Motifs

Can (John) Bruce, Ph.D.
Keck Foundation Biotechnology Lab
Bioinformatics Resource

Binding sites of restriction enzymes and transcription factors

- Restriction endonucleases:
  - EcoRI: G|AATTC
- Transcription factors:
  - glucocorticoid response element ("GRE"):

    5'-GGTACAnnnTGTTCT-3'

Two problems in motif analysis

- Given a collection of binding sites, develop a representation of those sites that can be used to search new sites and reliably predict where additional binding sites occur.
- Given a set of sequences known to contain binding sites for a common factor, but not knowing where the sites are, discover the location of the sites in each sequence and a representation of the protein.

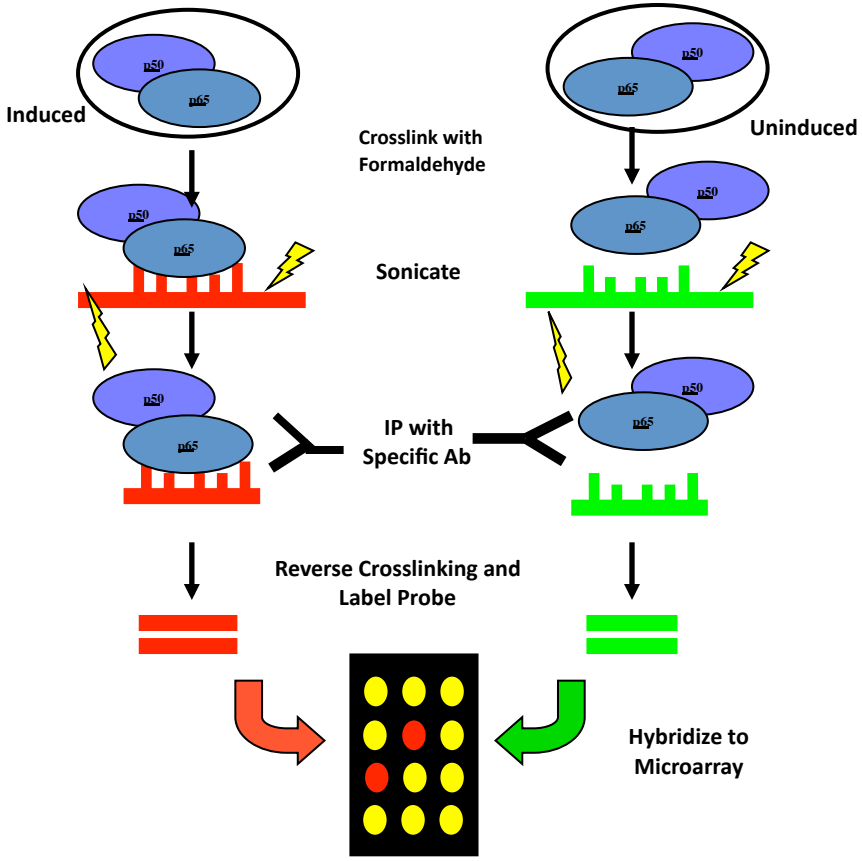
Motif Searching Tools

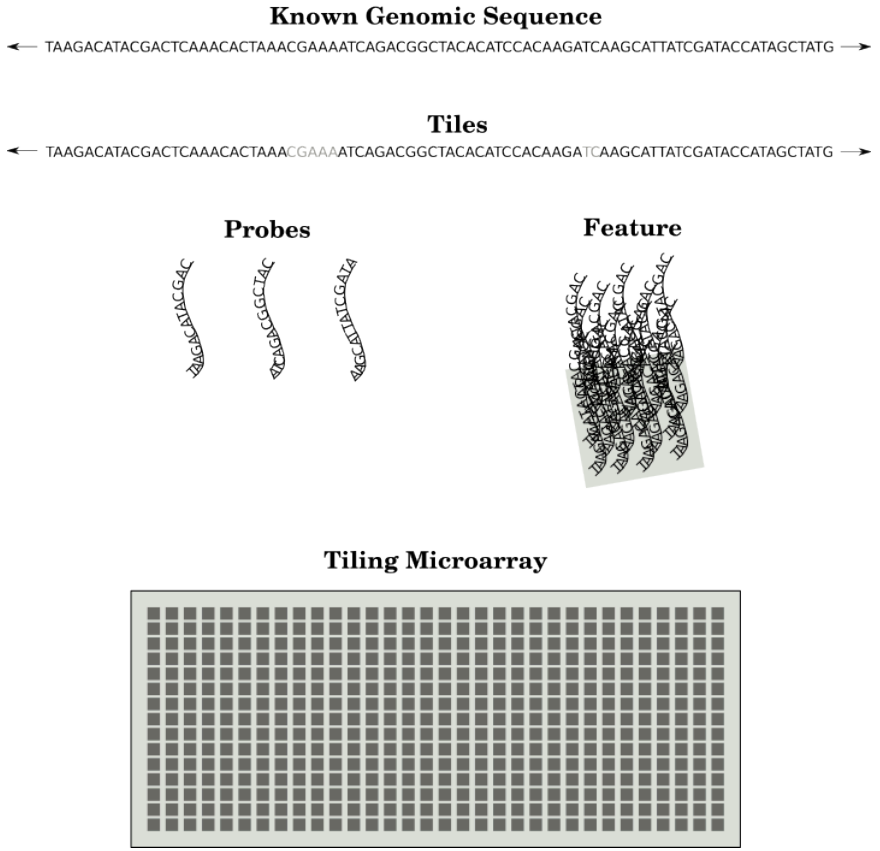
- Searching tools (pattern matching tools) take as input one or more sequences and a pattern. They decide whether the pattern matches the input sequence(s) and if so, where.
- Learning tools.
  - supervised pattern-recognition tool: take as input as set of sequences and discover a pattern that all of the sequences share.
  - unsupervised pattern-recognition tool: take as input as set of sequences and discover a pattern that some of the sequences share.

Finding TF-binding sequences

- ChIP-on-chip or ChIP-seq: Immunoprecipitate DNA-TF complexes, then either hybridize them to a microarray chip or sequence them.
- List promoter regions of co-regulated genes.
- SELEX: Systematic Evolution of Ligands by Exponential Enrichment (or in vitro selection). A library of random oligonucleotides are bound to a purified protein, then the bound ones are identified.

Chromatin Immunoprecipitation and Microarray (ChIP-chip)





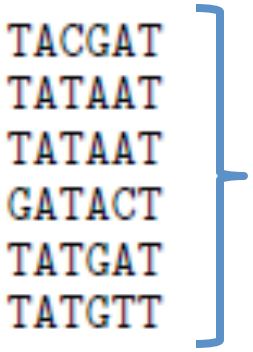
Consensus sequences

- *ATCGATYxxxRATCGAT or ATCGATYxxxxRATCGAT*
- This pattern may be written as a “regular expression”:

  `ATCGAT[TC].{3,4}[GA]ATCGAT`

- …. where ‘.’ means any character, {m,n} means preceding character repeated between m and n times, [ABCD] means one character that can be either A, B, C or D.
- Other conventions exist to represent sequence patterns.

How to define a consensus sequence



The -10 region of six promoters
(Pribnow, 1975)

```
TATAAT  consensus sequence
TATRNT  alternate consensus sequence
```

	With no mismatch		with 1 mismatch		with 2 mismatches	
TATAAT	2/6	1/4000 bp	3/6	1/200 bp	6/6	1/30 bp
TATRNT	4/6	1/200 bp	6/6	1/30 bp		

MATRICES

- A position frequency matrix (PFM) records the position-dependent frequency of each residue or nucleotide. PFMs can be experimentally determined or computationally discovered.
- A position weight matrix (PWM) contains log odds weights for computing a match score. A cutoff is needed to specify whether an input sequence matches the motif or not. PWMs are calculated from PFMs.

Pos	A	C	G	T	IUPAC
01	6	2	8	1	R
02	3	5	9	0	S
03	0	0	0	17	T
04	0	0	17	0	G
05	17	0	0	0	A
06	0	16	0	1	C
07	3	2	3	9	T
08	4	7	2	4	N
09	9	6	1	1	M
10	4	3	7	3	N
11	6	3	1	7	W

IUPAC Code	Meaning
G	G
A	A
T	T
C	C
R	G or A
Y	T or C
M	A or C
K	G or T
S	G or C
W	A or T
H	A or C or T
B	G or T or C
V	G or C or A
D	G or A or T
N	G or A or T or C

A Position Weight Matrix of log odds scores

A	−38	**19**	1	**12**	**10**	−48
C	−15	−38	−8	−10	−3	−32
G	−13	−48	−6	−7	−10	−48
T	**17**	−32	**8**	−9	−6	**19**

Fig. 2. Weight matrix represenation for −10 region of *E.coli* promoters. The boxed elements correspond to the consensus sequence TATAAT.

$m_{i,j} = log(p_{i,j} / b_i)$, where $p_{i,j}$ is the probability of observing symbol i at position j of the motif, and b_i is the probability of observing the symbol i in a background model.

Stormo, 1988

Information content

$$I_i = 2 + \sum_{b=A}^{T} f_{b,i} \log_2 f_{b,i}$$

I_i = Information content at position i; General case, where each base's composition = 25%

$$I_{\text{seq}}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

$I_{seq}(i)$=relative entropy at position i each base's composition = p_0

a

```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG
```

Eight known **ROX1** genomic binding sites in three *S. cerevisiae* genes.

b

```
YCHATTGTTCTC
```

Degenerate consensus sequence.

c

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

Frequencies of nucleotides at each position.

d

Frequencies of nucleotides at each position.

e

Sequence logo showing the frequencies scaled relative to the information content (measure of conservation) at each position.

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

f

Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*.

$$I_{seq}(i) = -\sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

Correlation between binding strength and homology score

Berg and Hippel (1987) showed by using statistical mechanics theory that the log of base frequencies should be proportional to the binding energy contribution of the bases.

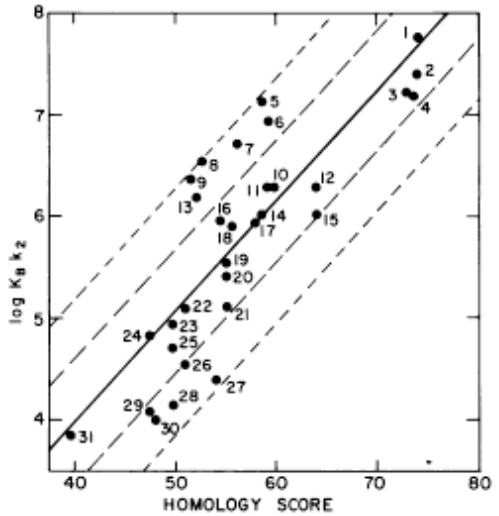


Figure 3. Correlation between log $K_B k_2$ and the homology score calculated for each promoter listed in Table 1. The solid line (linear least squares) has a slope of 0.1086 per homology score % and an intercept of -0.3634 with a correlation coefficient of 0.83. The dashed lines are drawn one (long dashes) and two (short dashes) standard deviations from the best fit line. 20 promoters fall within one standard deviation and eleven were between one and two standard deviations. The value of log $K_B k_2$ at the maximum homology score is 10.5 which would correspond to a value for $K_B k_2$ of 3.15×10^{10} $M^{-1}s^{-1}$.

Mulligan et al., 1984

Independence of bases within motif

- Limitation of position weight matrix is the assumption that the positions in the site contribute additively to the total binding activity.
- Statistical methods (e.g. neural networks) used to identify which pairs of sites are dependent on each other.

Correlated bases

(a)

(b)

1\2	C	G	T
A		0.19	
C		0.16	0.06
G	0.09		
T	0.44	0.06	

Fig. 2. (**a**) Sequence logo plot for the E2F sites predicted by the GMS-MP. The traditional consensus for the E2F motif is the one from positions 2 to 10. (**b**) The joint distribution of the position pair $(1, 2)$, which has been found to be significantly correlated by the GMS-MP.

Zhou and Liu, 2004

De novo discovery of motifs

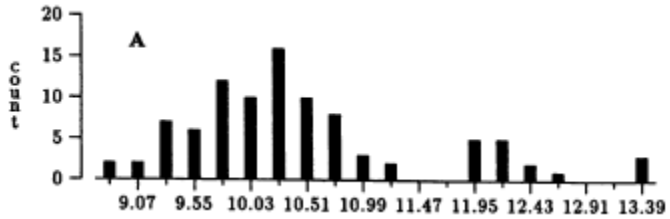
- MEME: uses expectation maximization
- Gibbs Sampler:
- PhyloGibbs: Uses pylogenetic information
- Weeder: Enumerates motifs
- … etc.

DISCOVERING SITES

Stormo and Hartzell, 1989

```
cole1      taatgtttgtgctggtTTTTGTGGCATCGGGCGAGAATagcgcgtggtgtgaaagactgtTTTTTTGATCGTTTTCACAAAAatggaagtccacagtcttgacag
ecoarabop  gacaaaaacgcgtaacAAAAGTGTCTATAATCACGGCAgaaaagtccacattgaTTATTTGCACGGCGTCACACTTtgctatgccatagcatttttatccataag
ecobglr1   acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaatAACTGTGAGCATGGTCATATTTttatcaat
ecocrp     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatGTATGCAAAGGACGTCACATTAccgtgcagtacagttgatagc
ecocya     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaAGGTGTTAAATTGATCACGTTTtagaccatttttcgtcgtgaaactaaaaaaacc
ecodeop    agtgaaTTATTTGAACCAGATCGCATTAcagtgatgcaaacttgtaagtagatttccttAATTGTGATGTGTATCGAAGTGtgttgcggagtagatgttagaata
ecogale    gcgcataaaaaacggctaaattcttgtgtaaacgattccacTAATTTATTCCATGTCACACTTttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr  gctccggcgggggttttttgttatctgcaattcagtacaAAACGTGATCAACCCCTCAATTttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ecolac     aacgcaatTAATGTGAGTTAGCTCACTCATtaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggAATTGTGAGCGGATAACAATTTcac
ecomale    acattaccgccaaTTCTGTAACAGAGATCACACAAagcgacggtgggggcgtagggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk    ggaggaggcgggaggatgagaacacggcTTCTGTGAACTAAACCGAGGTCatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgatttttatgtgcgca
ecomalt    gatcagcgtcgttttaggtgagttgttaataaagatttggAATTGTGACACAGTGCAAATTCagacacatataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa    gctgacaaaaaagattaaacataccttatacaagacttttttttcatATGCCTGACGGAGTTCACACTTgtaagttttcaactacgttgtagactttacatcgcc
ecotnaa    ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacGATTGTGATTCGATTCACATTTaaacaatttcaga
ecouxu1    cccatgagagtgaaatTGTTGTGATGTGGTTAACCCAAttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgCGGTGTGAAATACCGCACAGATgcgtaaggagaaaataccgcatcaggcgctc
trn9cat    CTGTGACGGAAGATCACTTCgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaAAATGAGACGTTGATCGGCACG
(tdc)      gatttttatactttaacttgttgatatttaaaggtatttaattgtaatаacgatactctggaaagtattgaaagttaATTTGTGAGTGGTCGCACATATcctgtt
```

1. Each sequence is 105 bases long and contains at least one CRP site.
2. The 86 20-long words of the first sequence constitute the first PWM. Each of these is compared with each of the 20-long words of the next sequence and the best match to each matrix kept as a two-sequence matrix.
3. Each of those is the compared with the 20-long of the next sequence and the best match to each matrix kept again. Repeat for all 18 sequences.
4. The total number of matrices is 94. Plot histogram of the information contents:



Filtering background sequences

- Many yeast promoters have unxpectedly common stretches of poly(A) and poly(T) sequences, and these can appear to be patterns sought by the motif searching program. But these patterns occur in many promoters, not just the ones that are co-regulated.
- In such cases, one approach is to identify the weight matrix that maximizes the probability of binding to the promoters in the collection, given the background of actual competing sites in the genome.

Two classes of motif discovery algorithms

- Multiple alignment methods.
  - Return PWM; use local search techniques such as Gibbs sampling or EM
- Deterministic combinatorial algorithms based on word frequency counts.
  - Search for various sized sequences exhaustively and evaluate significance.

Enumerative techniques

- dictionary-based methods count the number of occurrences of all n-mers in the target sequences, and calculate which ones are most overrepresented.
- a number of similar overrepresented words may be combined into a more flexible motif description.
- Alternatively, one can search the space of all degenerate consensus sequences up to a given length, for example, using IUPAC codes for 2-nucleotide or 3-nucleotide degenerate positions in the motif
- WEEDER describes a motif as a consensus sequence and an allowed number of mismatches, and uses an efficient suffix tree representation to find all such motifs in the target sequences

IUPAC Code	Meaning
G	G
A	A
T	T
C	C
R	G or A
Y	T or C
M	A or C
K	G or T
S	G or C
W	A or T
H	A or C or T
B	G or T or C
V	G or C or A
D	G or A or T
N	G or A or T or C

Consensus-based methods

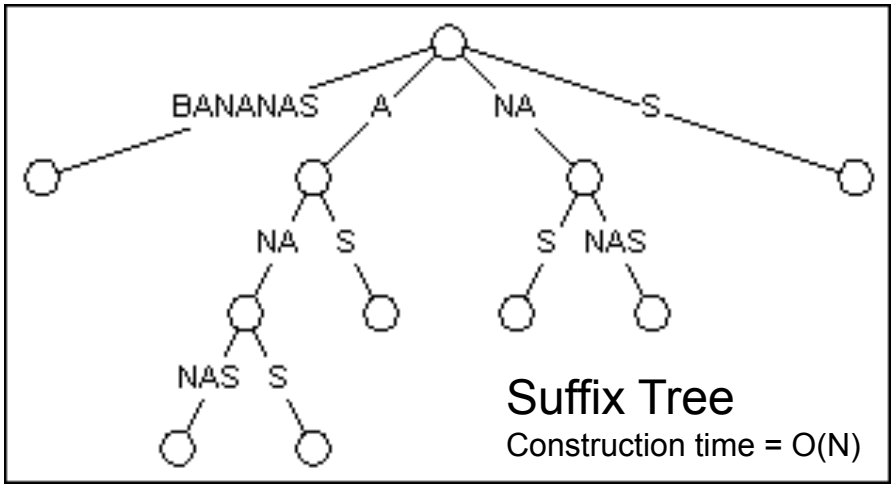
- Enumerate all the oligos of (or up to) a given length, in order to determine which ones appear, with possible substitutions, in a significant fraction of the input sequences, and finally to rank them according to statistical measure of significance.
- Drawbacks:
  - For motif length of m, there are 4^m candidates to enumerate. $O(4^m)$ execution time.
  - Too slow.
- Motif search can be accelerated by pre-processing the data in an indexing structure, such as a suffix tree.

Weeder

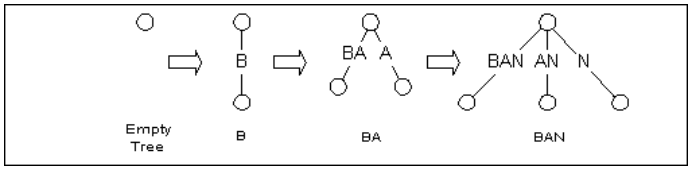
- Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurences (with substitutions) from input sequences.

Suffix Tree

- BANANAS
- ANANAS
- NANAS
- ANAS
- NAS
- AS
- S



• Starting from the root of the tree, each of the sufixes of BANANAS is found in the trie.
• Because of this organization, you can search for any substring of the word by starting at the root and following matches down the tree until exhausted.



Construction of a suffix tree

Probabilistic Approaches

- Expectation Maximization: Search the PWM space randomly
- Gibbs sampling: Search sequence space randomly.

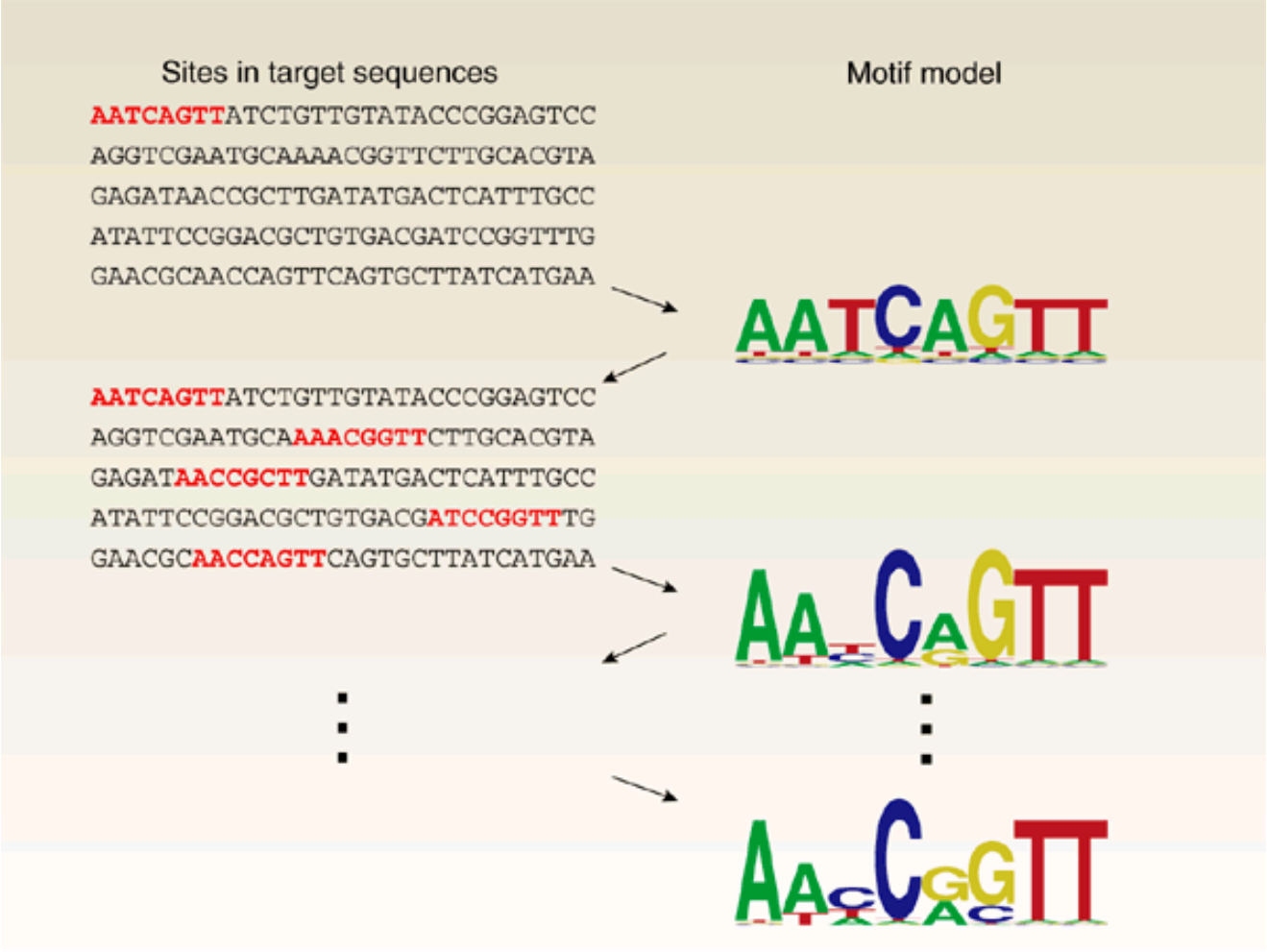
Expectation-Maximization (EM) algorithm

- Used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.
- EM alternates between performing
  - an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and
  - a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E step.
- The parameters found on the M step are then used to begin another E step, and the process is repeated.

Expectation Maximization method in motif finding

- The weight matrix for the motif is initialized with a single n-mer subsequence, plus a small amount of background nucleotide frequencies.
- Next, for each n-mer in the target sequences, we calculate the probability that it was generated by the motif, rather than by the background sequence distribution.
- Expectation maximization then takes a weighted average across these probabilities to generate a more refined motif model.
- The algorithm iterates between calculating the probability of each site based on the current motif model, and calculating a new motif model based on the probabilities.
- It can be shown that this procedure performs a gradient descent, converging to a maximum of the log likelihood of the resulting model.

How does EM algorithms work?



Starting from a single site, expectation maximization algorithms such as MEME[4] alternate between assigning sites to a motif (left) and updating the motif model (right). Note that only the best hit per sequence is shown here, although lesser hits in the same sequence can have an effect as well.

A sample problem for EM

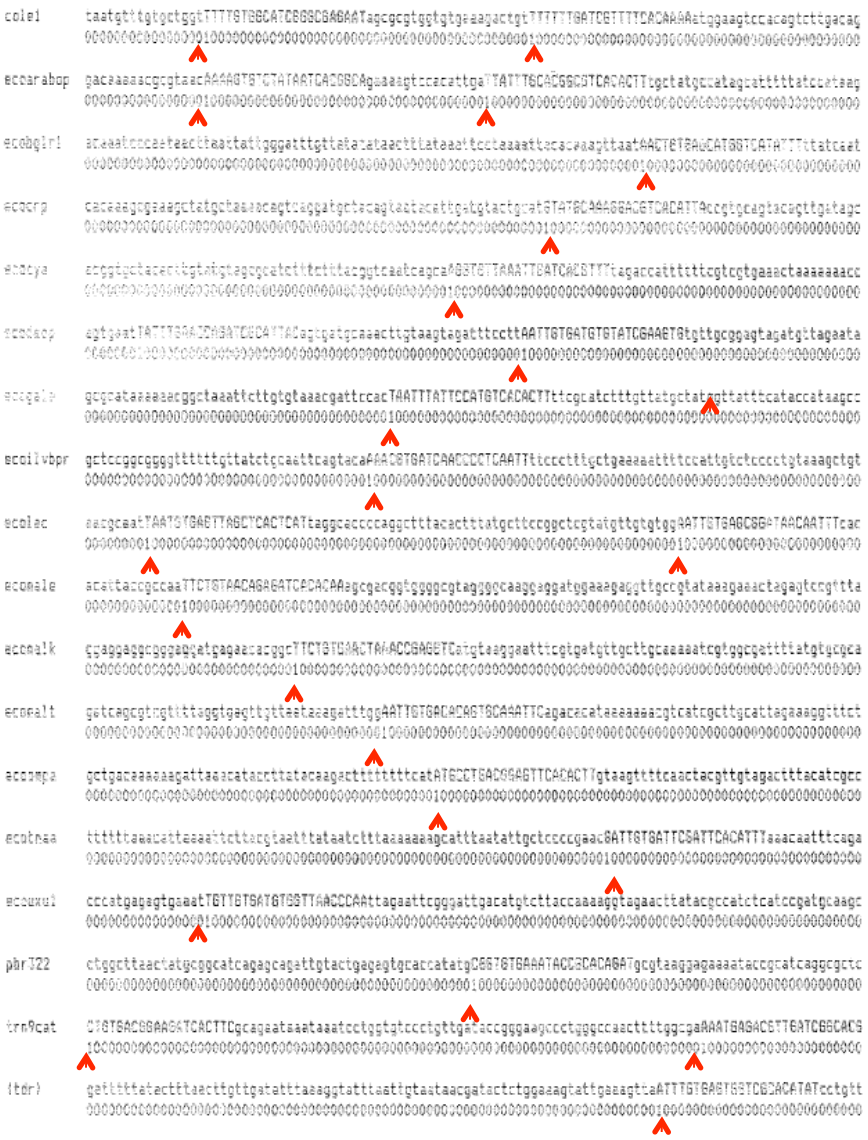
We know neither the PWM not the motif positions

Footprint Sites

Col E1 site 2	T T T T T T G A T C G T T T T C A C A A A A
Col E1 site 1	T T T T G T G G C A T C G G G C G A G A A T
ara site 2	T T A T T T G C A C G G C G T C A C A C T T
ara site 1	A A A A G T G T C T A T A A T C A C G G C A
Bgl R mut1	A A C T G T G A G C A T G G T C A T A T T T
crp	G T A T G C A A A G G A C G T C A C A T T A
cya	A G G T G T T A A A T T G A T C A C G T T T
deo P2 site 2	T T A T T T G A A C C A G A T C G C A T T A
deo P2 site 1	A A T T G T G A T G T G T A T C G A A G T G
gal	T A A T T T A T T C C A T G T C A C A C T T
ilv B	A A A C G T G A T C A A C C C C T C A A T T
lac site 2	T A A T G T G A G T T A G C T C A C T C A T
lac site 1	G A A T T G T G A G C G G A T A A C A A T T
mal E	T T C T G T A A C A G A G A T C A C A C A A
mal K	T T C T G T G A A C T A A A C C G A G G T C
mal T	A A T T G T G A C A C A G T G C A A A T T C
omp A	A T G C C T G A C G G A G T T C A C A C T T
tna A	G A T T G T G A T T C G A T T C A C A T T T
uxu AB	T G T T G T G A T G T G G T T A A C C C A A
Pbr P4	C G G T G T G A A A T A C C G C A C A G A T
cat	A A A A T G A G A C G T T G A T C G G C A C

A) Base frequencies in footprint sites

A	8	10	9	2	0	0	4	15	8	5	3	10	3	7	1	2	15	4	14	4	7	6
C	1	0	3	2	1	1	0	1	5	8	5	1	4	3	2	18	1	15	1	7	1	3
G	3	3	3	0	14	2	15	3	2	5	6	5	10	6	3	0	4	1	5	4	0	1
T	9	8	6	17	6	18	2	2	6	3	7	5	4	5	15	1	1	1	1	6	13	11



Lawrence and Reilly, 1990

E step

$$\log L = N\sum_{j=1}^{J}\sum_{b=A}^{T} f_{b,j}\log_e(\rho_{b,j}) +$$

$$N(L-J)\sum_{b=A}^{T} f_{b,0}\log_e(\rho_{b,0}),$$

•Let's say we are at the q^{th} iteration. From the previous steps, we have an estimate of the population frequency estimates.

• We calculate the probability of observing the data in each sequence assuming the site starts in each of the possible L-J+1 positions.

•Using these probabilities as weights, add across the positions to find the expected number of the bases at each position in the site.

•E.g., assume that there is an A in the 1^{st} position of the window that starts at position 50 in the third sequence. If the probability that the site starts at position 50 in the third sequence is 0.01: add 0.01 A's to the accumulating expected number of As in the first position of the site.

M step

- Maximum likelihood estimates for the population frequencies are just the sample frequencies when complete data is available.
- Substitute into equation the expected number of bases for each position in the site from the E step for the (unavailable) observed number of bases.

$$\hat{\rho}_{b,0} = f_{b,0} = n_{b,0}/(N[L - J])$$

$$\hat{\rho}_{b,j} = f_{b,j} = n_{b,j}/N.$$

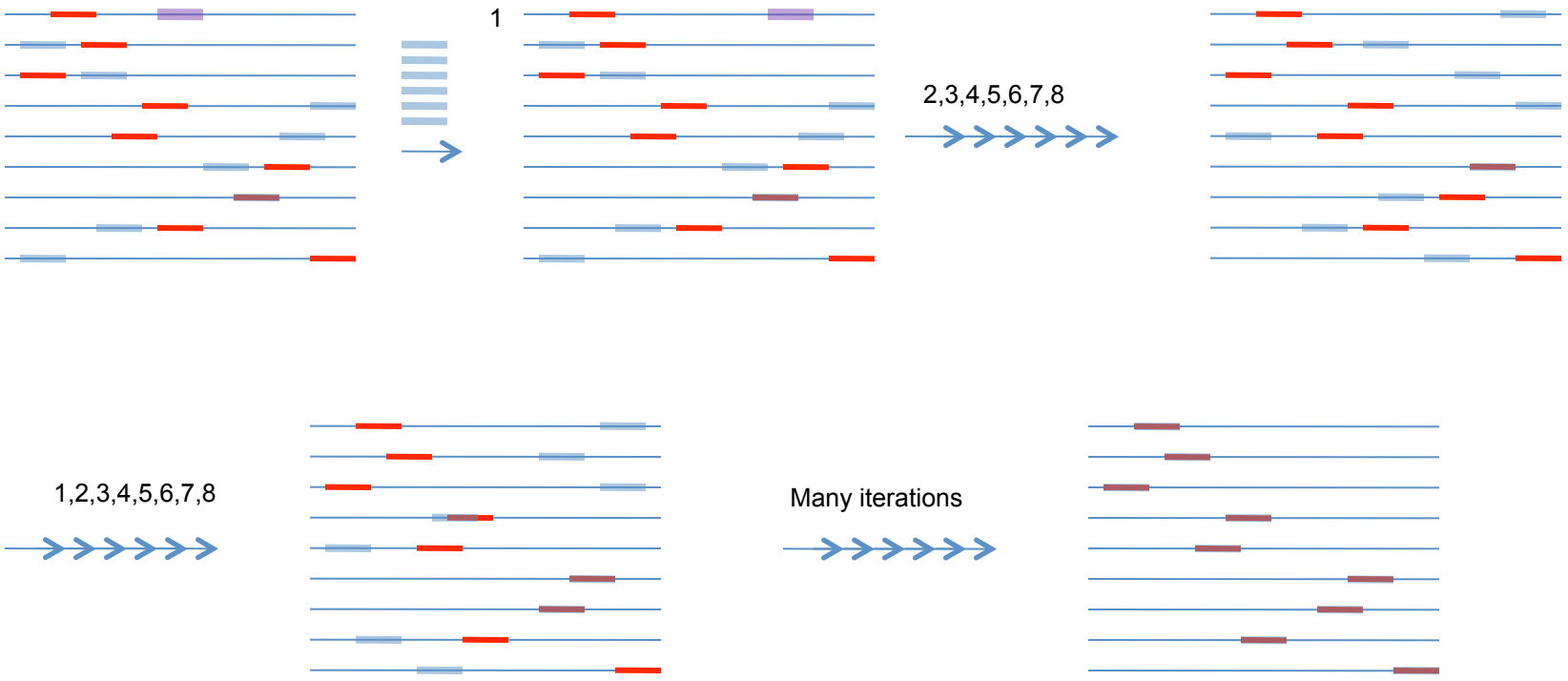
MEME

- Subsequences which occur in the input DNA sequence are used as the starting points from which EM converges iteratively to locally optimal motifs. This increases the likelihood of finding globally optimal motifs.
- Multiple occurrences of a motif are allowed. Algorithm is allowed to ignore sequences with no appearance of a shared motif. So, more resistance to noisy data.
- Motifs are probabilistically erased after they are found, so more than one motif can be found.

Gibbs sampling

- An algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables. The purpose of such a sequence is to approximate the joint distribution, or to compute an integral (such as an expected value).
- Gibbs sampling is applicable when the joint distribution is not known explicitly, but the conditional distribution of each variable is known. The Gibbs sampling algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables.

Gibbs sampling illustration



Gibbs sampler (1)

- We have N sequences S_1, …, S_N and we seek within each sequence mutually similar segments of width W.
- The algorithm maintains two evolving data structures:
  - a PWM consisting of variables $q_{i,A}$, … $q_{i,T}$ and a probabilistic description of "background frequencies" p_A, … , p_T .
  - the alignment, a set of positions a_k, for k from 1 to N.
- The objective is to identify the most probable common pattern. This is obtained by locating the alignment that maximizes the ratio of the corresponding pattern probability to background probability.

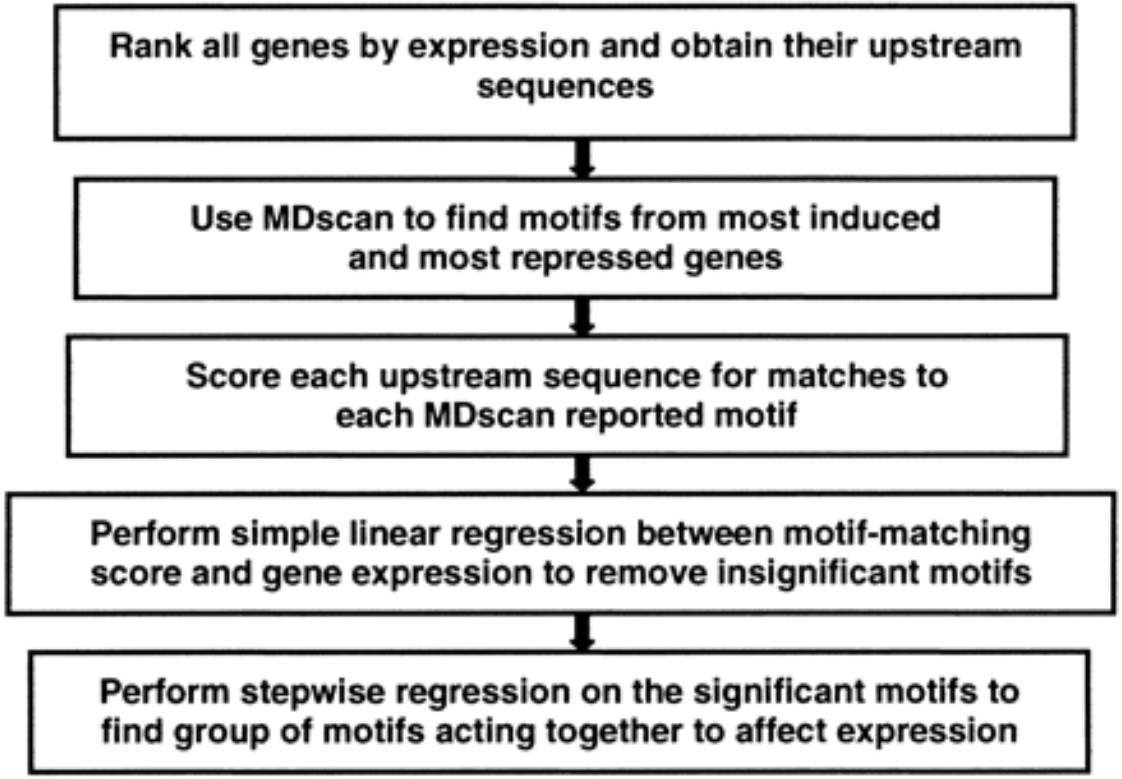
Lawrence et al., 1993

Gibbs Sampler (2)

- Initiate algorithm by choosing random starting positions within the various sequences. It then proceeeds through any iterations to execute the following two steps
  1. One of the N sequences, z is chosen either at random or in specified order. The patterns description and background frequencies are then calculated from the current positions a_k in all sequences k in 1 … N excluding z.
  2. Sampling step. Every segment of width W within sequence z is considered as a possible instance of the pattern.
     - The probabilities Q_x of generating each segment x according to the current pattern probabilities $q_{i,j}$ are calculated, as are the probabilities P_x of generating these segments by the background probabilities p_j.
     - The weight $A_x=Q_x/P_x$ is assigned to segment x, and with each segment so weighted, a random one selected. Its position becomes the new a_k.
- The more accurate the determination of its location in step #1, the more accurate the determination of this location in step #2. Once some correct a_k have been selected by chance, $q_{i,j}$ begin to reflect, albeit imperfectly, a pattern present in other sequences.

INTERPRETING THE BIOLOGICAL ROLE OF MOTIFS

Regression-based techniques to identify motifs



$$S_{mg} = \log_2 \left[\sum_{x \in X_{wg}} \Pr(x \text{ from } \theta_m) / \Pr(x \text{ from } \theta_0) \right]$$

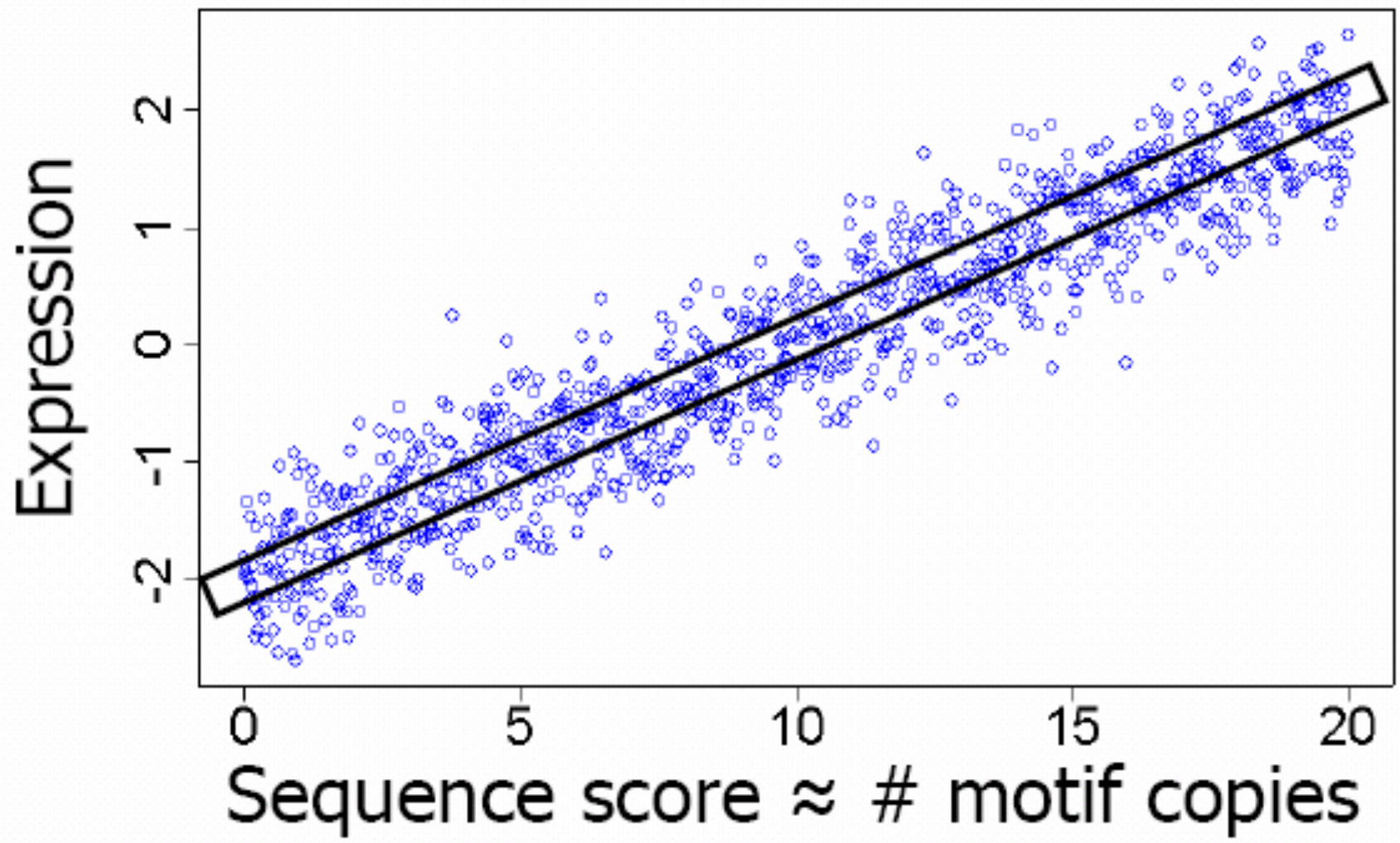
$$Y_g = \alpha + \beta_m S_{mg} + \varepsilon_g$$

Conlon et al. 2003

MDscan Motif Finding Algorithm

- Uses 100 highest expressed genes, finds 30 candidate motifs for each width [5,15]
- Confirms motifs using 500 highest expressed genes
- Repeat for lowest expressed genes

Single Motif Regression



Linear Regression Model

For each motif:

$$Y_g = \alpha + \beta_m S_{mg} + e_g$$

where

Y_g = $\log_2$-ratio of expression

β_m = regression coefficient

S_{mg} = sequence score

e_g = error

Over-expressing a Transcription Factor

Known binding site: TCTATTGTTT

Motif Regressor (p-value)	AlignAce	MEME
TCTATTGTT (<1e-16)	AAAAAAAAAAAAAAAAAAAAAAA	TTTTTTCTTTT
TT**TCTATTGT** (<1e-16)	AAAAAAAAAAAG	TTCCGCGGA
CTATTGTTTTC (<1e-16)	AAGGAAAAAAAGAAAAAAAAA	
ACT**TCTATTGT** (<1e-16)	AAAAAAAAGAAAAGAAAAAAA	
TT**TCTATTGTTT**T (<1e-16)	AAGGAAAAAAGAAA	
TT**TCTATTGTTT**TT (<1e-16)	AAGAAAAAAA	
CTATTGTT (1.11e-16)	GCGCCCCGGA	
ATTGT (1.20e-14)	GAGCGCTCATGCCGCTGTTTT	
GGTGGC (1.38e-11)	AAAATAAAAAAAAAAAAA	
TATTGTT (1.04e-10)	CTGCGGAAAA	

Erin M. Conlon

Multiple Regression Model to Determine Motifs Working Together

$$Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + e_g$$

where

Y_g = $\log_2$-ratio of expression

β_m = regression coefficient

S_{mg} = sequence score

M = subset of significant motifs

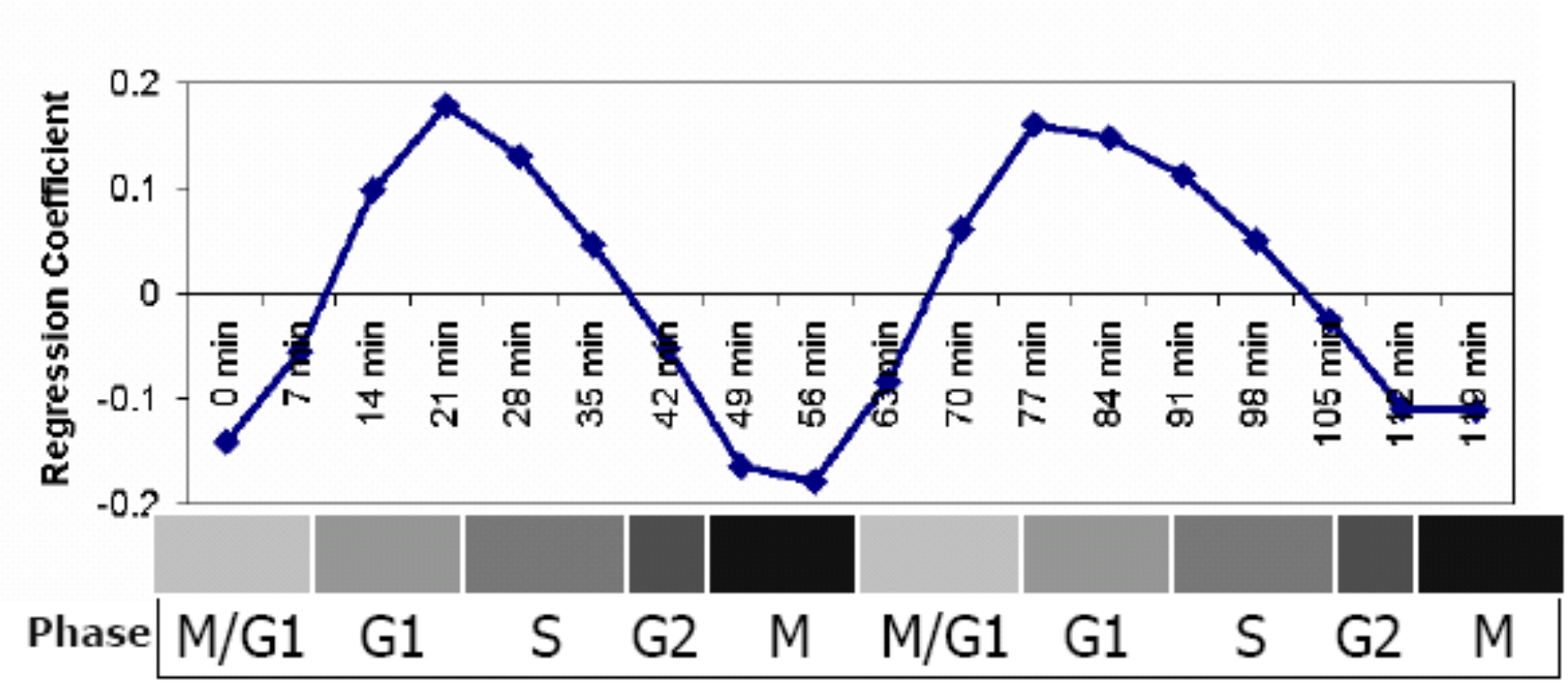
e_g = error

Multiple Motifs Influencing Expression

Motif #	Motif sequence logo	Known motif	Motif coefficient
1		MET4	0.107
2		PHO4	0.088
3, 11, 20		M3A	-0.09
4			-0.08
5, 22		RAP1	-0.1
6, 13, 18		STRE	0.084
7			-0.064
8, 21			-0.072
9, 19			0.068
10			-0.057
12			0.045
14, 15, 23		GCN4	0.059
16			-0.056
17		URS1	0.057
24, 25		M3B	-0.08

Erin M. Conlon

Motif: ACGCGTCGCG



Erin M. Conlon

WHICH ALGORITHM TO USE?

Some motif search tools

Multi-purpose packages

TAMO	TAMO integrates several motif discovery programs. It includes support for motif scanning, scoring, evaluation of statistical significance, clustering, comparison, input/output, conversion between different motif representations, and visualization. http://fraenkel.mit.edu/webtamo/
BEST	BEST is a suite of four motif discovery tools integrated in a graphical user interface. BEST incorporates the BioOptimizer tool used to rank and improve the predictive power of the discovered motifs. http://webster.cs.uga.edu/~che/BEST/
TOUCAN2	TOUCAN2 provides an interface to the Ensembl and EMBL databases of sequence and annotation. It incorporates tools for sequence alignment, motif discovery, and scanning. http://homes.esat.kuleuven.be/~saerts/software/toucan.php
Expander	Expander is a tool for analyzing expression data. It can cluster genes, identify over-represented functional categories in clusters, and scan corresponding promoter regions for motifs. http://www.cs.tau.ac.il/~rshamir/expander/
MDScan **BioProspector** **Compare-Prospector**	MDScan uses ChIP-chip enrichment ratio data to help the motif search. BioProspector is a Gibb's sampling program. CompareProspector incorporates comparative genomics, biasing the search to regions of high conservation. http://seqmotifs.stanford.edu
Consensus **PhyloCon**	The Consensus program finds motifs in a set of unaligned sequences. PhyloCon builds on this framework by modeling conservation across orthologous genes from multiple species. http://ural.wustl.edu/
Weeder	An enumerative motif discovery program that performed well in a recent comparative analysis of fourteen algorithms. http://www.pesolelab.it/
MEME	The popular EM-based motif discovery program. Part of the MEME/MAST system for motif discovery and search. http://meme.sdsc.edu/meme/intro.html
AlignACE	A Gibbs sampling algorithm that can identify multiple motifs in a sequence set using an iterative masking procedure. http://atlas.med.harvard.edu/

Motif Scanning

Ahab	The Ahab webserver allows users to scan for motifs in a set of sequences. Motifs may be user-specified or selected from a database of pre-defined matrices. http://gaspard.bio.nyu.edu/Ahab.html
Clover	Clover identifies overrepresented motifs in a set of sequences, based on a pre-compiled library of motif matrices. http://zlab.bu.edu/clover/
MAST	MAST allows users to scan sequence databases for matches to motifs. It produces detailed annotations and figures for matches in the input sequences. http://meme.sdsc.edu/meme/intro.html
Monkey	Monkey analyzes multiple sequence alignments to identify evolutionarily conserved matches to a motif. http://rana.lbl.gov/~alan/Monkey.htm

Motif Discovery Programs

Databases

cisRED **ORegAnno**	cisRED is a database of conserved motifs and motif patterns obtained by genome scale motif discovery. ORegAnno is a database of regulatory sites curated from the scientific literature. http://www.cisred.org/ http://www.oreganno.org/
UCSC Genome Browser	Online repository of genomic sequence, multiple sequence alignments, and annotation data. The browser includes tracks for identifying conserved transcription factor binding sites. http://genome.ucsc.edu/
ENSEMBL	Another online genomic sequence repository. Includes online tools for data mining as well as BLAST searches. http://www.ensembl.org/index.html
TRANSFAC	Commercial database of transcription factors, binding sites, and motifs. Includes several tools for motif scanning in sequence. http://www.gene-regulation.com/
JASPAR	Curated public database of transcription factor binding specificities represented as PWMs. http://jaspar.cgb.ki.se/

An assessment of motif discovery tools

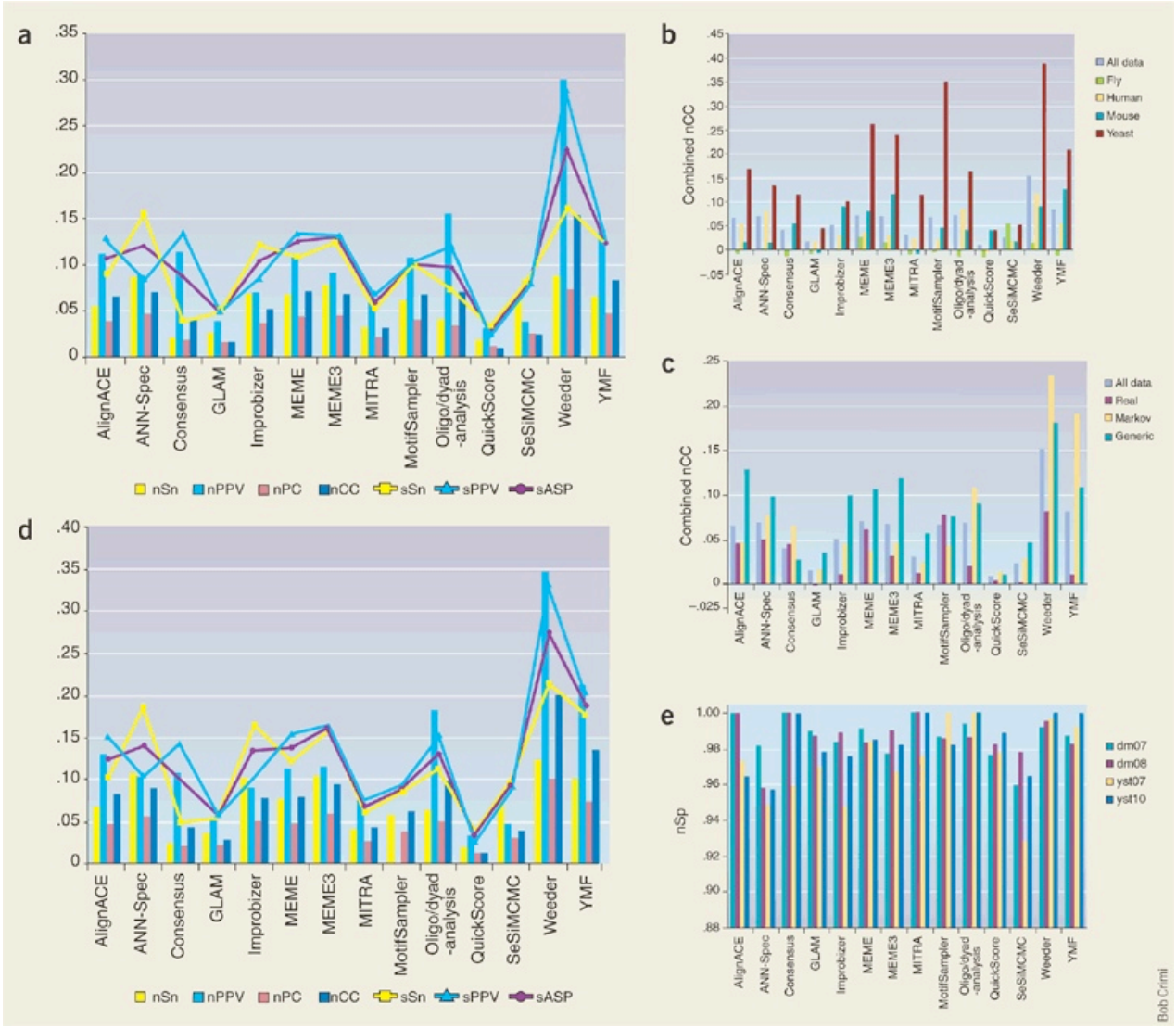


Figure 1. Representative statistics comparing the accuracy of the 13 tools assessed in this analysis. (**a**) Combined measures of correctness over all 56 data sets. See Tables 1 and 2 for details on the individual tools, Methods section for an explanation of data set types and Box 2 for definitions of all statistics. (**b**) Correlation coefficient (*nCC*) by species. (**c**) Correlation coefficient (*nCC*) by data set type. (**d**) Combined measures of correctness over generic and Markov data sets. (**e**) Specificity (*nSp*) on four negative control data sets.

Tompa et al., 2005

Pairs of motif finding tools work better than individual ones.

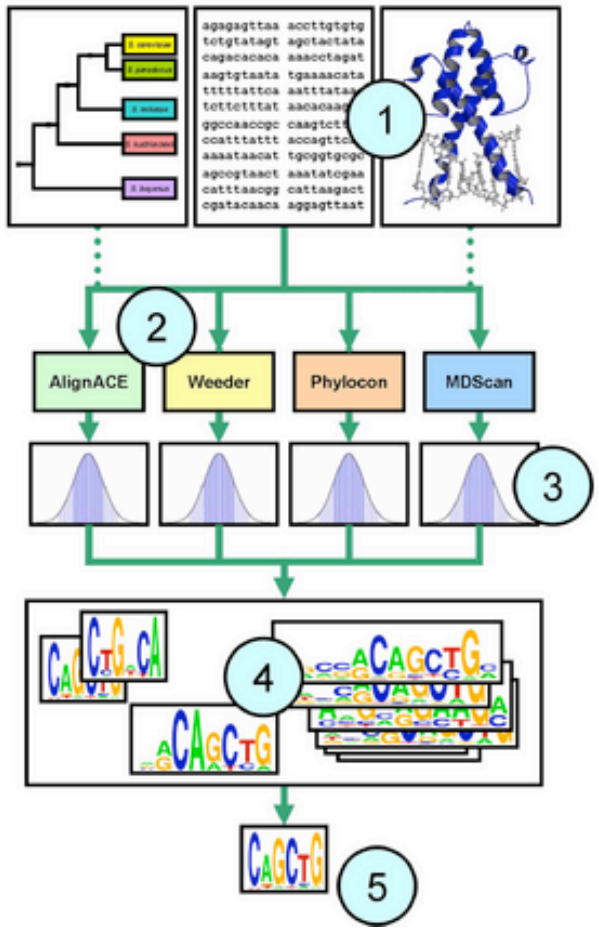
Table 3 Correlation coefficient (*nCC*) for all pairs of tools[a]

	Quick score	GLAM	SeSi MCMC	MITRA	Consen	Improb	Align ACE	Motif sampler	MEME3	MEME	Oligo/ dyad	ANN-Spec	YMF	Weeder
QuickScore	0.009	0.020	0.042	0.030	0.025	0.052	0.068	0.072	0.072	0.074	0.038	0.064	0.061	0.084
GLAM	0.031	0.016	0.060	0.037	0.039	0.068	0.066	0.084	0.088	0.086	0.052	0.082	0.090	0.113
SeSiMCMC	0.049	0.059	0.024	0.068	0.042	0.083	0.071	0.091	0.081	0.088	0.058	0.103	0.104	0.092
MITRA	0.042	0.041	0.072	0.031	0.054	0.082	0.084	0.097	0.106	0.105	0.070	0.101	0.103	0.131
Consensus	0.067	0.060	0.075	0.053	0.042	0.077	0.079	0.109	0.084	0.077	0.074	0.082	0.081	0.098
Improbizer	0.065	0.069	0.083	0.077	0.056	0.052	0.089	0.117	0.096	0.098	0.083	0.112	0.091	0.117
AlignACE	0.088	0.084	0.089	0.090	0.085	0.111	0.068	0.097	0.102	0.091	0.088	0.091	0.115	0.119
MotifSampler	0.071	0.092	0.107	0.097	0.077	0.103	0.099	0.068	0.112	0.119	0.103	0.127	0.130	0.134
MEME3	0.089	0.094	0.092	0.102	0.074	0.102	0.093	0.124	0.069	0.106	0.094	0.129	0.126	0.114
MEME	0.091	0.090	0.100	0.102	0.077	0.091	0.095	0.120	0.100	0.073	0.104	0.123	0.121	0.121
Oligo/dyad	0.073	0.088	0.111	0.088	0.082	0.082	0.099	0.136	0.119	0.112	0.071	0.106	0.107	0.130
ANN-Spec	0.085	0.091	0.111	0.094	0.090	0.100	0.085	0.122	0.114	0.110	0.089	0.074	0.118	0.117
YMF	0.094	0.095	0.112	0.101	0.093	0.100	0.114	0.146	0.121	0.129	0.092	0.131	0.084	0.137
Weeder	0.164	0.169	0.162	0.167	0.157	0.171	0.166	0.186	0.168	0.164	0.173	0.167	0.167	0.156

[a]The primary tool is listed in the row header and the secondary tool in the column header. The score shown for the same tool on both axes (that is, along the main diagonal) is the individual *nCC* score from **Figure 1.** Numerical values are categorized by color, ranging from dark blue (poorer predictions) to red (better predictions).

Tompa et al., 2005

Motif Discovery Workflow



1. **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2. **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3. **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4. **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

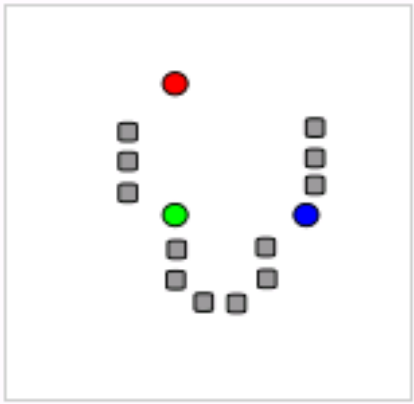
5. **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

MacIsaac & Fraenkel, 2006

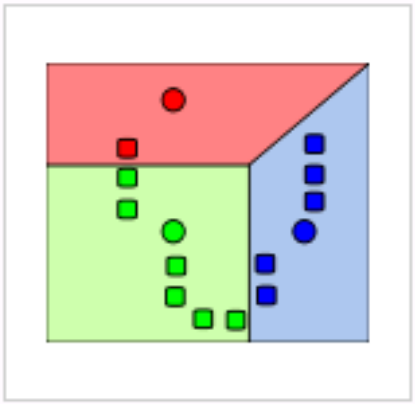
How to cluster motifs?

- TATAATTA TTACGTAA
- TATTATTA TTATATAA
- ATAATTAAG TATTTAAA
- TATTAT ATTATTTAA
- TATCATT TTATCTAA
- TAATT GCCTTACCTAA

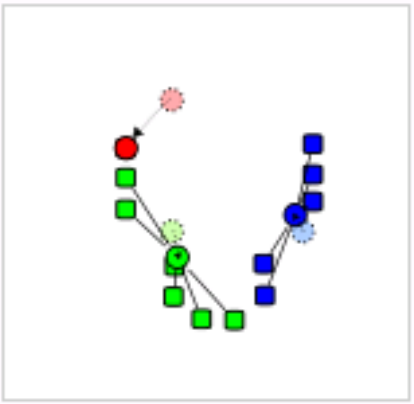
K-means clustering



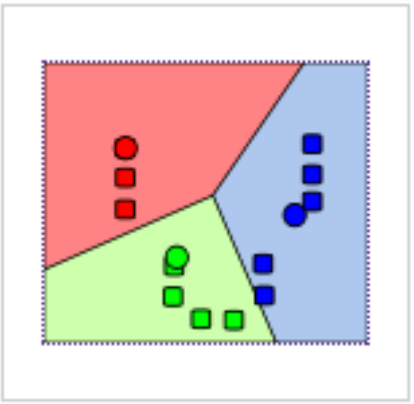
Shows the initial randomized point and a number of points.



Points are associated with the nearest initial randomized point.



Now the intial randomized points are moved to the center of their respective clusters(the centroids).



Steps 2 & 3 are repeated until a suitable level of convergence has been reached.

Clustering motifs

A	0.05	0.80	0.05	0.80	0.75	0.05
C	0.10	0.10	0.10	0.10	0.10	0.05
G	0.10	0.05	0.10	0.05	0.10	0.10
T	0.75	0.05	0.75	0.05	0.05	0.80

A	0.10	0.80	0.05	0.80	0.80
C	0.05	0.05	0.10	0.05	0.10
G	0.05	0.10	0.05	0.10	0.05
T	0.80	0.05	0.80	0.05	0.05

A	0.05	0.80	0.05	0.80	0.75	0.05
C	0.10	0.10	0.10	0.10	0.10	0.05
G	0.10	0.05	0.10	0.05	0.10	0.10
T	0.75	0.05	0.75	0.05	0.05	0.80

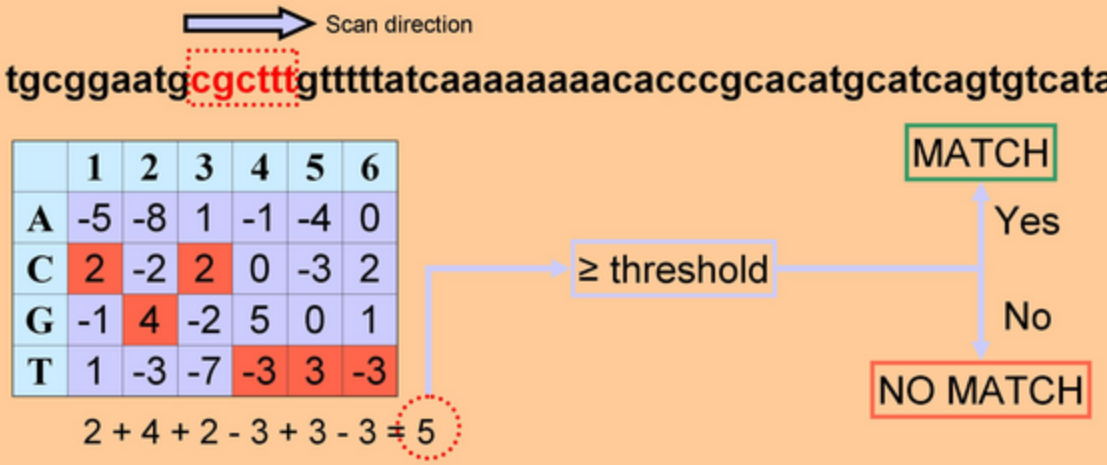
A	0.10	0.80	0.05	0.80	0.80
C	0.05	0.05	0.10	0.05	0.10
G	0.05	0.10	0.05	0.10	0.05
T	0.80	0.05	0.80	0.05	0.05

What is the “distance” between these motifs?

HOW TO PREDICT BINDING SITES

Scanning for Motifs with PWMs

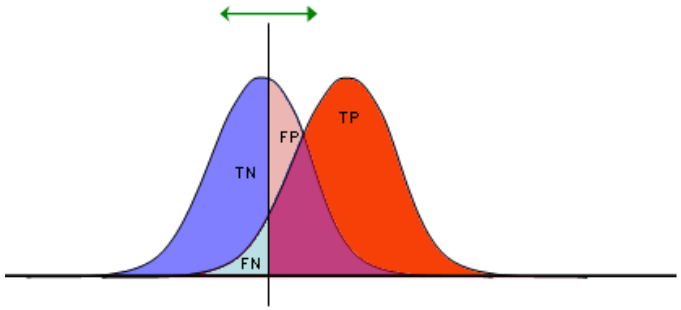
Position Weight Matrices define an additive scheme for scoring sequence. Often, the weights are simply log likelihood ratios of observing a nucleotide in a binding site relative to genomic background. Sequences are scanned by scoring every site, on both the forward and reverse complement strands, and identifying matches as shown in the schematic below:



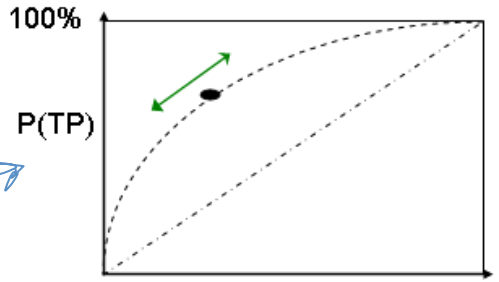
A particular site is evaluated by adding up the entries from the scoring matrix at each position, and comparing the sum to a match threshold. For log ratio PWMs, an empirically chosen threshold of 60% of the maximum positive score has been used by Harbison et al. and is approximately equal to cutoffs determined by the principled cross-validated method presented in MacIsaac et al. More sophisticated algorithms developed specifically for motif scanning are described briefly in Figure 3.

MacIsaac & Fraenkel, 2006

Receiving Operating Curve (ROC)



TP	FP
FN	TN
1	1



Regulated genes

Unregulated genes

Cross-validation

- the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis.

Sequences chosen with uniform probability

If motif is "incorrect", the positive sequences are "randomly selected" from among intergenic sequences, without any correlation or bias toward sequences containing the incorrect motif.

$$P_{hyper}(k \mid n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Where n is the number of positive sequences, N is the total number of sequences (positive and negative) and K is the number of sequences in which the word m occurs.

The p-value for the null hypothesis being true is the sum of the the probability distribution for k' >= k.

$$\text{p-value}(k) = \sum_{k'=k}^{n} P_{hyper}(k' \mid n, K, N)$$

Other measures of motif quality

- **Group specificity (or site specificity).** The probability of having this many target sequences containing the site (or this many sites within the target sequences), considering the prevalence of the motif throughout the genome.
- **Sequence specificity.** Emphasizes both the number of sequences with binding sites, and the number of sites per sequence.
- **Positional bias or uniformity.** Measures how uniform the binding site locations are distributed, with respect to the transcription start site of the gene. Real transcription factor binding sites often (but not always) show a marked preference for a specific region upstream of the genes they regulate.

False positives

- Some sites conform to the sequence identified by motif searching programs but *in vivo* do not bind to the protein.
- These sites are probably not available to be bound due to the conformation of the chromatin

Practical guidelines

Given the rates of false positives and false negatives, any of these motif discovery tools should be used with caution, and their results should be examined carefully. Here are some useful guidelines for applying them effectively.

1. If possible, remove spurious patterns from the target sequences. For example, using RepeatMasker (http://www.repeatmasker.org/).
2. Use multiple motif prediction algorithms.
3. Run probabilistic algorithms multiple times—you may not get the best scoring motif on the first run.
4. If possible, ask for multiple motifs to be returned—the highest scoring one may not be the most biologically relevant.
5. If necessary, try a range of motif widths and expected number of sites (some tools will automatically optimize these parameters for you).

Practical guidelines (2)

6. If needed, filter out motifs with biologically implausible distribution of information content (see the “block filtering” approach by Huber and Bulyk).
7. Combine similar motifs, for example by calculating their similarity using AlignACE, clustering them, and taking the best representative from each cluster.
8. Use AlignACE to match up with known motifs for the organism.
9. Evaluate the resulting motifs based on group specificity, set specificity, positional bias, etc.

Lately, a few packages have become available that combine multiple motif discovery algorithms, plus pre- and post-processing and analysis. Examples include MultiFinder and RgS-Miner.