Course organization

- Introduction (Week 1-2)
  - Course introduction
  - A brief introduction to molecular biology
  - A brief introduction to sequence comparison
- Part I: Algorithms for Sequence Analysis (Week 3 - 11)
  - Chapter 1-3, Models and theories
    - » Probability theory and Statistics (Week 4)
    - » Algorithm complexity analysis (Week 5)
    - » Classic algorithms (Week 6)
    - » Lab: Linux and Perl
  - Chapter 4, Sequence alignment (week 7)
  - Chapter 5, Hidden Markov Models (week 8)
  - Chapter 6. Multiple sequence alignment (week 10)
  - **Chapter 7. Motif finding (week 11)**
  - **Chapter 8. Sequence binning (week 11)**
- Part II: Algorithms for Network Biology (Week 12 - 16)

Chapter 7
Motif Finding

Chaochun Wei

Fall 2014

Contents

1. Reading materials
2. Motif finding
    - Motif
    - WMM
    - Motif finding methods

Reading materials

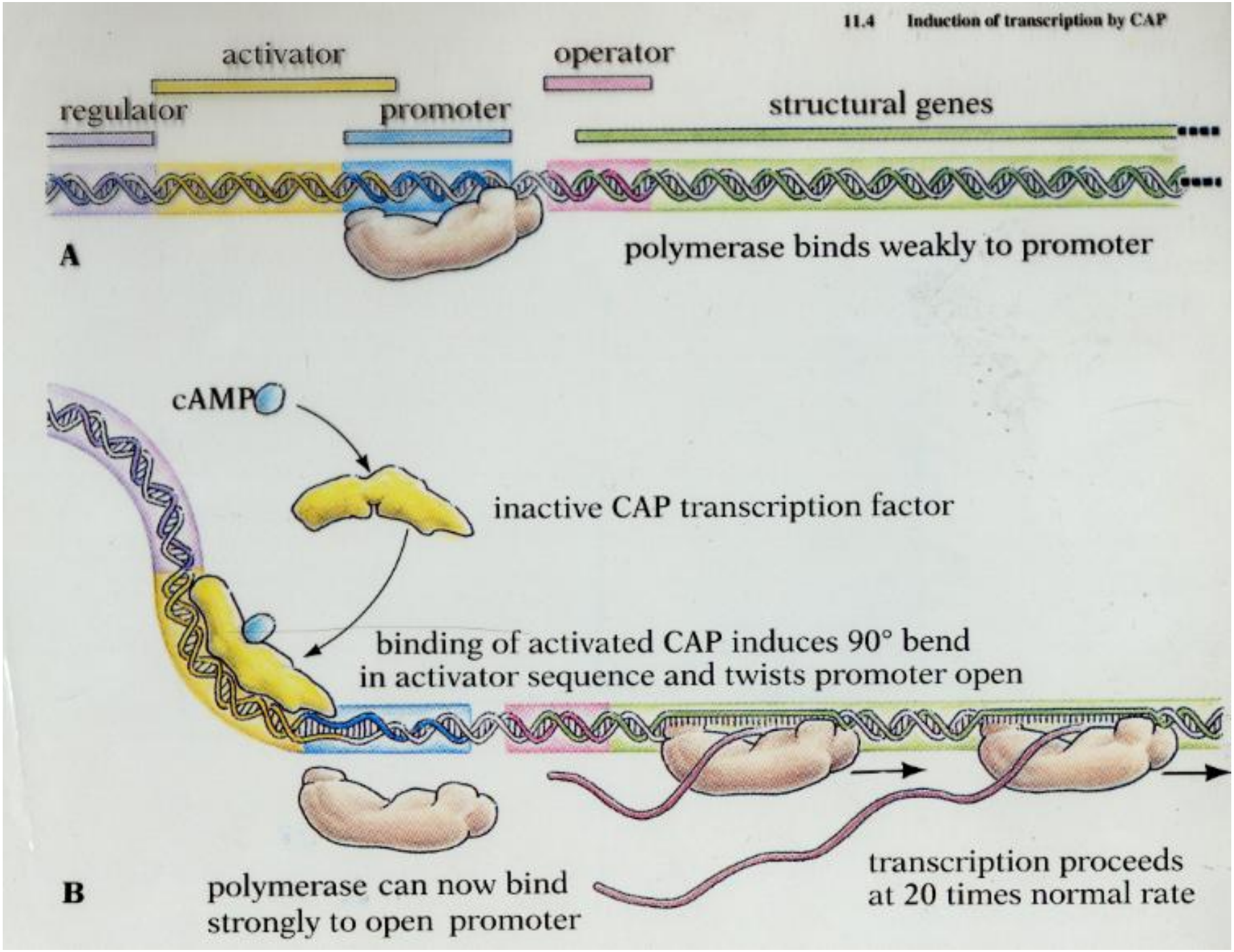
- Tompa et al (2005), “Assessing computational tools for the discovery of transcription factor binding sites”, Nature Biotechnology, 23(2):137-144

Sequence Motifs

- Motif: subsequence with some specific function
- May be in DNA, RNA, protein
- Function may be context dependent
  - Ribosome binding site must be transcribed
  - RNA, protein motifs may depend on structure
  - May be gapped or ungapped
- Use model to search for (predict) new sites
  - Models may be simple sequences (regular expressions) or probabilistic patterns
- Modeling approach depends on data available
  - Quantitative/qualitative

Motifs: DNA binding sites

- Gene regulation often controlled by proteins (transcription factors) that bind to specific DNA sequences to activate or repress transcription
- Binding sites are usually short (6-30bp) and typically ungapped, probabilistic patterns
- Would like to have a quantitative model that predicts binding affinity and/or occupancy


11.4 Induction of transcription by CAP
activator
operator
regulator
promoter
structural genes
A
polymerase binds weakly to promoter
cAMP
inactive CAP transcription factor
binding of activated CAP induces 90° bend
in activator sequence and twists promoter open
B
polymerase can now bind
strongly to open promoter
transcription proceeds
at 20 times normal rate

Weight Matrix Model

A:	-8	10	-1	2	1	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	10	-6	9	0	-1	11

Weight Matrix Model

-24

….A **C** **T** **A** **T** **A** **A** T G T.

A:	-8	10	**-1**	2	**1**	**-8**
C:	**-10**	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	**10**	**-6**	9	**0**	-1	11

Weight Matrix Model

43

….A C **T A T A A T** G T…

A:	-8	**10**	-1	**2**	**1**	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	**10**	-6	**9**	0	-1	**11**

How to get optimal models

- Quantitative binding data
  - Can do regression to get best fit
  - Can test different models
    - Must take into account that a more complex model can always give better fit, but is it useful?
- Qualitative binding data – set of sites
  - Log-odds methods
  - Can use minimum volume method (QP)

A.

A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

N(b,i)

B.

A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

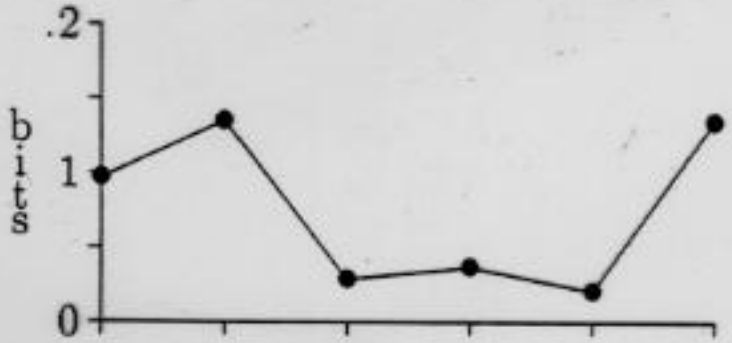
F(b,i)

C.

A	-2.76	1.82	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

$S(b,i) = \log[F(b,i)/P(b)]$

D.



$I(i) = \sum_{b} F(b,i)S(b,i)$

Likelihood Ratio Statistics Primer

Given two probability distributions P_i *and* Q_I

$$\sum P_i = \sum Q_i = 1$$

And some data, D_i, which is number of times each type *i* is observed in *N* total observations

The Likelihood Ratio of the data being from distribution Q_i versus P_i is:

$$\mathbf{LR} = \prod (Q_i/P_i)^{Di}$$

And the log-Likelihood Ratio is

$$\mathbf{LLR} = \sum D_i \ln (Q_i/P_i)$$

LLR = $\sum D_i \ln (Q_i/P_i)$

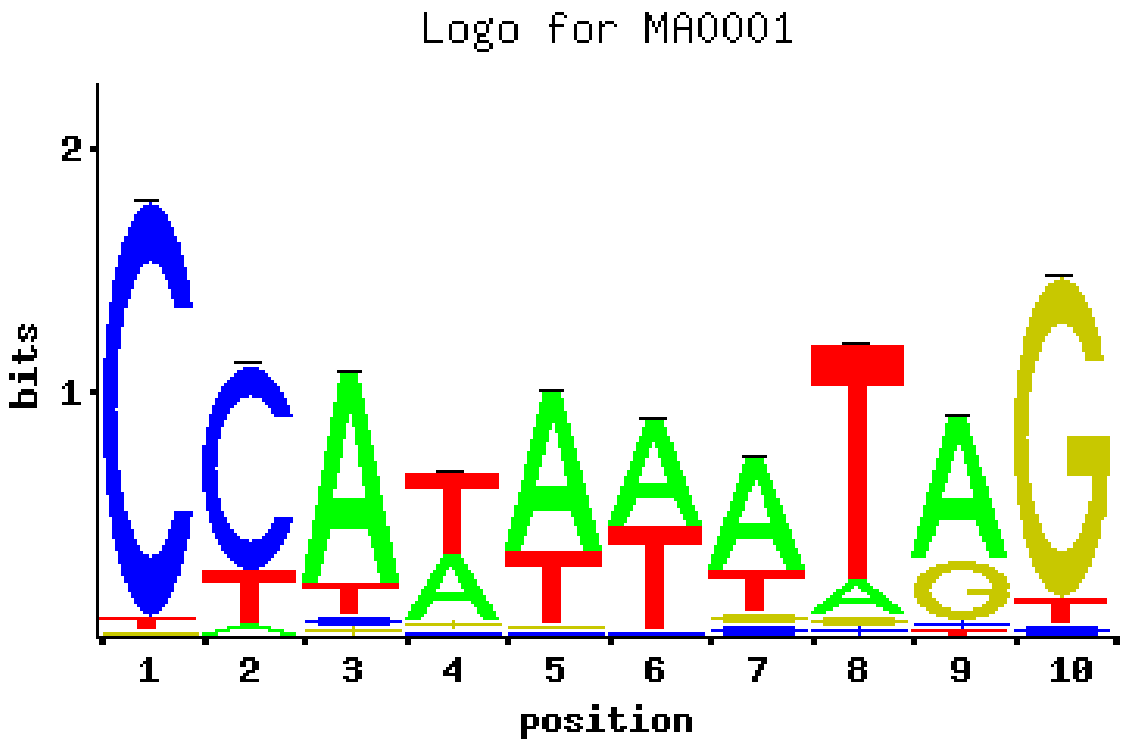
Maximum likelihood distribution is $Q_i = D_i/N$

So max **LLR** = $N \sum Q_i \ln (Q_i/P_i)$

$\sum Q_i \ln (Q_i/P_i) \geq 0$

≡ Information Content
Relative Entropy
Kullbach-Liebler Distance

Related to G-statistic and χ^2



A	0	3	79	40	66	48	65	11	65	0
C	94	75	4	3	1	2	5	2	3	3
G	1	0	3	4	1	0	5	3	28	88
T	2	19	11	50	29	47	22	81	1	6

Motif Discovery

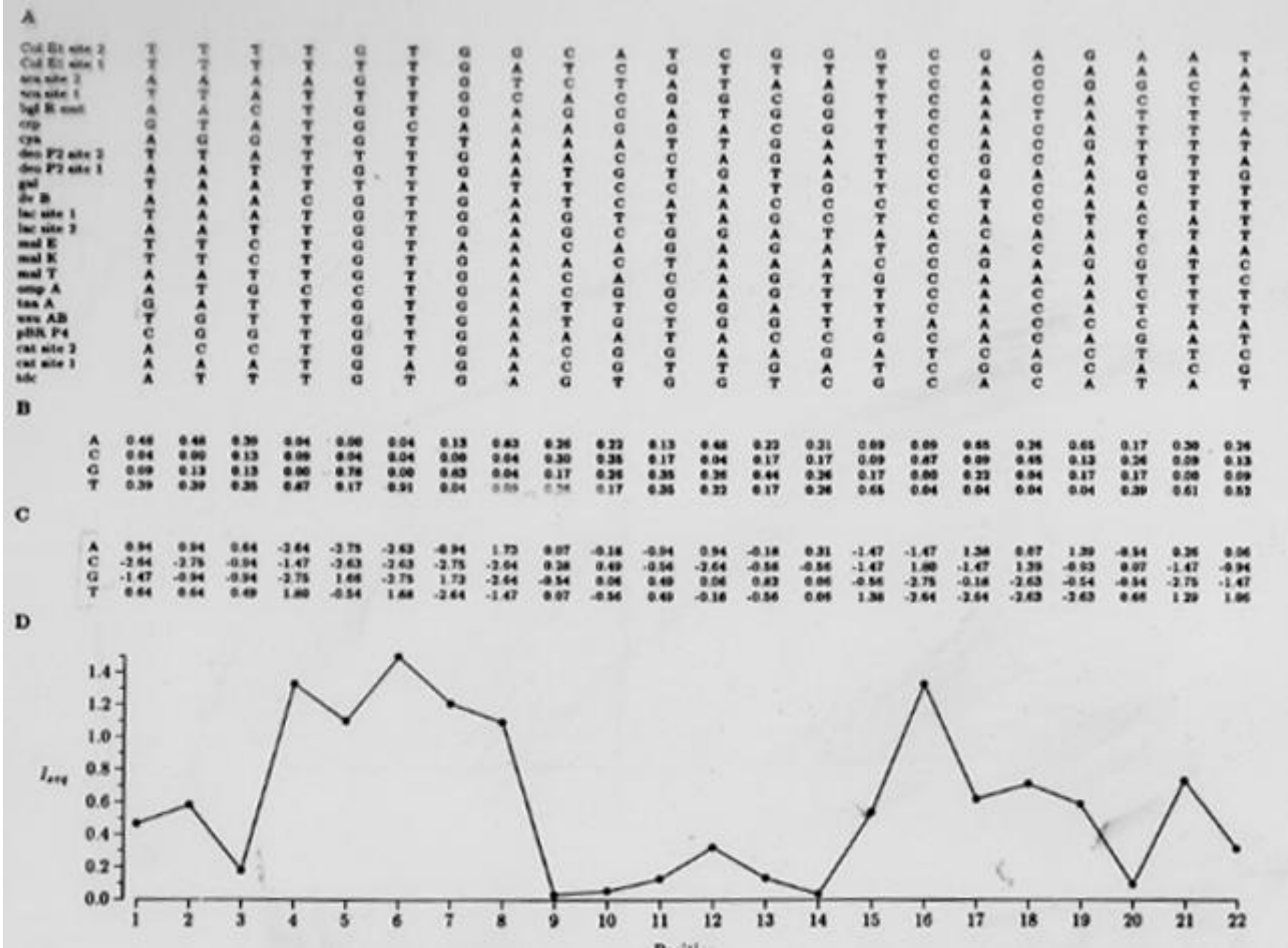
Find significant motifs in long sequences

- Types of data
  - Co-regulated promoters
  - Segments bound to specific proteins
  - Phylogenetically conserved segments
- Algorithm classes – search space
  - Pattern searches – consensus motifs
  - Alignment searches – PWM motifs
- Objective Functions

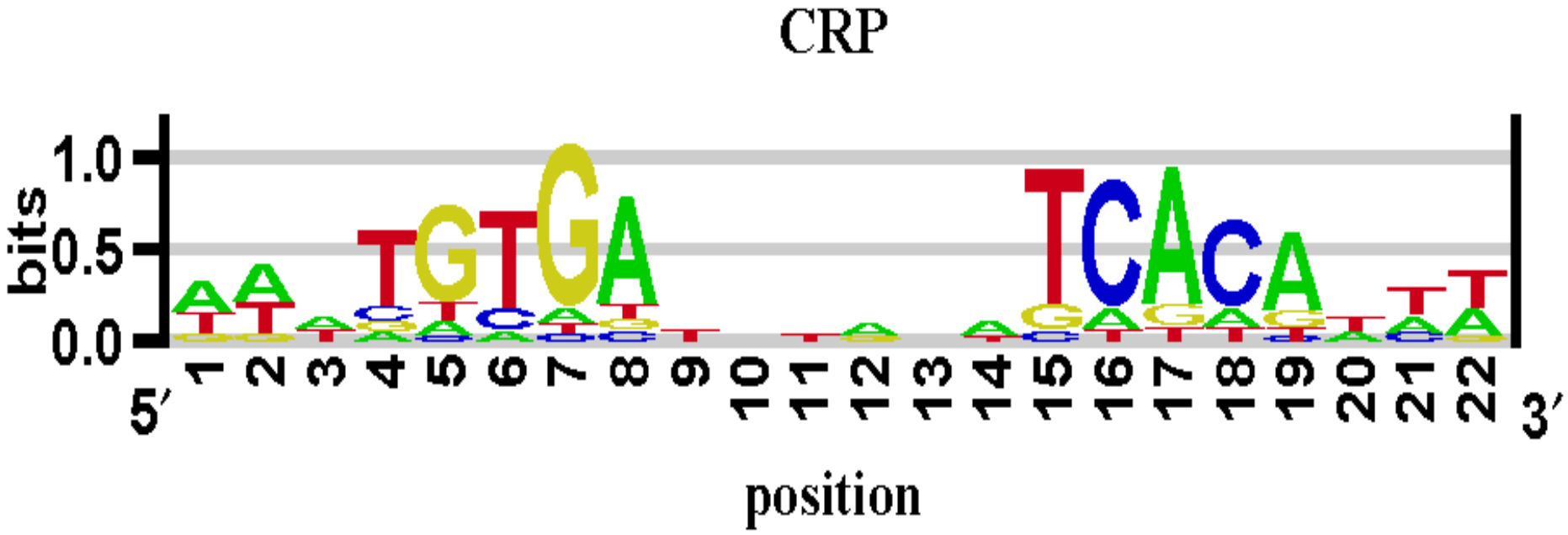
Example (small) dataset

```
CE1CG
\TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG\
ECOARABOP
\GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG\
ECOBGLR1
\ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAATTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT\
ECOCRP
\CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTACAGTAATACATTGATGTACTGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGC\
ECOCYA
\ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGTGTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAACC\
ECODEOP2
\AGTGAATTATTTGAACCAGATCGCATTACAGTGATGCAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATA\
ECOGALE
\GCGCATAAAAAACGGCTAAATTCTTGTGTAAACGATTCCACTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAGCC\
ECOILVBPR
\GCTCCGGCGGGGTTTTTTGTTATCTGCAATTCAGTACAAAACGTGATCAACCCCTCAATTTTCCCTTTGCTGAAAAATTTTCCATTGTCTCCCCTGTAAAGCTGT\
ECOLAC
\AACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAC\
ECOMALBA
\ACATTACCGCCAATTCTGTAACAGAGATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTA\
ECOMALBA
\GGAGGAGGCGGGAGGATGAGAACACGGCTTCTGTGAACTAAACCGAGGTCATGTAAGGAATTTCGTGATGTTGCTTGCAAAAATCGTGGCGATTTTATGTGCGCA\
ECOMALT
\GATCAGCGTCGTTTTAGGTGAGTTGTTAATAAAGATTTGGAATTGTGACACAGTGCAAATTCAGACACATAAAAAAACGTCATCGCTTGCATTAGAAAGGTTTCT\
ECOOMPA
\GCTGACAAAAAAGATTAAACATACCTTATACAAGACTTTTTTTTCATATGCCTGACGGAGTTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTACATCGCC\
ECOTNAA
\TTTTTTAAACATTAAAATTCTTACGTAATTTATAATCTTTAAAAAAAGCATTTAATATTGCTCCCCGAACGATTGTGATTCGATTCACATTTAAACAATTTCAGA\
ECOUXU1
\CCCATGAGAGTGAAATTGTTGTGATGTGGTTAACCCAATTAGAATTCGGGATTGACATGTCTTACCAAAAGGTAGAACTTATACGCCATCTCATCCGATGCAAGC\
PBR322
\CTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTC\
TRN9CAT
\CTGTGACGGAAGATCACTTCGCAGAATAAATAAATCCTGGTGTCCCTGTTGATACCGGGAAGCCCTGGGCCAACTTTTGGCGAAAATGAGACGTTGATCGGCACG\
TDC
\GATTTTTATACTTTAACTTGTTGATATTTAAAGGTATTTAATTGTAATAACGATACTCTGGAAAGTATTGAAAGTTAATTTGTGAGTGGTCGCACATATCCTGTT\
```

Example Output



Example Output



Types of Motifs

1. Motif: Consensus Sequence pattern
   - May include degenerate bases and allow for mismatches
   - *Search space is over possible patterns*

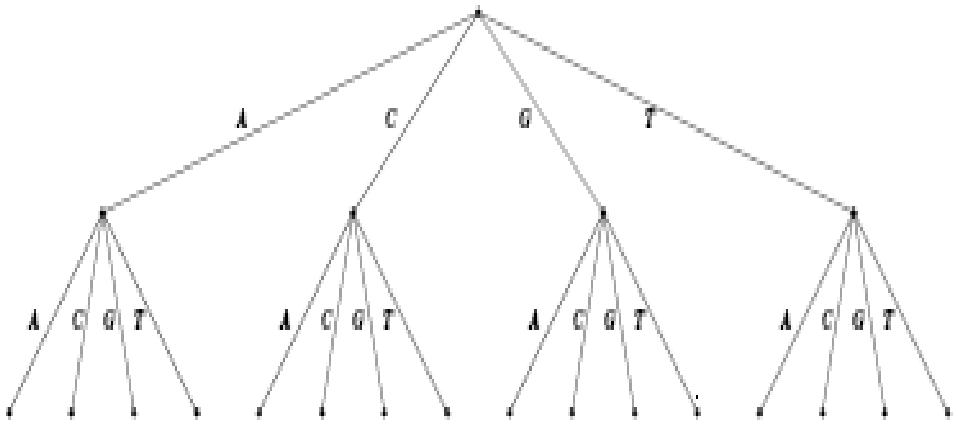
2. Weight Matrix (PWM, Profile, PSSM)
   - Might go to higher order models
   - *Search space is over possible alignments*

Pattern based algorithms

- Motif length l, mismatches m; N seqs, L long
- 4^l patterns, search for most common (or most significant) allowing up to m mismatches
  - P-value from background distribution
  - Can allow for m mismatches
  - Can allow degenerate positions
  - Can just search using existing l-mers

- Can use suffix tree for efficient search of patterns allowing mismatches

Suffix Trees



All sequneces included in O(NL) time and space

Each node can be labeled with N-bit string to indicate which sequences contain specific l-mer

Each pattern can be searched for in linear time, can find all patterns that match criteria of occurring in at least k/N seqs

Significance of each matching Pattern can be found from an Expectation based on background

Can allow m mismatches:
At each string keep track of number of mismatches to pattern being considered and continue down branches until that is exceeded.
Can speed up by requiring the mismatches to be spaced; i.e. not all at beginning of string

WEEDER program: good performance on benchmark tests.

Alignment (Profile) based methods

- Greedy algorithm (Consensus)
- Expectation Maximization (MEME)
- Gibbs Sampler
- Regression (MatrixReduce)

- Can use phylogenetic conservation

Greedy Algorithm (Consensus)

- Simple version: assume every sequence contains at least one true binding site
- Using each l-mer find best match to generate 2-seq alignments
- Using top K PWMs to search remaining sequences to include a new sequence
- Repeat until all seqs contribute
  - Or objective function is maximized (IC, p-value)

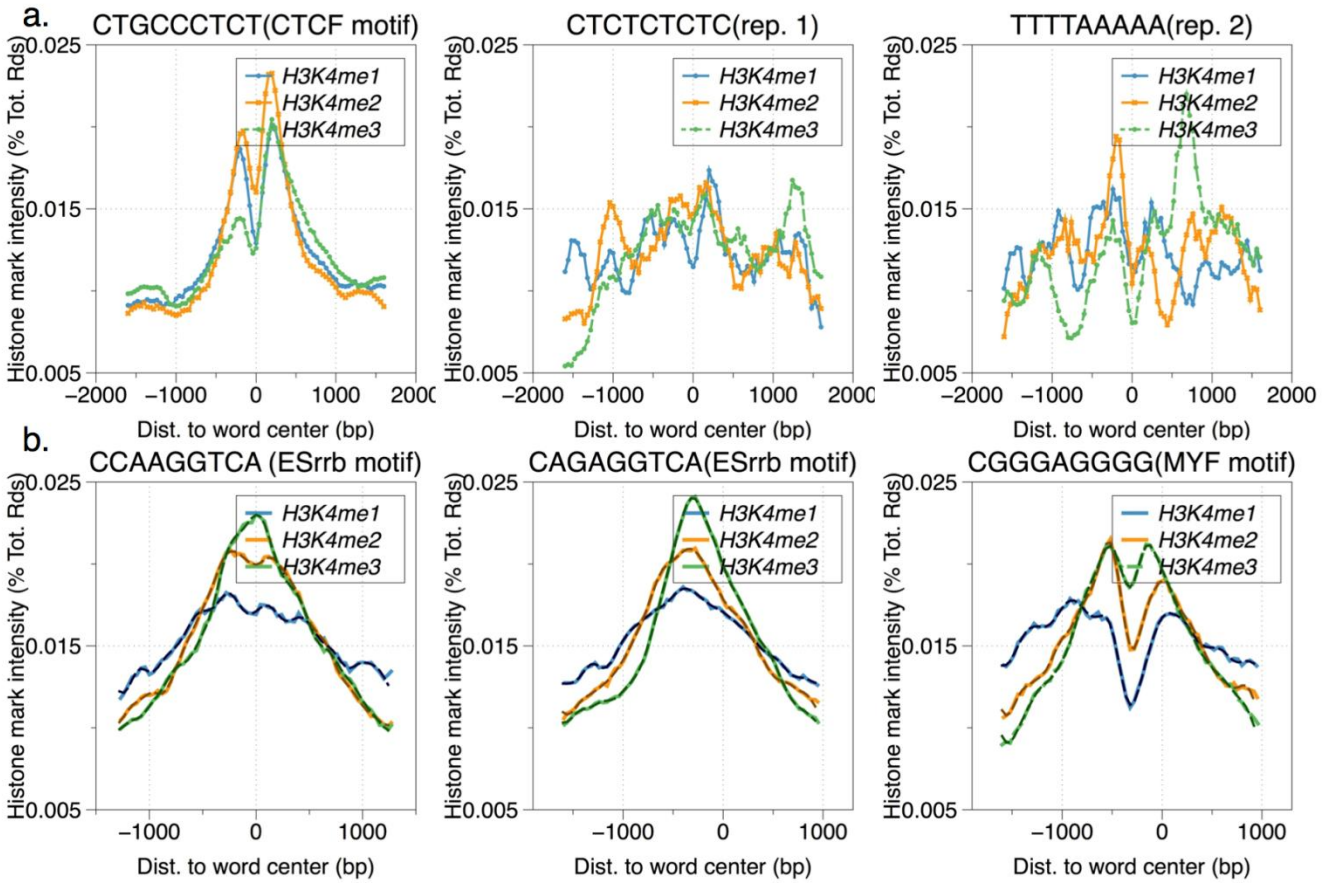
Expectation Maximization (MEME)

- Initial PWM (at random or from average over all potential sites)
- Using current PWM determine probability of all positions being sites
- Re-estimate PWM based on those probabilities
- Continue until convergence – always convergences
- Objective is LLR

Gibbs Sampling

- Similar to EM, but some important differences
- At each iteration pick one site on each seq, chosen by its probability, to update PWM
- Not guaranteed to converge, but tends to increase objective (IC) and plauteau
- Can escape local optima
  - Other MCMC algorithms
    - Metropolis
    - Simulated annealing

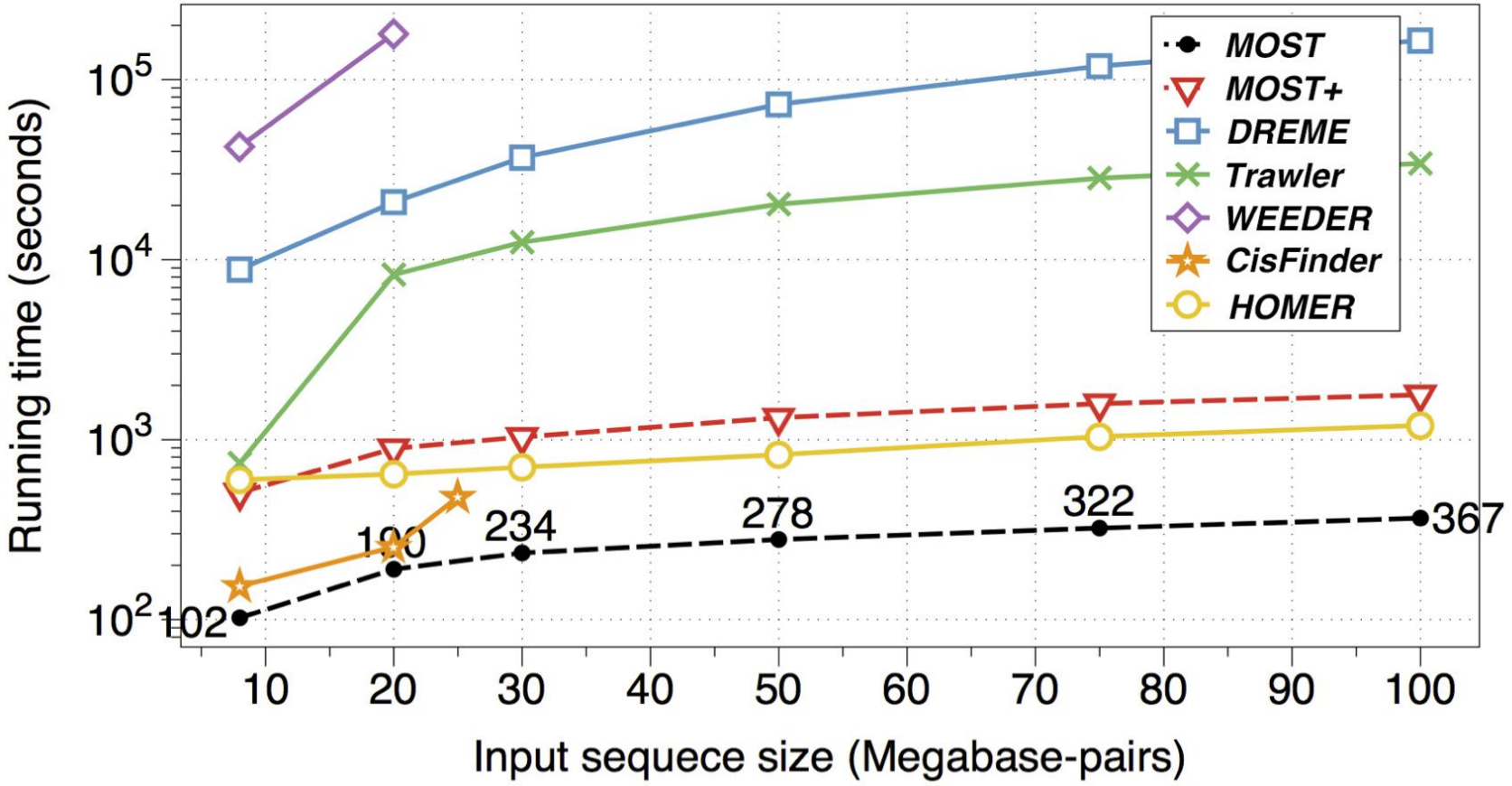
Motif finding by combining genomic sequence and heterogeneous genome-wide signatures



Distributions of several highly enriched word instances found in CTCF and ESrrb's ChIP-seq data set:

Zhang, Y.Z., He, Y.P., Zheng, G.Y., Wei, C.C., BMC Genomics, in press

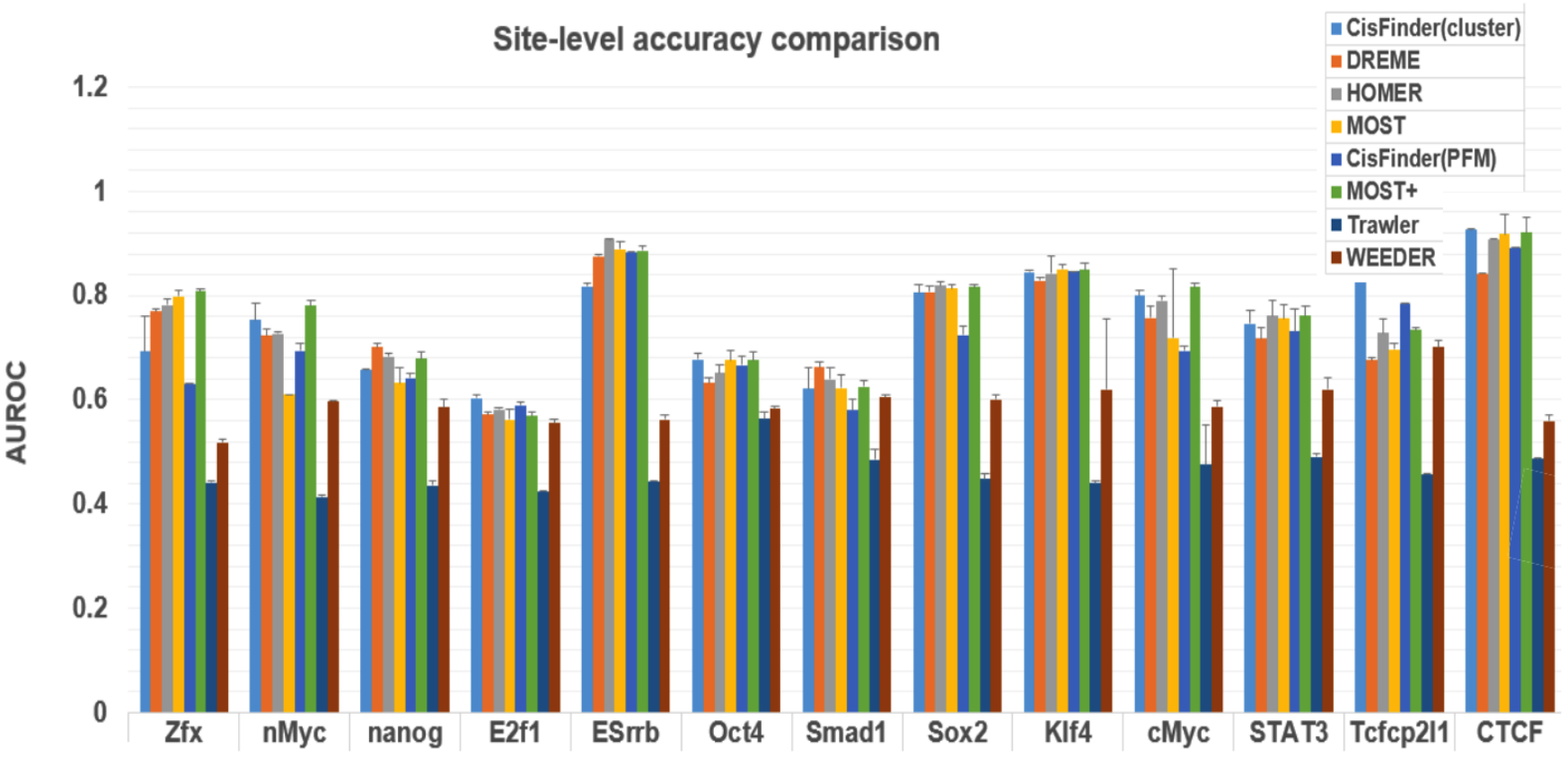
Motif finding by combining genomic sequence and heterogeneous genome-wide signatures



Comparison of different motif finding methods

Zhang, Y.Z., He, Y.P., Zheng, G.Y., Wei, C.C., BMC Genomics, in press

Motif finding by combining genomic sequence and heterogeneous genome-wide signatures



Comparison of different motif finding methods

Zhang, Y.Z., He, Y.P., Zheng, G.Y., Wei, C.C., BMC Genomics, in press

Acknowledgement

Most of the slides in this chapter were provided by Prof. Gary Stormo.