \le i \le L - l$ **Prior:** $P(i) = \frac{1}{L - l + 1}$
- **Likelihood:** Combination of the probability of the site and the other bases under a `background' model.



Site:
$$P(s_{[i,l]}|w) = \prod_{k=1}^{l} w^k_{s_{i+k}}$$

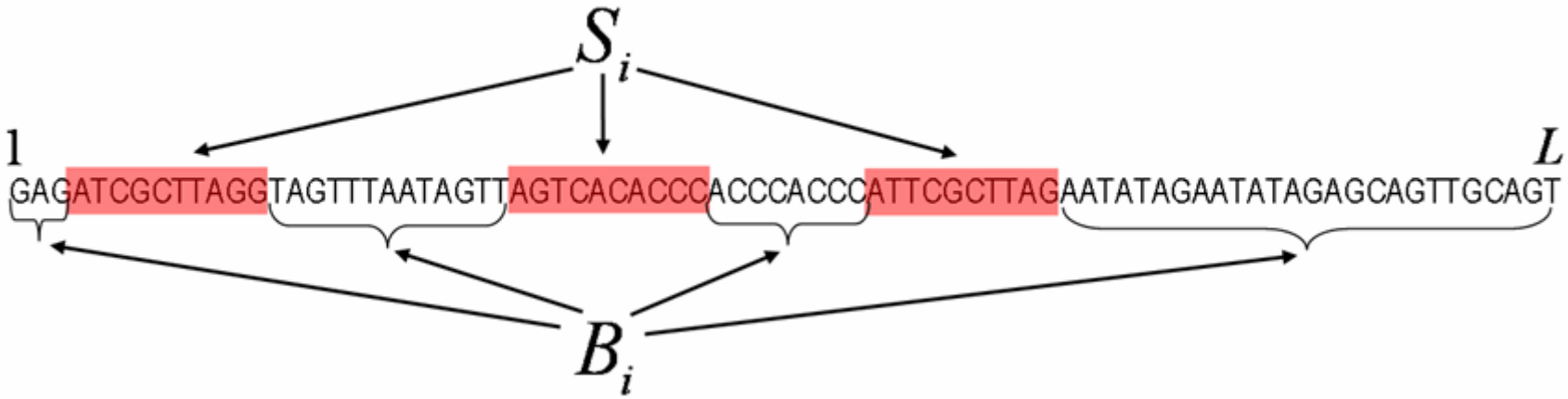
Left background:
$$P(s_{[0,i]}|b) = \prod_{k=1}^{i} b_{s_k}$$

Right background:
$$P(s_{[i+l,L-i-l]}|b) = \prod_{k=i+l+1}^{L} b_{s_k}$$

Posterior that site occurs at position *i*
$$P(i|D) = \frac{P(D|i)}{\sum_{j=0}^{L-l} P(D|j)}$$

Finding multiple sites for a known WM

Arbitrary site configurations: $i = (i_1, i_2, \cdots, i_n)$



Likelihood:

$$P(D|i) = \left[\prod_{\sigma \in B_i} b_\sigma\right] \prod_{s \in S_i} P(s|w)$$

Prior: Assume there is a constant probability π per position for a site to occur.

$$P(i) \propto \pi^{n(i)}(1-\pi)^{L-ln(i)}$$

Posterior:

$$P(i|D) = \frac{P(D|i)\pi^{n(i)}(1-\pi)^{L-ln(i)}}{\sum_j P(D|j)\pi^{n(j)}(1-\pi)^{L-ln(j)}}$$

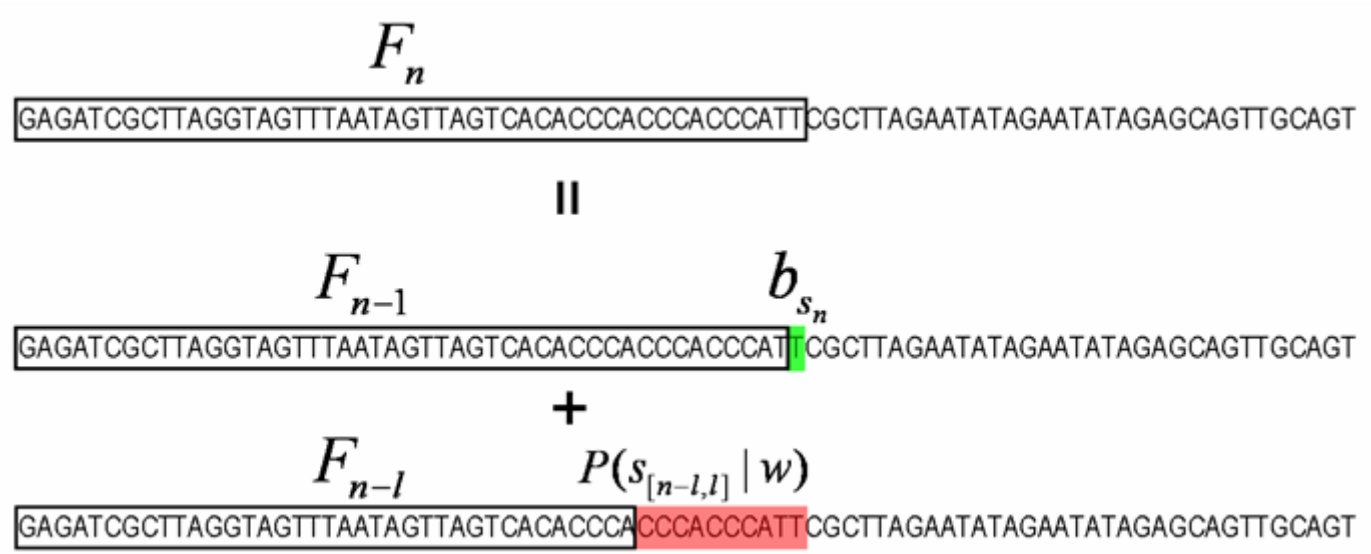
Note: In the denominator we have to sum over all possible binding site configurations.

Finding multiple sites for a known WM

Partition sum: $P(D) = \sum_j P(D \mid j) P(j)$

$F_n =$ Partition sum up to position *n* in the sequence.

Recursion relation: $F_n = F_{n-1}(1-\pi) b_{s_n} + F_{n-l} \pi P(s_{[n-l,l]} | w)$



Similarly for the partition sum from position n to the end:

$$R_n = b_{s_n}(1-\pi) R_{n+1} + P(s_{[n-1,l]} | w) \pi R_{n+l}$$

Finding multiple sites for a known WM

Partition sum: $P(D) = F_L = R_1$

Probability to find a site at position *n* independent of everything else.
{*n*} = Set of all configurations that have a site at position *n.*

$$P(\{n\} \mid D) = \sum_{i \in \{n\}} P(i \mid D) = \frac{F_n P(s_{[n,l]} \mid w) \pi R_{n+l+1}}{F_L}$$

Expected total number of sites:

$$\langle n \rangle = \sum_{n=0}^{L-l} P(\{n\} \mid D)$$

Summary:
Given a sequence *s,* a WM *w*, and a prior *π* we can determine, in linear time, the probability for the site to occur at any given position, and the expected total number of sites in the sequence.

Note: In complete analogy one could for a given energy matrix, and a given concentration of the TF, calculate the partition sum, the fraction of time the TF is bound at each position, and the expected number of TFs bound.

Optimizing the prior

What if we do not know the prior *π*?

Ideally we would write the probability as an explicit function of *π:*

$$P(D) = P(D \mid \pi) \text{ and } P(\pi \mid D) = \frac{P(D \mid \pi)P(\pi)}{\int_0^1 P(D \mid \pi)P(\pi)d\pi}$$

Unfortunately the integral can not be easily calculated.

Instead, we can optimize with respect to *π*:

$$\frac{d\log(P(D))}{d\pi} = \sum_i \frac{P(D \mid i)}{P(D)} \frac{dP(i)}{d\pi} = \sum_i \frac{P(D \mid i)}{P(D)} \frac{d\left[\pi^{n(i)}(1-\pi)^{L-n(i)l}\right]}{d\pi} =$$

$$\sum_i \frac{P(D \mid i)}{P(D)} P(i)\left[\frac{n(i)}{\pi} - \frac{L-n(i)l}{1-\pi}\right] = \frac{\langle n \rangle}{\pi} - \frac{L-\langle n \rangle l}{1-\pi}$$

Optimal value of *π obeys* :

$$\pi = \frac{\langle n \rangle}{\langle n \rangle + \left(L - \langle n \rangle l\right)}$$

Expectation-Maximization procedure:

1. Start with some value of *π.*
2. Calculate <n>.
3. Set *π* according to the equation on the left.
4. Go to step 2.

BIOZENTRUM

SIB

Finding multiple sites for multiple WMs

All calculations described so far easily generalize to an arbitrary number of WMs.

Let π_w denote the prior for WM *w* (for simplicity we can consider the background one of the WMs which happens to have length 1).

Recursion relation: $$F_n = \sum_w F_{n-l_w} \pi_w P(s_{[n-l_w, l_w]} | w)$$

Equations determining the optimum prior:

$$\text{constant} = \frac{d \log(Z)}{d\pi_w} = \frac{d \log(F_L)}{d\pi_w} = \frac{\langle n(w) \rangle}{\pi_w}$$

Update equation for the Expectation Maximization:

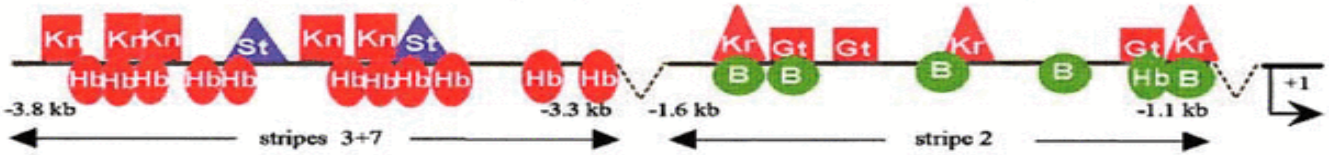
$$\pi_w = \frac{\langle n(w) \rangle}{\sum_{\tilde{w}} \langle n(\tilde{w}) \rangle}$$

Discovering regulatory modules

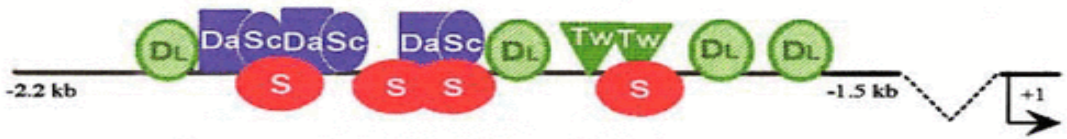
Drosophila

(*G*) *ftz* zebra-stripe element

(*H*) *eve* stripe 3+7 and 2 elements



(*I*) *rho* lateral neurectoderm stripe element



(*J*) *kni* posterior element

(*K*) PBX element of ubx gene

(from Arnone, M. I. and Davidson, E. H., *Development*, **124**(10):1851-64, 1997.)

Discovering regulatory modules

- Gather sets of TFs for which binding sites are known and that (ideally) are also known to interact in regulatory modules.
- Search genome-wide for relatively short sequence segments, i.e. 200-500 bp, that have a surprisingly high concentration of sequences that 'match' these WMs.

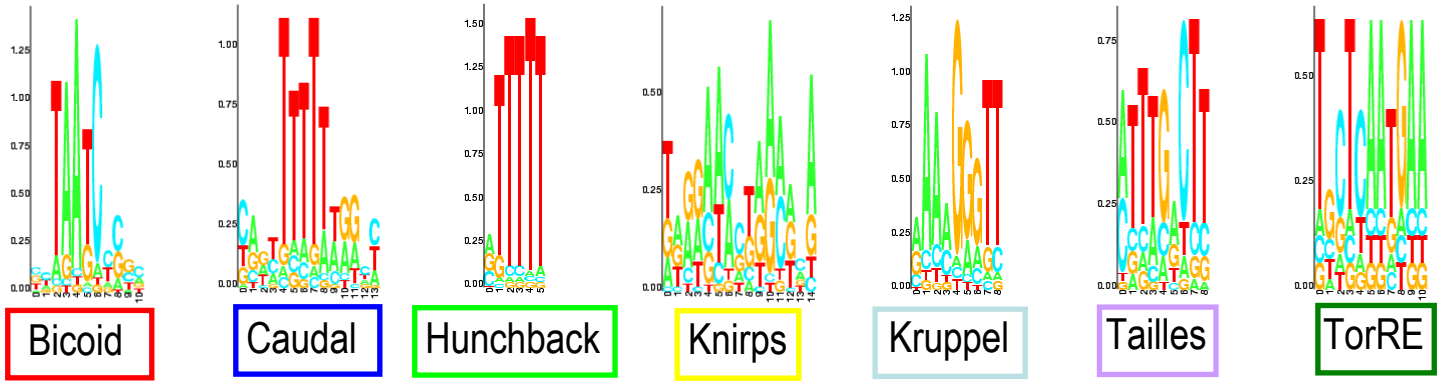
  - Berman et al., *PNAS* (2002) **99** 757-762
    (*first publication presenting the idea using simple filtering methods)*

  - **Ahab**: Rajewsky N, Vergassola M, Gaul U, Siggia ED, *BMC Bioinformatics* (2002) **3** 30
    (*presentation of algorithm with applications to Fly*)

  - **Smash**: Zavolan M, Rajewsky N, Socci N, Gaasterland T, *ICSSB* (2003)
    (*essentially same algorithm with applications to human/mouse*)

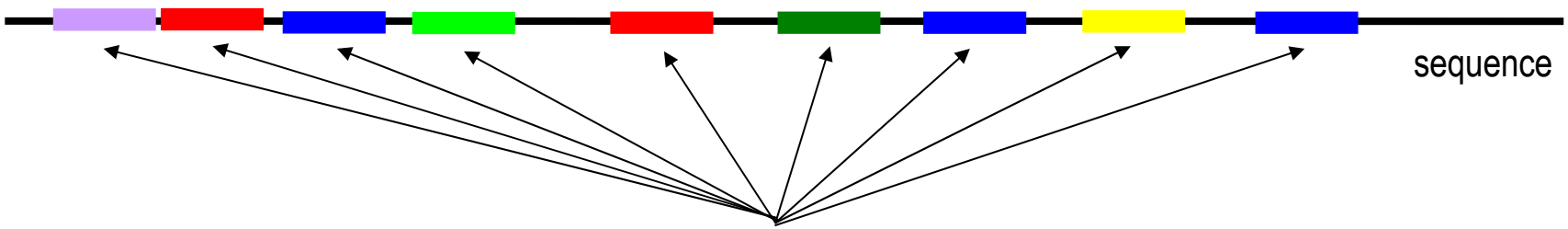
  - **Stubb**: Sinha S, van Nimwegen E, Siggia ED, *ISMB* (2003)
    (*extensions taking multiple species and correlations in neighboring sites into account***.**)

Discovering regulatory modules



Set of Weight Matrices for the gap gene transcription factors known to be involved in early body-patterning in fly.

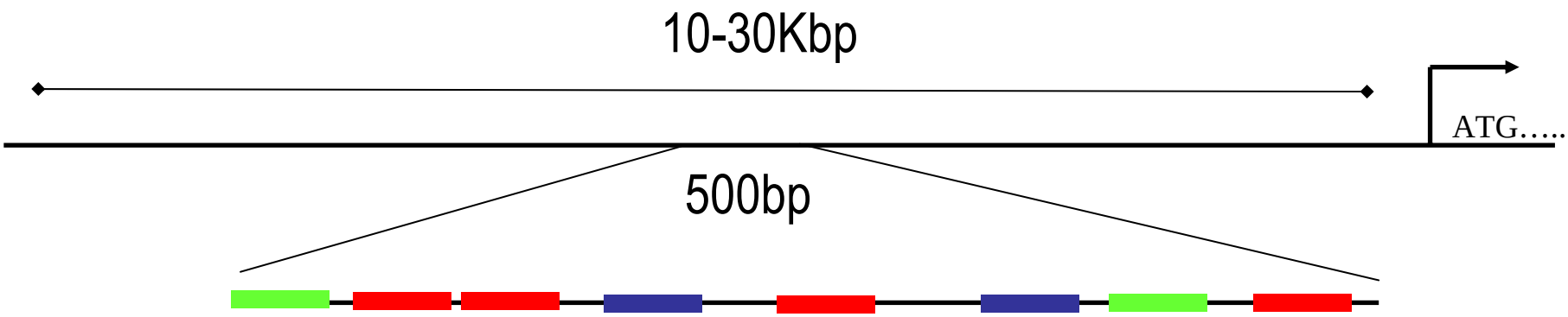
Given a sequence we want to consider all ways in which the sequence can be “parsed” into binding sites for the set of TFs and assign probabilities.



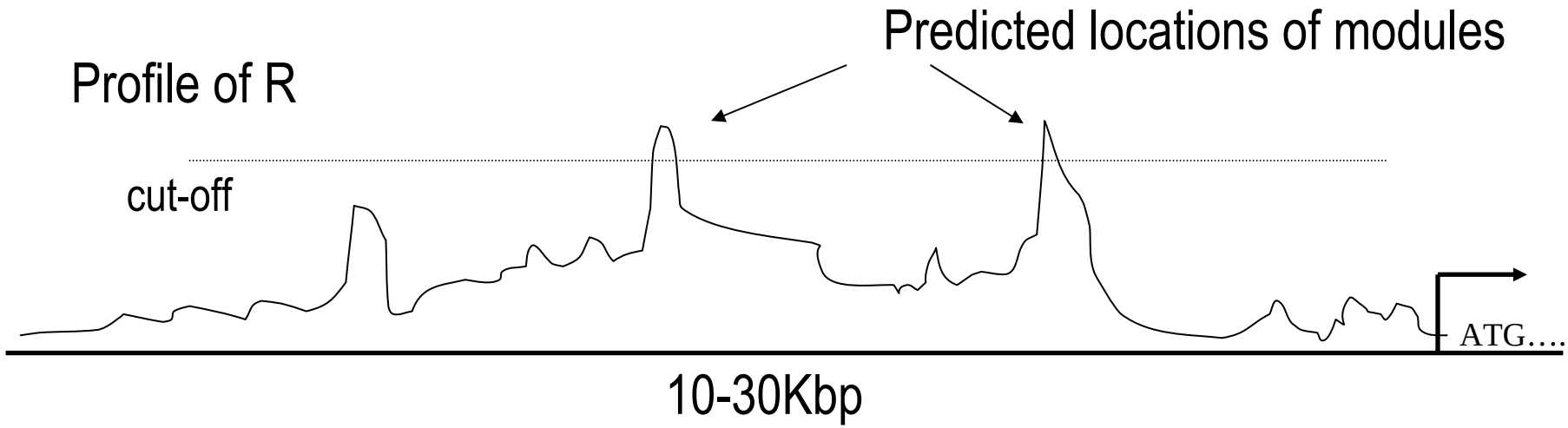
A `configuration’ ρ of the sequence S in terms of hypothesized binding sites.

$P(S \mid \rho)$ Probability of the observed sequence given the parse.

Discovering regulatory modules



- Slide a window of length 500 over the sequence and calculate $R = \frac{P(s \mid \{w\})}{P(s \mid \mathrm{bg})}$ at each position. This gives a profile of *R*.



- A predicted module occurs at every window where *R* exceeds a cut-off.

Discovering regulatory modules

Open access, freely available online PLOS BIOLOGY

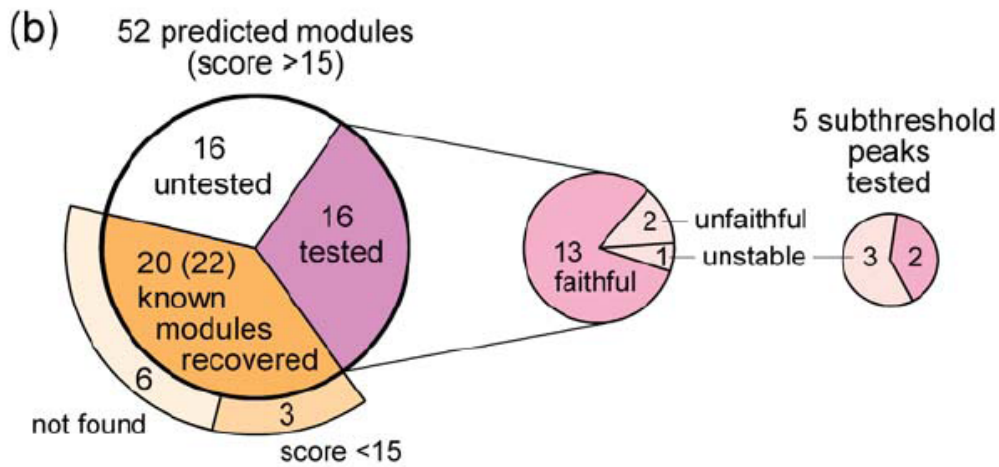
Transcriptional Control in the Segmentation Gene Network of *Drosophila*

Mark D. Schroeder[1☯], Michael Pearce[1☯], John Fak[1☯], HongQing Fan[1], Ulrich Unnerstall[1], Eldon Emberly[2], Nikolaus Rajewsky[2¤], Eric D. Siggia[2], Ulrike Gaul[1*]

1 Laboratory of Developmental Neurogenetics, Rockefeller University, New York, New York, United States of America, 2 Center for Studies in Physics and Biology, Rockefeller University, New York, New York, United States of America

Citation: Schroeder MD, Pearce M, Fak J, Fan HQ, Unnerstall U, et al. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2(9): e271.

- Ahab given a set of 9 transcription factors: Bcd, Hb, Cad, TorRE, D-stat, Kr, Kni, Gt, Tll
- Run on upstream regions of 29 genes with gap and pair-rule patterns (750,000bp total).



Inferring a WM from a set of sites

Alignment of known **fruR** binding sites:

```
AAGCTGAATCGATTTTATGATTTGGT
AGGCTGAATCGTTTCAATTCAGCAAG
CTGCTGAATTGATTCAGGTCAGGCCA
GTGCTGAAACCATTCAAGAGTCAATT
GTGGTGAATCGATACTTTACCGGTTG
CGACTGAAACGCTTCAGCTAGGATAA
TGACTGAAACGTTTTTGCCCTATGAG
TTCTTGAAACGTTTCAGCGCGATCTT
ACGGTGAATCGTTCAAGCAAATATAT
GCACTGAATCGGTTAACTGTCCAGTC
ATCGTTAAGCGATTCAGCACCTTACC
```

} *S*

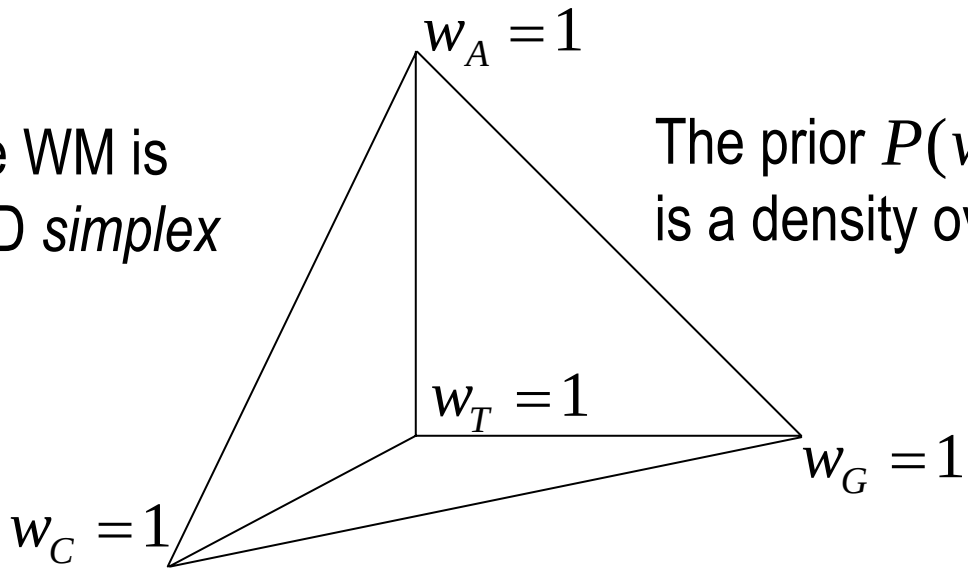
Probability of the sequences *S* given a WM *w:*

$$P(S \mid w) = \prod_{s \in S} P(s \mid w) = \prod_{s \in S} \left[\prod_{i=1}^{l} w_{s_i}^{i} \right] = \prod_{i=1}^{l} \left[\prod_{\alpha} \left(w_{\alpha}^{i} \right)^{n_{\alpha}^{i}} \right]$$

To calculate P(*w*|*S*) we need a prior P(*w*).

Prior over WMs

A column of the WM is a point in the 4D *simplex*



The prior $P(w_A, w_C, w_G, w_T)$ is a density over the simplex

The family of *Dirichlet* priors:

$$P(w)dw = \frac{\Gamma(4\gamma)}{[\Gamma(\gamma)]^4} \prod_{\alpha} (w_\alpha)^{\gamma-1} dw$$

- The case $\gamma = 1$ corresponds to the uniform prior.
- For $\gamma < 1$ more weight is on the corners and edges of the simplex, i.e. one expects a distribution heavily biased to one or two bases.
- For $\gamma > 1$ more weight is on the middle of the simplex, i.e. one expects all bases to have equal probabilities (unlikely to apply in practice).

Inferring a WM from a set of sites

Likelihood time prior:

$$P(S \mid w)P(w) = \prod_{i=1}^{l}\left[\Gamma(4\gamma)\prod_{\alpha}\frac{\left(w_{\alpha}^{i}\right)^{n_{\alpha}^{i}}}{\Gamma(\gamma)}\right]$$

Posterior:

$$P(w \mid S) = \frac{P(S \mid w)P(w)}{P(S)}$$

The normalization constant is:

$$P(S) = \int P(S \mid w)P(w)dw = \prod_{i=1}^{l}\left[\frac{\Gamma(4\gamma)}{\Gamma(n+4\gamma)}\prod_{\alpha}\frac{\Gamma(n_{\alpha}^{i}+\gamma)}{\Gamma(\gamma)}\right]$$

Note: This is the probability all sequences in *S* come from one WM, irrespective of what that WM is.

What is the probability that a new sequence *s* derives from the *same* matrix as the sites *S*?

$$P(s \mid S) = \int P(s \mid w)P(w \mid S)dw = \frac{P(s,S)}{P(S)} = \prod_{i=1}^{l}\frac{n_{s_i}^{i}+\gamma}{n+4\gamma} \equiv \prod_{i=1}^{l}\tilde{w}_{\alpha}^{i}$$

Note: The probability P(*s*|*S*) is like P(*s*|*w*) for a given WM, if we set the WM entries as above.

Probability sequences deriving from same WM

- With the Dirichlet prior the integral becomes:

$$P(S) = \prod_{i=1}^{l}\left[\frac{\Gamma(4\gamma)}{[\Gamma(\gamma)]^4}\int\prod_{\alpha}(w_\alpha)^{n_\alpha^i+\gamma-1}dw\right] = \prod_{i=1}^{l}\left[\frac{\Gamma(4\gamma)}{\Gamma(n+4\gamma)}\prod_{\alpha}\frac{\Gamma(n_\alpha^i+\gamma)}{\Gamma(\gamma)}\right]$$

- This gives the probability to obtain the sequences *S* from a common WM.
- Now we can calculate all kinds of things like:

1. Given a set *S* and one more sequence *s*, what is the probability that *s* comes from the same WM as the WM that the sequences in *S* came from.
2. Given two sets S_1 and S_2 what is the probability to obtain them both from a single WM versus the probability to obtain them when sampling from two independent WMs.

Probability sequences deriving from same WM

Generalization to two sets of sequences S_1 and S_2 .

$$n_\alpha^i = \text{number of times } \alpha \text{ occurs at position } i \text{ in } S_1$$

$$m_\alpha^i = \text{number of times } \alpha \text{ occurs at position } i \text{ in } S_2$$

Again assuming a prior π for the probability that S_1 and S_2 come from the same WM the posterior that S_1 and S_2 come from the same WM becomes:

$$P(S_1 \text{ same WM as } S_2) = \frac{P(S_1,S_2)\pi}{P(S_1,S_2)\pi + P(S_1)P(S_2)(1-\pi)} = \frac{x\pi}{x\pi + 1 - \pi},$$

$$x = \prod_{i=1}^{l}\left[\frac{\Gamma(n+4\gamma)\Gamma(m+4\gamma)}{\Gamma(n+m+4\gamma)\Gamma(4\gamma)}\prod_\alpha \frac{\Gamma(n_\alpha^i + m_\alpha^i + \gamma)\Gamma(\gamma)}{\Gamma(n_\alpha^i + \gamma)\Gamma(m_\alpha^i + \gamma)}\right]$$

One can generalize this to arbitrary partitions of sequences into subsets: PROCSE

Probability sequences deriving from same WM

- Using Stirling's approximation we can rewrite *P(S)* for $\gamma = 1$ as:

$$P(S) = \left[\binom{n+3}{3} \right]^{-1} \exp\left(-n \sum_{i=1}^{l} H\left[\frac{n_A^i}{n}, \frac{n_C^i}{n}, \frac{n_G^i}{n}, \frac{n_T^i}{n} \right] \right)$$

$$H\left[\frac{n_A^i}{n}, \frac{n_C^i}{n}, \frac{n_G^i}{n}, \frac{n_T^i}{n} \right] = -\sum_{\alpha} \frac{n_\alpha^i}{n} \log\left[\frac{n_\alpha^i}{n} \right]$$

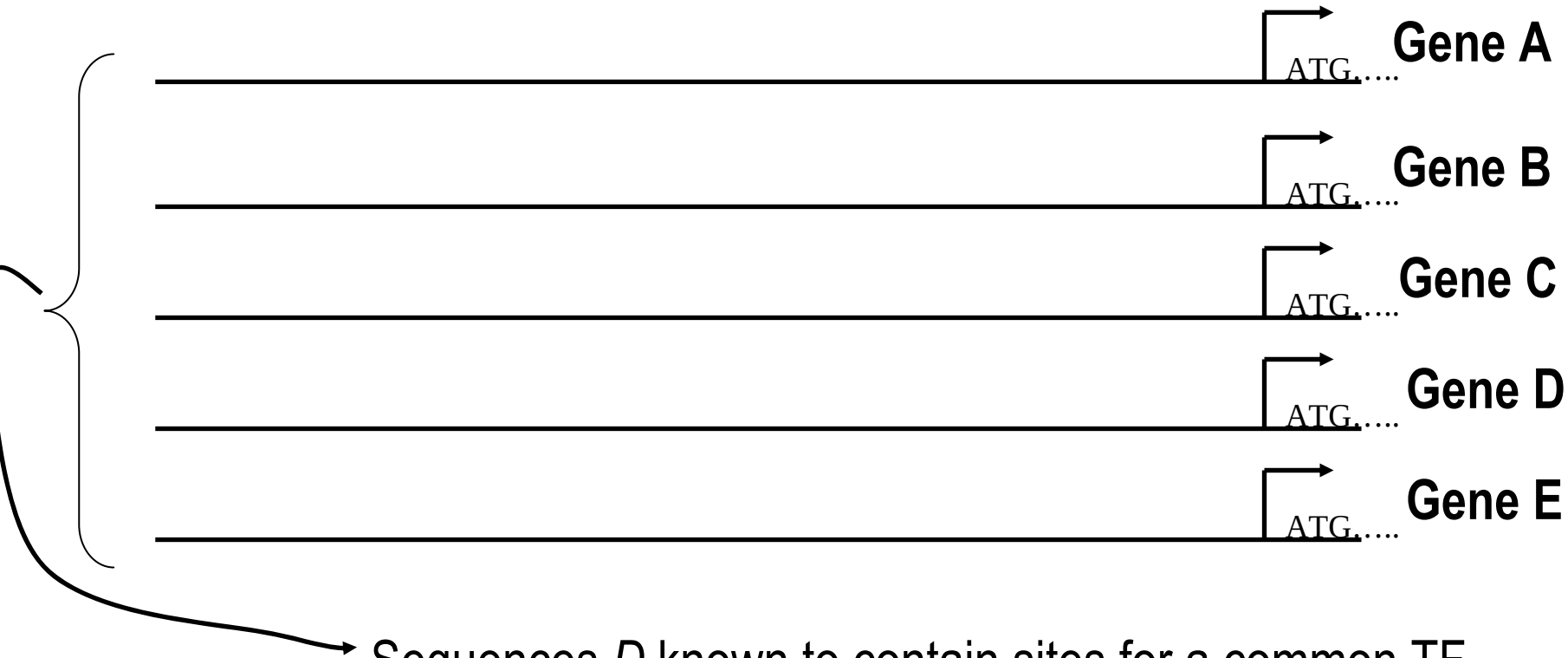
Number of possible base counts

How many different ways the observed count can be realized.

Probability to observe *S* when each sequence *s* stems from a different WM:

$$P(S) = \left[\frac{\Gamma(4\gamma)\Gamma(\gamma+1)}{\Gamma(4\gamma+1)\Gamma(\gamma)} \right]^l = \left(\frac{1}{4} \right)^l$$

Inferring a WM from a set of sequences



Expectation-Maximization approach (MEME, Bailey and Elkan 1994):

Maximize the partition sum with respect to the weight matrix entries.

Inferring a WM from a set of sequences

Recollect the derivative of the partition sum of the data P(D) wrt the prior:

$$\frac{d\log(P(D))}{d\pi} = \frac{\langle n \rangle}{\pi} - \frac{L - \langle n \rangle l}{1-\pi}$$

Similarly, the derivative wrt to a component of the WM is given by:

$$\frac{d\log[P(D|w)]}{dw_\alpha^k} = \frac{\langle n_\alpha^k \rangle}{w_\alpha^k}$$

where $\langle n_\alpha^k \rangle$ is the expected number of sites for *w* with letter *α* at position *k*.

We can thus optimize the WM by Expectation maximization, using update equations:

$$w_\alpha^k = \frac{\langle n_\alpha^k \rangle}{\langle n \rangle}$$

MEME approach

1. Choose a random segment and use its sequence to seed a weight matrix.

$$w_G^5 = 0.5, w_A^5 = w_C^5 = w_T^5 = 0.167$$

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …Gene A

ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …Gene B

CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …Gene C

TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …Gene D

ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …Gene E

ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …Gene E

GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …Gene F

MEME approach

1. Choose a random segment and use its sequence to seed a weight matrix.
2. For each position *i* in each sequence *s*, calculate the probability P(*s*,*i*) that this sequence is a binding site for the WM.

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …Gene A

ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …Gene B

CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …Gene C

TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …Gene D

ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …Gene E

ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …Gene E

GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …Gene F

MEME approach

1. Choose a random segment and use its sequence to seed a weight matrix.
2. For each position *i* in each sequence *s*, calculate the probability P(*s*,*i*) that this sequence is a binding site for the WM.
3. Construct a new WM by averaging the potential sites at all possible positions (*s*,*i*), weighing each potential site with the probability that it is a site for the previous WM.

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …Gene A

ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …Gene B

CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …Gene C

TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …Gene D

ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …Gene E

ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …Gene E

GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …Gene F

MEME approach

1. Choose a random segment and use its sequence to seed a weight matrix.
2. For each position *i* in each sequence *s*, calculate the probability P(*s*,*i*) that this sequence is a binding site for the WM.
3. Construct a new WM by averaging the potential sites at all possible positions (*s*,*i*), weighing each potential site with the probability that it is a site for the previous WM.
4. This is iterated until the WM does not longer change

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …Gene A

ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …Gene B

CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …Gene C

TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …Gene D

ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …Gene E

ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …Gene E

GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …Gene F

MEME approach

1. Choose a random segment and use its sequence to seed a weight matrix.
2. For each position *i* in each sequence *s*, calculate the probability P(*s*,*i*) that this sequence is a binding site for the WM.
3. Construct a new WM by averaging the potential sites at all possible positions (*s*,*i*), weighing each potential site with the probability that it is a site for the previous WM.
4. This is iterated until the WM does not longer change
5. The best WM over many seeds is reported.

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …Gene A

ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …Gene B

CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …Gene C

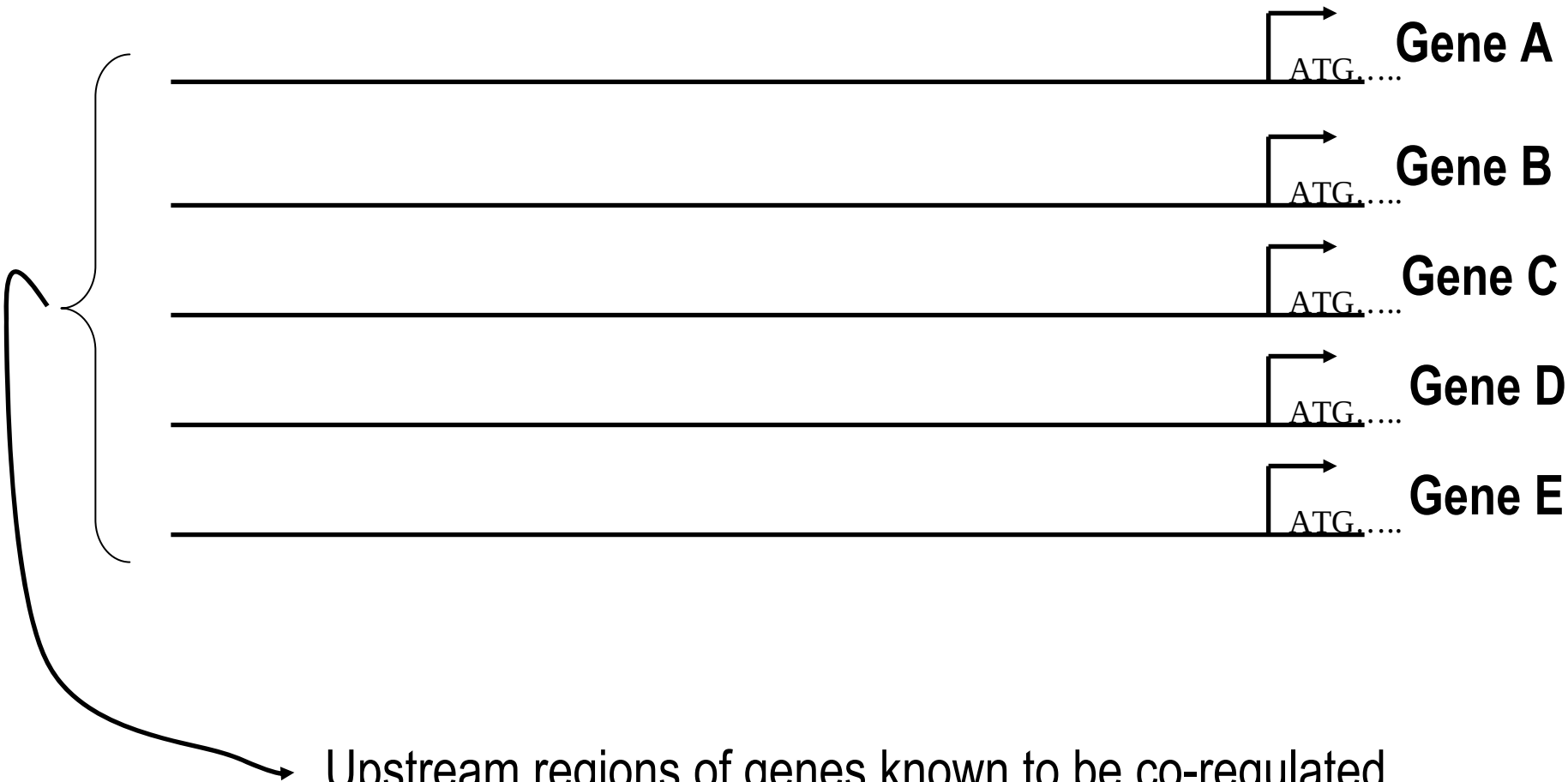
TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …Gene D

ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …Gene E

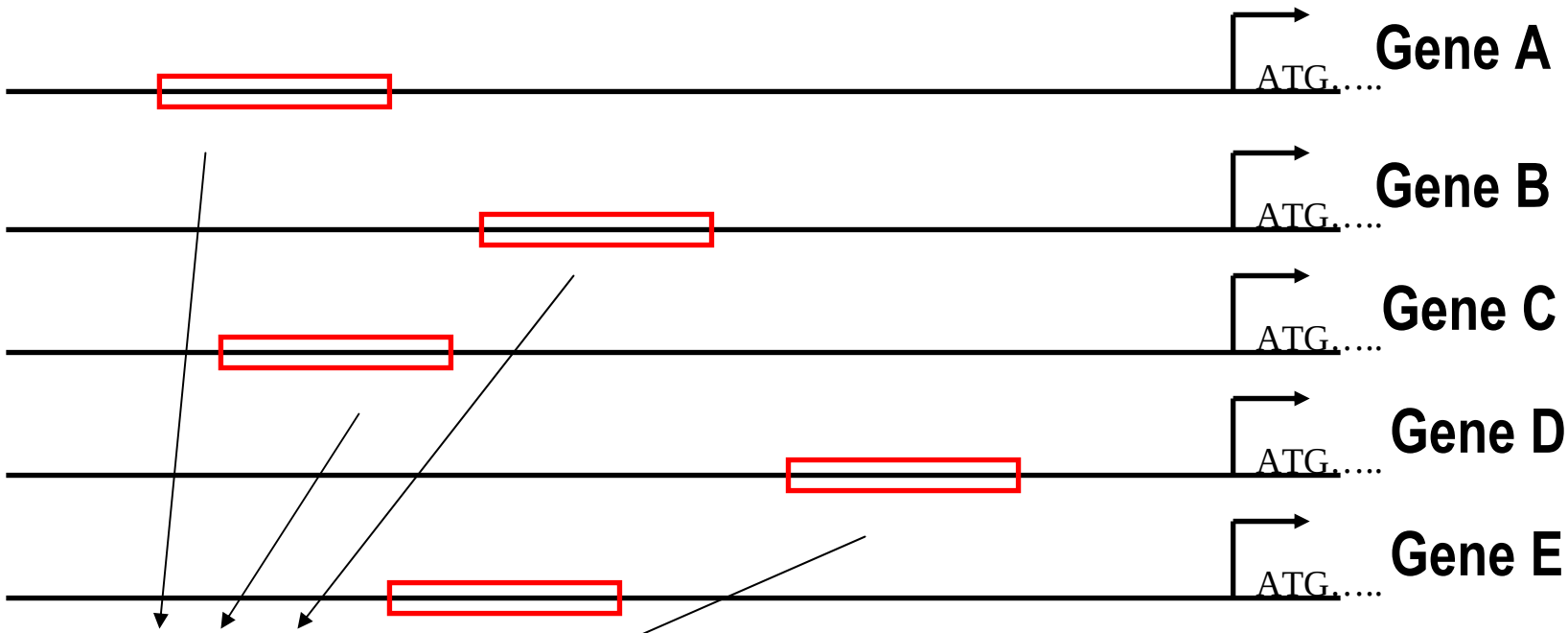
ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …Gene E

GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …Gene F

Gibbs Sampler (*J. Liu and C. Lawrence*): Finding motifs upstream of genes that are known to be co-regulated



Gibbs Sampler (*J. Liu and C. Lawrence*): Finding motifs upstream of genes that are known to be co-regulated

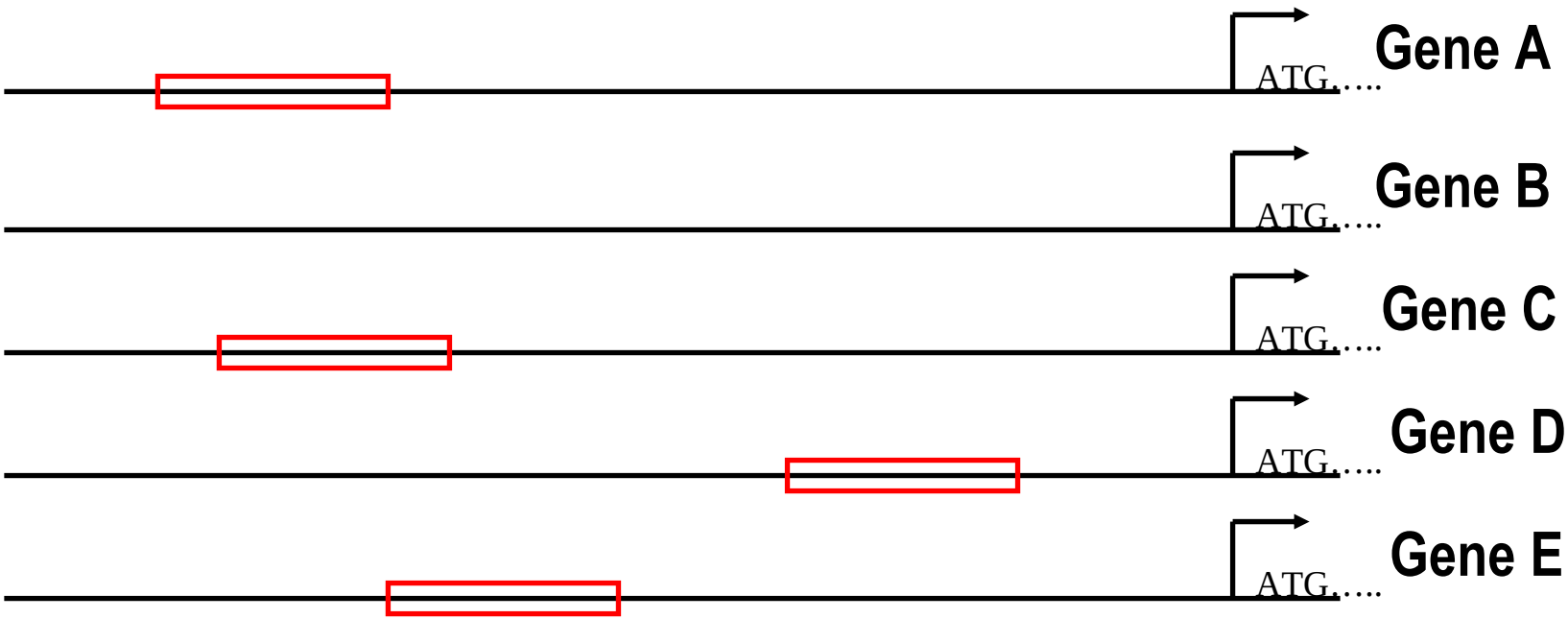


Alignment

```
acgtaacagttga
tcattggctagtg
tgagctagattat
aaagcgtagctag
ggctagcatggaa
gcattactatcaa
ccctttatatcta
```

- **Configuration:** assignment of windows to the sequences. This defines an alignment of sites.
- **Score:** Probability P(*S*) that all sites were drawn from a common weight matrix.

Gibbs Sampler (*J. Liu and C. Lawrence*): Finding motifs upstream of genes that are known to be co-regulated

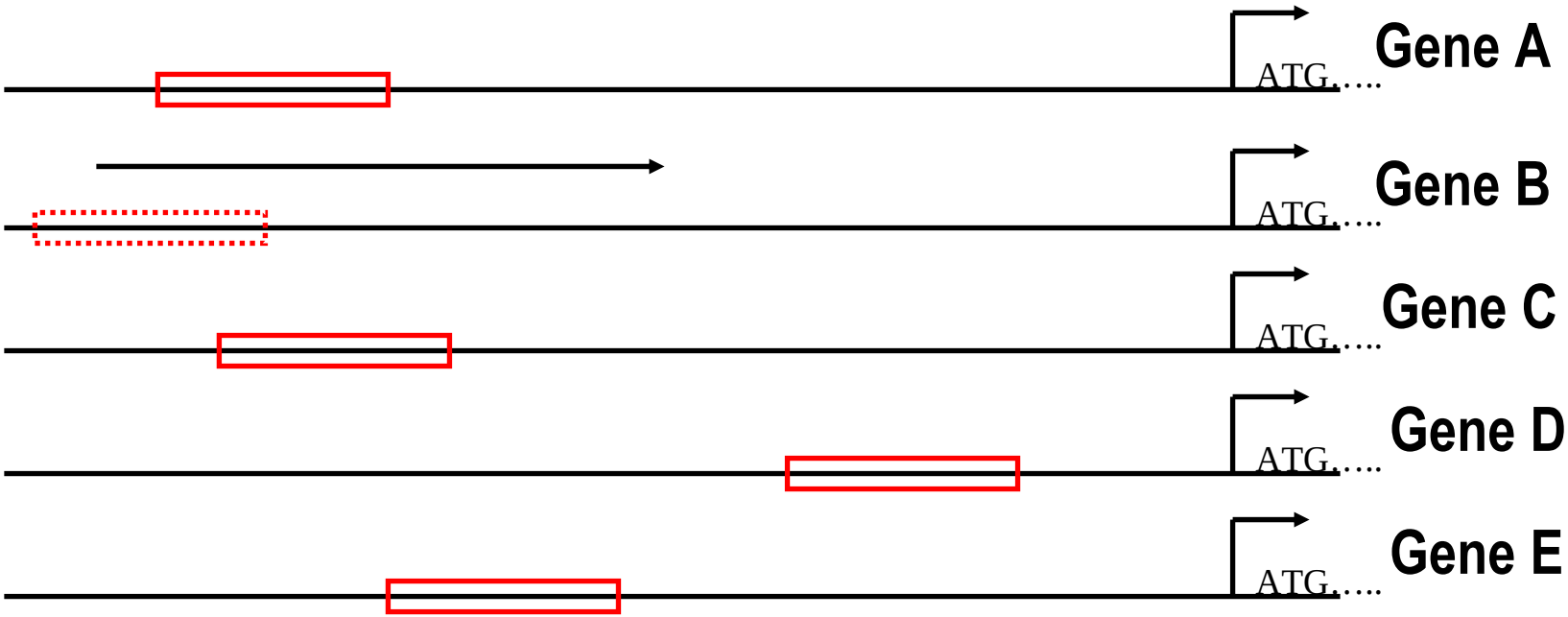



```
acgtaacagttga
tcattggctagtg
tgagctagattat

ggctagcatggaa
gcattactatcaa
ccctttatatcta
```

Remove a randomly chosen window.

Gibbs Sampler (*J. Liu and C. Lawrence*): Finding motifs upstream of genes that are known to be co-regulated



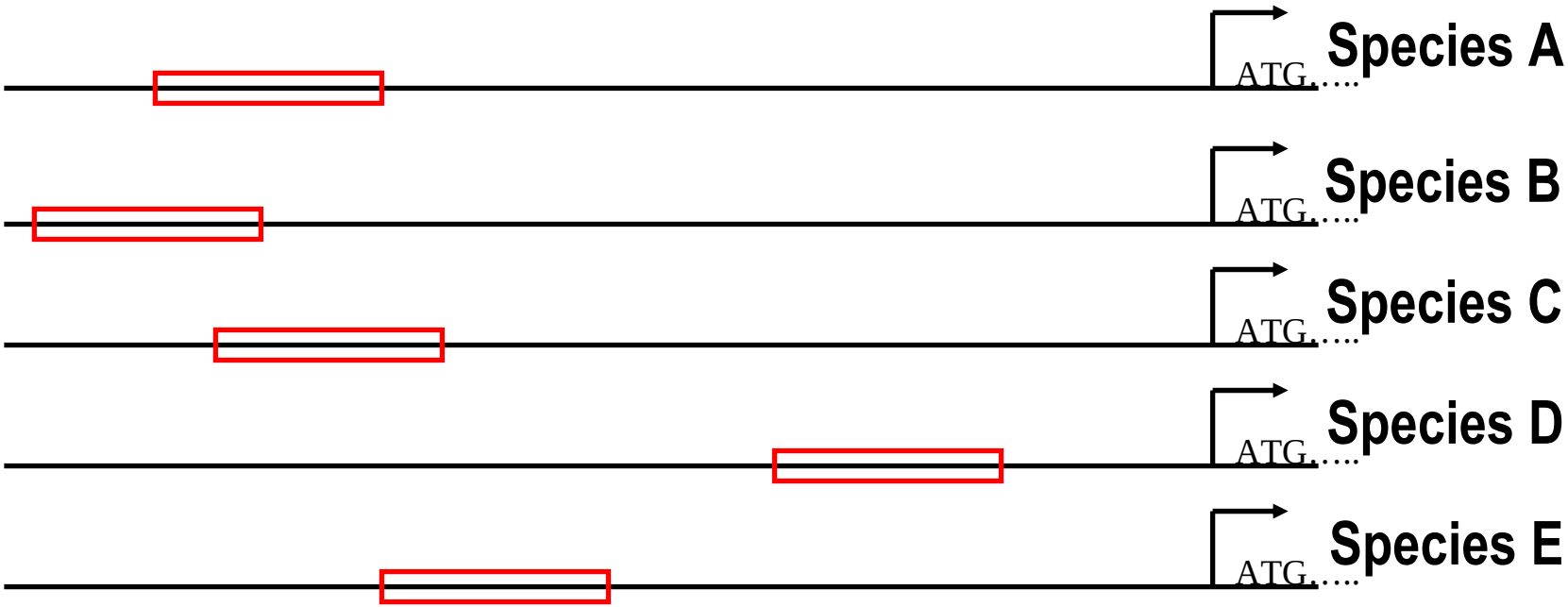
acgtaacagttga
tcattggctagtg
tgagctagattat
aactagtgtcaag
ggctagcatggaa
gcattactatcaa
ccctttatatcta

Scan along the sequence.

Probability to put window down at each position is

$$\frac{P(S,s)}{P(S)} = \prod_{i=1}^{L} \frac{n_{s_i}^{i} + \gamma}{n + 4\gamma}$$

Gibbs Sampler (*J. Liu and C. Lawrence*): Finding motifs upstream of genes that are known to be co-regulated




```
acgtaacagttga
tcattggctagtg
tgagctagattat
aactagtgtcaag
ggctagcatggaa
gcattactatcaa
ccctttatatcta
```

Now what?

Finding binding sites in multiple alignments of phylogenetically related sequences

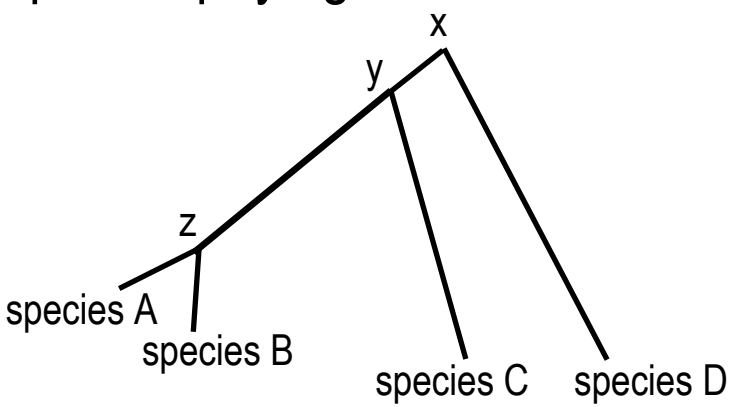
Potential site in aligned sequence block:

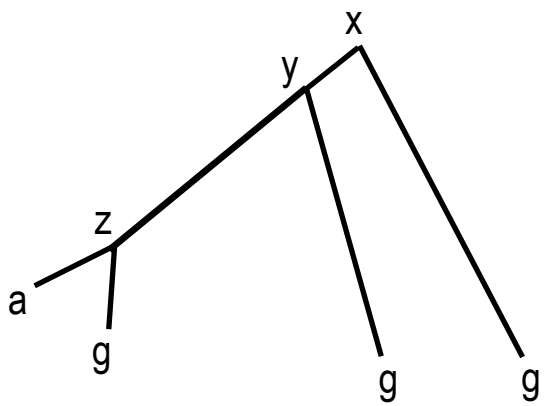
```
species A   acgtaactagtga
species B   acgttgctagatg
species C   tcgttgctataat
species D   aggtagcgagaag
```

S

Probability $P(S \mid w)$ of the set of bases *S* in each column is independent of the other columns.

Species phylogenetic tree



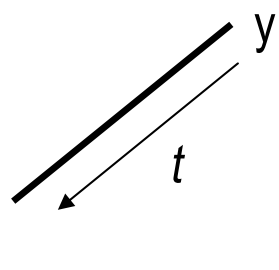


$P(S \mid w)$ is the probability to observe the set of bases *S* at the leafs given the phylogenetic tree and the WM *w*.

Site evolution along a single branch

Probability along a single branch:

$P(x \mid y,t,w)$ probability to end up with base *x* after time t, starting from base *y*.



General time evolution:

$$\frac{dP(x \mid y,t,w)}{dt} = \sum_z \mu(x \leftarrow z \mid w)P(z \mid y,t,w) - \mu(z \leftarrow x \mid w)P(x \mid y,t,w)$$

Assumptions:

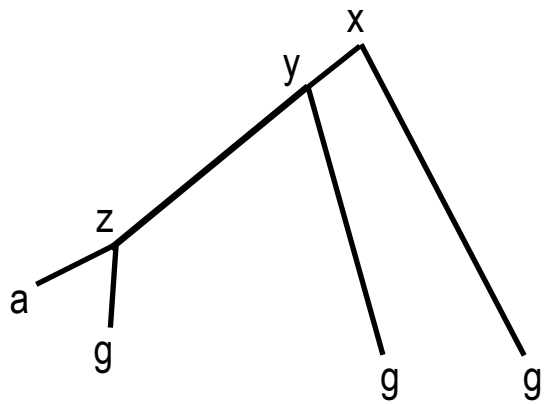
$$P(x \mid y,\infty,w) = w_x \qquad \mu(x \leftarrow y \mid w) = \mu(x \mid w)$$

Solution:

$$\mu(x \mid w) = w_x \mu \qquad P(x \mid y,t,w) = \delta_{xy} e^{-\mu t} + w_x \left(1 - e^{-\mu t}\right)$$

In terms of no-mutation probability *q*: $P(x \mid y,q,w) = \delta_{xy} q + w_x (1-q)$

Probability of an alignment column



The probability $P(S \mid w)$ of the bases at the leafs is the product over the probabilities of each of the branches, summed over the possible bases at the internal nodes:

$$P(S \mid w) = \sum_{x,y,z} w_x (\delta_{xy} q_{xy} + (1 - q_{xy}) w_y)(\delta_{xg} q_{xD} + (1 - q_{xD}) w_g)(\delta_{yg} q_{yC} + (1 - q_{yC}) w_g) \cdots$$

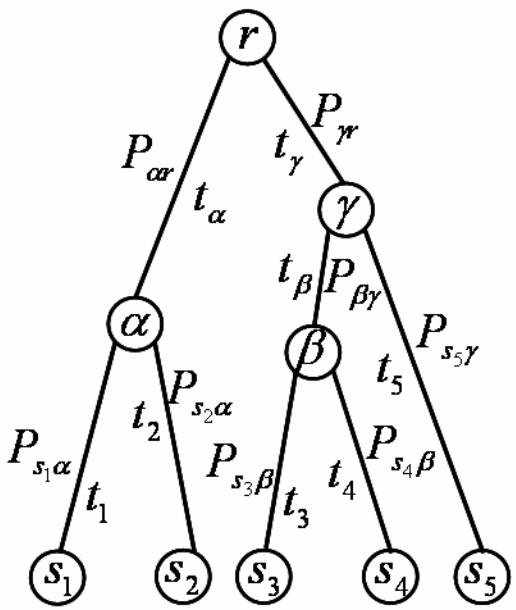
and similarly the probability given the background:

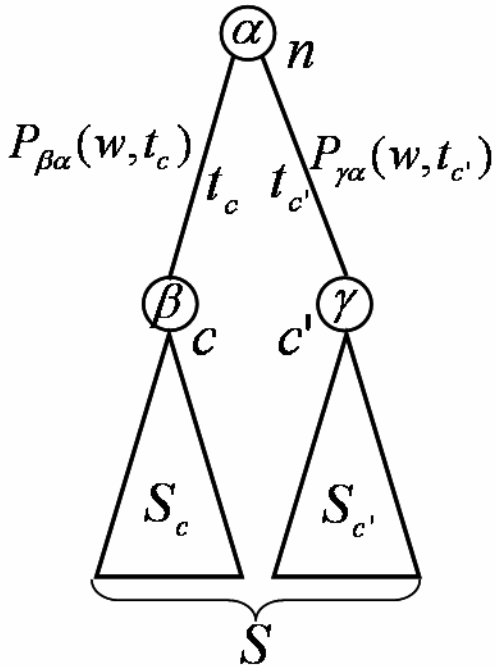
$$P(S \mid b) = \sum_{x,y,z} b_x (\delta_{xy} q_{xy} + (1 - q_{xy}) b_y)(\delta_{xg} q_{xD} + (1 - q_{xD}) b_g)(\delta_{yg} q_{yC} + (1 - q_{yC}) b_g) \cdots$$

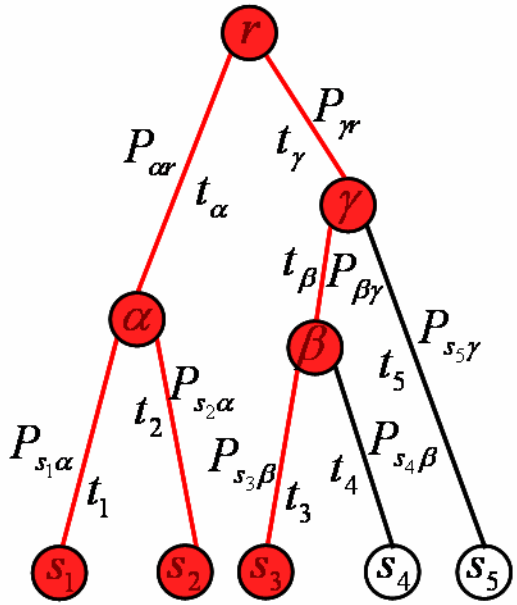
Note:

It seems that to evaluate this expression we have to sum over the states at all internal nodes, so that the number of terms would grow exponentially with the number of species.
Luckily, there is again a recursion relation (Felsenstein, 1981)

Probability of an alignment column







At every node the probability of the data under that node, given the node is α is given by:

$$D_\alpha(n,w) = \prod_{m \in c(n)} \left[\sum_\beta P_{\alpha\beta}(w,t_m) D_\beta(m,w) \right]$$

The probability of the alignment column given the tree and the WM is given by summing over the base at the root:

$$P(S|T,w) = \sum_\alpha w_\alpha D_\alpha(r,w)$$

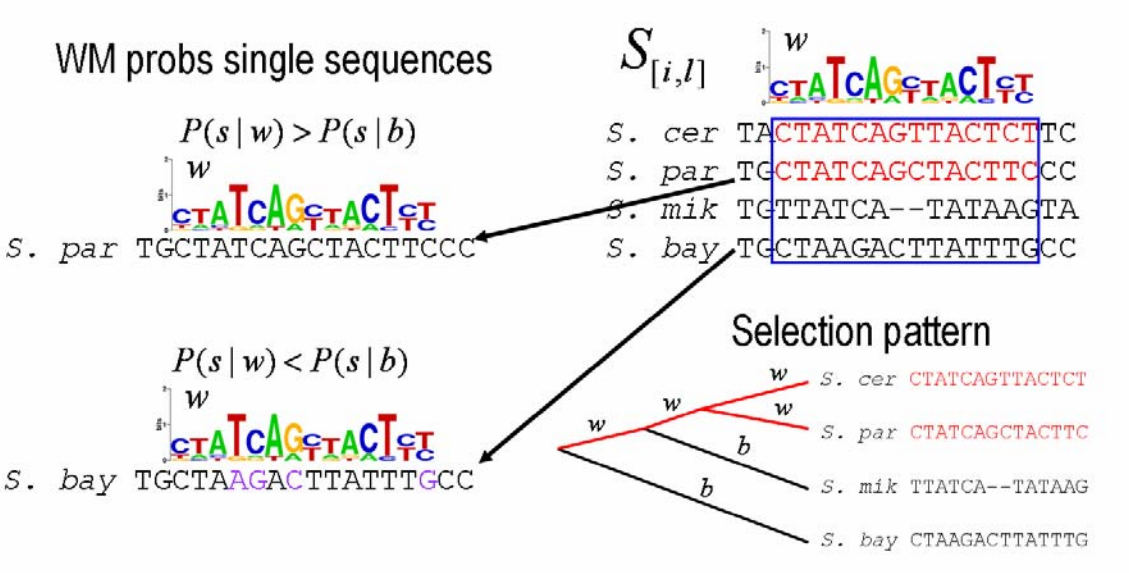
Finding binding sites in multiple alignments of phylogenetically related sequences

MotEvo

Using the Felsenstein algorithm we can easily calculate the probability P(*S*|*T*,*w*) of an alignment column given the WM *w* and the phylogenetic tree *T* and the evolution model for each branch.

Now we can apply the same recursion relation to calculate the partition sum of binding site configurations to an alignment:

$$F_n = F_{n-1} P(S_{[n-1,1]}|T,b)\pi_{\mathrm{bg}} + \sum_w F_{n-l_w} P(S_{[n-l_w,l_w]}|T,w)\pi_w$$



Ab initio discovery of regulatory sites

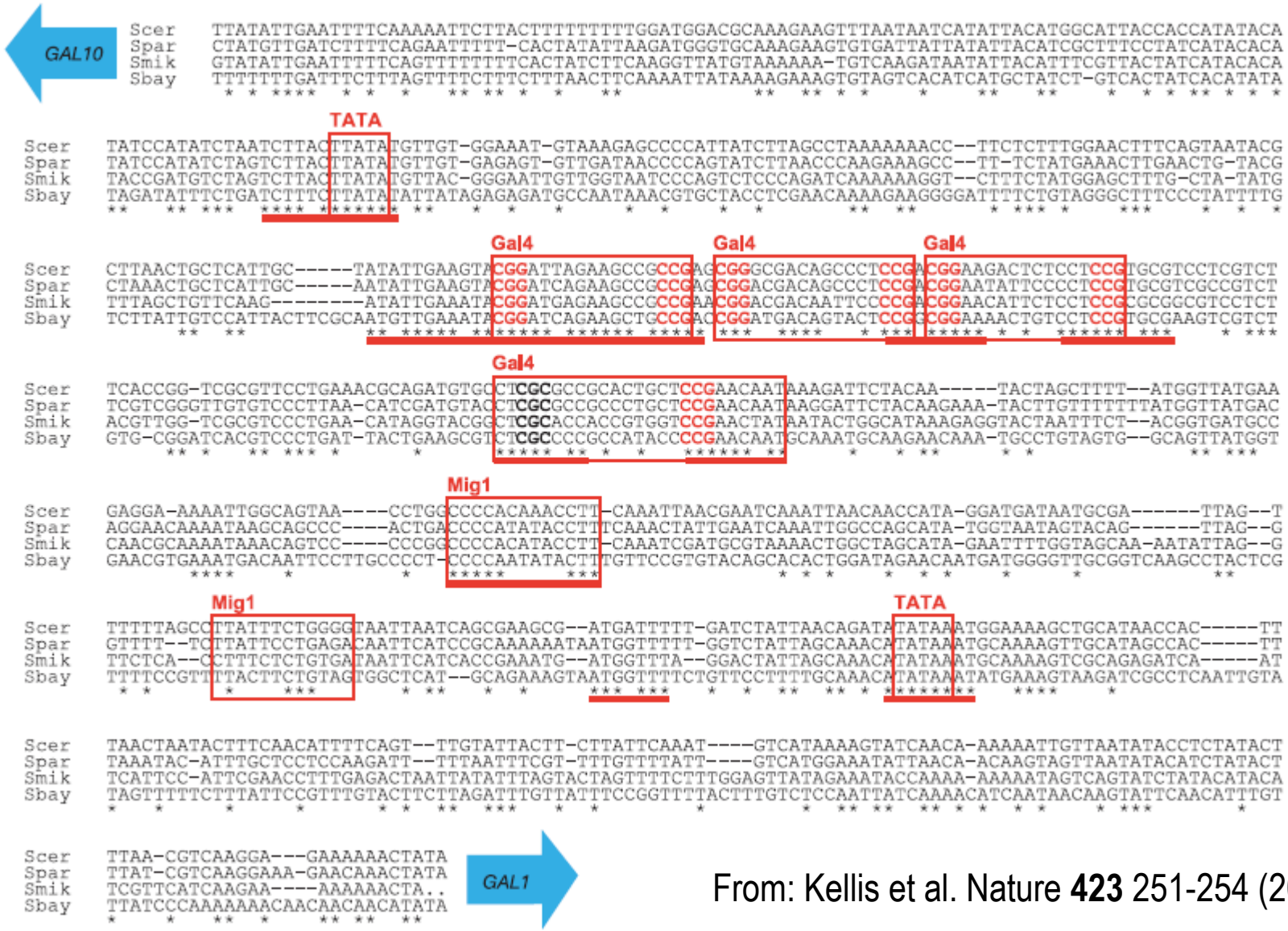
General Approaches:

1. Collect sets of (intergenic) sequences that are thought to contain binding sites for a common regulatory factor. Examples:
   - Upstream regions of co-regulated genes.
   - Sequence fragments pulled down with ChrIP

   then search for overrepresented short sequence motifs.

2. Phylogenetic footprinting: create multiple alignments of orthologous intergenic sequences and identify sequence segments more conserved than “average”.

Phylogenetic Footprinting



From: Kellis et al. Nature **423** 251-254 (2003)

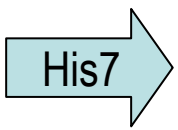
Phylogenetic Footprinting

```
Scer ATGTTTTTTTAATGATATATGTAACGTACATTCTTTCCTCTACCACTGCCAATTCGGTATTATTTAATTGTGTTTAGCGCTATTTAC
Spar -ATGTTTTTTAATGATATATGTAACGTACATTCTTC---CTACTGCTACCAAGTCGGTATTATTTAATTGTGTTTAGCGCTATTTAC
Smik ------------------------TCTTTTCTCTA--CCACTACTACCAATTCGGTATTATTTAATTGTGTTTAGCACTATTTAC
Sbay --ATGTTCTTAATGATATATATAACGTACATTTTTT---CCTCTACTAGCCAATCGGTGTTATTTAATTGTGTTTAGCTCTATTTAC
                              *   ** *     *  *  **  * * ***** ****************** ********

Scer TAATTAACTAGAAACTCAATTTTTAAAGGCAAAGCTCGCTGACCT--TTCACTGATTTCGTGGATGTTATACTATCAGTTACTCTTC
Spar CCACTAACTAGAAACTCGATTTTTAAAGGCAAAATTCAGTGTCCT--TTCACTAGTTTTGCAGATGTCCTGCTATCAGCTACTTCCC
Smik TCACTAAC-AAAAACTCAATTTTGAAGGGCTGA-TTAAATATCCTCCTTTAATAGTTTTGCGCTTAGCCTGTTATCA--TATAAGTA
Sbay TCACTTAACAAAAAAACCAACTTCAAAAGTATAATACAATAATTTC-TCCGTTGATCTTGTGAACTACATGCTATCACTTATTTGCC
       * * *  * ***  * *  ** **  *   *      *    *  *    *  * * *         *  *****  **
                                GCN4                                ABF1
Scer TGCAAAAAAAAA-----------TTGAGTCATATCGTAGCTTTGGGATTATTTTTCT-CTCTCTCCACGGCTAATTAGGTGATCATG
Spar TGCAGAAAAGAAAAATA-----TTTGAGTCATATCATCGTCTAGGAAGTGTTTTTCT-CTCTCTCCACGGATAGTTAAGTGATCATG
Smik TACAAAAAGAGAATAT------TTTGAGTCATATCATCGCCTAGGAAGTATTTTTTTTCTCTCTTCACGGTTAATTAGGTGATTTCT
Sbay TGTAAAAAGAAAATCGTTTCGTTTTGAGTCATATCATGTTCTCATAA-TATTTTTTT--TTCCTTAGCGATTAA------------
     *  * ***   *           ************ *    *    * * ***** *  *  **   **  **
                           GCN4
Scer AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik GAAAAACGAAAAATTCATG-GAAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTCGACAAAAATCTG
Sbay GAAAAATAAAAAGTGATTG-GAAAAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA
      *****  **** *   ** ***********    **  ************     * * **  * **    *       *   **

Scer CTCAATCAGG-TTTTAAAAGAAAAGAGGCA-GCTATTGAAGTAGCAGT-ATCCAGTTTAGGTTTTTTAATTATTTACAAGTAAA-GA
Spar CTCAATCAA--GTTTAATAGAAGAAAGAGG-AAGGTTGAGATAGGTAT-ATCCAGTTTAGGTTTC--AATTATTTAATAATAAA-GG
Smik CAATATTCATTATTCAAAACTCAAAAGAAG-AAGGTTCGAATTGGTGT-GTCCAGTTTAGGCTCT--AATTGTTGAATAATAAAAGG
Sbay TCCACCACAA-ATTGAAGGTGAGGAAGAAACAAAGTTAAAGCAAGAATCGGCTTGTGTCTTTTTT--GATTGCGTATT--TGAAAGG
               ** **       **        **          *   *  ** *    *     ***    *    * ** *

Scer AAAAGAGA--------------
Spar TAAAGAA---------------
Smik CGAAGAAATAACGATCCAAAAA
Sbay TAAAGGAATACAACAAAAA---
       ***
```



Ab initio identification of regulatory motifs

General Approaches:

1. Collect sets of (intergenic) sequences that are thought to contain binding sites for a common regulatory factor. Examples:
   - Upstream regions of co-regulated genes.
   - Sequence fragments pulled down with ChrIP

   then search for overrepresented short sequence motifs.

2. Phylogenetic footprinting: create multiple alignments of orthologous intergenic sequences and identify sequence segments more conserved than "average".

Wish: combine these two approaches into a single procedure.

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.

```
AGAAGAAAGTAAttcttATGAGAAAATTTGCGGGAGTTCTTTGCCAGTGAGATAAAGtttttttt-------AATTTTAATCAACACAAAATACACATATTTATATAAACTGacgaaata-
TGAAAAAAGTAAccttcATGAGATATATTGCGGAAGTCCATTACCAGTAAGTTAGAGttagaaaatttcgatcgacacaatttatacttcgatatatactggcaaaaaa------------
tgggaggaaaaaaaccattacctgtatgaaaaagattgcaaggattccttttgttagtgaactgaactTTAGGGATTTTAATCAACACAGTATATACATATatctttgtatactgacaaata
agtaagctatatgaaaagtttcctttagcagtaaatttagagc-----------------------TTAGGAATTTTGATCAAGACACAATATATATAGCTTTATATATTGtcaaata--
```

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

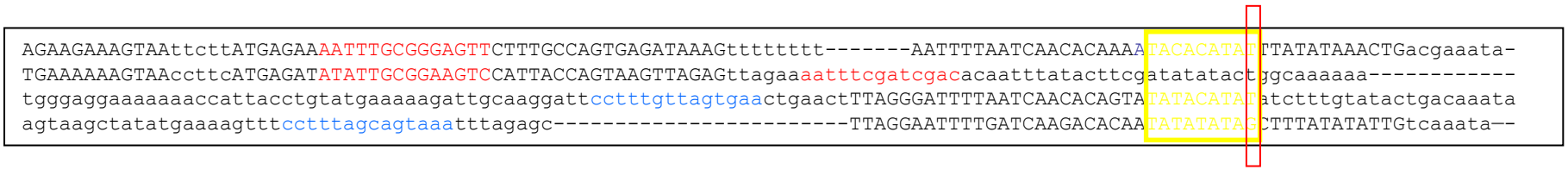
- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.

Sites in aligned region scored according to phylogeny

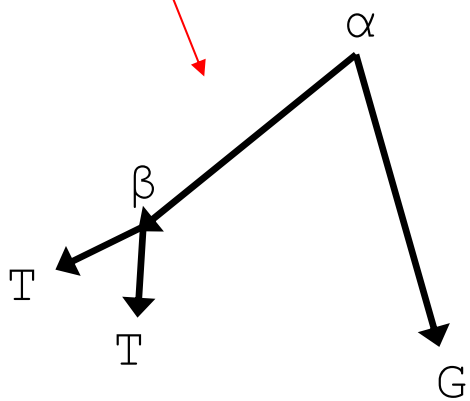
PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.



- Calculate the probability of the observed set of bases given the evolutionary model.
- The evolutionary model assumes that substitution rates are biased by the weight matrix of the motif.
- Unknown bases at internal nodes are summed over.



PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.

S

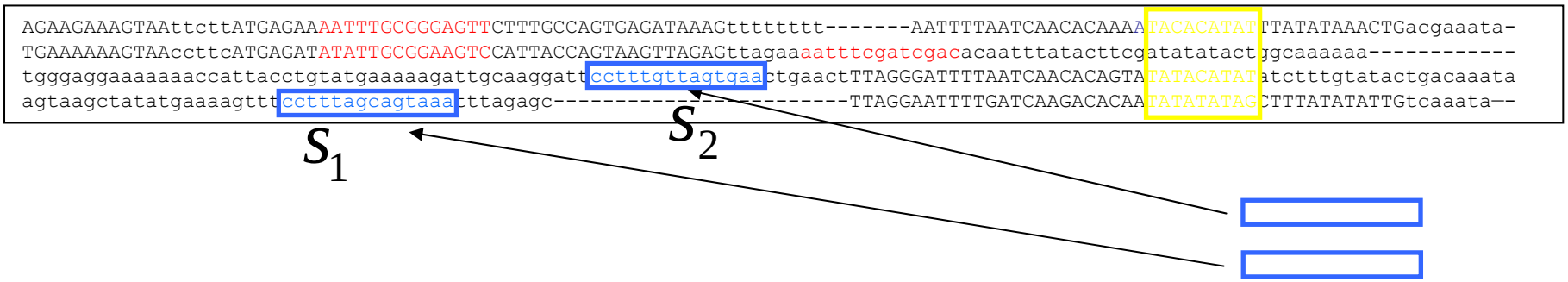
Sites in aligned region scored according to phylogeny

$$R(S) = \frac{P(S)}{P(S \mid b)} = \frac{\int P(S \mid w) P(w) dw}{P(S \mid b)}$$

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.



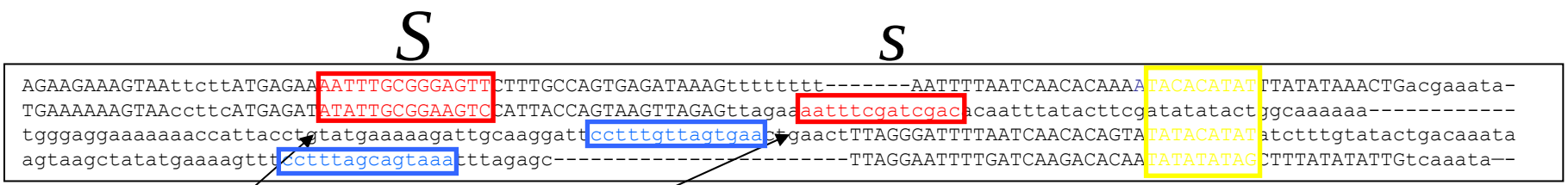
Scored according to independent model.

$$R(s_1, s_2) = \frac{P(s_1, s_2)}{P(s_1 \mid b)P(s_2 \mid b)} = \frac{\int P(s_1 \mid w)P(s_2 \mid w)P(w)dw}{P(s_1 \mid b)P(s_2 \mid b)}$$

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.



$$R(S,s) = \frac{P(S,s)}{P(S \mid b)P(s \mid b)} = \frac{\int P(S \mid w)P(s \mid w)P(w)dw}{P(S \mid b)P(s \mid b)}$$

Phylogeny based score combined with site for same WM at phylogenetically unrelated position.

PhyloGibbs

Siddharthan R, Siggia ED, van Nimwegen E, PLoS Comput Biol 2005 1(7) e67

- **Input:** Multiple alignments of orthologous intergenic sequences plus species tree.
- **Search:** MCMC through all 'binding site configurations' to the input sequences.
- **Score configuration:** Probability of the observed sequences assuming that all binding sites for a common motif are constrained by and evolving according to a common PSWM.

Score of the binding site configuration *C* is product of the scores for each 'color':

$$R(D \mid C) = R(S_{\text{yellow}})R(s_{1\,\text{blue}}, s_{2\,\text{blue}})R(s_{\text{red}}, S_{\text{red}})$$

Identifying TF binding sites in species with different degrees of evolutionary relatedness: The **phylogibbs** algorithm

Posterior probability configuration *C*: $P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)}$

$P(C) = \text{prior on configuration } C.$

Motif finding strategy:

- Anneal: Sample all configurations according to $P(C \mid D)^{\beta}$ slowly increase β.
- Take configuration *C** at end of anneal as reference configuration.
- Sample all configurations according to $P(C \mid D)$ and track the average membership of each site group of the reference configuration.

Results on gene groups identified by ChIP-on-chip

letters to nature

Nature **431**, 99 - 104 (02 September 2004); doi:10.1038/nature02800

Transcriptional regulatory code of a eukaryotic genome

CHRISTOPHER T. HARBISON, D. BENJAMIN GORDON, TONG IHN LEE, NICOLA J. RINALDI, KENZIE D. MACISAAC, TIMOTHY W. DANFORD, NANCY M. HANNETT, JEAN-BOSCO TAGNE, DAVID B. REYNOLDS, JANE YOO, EZRA G. JENNINGS, JULIA ZEITLINGER, DMITRY K. POKHOLOK, MANOLIS KELLIS, P. ALEX ROLFE, KEN T. TAKUSAGAWA, ERIC S. LANDER & DAVID K. GIFFORD

Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, Massachusetts 02139, USA
Broad Institute, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA
These authors contributed equally to this work

- ChIP-on-chip for 203 yeast DNA binding proteins.
- Ran 6 different *ab initio* motif finding algorithms on upstream regions that were pulled down by a given protein.
- Defined motifs for 116 DNA binding proteins.
- We focus on 45 TFs that have between 3 and 25 annotated sites.
- In 21 cases all computational methods failed to identify a motif and the reported motif simply copies the motif reported in the literature.

Results on gene groups identified by ChIP-on-chip

Results PhyloGibbs: Motif matching literature was found on 16 out of 21

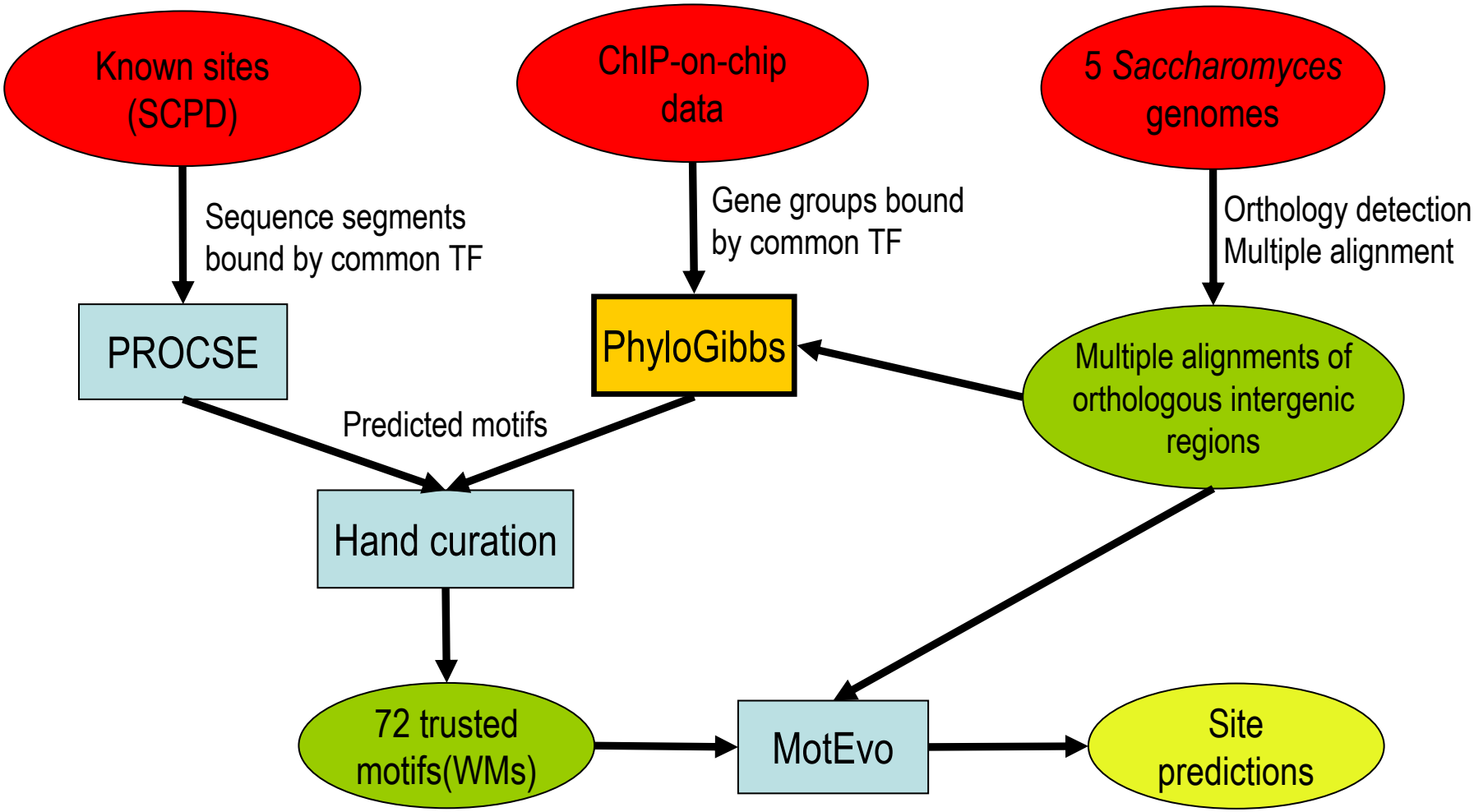
Name	anneal WM	track WM	anneal sites	track sites	num. annotated [23]
GAL80*	0.88	0.89	3/4	2.73/3.41	3
GCR1+	1.00	1.00	6/10	5.17/7.96	7
HAP2	1.00	1.00	15/15	4.5/6.72	21
HAP3	1.00	1.00	10/10	7.14/8.21	13
MET32*	1.00	1.00	9/11	7.33/9.77	13
MSN4*	1.0	1.0	14/23	12.41/20.68	21
RGT1+	0.42	0.67	9/13	8.27/12.35	12
RTG3*	1.0	1.0	3/7	2.31/5.77	5
PUT3+	0.03	0.0	3/3	2.73/2.73	4
MET31*	1.0	1.0	2/3	1.84/2.52	5
ADR1*	0.92	0.92	1/9	0.19/6.21	13
MAC1	0.77	0.69	1/4	0.85/2.67	5
HAP5	1.0	0.0	10/15	0.99/10.97	21
SKO1	1.0	0.0	2/4	0.91/4.13	7
GZF3	0.74	0.0	0/3	0/0.83	3
RLM1*	0.74	0.0	0/8	0/3.64	9
DAL80	0.03	0.0	0/6	0/5.75	9
MOT3	0.0	0.0	0/10	0/11.69	11
ROX1	0.0	0.0	0/10	0/6.65	12
YAP6	0.0	0.0	0/4	0/2.5	3
YOX1	0.0	0.0	1/2	0.91/1.82	3

Comparing literature with ChIP-on-chip data

TF	literature targets	ChIP-on-chip targets	overlap	PhyloGibbs on lit. targets
GCR1	8	4	1	found literature motif
MET31	5	4	0	found literature motif
MAC1	4	5	1	found literature motif
SKO1	5	6	0	found literature motif
RLM1	6	9	1	found literature motif
GZF3	3	3	0	found literature motif
ADR1	3	10	1	found literature motif
DAL80	4	8	0	found literature motif
MOT3	3	8	0	literature motif not found
ROX1	3	11	0	found literature motif
YAP6	0	3	0	-
YOX1	28	3	1	found literature motif

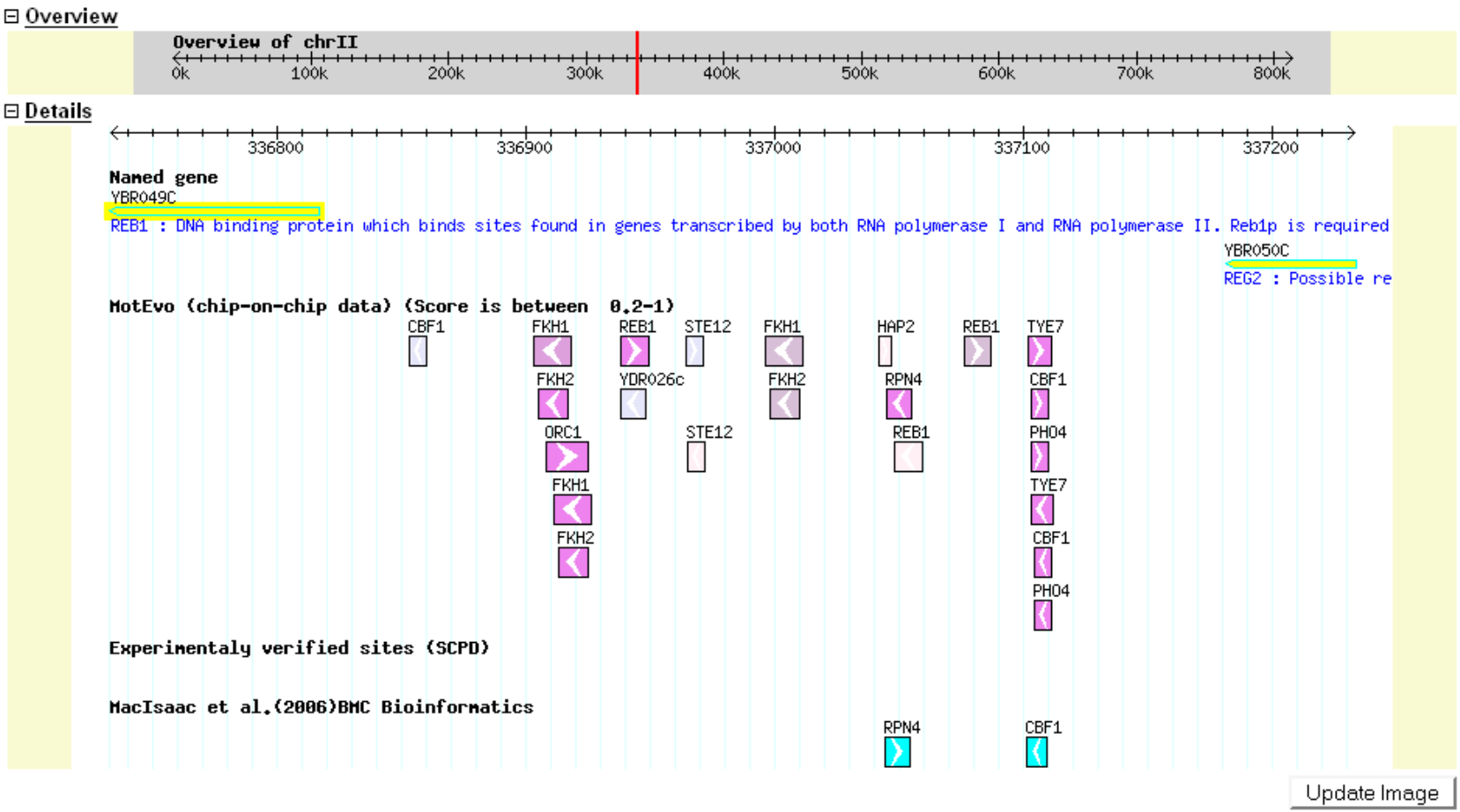
- The target genes in the literature show little overlap with targets predicted through ChIP-on-chip.
- In 3 of 4 cases where the known motif was not found on the ChIP targets it *was* found when PhyloGibbs was run on the literature targets.

Genome-wide site annotation for *S. cerevisae*



Genome-wide regulatory site annotation for Yeast.

Site annotations implemented in Gbrowse (**www.swissregulon.unibas.ch**):



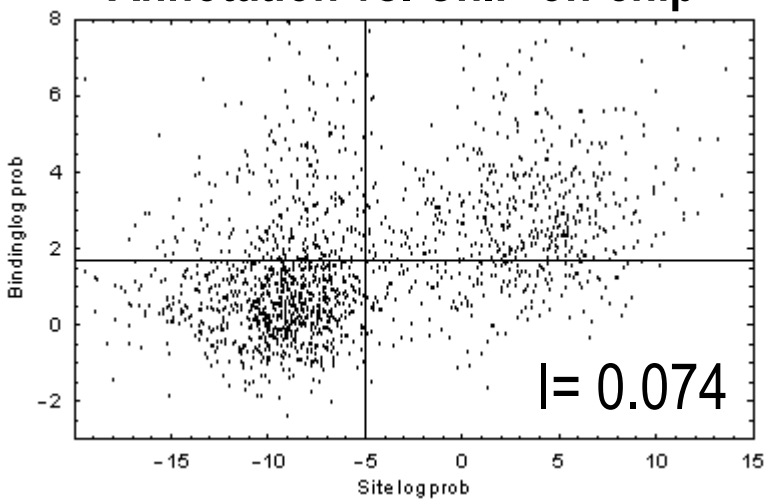
80 TFs

37,000 sites with posterior probability larger than 0.1
12,000 sites with posterior probability larger than 0.5.

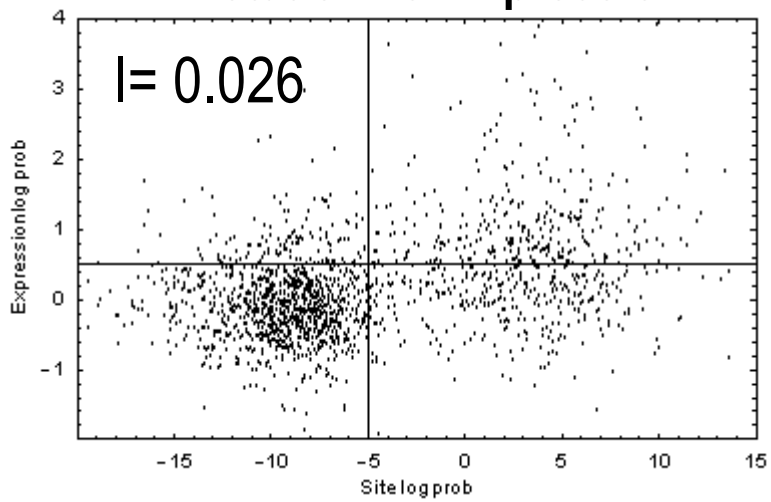
Annotation reliability:

Comparison with ChIP-on-chip and expression data

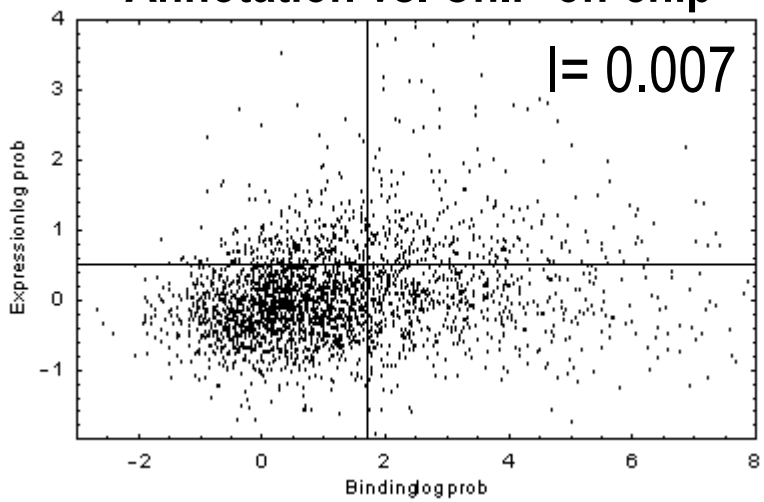
Annotation vs. ChIP-on-chip



Annotation vs. Expression



Annotation vs. ChIP-on-chip



ABF1

Collaboration with:
Ulrich Schlecht, Michael Primig
Biozentrum, University of Basel

Annotation reliability:
Comparison with literatture sites

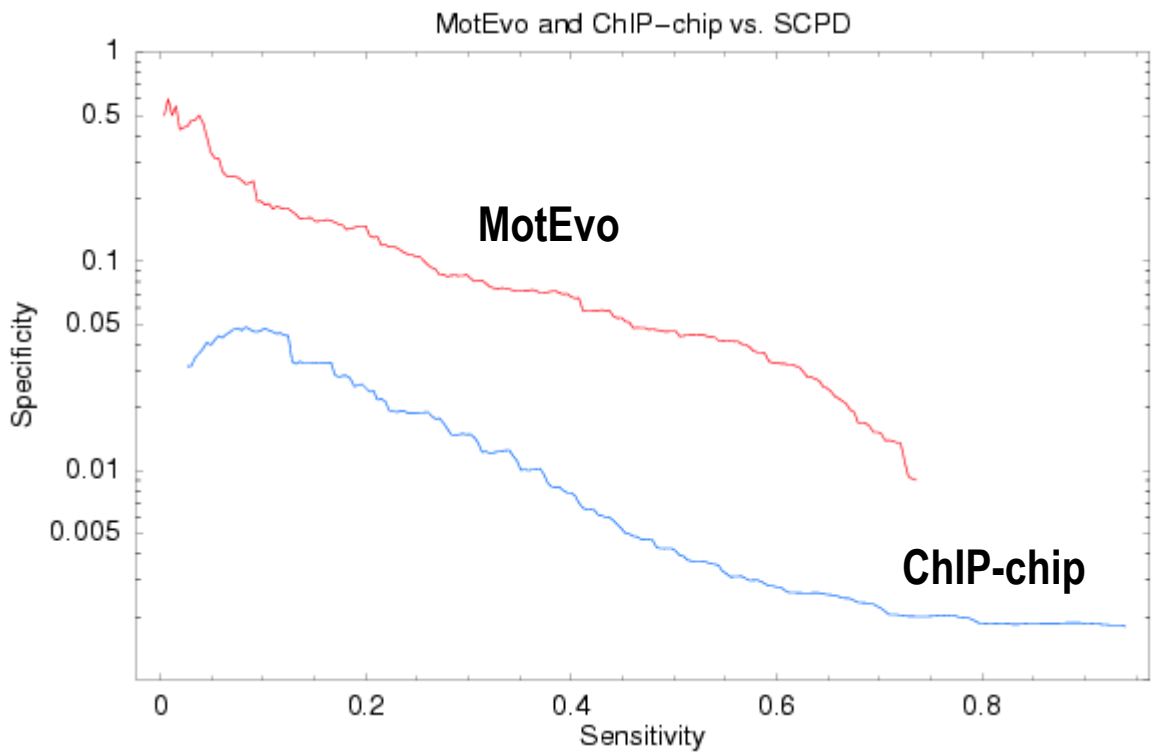
SCPD
The Promoter Database of *Saccharomyces cerevisiae* (Michael Zhang, Cold Spring Harbor)

After 'clean up' 437 known binding sites upstream of 200 cerevisiae genes.

- For each TF with at least 1 known site, determine the probability for each intergenic region to have at least 1 site.
- Using, the Harbison Chip-chip data, determine for each intergenic region the p-value that the region is not bound.
- Compare the predictions of MotEvo and Chip-chip taking literature sites as reference.

Sensitivity: Fraction of regions with known sites that are predicted to indeed have a site (for the right TF).
Specificity: Fraction of all regions predicted to contain a site that are found in the literature (for the right TF).

Annotation reliability:
Comparison with literatture sites



To hit the same number of literature regions ChIP-chip has to predict 10-20 times as many regions.

Network topology

letter

Topological and causal structure of the yeast transcriptional regulatory network

Nabil Guelzim[1,2], Samuele Bottani[3], Paul Bourgine[2] & François Képès[1]

Published online: 22 April 2002, DOI: 10.1038/ng873

NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

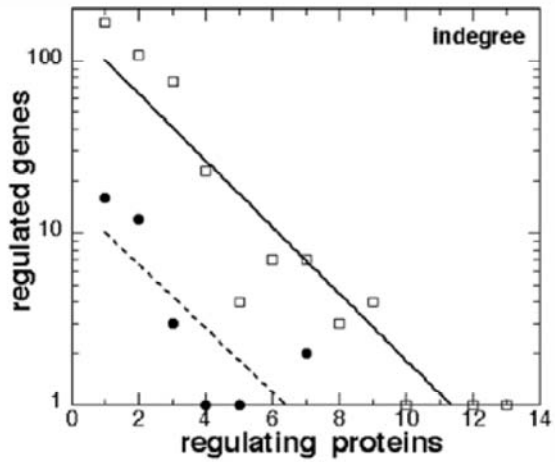
Albert-László Barabási & Zoltán N. Oltvai‡*

letter

Network motifs in the transcriptional regulation network of *Escherichia coli*

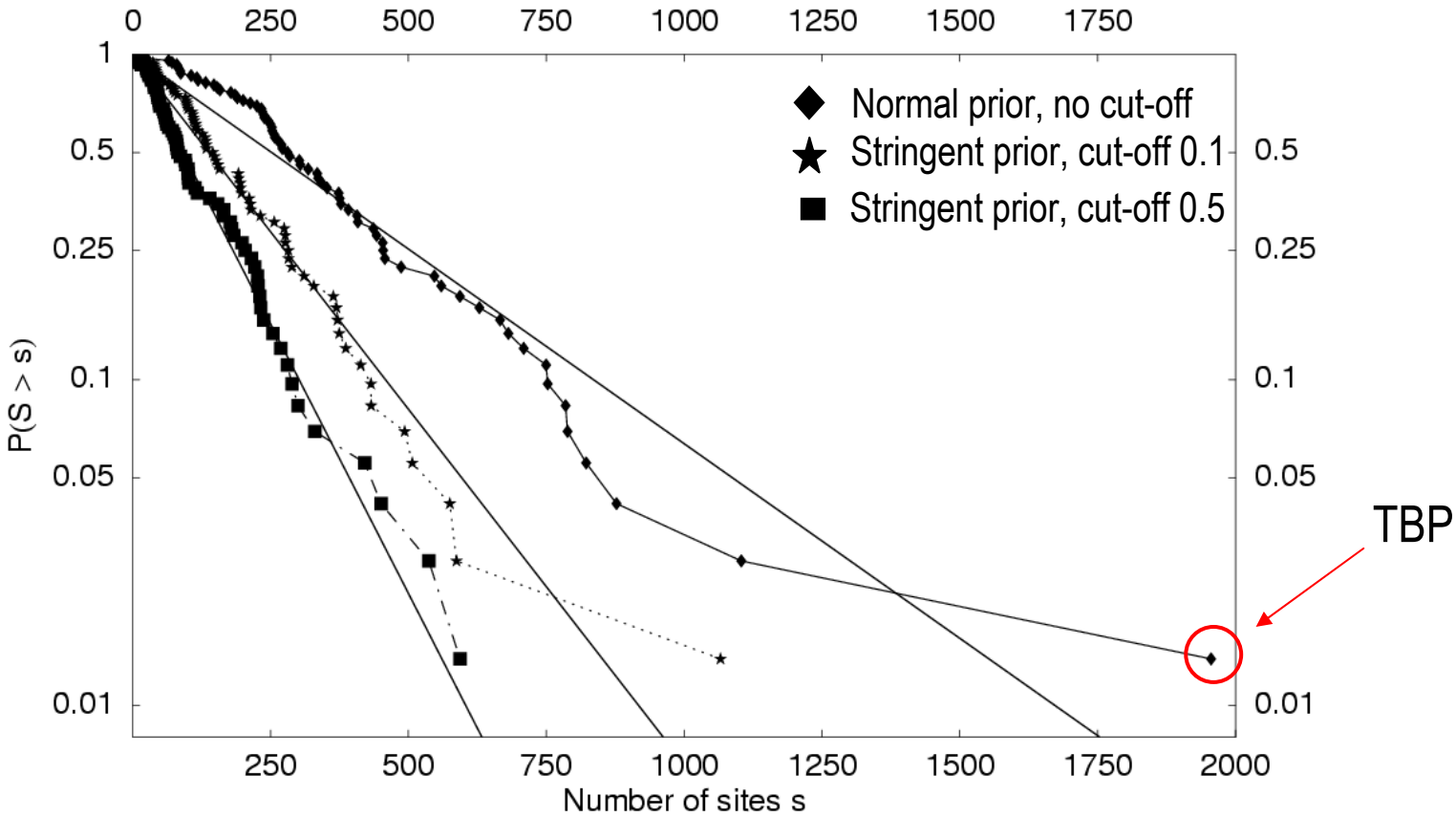
Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

Published online: 22 April 2002, DOI: 10.1038/ng881



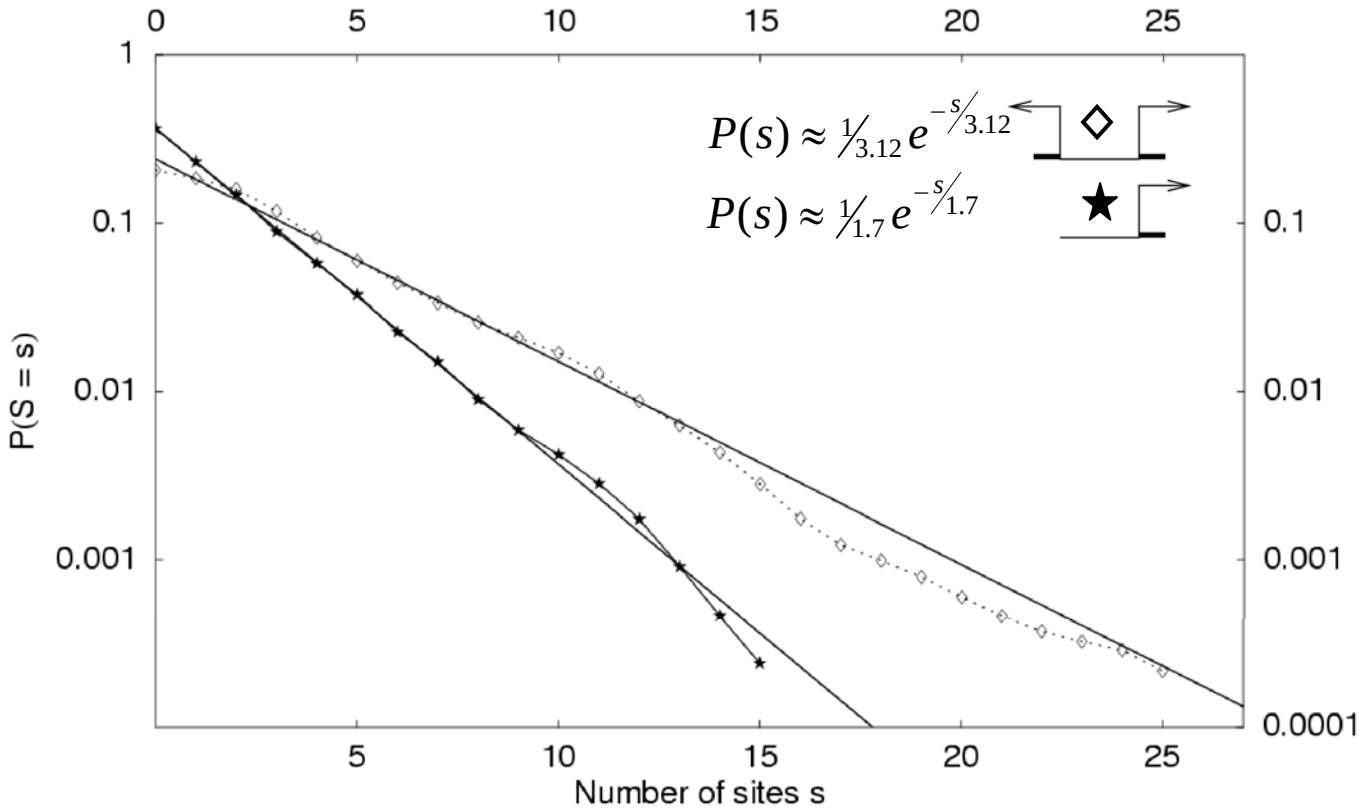
Transcription regulatory networks of Yeast and *E. coli* have been claimed to be *scale free*, i.e. with power-law distributed in- and out-degrees.

Network topology
Outdegree: Number of sites per motif



Number of sites per motif is exponentially distributed.

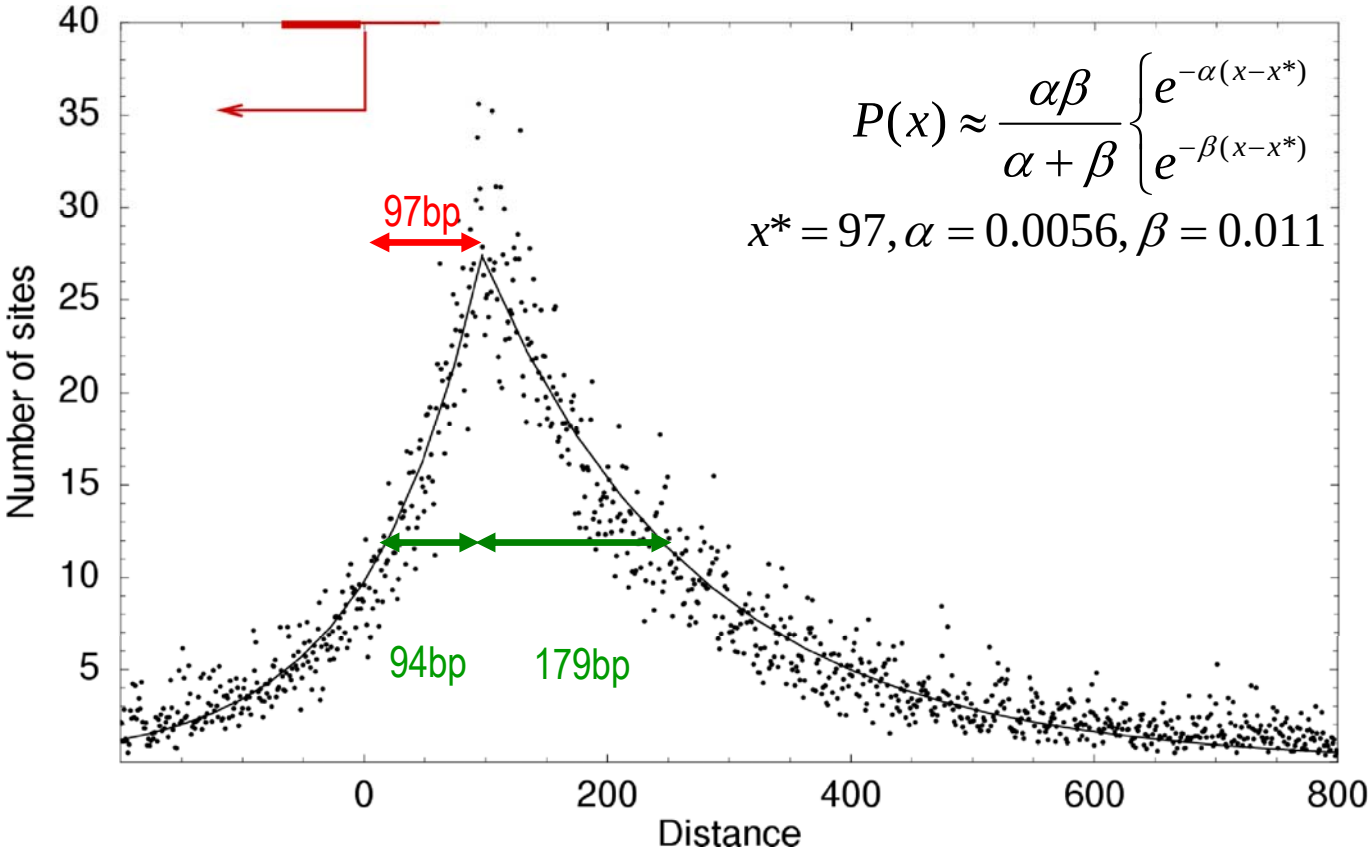
Network topology
Indegree: Number of sites per intergenic region



Number of sites per intergenic region is also exponentially distributed.

Almost precisely twice as many sites in divergently transcribed regions

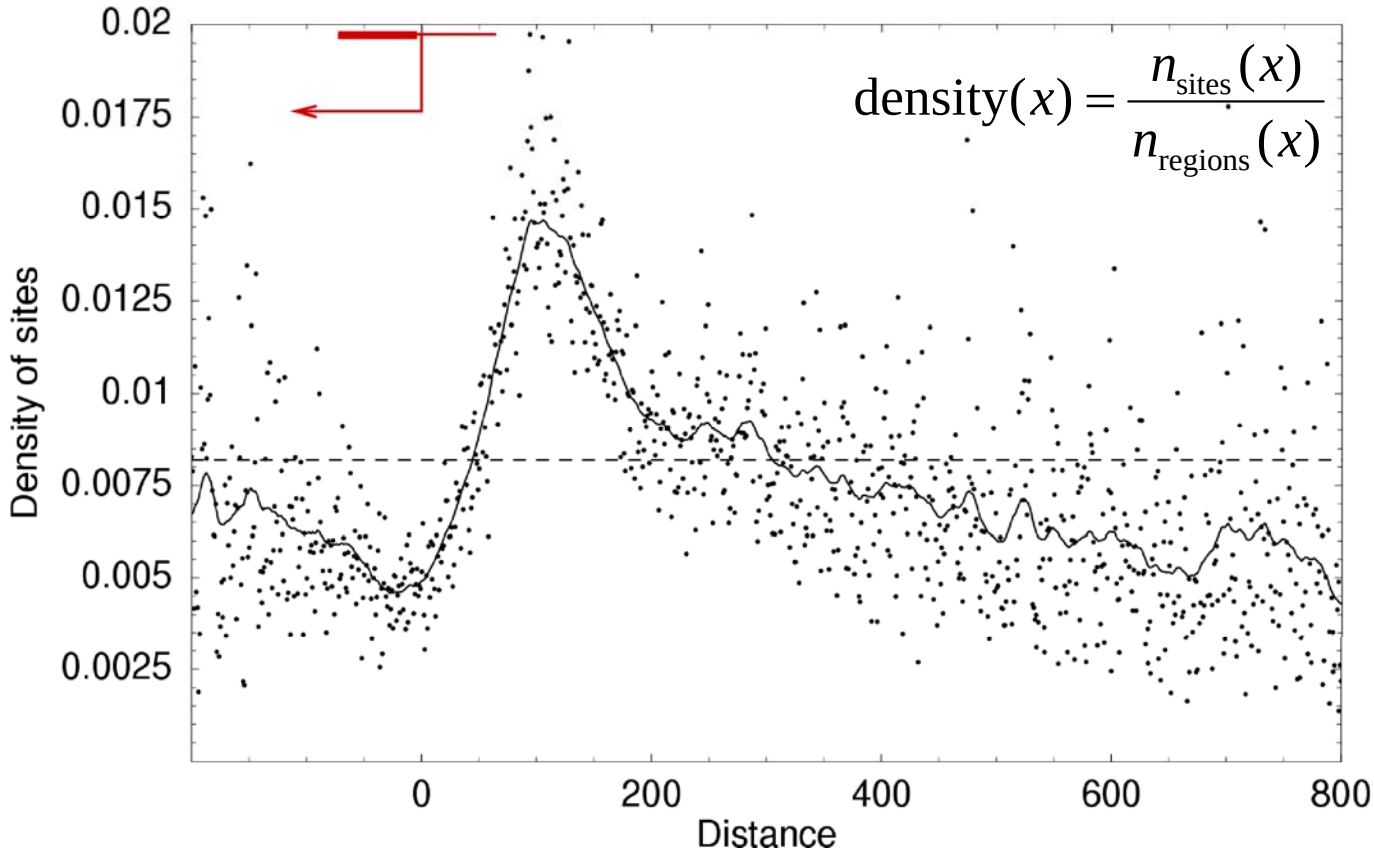
Site Locations: distances to Transcription Start Sites



Sharp peak at 97bps upstream of TSS

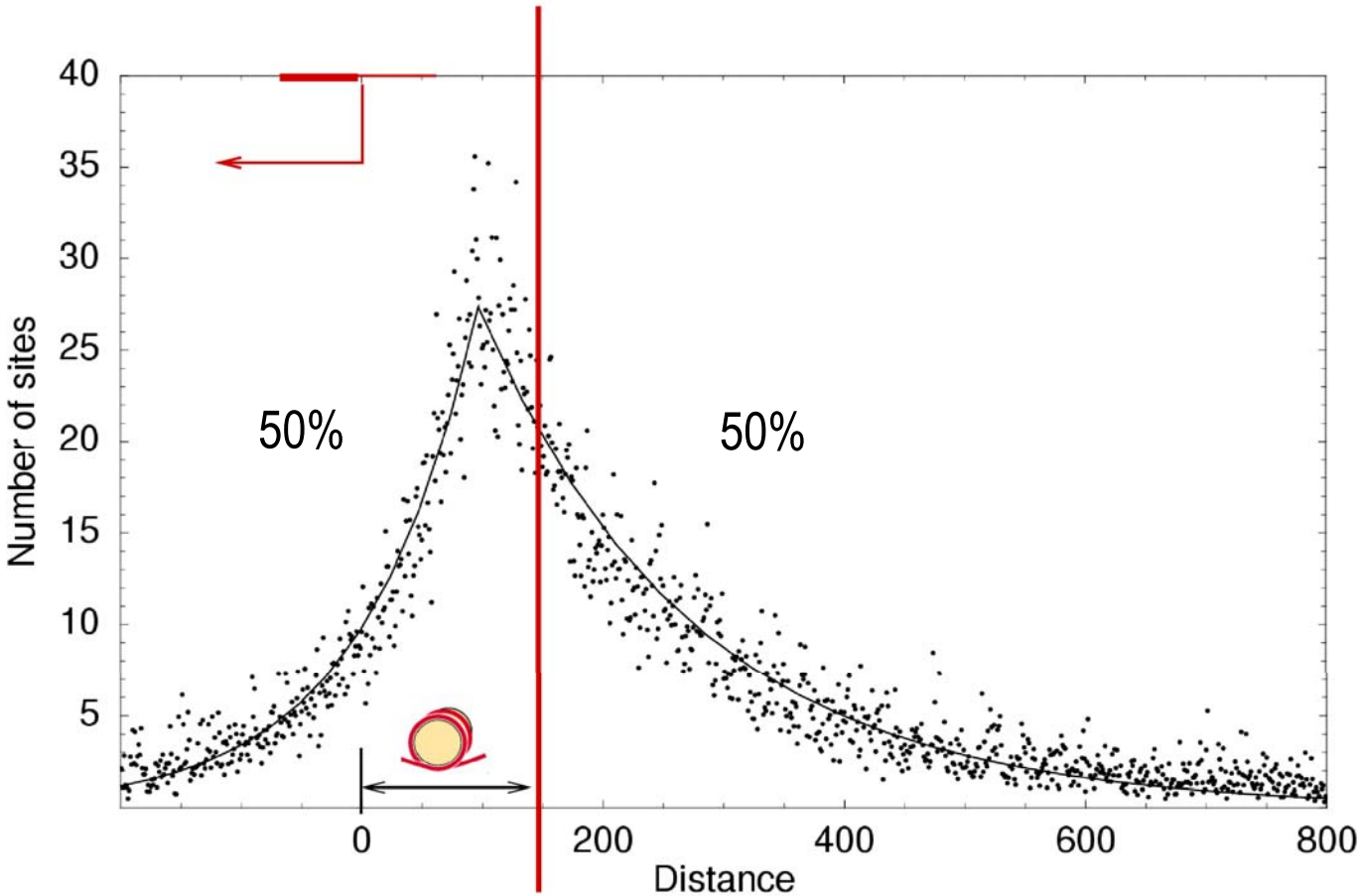
TSS sites from: Zhang Z, Dietrich FS, *Nucl. Acids Res.* (2005)
David L, Huber W, Granovskaia et al. *PNAS* (2006)

Site Locations: distances to Transcription Start Sites



Peak also found in site *density* profile.
Density Minimum around TSS.

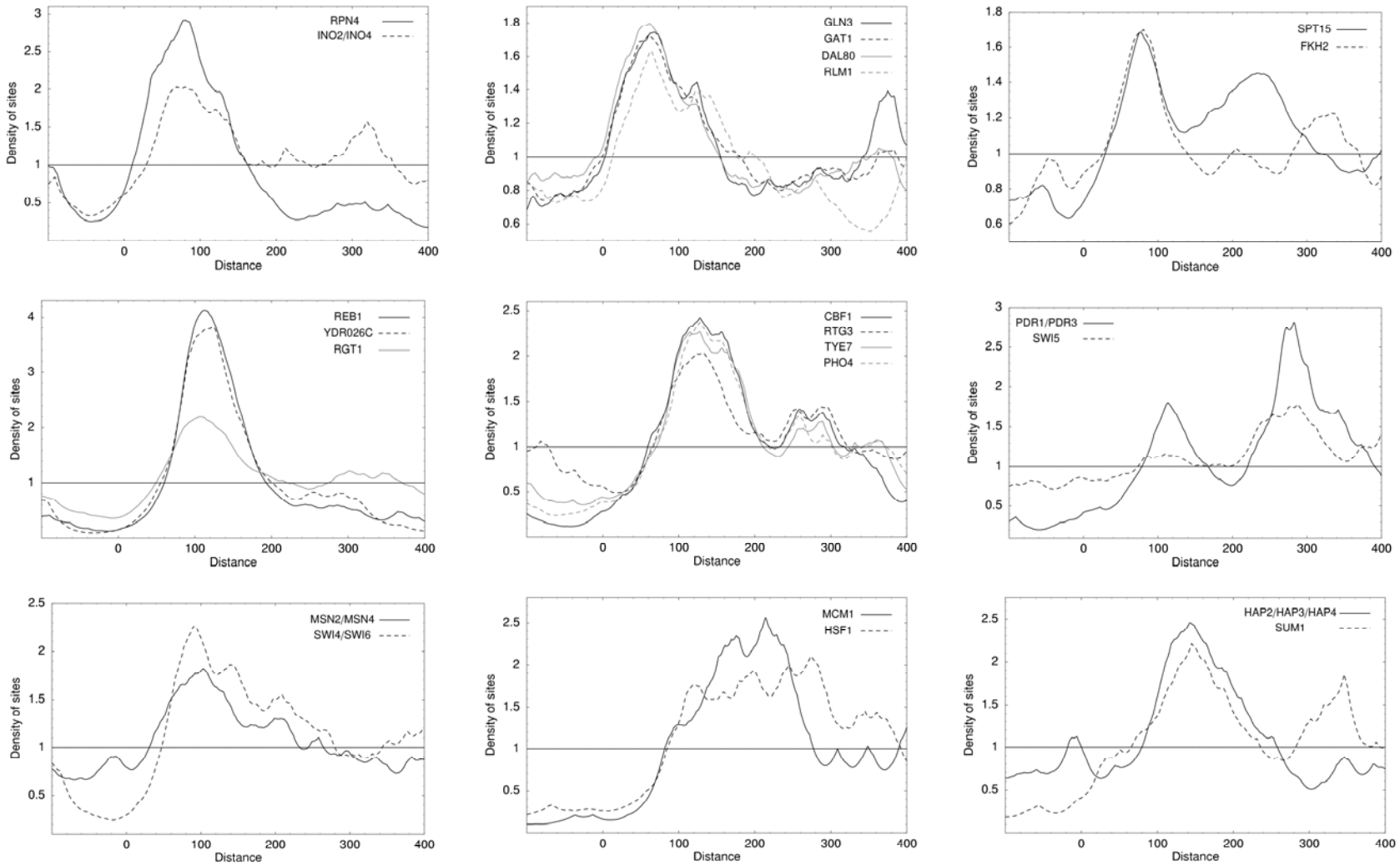
Site Locations:
distances to Transcription Start Sites



Median at 146 bp

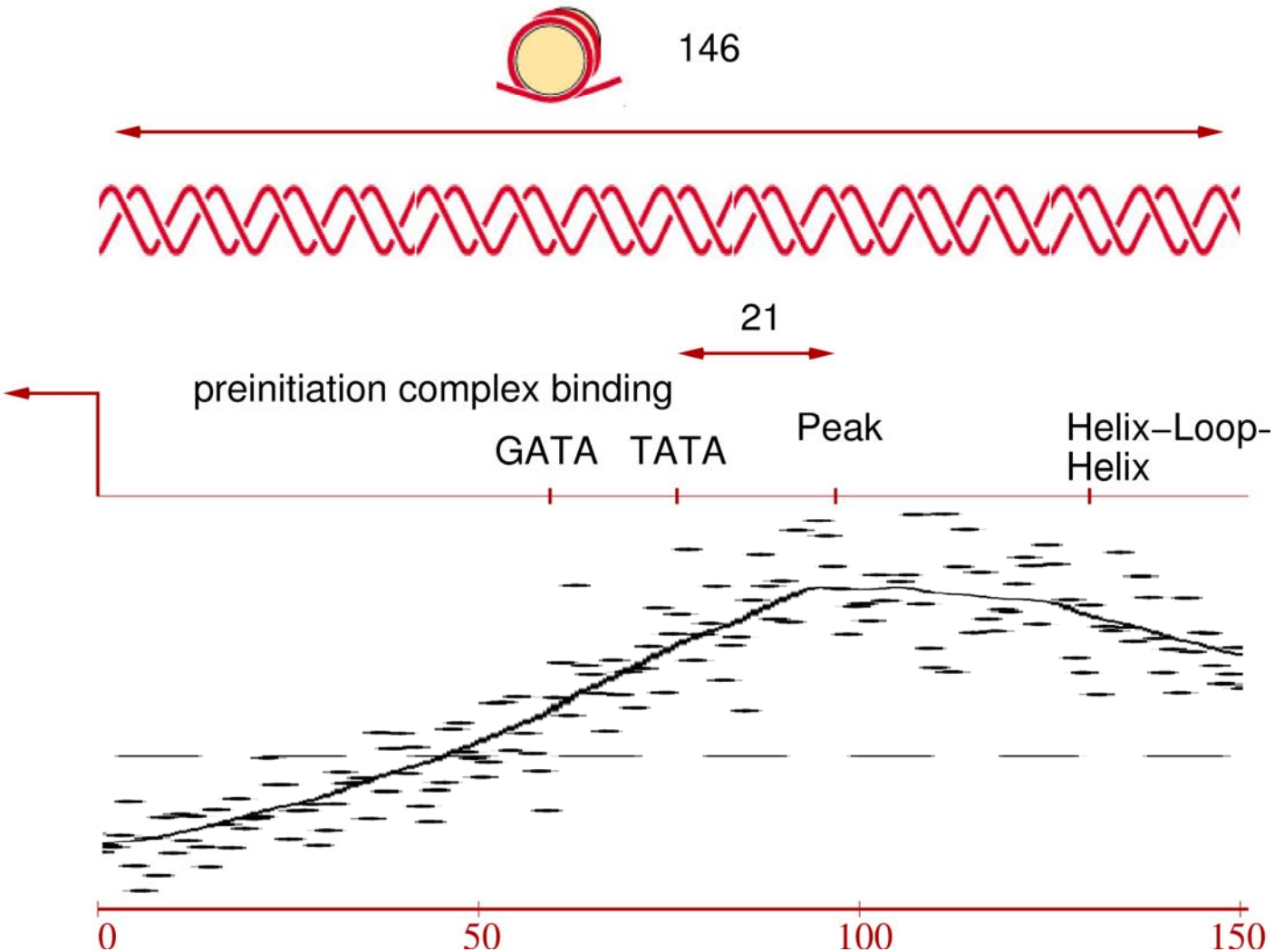
One nucleosome wraps 147bp of DNA

Site Locations:
Distance distributions for individual factors

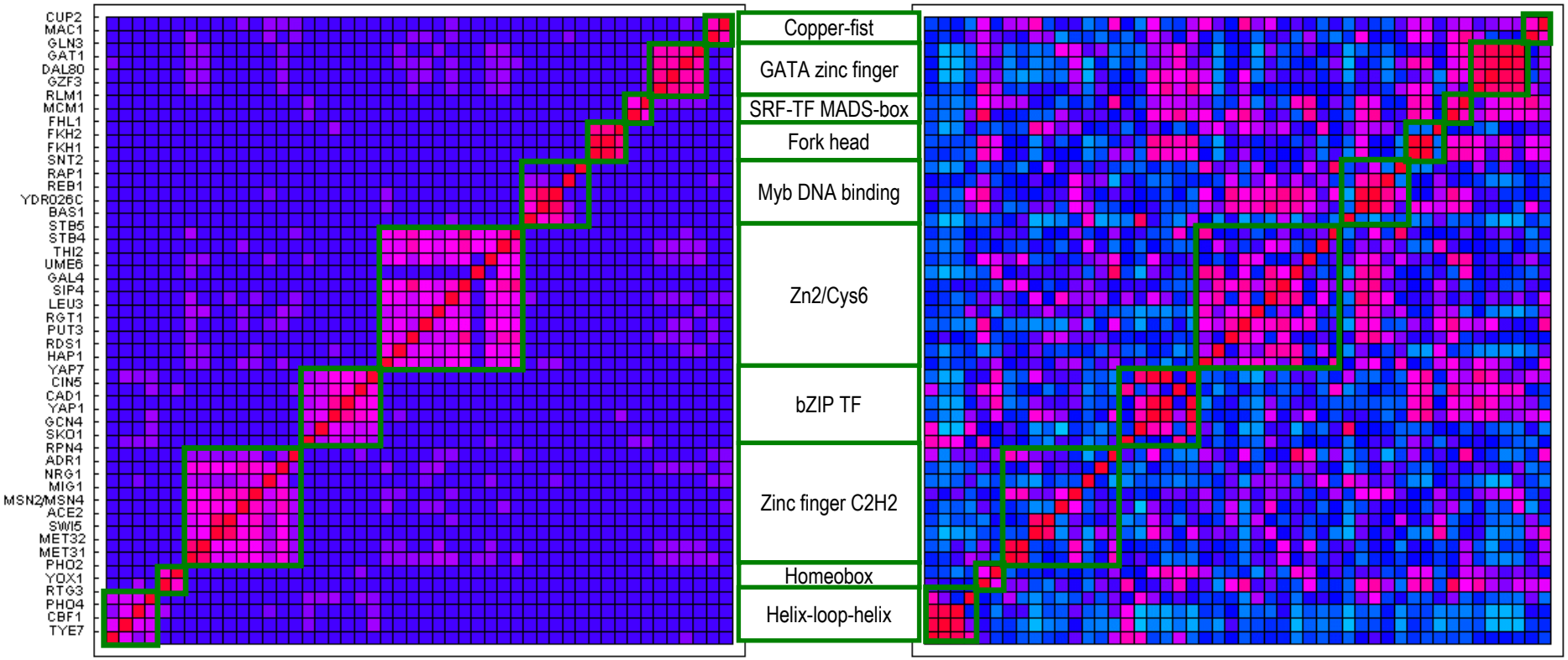


There seem to be classes of TFs with similar density profiles

Site Locations:
Distance distributions for individual factors



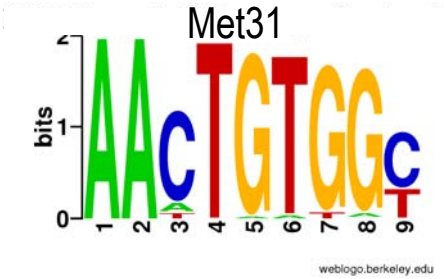
Similar DNA binding domains bind similar motifs

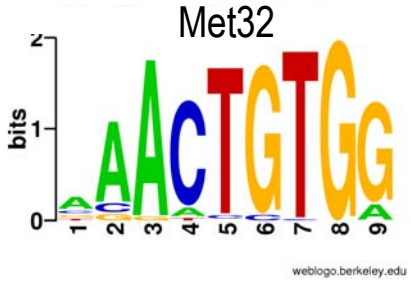


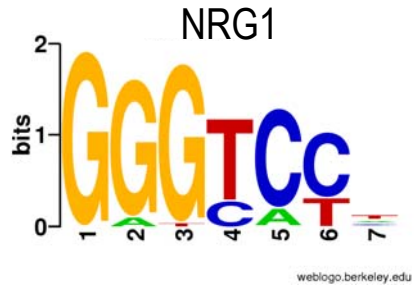
Pairwise similarity of DNA binding domains (BLAST score)

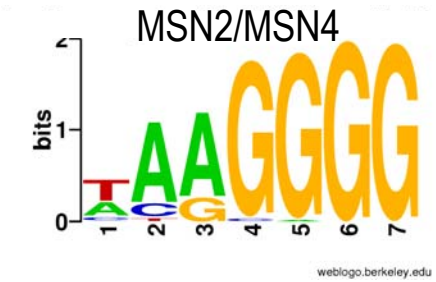
Pairwise similarity of DNA binding motifs.

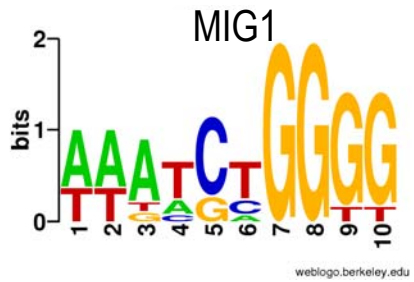
Motifs for Zinc finger-C2H2

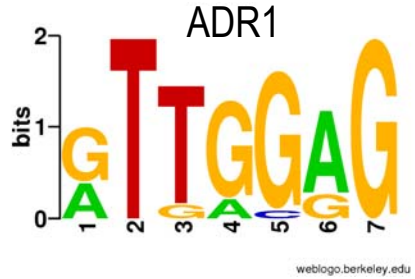


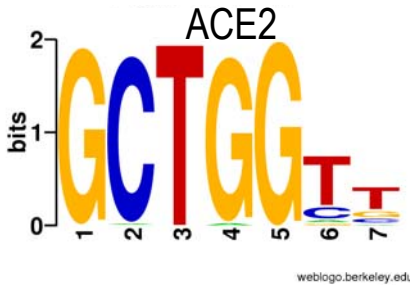


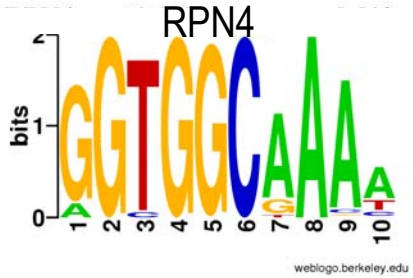












Motifs for this class of TFs have been computationally modeled:

1: PLoS Comput Biol. 2005 Jun;1(1):e1. Epub 2005 Jun 24.

Ab initio prediction of transcription factor targets using structural knowledge.

Kaplan T, Friedman N, Margalit H.

Those that have worked on the presented material

Ionas Erb
MotEvo

Mikhail Pachkov
SwissRegulon

Nacho Molina
Ortholog identification
Phylogeny reconstruction
Multiple alignment

PhyloGibbs

Rahul Siddharthan (IMSc, Chennai, India)
Eric D. Siggia (Rockefeller University, New York, USA)