CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 8: Transcription-Factor Interaction

Wong Limsoon

National University of Singapore

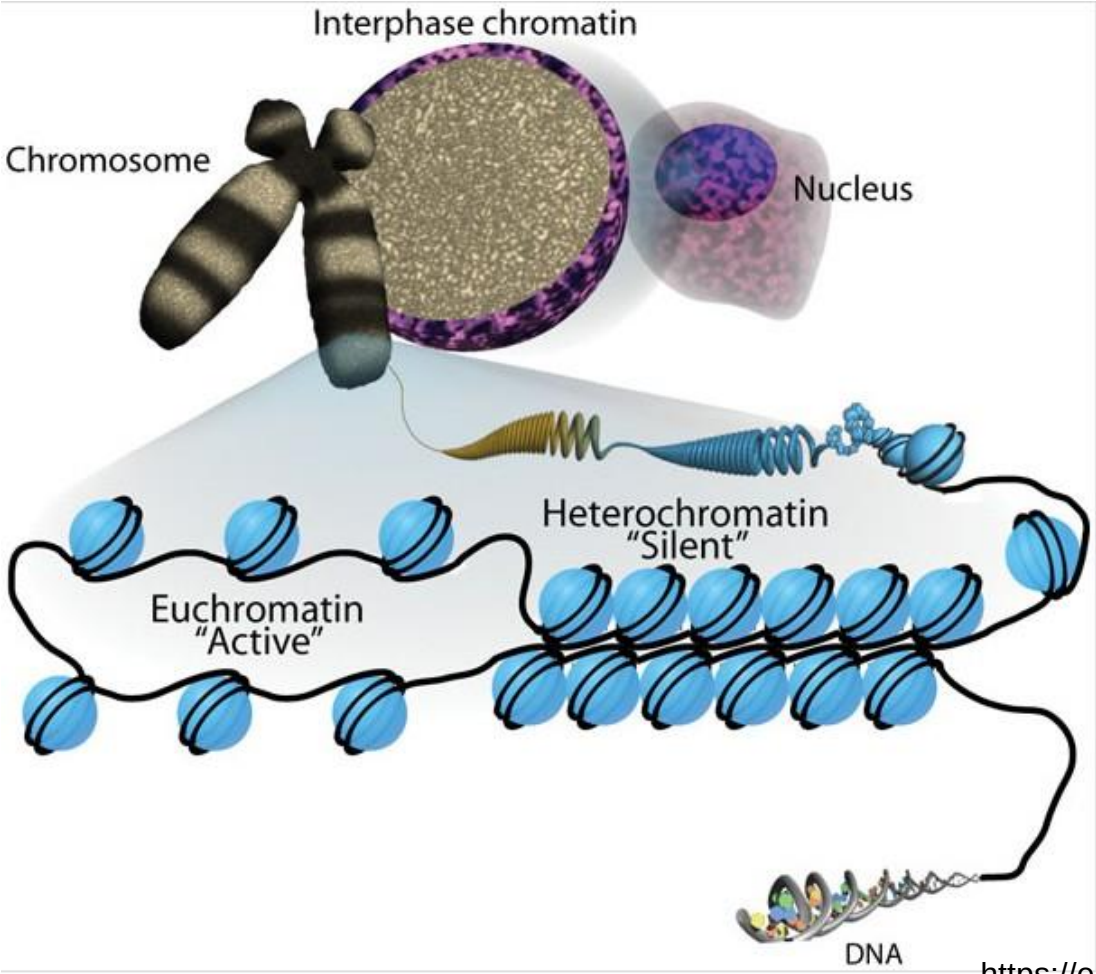
Outline

- **Gene regulation**
  - Chromatin organization, transcription factor (TF), TF binding site (TFBS), and histone code
- **TFBS discovery**
  - TFBS representation, TFBS databases, MEME
- **TF target-gene identification**
  - Gene expression, ChIP-x, BETA
- **TF-TF interactions**

Gene regulation

NUS
National University
of Singapore

Chromatin organization



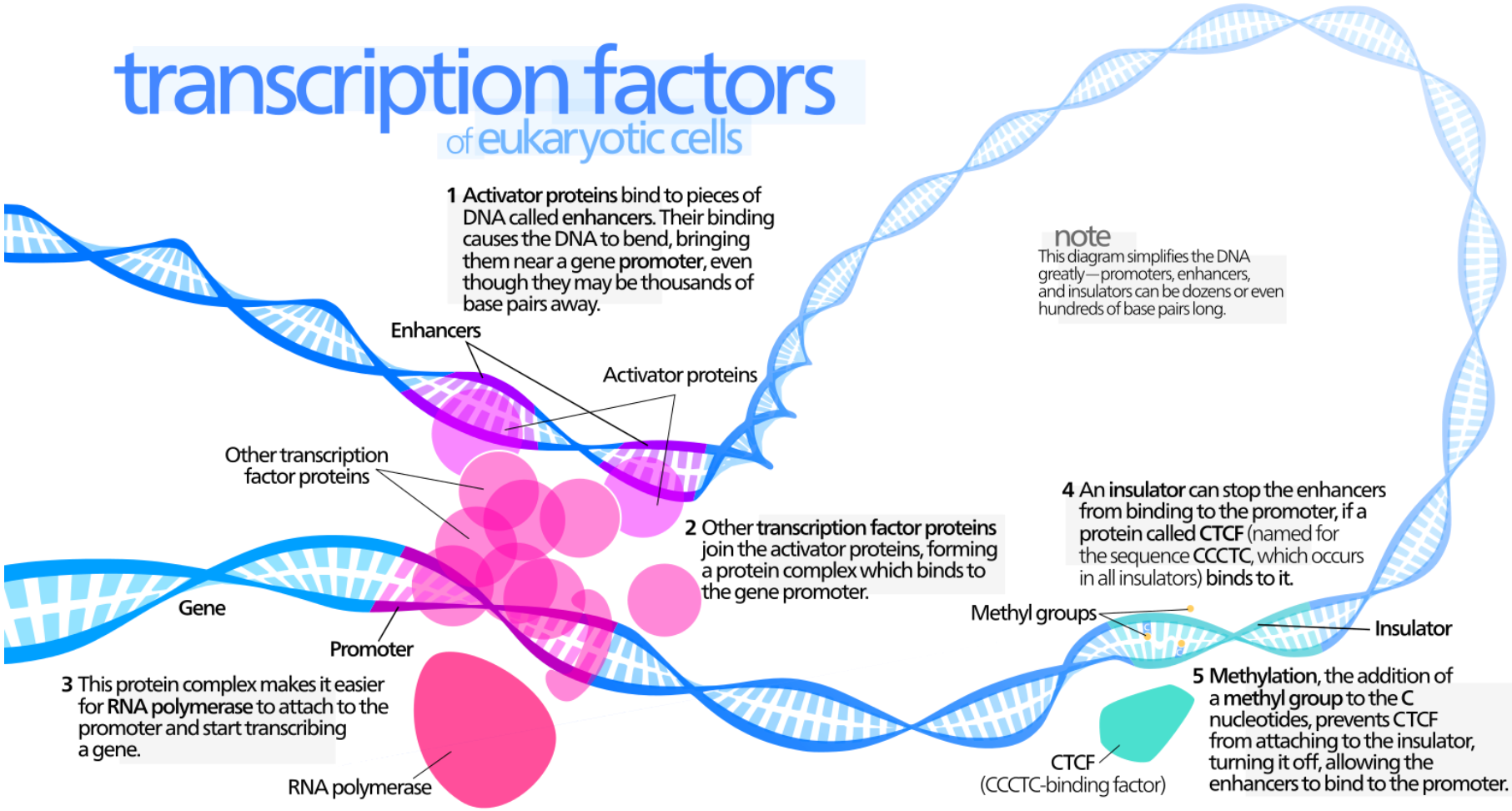
The basic unit of chromatin organization is the nucleosome, which comprises 147 bp of DNA wrapped around a core of histone proteins

Euchromatin (loose or open chromatin) structure is permissible for transcription.

Heterochromatin (tight or closed chromatin) is more compact and refractory to factors that need to gain access to the DNA template.

https://en.wikipedia.org/wiki/Chromatin_remodeling

https://en.wikipedia.org/wiki/Transcription_factor



- ~10% of genes in the human genome code for transcription factors (TFs)

- Genes are often flanked by several binding sites for distinct TFs, and efficient expression of each of these genes requires the cooperative action of several different TFs

- Combinatorial use of a subset of the ~2000 TFs easily accounts for the unique regulation of each gene in the human genome during development

Mechanisms

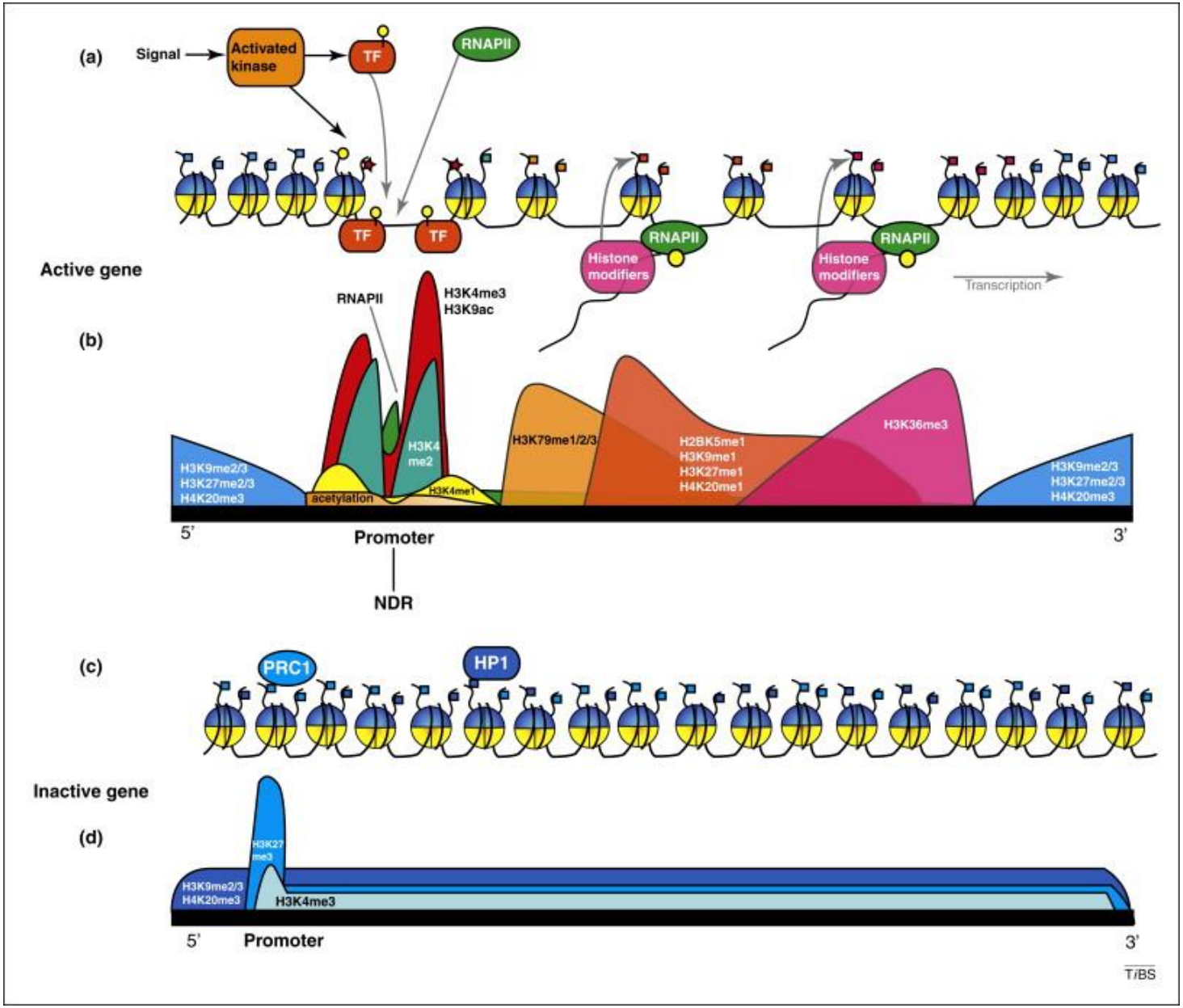
- **TFs bind to enhancer or promoter regions of DNA adjacent to the genes that they regulate**
- **Depending on the TF, the transcription of the adjacent gene is either up- or down-regulated via:**
  - Stabilize or block binding of RNA polymerase to DNA
  - Catalyze the acetylation or deacetylation of histones

- **Histone acetyltransferase (HAT) activity**
  - Acetylates histones➔weakens association of DNA w/ histones➔DNA more accessible to transcription➔transcription up
- **Histone deacetylase (HDAC) activity**
  - Deacetylates histones➔strengthens association of DNA w/ histones➔DNA less accessible to transcription➔ transcription down

Histone marks

Type of modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]	activation[7]	activation[7]
di-methylation	activation	repression[3]		repression[3]	activation[8]		
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]		repression[3]
acetylation		activation[9]	activation[9]				

- H3K4me3 is found in actively transcribed promoters
- H3K9me3 is found in constitutively repressed genes
- H3K27me is found in facultatively repressed genes
- H3K36me3 is found in actively transcribed gene bodies
- H3K9ac is found in actively transcribed promoters
- H3K14ac is found in actively transcribed promoters



http://www.cell.com/cms/attachment/610399/4879518/gr1.jpg

TFBS discovery

NUS
National University
of Singapore

Representations of TF binding sites

- **Position-specific frequency matrix (PSFM, PWM)**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1	0	1	5	32	5	35	23	34	14	43	13	34	4	52	3
C	50	1	0	1	5	6	0	4	4	13	3	8	17	51	2	0
G	0	0	54	15	5	5	12	2	7	1	1	3	1	0	1	52
T	5	55	1	35	14	40	9	27	11	28	9	32	4	1	1	1
Sum	56	56	56	56	56	56	56	56	56	56	56	56	56	56	56	56

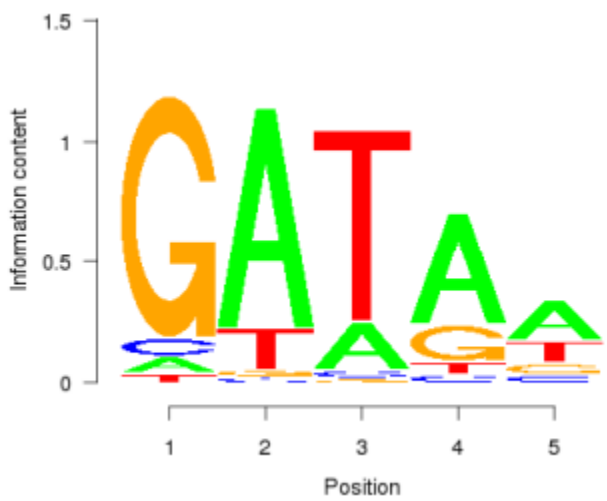
PSFM for the transcriptional repressor LexA as derived from 56 LexA-binding sites stored in Prodoric

- **Consensus sequence**
  - What is the consensus sequence for the TF binding site (TFBS) above?

https://en.wikipedia.org/wiki/Sequence_logo

Sequence logo

- **PSFM of TFBS is often visualized using "sequence logo"**



The information content (y-axis) of position i is given by:[2]

for amino acids, $R_i = \log_2(20) - (H_i + e_n)$
for nucleic acids, $R_i = \log_2(4) - (H_i + e_n)$

where H_i is the uncertainty (sometimes called the Shannon entropy) of position i

$$H_i = -\sum f_{a,i} \times \log_2 f_{a,i}$$

Here, $f_{a,i}$ is the relative frequency of base or amino acid a at position i, and e_n is the small-sample correction for an alignment of n letters. The height of letter a in column i is given by

$$\text{height} = f_{a,i} \times R_i$$

The approximation for the small-sample correction, e_n, is given by:

$$e_n = \frac{1}{\ln 2} \times \frac{s-1}{2n}$$

where s is 4 for nucleotides, 20 for amino acids, and n is the number of sequences in the alignment.

TFBS databases

Name	Organisms	Source
JASPAR	Vertebrates, Plants, Fungi, Flies, and Worms	Expert curation with literature support
CIS-BP	All Eukaryotes	Experimentally derived motifs and predictions
CollecTF	Prokaryotes	Literature curation
RegPrecise	Prokaryotes	Expert curation
RegTransBase	Prokaryotes	Expert/literature curation
RegulonDB	Escherichia coli	Expert curation
PRODORIC	Prokaryotes	Expert curation
TRANSFAC	Mammals	Expert/literature curation
TRED	Human, Mouse, Rat	Computer predictions, manual curation
DBSD	Drosophila species	Literature/Expert curation
HOCOMOCO	Human	Literature/Expert curation

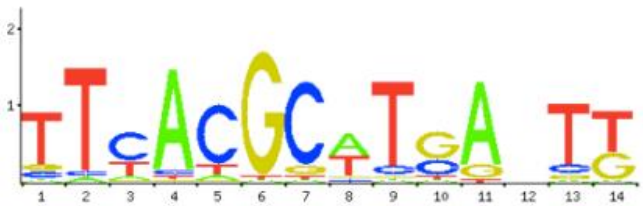
https://en.wikipedia.org/wiki/DNA_binding_site

Jaspar example: Pax6

Summary page for ID: MA0069 NAME: Pax6 from the JASPAR_CORE database ?

DATA	
name	Pax6
class	PAIRED
species	Homo sapiens
total_ic	13.7960
sysgroup	vertebrate
acc	P26367
type	SELEX
medline	8132558
comment	

SEQUENCE LOGO ?



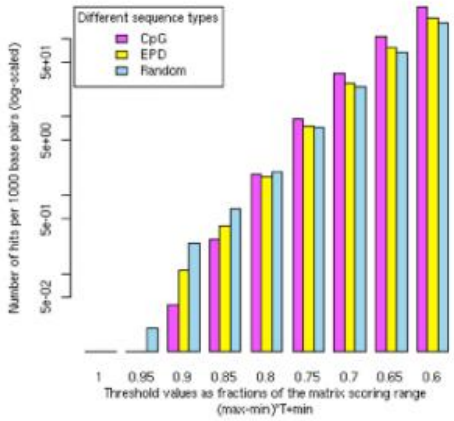
Make a SVG logo ?

FREQUENCY MATRIX ?

```
A  [ 2  2  4 39  3  1  1 21  1  2 36 11  1  1 ]
C  [ 4  2 26  2 34  0 37  2  4 14  0 11  5  0 ]
G  [ 4  0  1  1  1 41  4  2  1 25  6 13  3 17 ]
T  [33 39 12  1  5  1  1 18 37  2  1  8 34 25 ]
```

Reverse complement ?

EXPECTED PREDICTIONS/BP ?



Should you do a whole-genome scan (using such a TFBS motif) to see where a TF binds?

Table of number of hits per 1000 base pairs for each sequence type

Threshold	CpG	EPD	Random
1	0	0	0
0.95	0	0	0.01
0.9	0.03	0.1	0.23
0.85	0.26	0.39	0.67
0.8	1.85	1.71	1.94
0.75	9.36	7.44	7.19
0.7	35.2	26.42	23.61
0.65	105.31	76.35	67.07
0.6	253.49	183.08	159.38

http://bioinfo.cnio.es/files/training/Fourth_Sequence_Analysis_2011/TFBSdetection_2011.pdf

The high-quality transcription factor binding profile database

Browse the JASPAR_CORE database right away!

Click here to select all TFBS

http://bioinfo.cnio.es/files/training/Fourth_Sequence_Analysis_2011/TFBSdetection_2011.pdf

TFBS discovery: General strategy

- **Identify target genes of the TF**
- **Extract promoters or putative binding sites**
- **Align and look for enriched patterns**

Reference Genome		Sequences of interest		
Seq. oligo	expected frequency	Seq. oligo	observed frequency	
AAAAAA	0.00024	AAAAAA	0.00023	
AAAAAC	0.00030	AAAAAC	0.00031	
AAAAAG	0.00031	AAAAAG	0.00125	***
AAAAAT	0.00024	AAAAAT	0.00018	
AAAACC	0.00028	AAAACC	0.00026	
…		…		

http://bioinfo.cnio.es/files/training/Fourth_Sequence_Analysis_2011/TFBSdetection_2011.pdf

- **Popular tool: MEME**

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538909/

Nucleic Acids Res. 2006 Jul 1; 34(Web Server issue): W369–W373.
Published online 2006 Jul 14. doi: 10.1093/nar/gkl198

PMCID: PMC1538909

MEME: discovering and analyzing DNA and protein sequence motifs

Timothy L. Bailey,* Nadya Williams,[1] Chris Misleh,[1] and Wilfred W. Li[1]

Author information ▶ Article notes ▶ Copyright and License information ▶

ABSTRACT

Go to:

MEME (Multiple EM for Motif Elicitation) is one of the most widely used tools for searching for novel 'signals' in sets of biological sequences. Applications include the discovery of new transcription factor binding sites and protein domains. MEME works by searching for repeated, ungapped sequence patterns that occur in the DNA or protein sequences provided by the user. Users can perform MEME searches via the web server hosted by the National Biomedical Computation Resource (http://meme.nbcr.net) and several mirror sites. Through the same web server, users can also access the Motif Alignment and Search Tool to search sequence databases for matches to motifs encoded in several popular formats. By clicking on buttons in the MEME output, users can compare the motifs discovered in their input sequences with databases of known motifs, search sequence databases for matches to the motifs and display the motifs in various formats. This article describes the freely accessible web server and its architecture, and discusses ways to use MEME effectively to find new sequence patterns in biological sequences and analyze their significance.

MEME: Main idea

```
MEME(dataset, W, NSITES, PASSES) {
for i = 1 to PASSES {
    for each subsequence in dataset {
        run EM for 1 iteration with starting point derived from this sequence
        }
    choose model of the motif with highest likelihood
    run EM to convergence from starting point which generated that model
    print converged model of that motif
    Erase appearances of that motif from dataset
    }
```

Bailey & Elkan, *Machine Learning Journal*, 21:51-83, 1995

```
EM (dataset, W) {
    choose starting point (ρ)
    do {
        reestimate z from ρ
        reestimate ρ from z
    } until (change in ρ < ε)
    return
}
```

ρ = matrix of letter probability $\rho_{i,j}$

z = matrix of offset probability $z_{i,j}$

Choose a substring (e.g. TATAAT) in a sequence as a starting point

Initialize ρ

letter	*position in motif*					
	1	2	3	4	5	6
A	0.17	0.5	0.17	0.5	0.5	0.17
C	0.17	0.17	0.17	0.17	0.17	0.17
G	0.17	0.17	0.17	0.17	0.17	0.17
T	0.5	0.17	0.5	0.17	0.17	0.5

Estimate z: Use ρ to find best offsets for the substring

Estimate ρ based on these offsets

Compute log likelihood of ρ

$$\log(likelihood) = N \sum_{j=1}^{W} \sum_{l \in \mathcal{L}} f_{lj} \log(\rho_{lj}) + N(L-W) \sum_{l \in \mathcal{L}} f_{l0} \log(\rho_{l0}) + N \log\left(\frac{1}{L-W+1}\right)$$

where N is the number of sequences in the dataset, L is the length of the sequences, W is the length of the shared motif, $\mathcal{L}$ is the alphabet of the sequences, ρ_{lj} is the (unknown) probability of letter l in position j of the motif, ρ_{l0} is the (unknown) probability of letter l in all non-motif positions, f_{lj} is the observed frequency of the letter l in position j of the motif, and f_{l0} is the observed l in all non-motif positions of the sequences.

MEME: Sample results

Table 1. Overview of the contents of the datasets.

dataset	*samples*	*average length of samples*	*proven CRP sites*	*proven LexA sites*
CRP	18	105	18	0
LexA	16	192	1	11
CRP/LexA	34	150	19	11
promoter	231	58	NA	NA

Table 2. The models found by each pass of MEME on the CRP/LexA dataset can be visually summarized by the consensus sequence derived from the ρ matrix by choosing the letter with the highest probability. The values of information content and log(*likelihood*) give a qualitative idea of the statistical significance of the model. Higher values imply the model is more significant. The models found for LexA and CRP on passes 1 and 2 of MEME have considerably higher log(*likelihood*) and information content than the models found on later passes. Note that $W = 20$ and $NSITES = 17$.

pass	*starting subsequence*	*final consensus*	I_{model}	log(*likelihood*)
1	TACTGTATATAAAACCAGTT	TACTGTATATATATACAGTA	13.206	-435.174
2	TTATTTGCACGGCGTCACAC	TTTTTTGATCGGTTTCACAC	9.087	-515.837
3	ATTATTATGTTGTTTATCAA	TTTATTTTGATGTTTATCAA	6.527	-539.083
4	TGCGTAAGGAGAAAATACCG	TGCGTAAGAAGTTAATACTG	7.912	-531.419
5	CAAATCTTGACATGCCATTT	CAAATATGGAAAGGCCATTT	8.027	-533.662

This matches LexA motif

This matches CRP motif

Table 3. Values of z_{ij} for the model found by MEME in pass 1 on the CRP/LexA dataset at the positions of the known LexA sites. Virtually all of the known sites have very high values of z_{ij} compared to the rest of the positions in the samples. The table shows the positions of the known sites (*site 1*, *site 2* and *site 3*) and the values of z_{ij} of the model at those positions. All other positions have values of z_{ij} below 0.17. Although the site at position 112 in the colicin E1 sequence has z_{ij} value only 0.05, this is one of the four highest z_{ij} values for this sequence. No proven sites are known for *himA* and *uvrC* and z_{ij} for all positions in those samples was very low, less than 0.0001.

sample	*site 1*	z_{ij}	*site 2*	z_{ij}	*site 3*	z_{ij}
cloacin DF13	97[a]	0.998684				
colicin E1	97	0.948441	112	0.051543		
colicin Ia	99[a]	0.998709				
colicin Ib	99[a]	0.990472				
recA	71	0.999987				
recN	71	0.999988	93	0.865704	111[a]	0.134281
sulA	85[a]	0.999990				
umuDC	91	0.999931				
uvrA	60	0.987786				
uvrB	71	0.999972				
uvrD	102	0.998539				
colicin A	34[a]	0.683563	48[a]	0.314723		
lexA	76	0.999982	55	0.999933		
mucAB	49[a]	0.999978				
himA						
uvrC						

[a] Indicates site known only by sequence similarity to known sites.

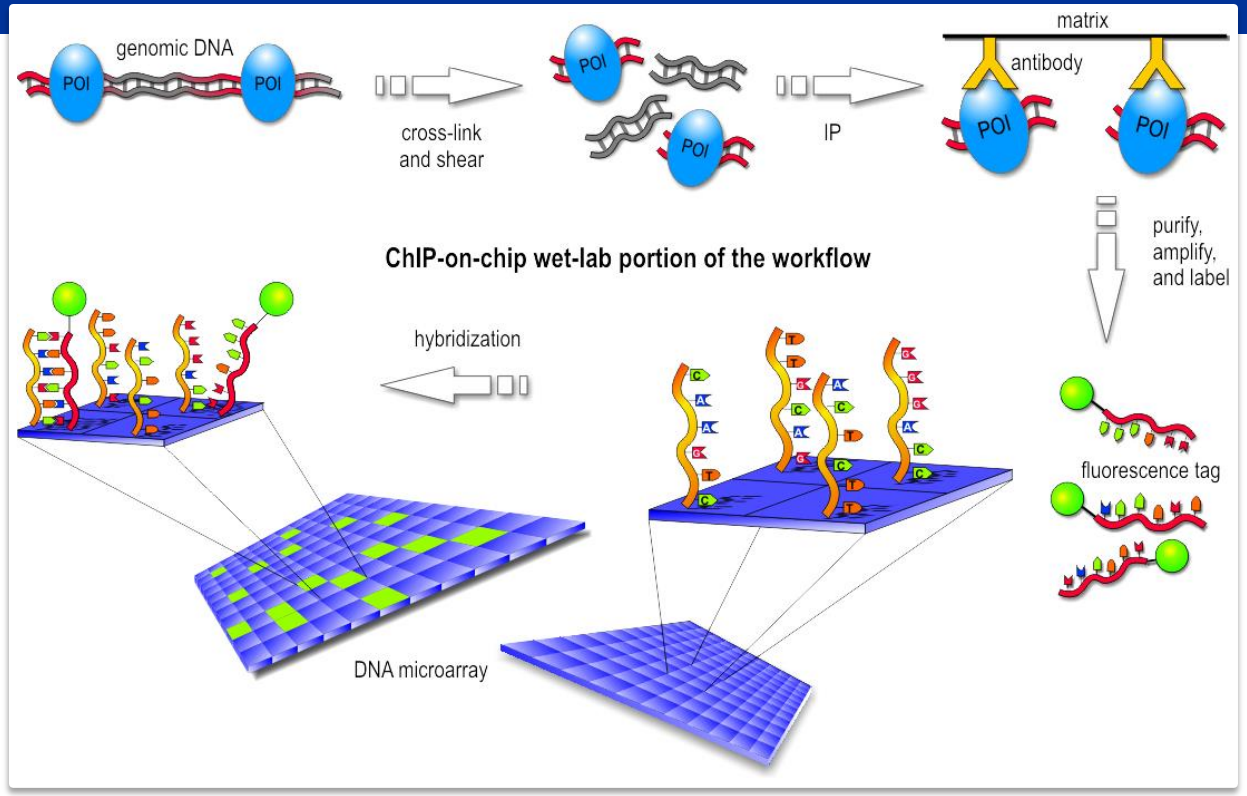
TF target-gene identification

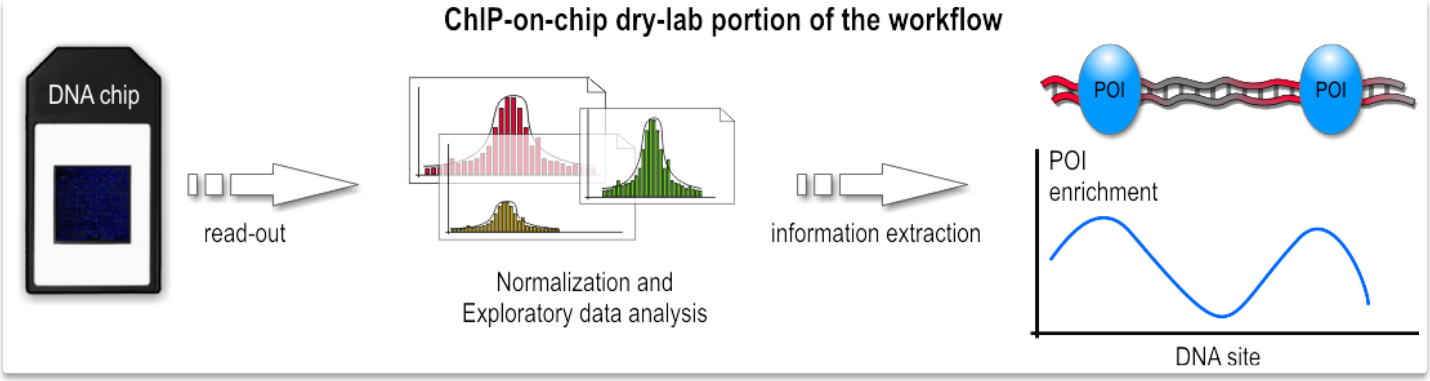
NUS
National University
of Singapore

The converse: How to find target genes of a given TF?

- **Gene expression data from TF-perturbation expt**
- **ChIP-chip = Chromatin immunoprecipitation + DNA microarray**
- **ChIP-seq = ChIP + massively parallel sequencing**
- **Popular tool: BETA**
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4135175/
- **Also,** DNAse-seq **is a laboratory method for identifying accessible DNA regions (i.e. open chromatin)**

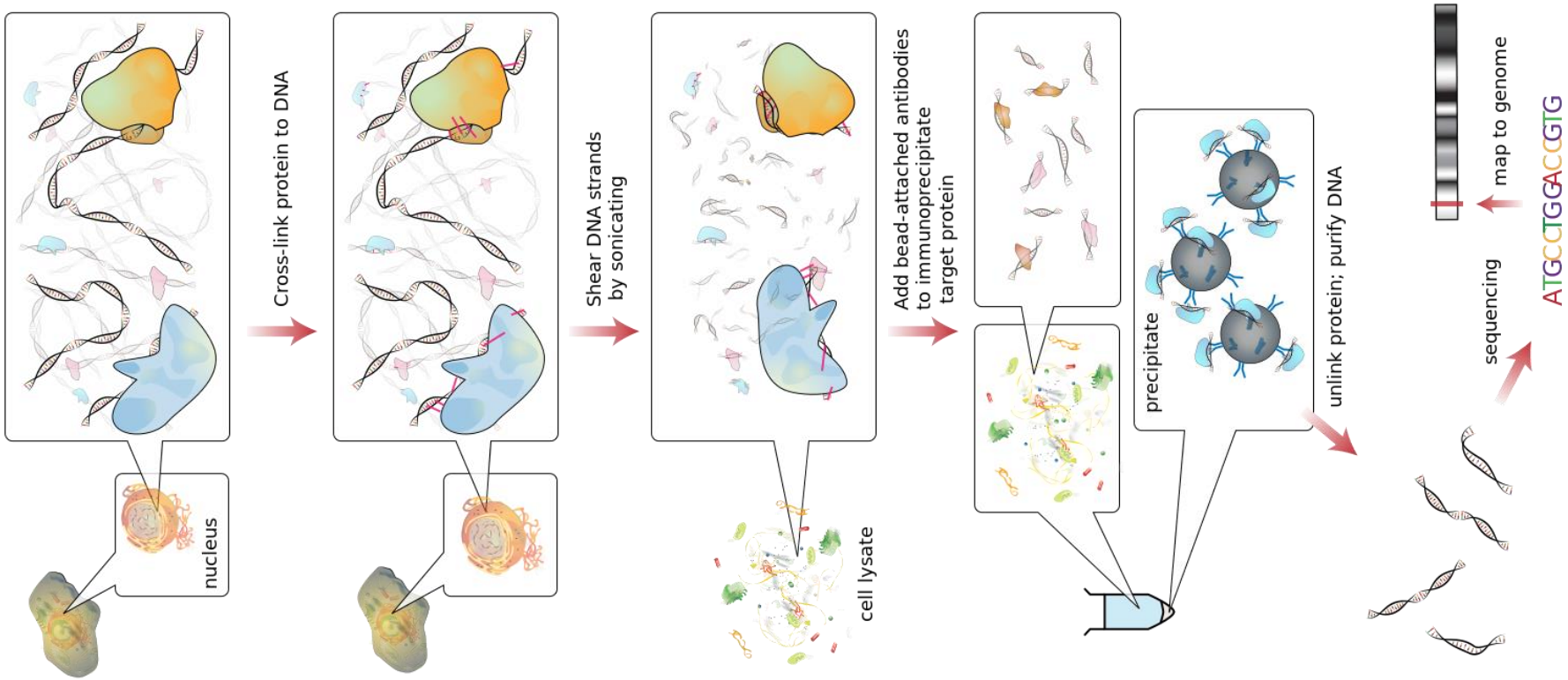
ChIP-chip





https://en.wikipedia.org/wiki/ChIP-on-chip

ChIP-seq

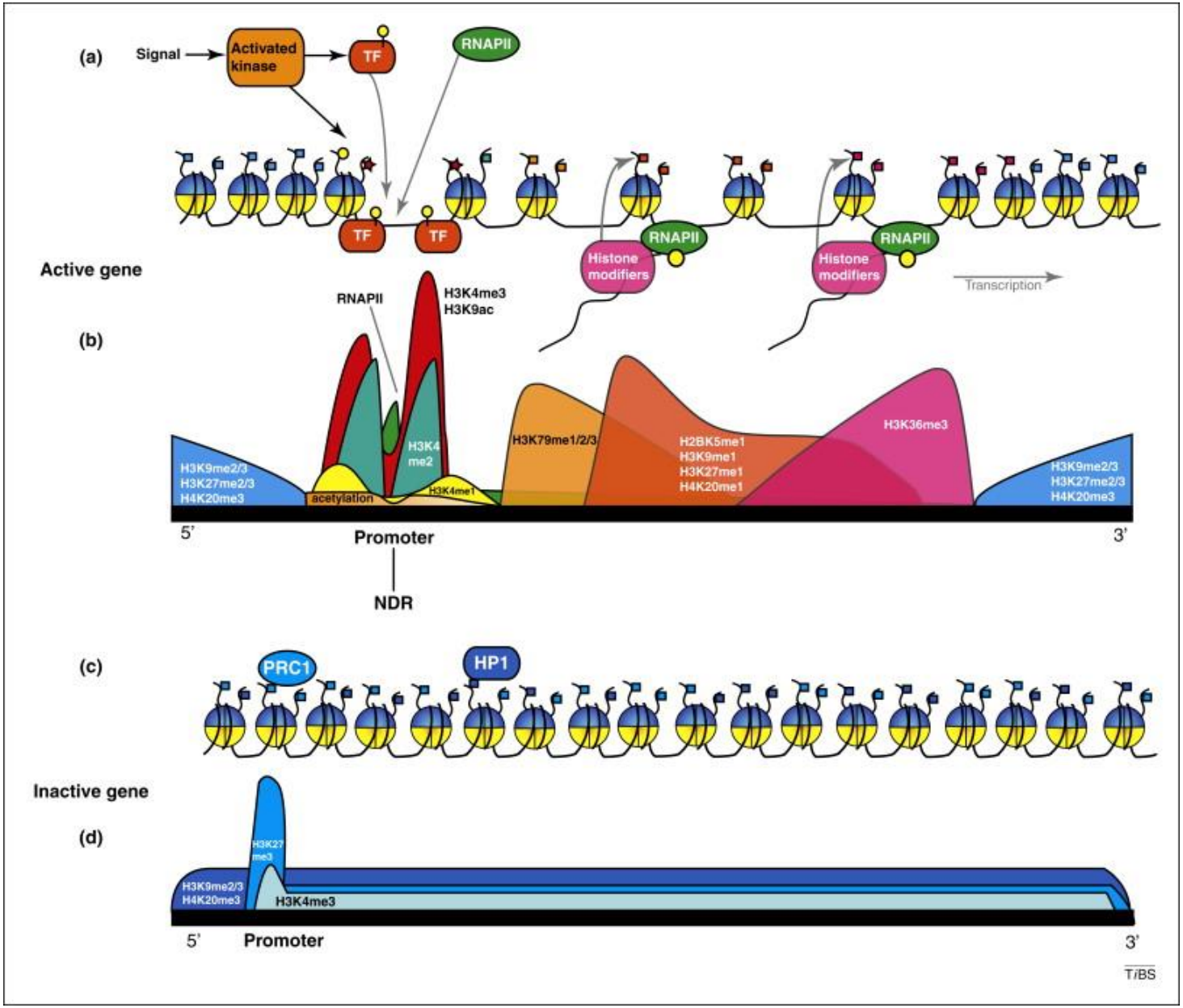


https://en.wikipedia.org/wiki/ChIP-sequencing

Histone marks

Type of modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]	activation[7]	activation[7]
di-methylation	activation	repression[3]		repression[3]	activation[8]		
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]		repression[3]
acetylation		activation[9]	activation[9]				

- H3K4me3 is found in actively transcribed promoters
- H3K9me3 is found in constitutively repressed genes
- H3K27me is found in facultatively repressed genes
- H3K36me3 is found in actively transcribed gene bodies
- H3K9ac is found in actively transcribed promoters
- H3K14ac is found in actively transcribed promoters



http://www.cell.com/cms/attachment/610399/4879518/gr1.jpg

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4135175/

Nat Protoc. Author manuscript; available in PMC 2014 Aug 18.
Published in final edited form as:
Nat Protoc. 2013 Dec; 8(12): 2502–2515.
Published online 2013 Nov 21. doi: 10.1038/nprot.2013.150

PMCID: PMC4135175
NIHMSID: NIHMS607566

Target analysis by integration of transcriptome and ChIP-seq data with BETA

Su Wang,[1] Hanfei Sun,[1] Jian Ma,[1] Chongzhi Zang,[2] Chenfei Wang,[1] Juan Wang,[1] Qianzi Tang,[1] Clifford A Meyer,[2] Yong Zhang,[1] and X Shirley Liu[2]

Author information ▶ Copyright and License information ▶

Abstract

Go to:

The combination of ChIP-seq and transcriptome analysis is a compelling approach to unravel the regulation of gene expression. Several recently published methods combine transcription factor (TF) binding and gene expression for target prediction, but few of them provide an efficient software package for the community. Binding and expression target analysis (BETA) is a software package that integrates ChIP-seq of TFs or chromatin regulators with differential gene expression data to infer direct target genes. BETA has three functions: (i) to predict whether the factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. Here we describe the implementation and features of BETA to demonstrate its application to several data sets. BETA requires ~1 GB of RAM, and the procedure takes 20 min to complete. BETA is available open source at http://cistrome.org/BETA/.

BETA: Main idea

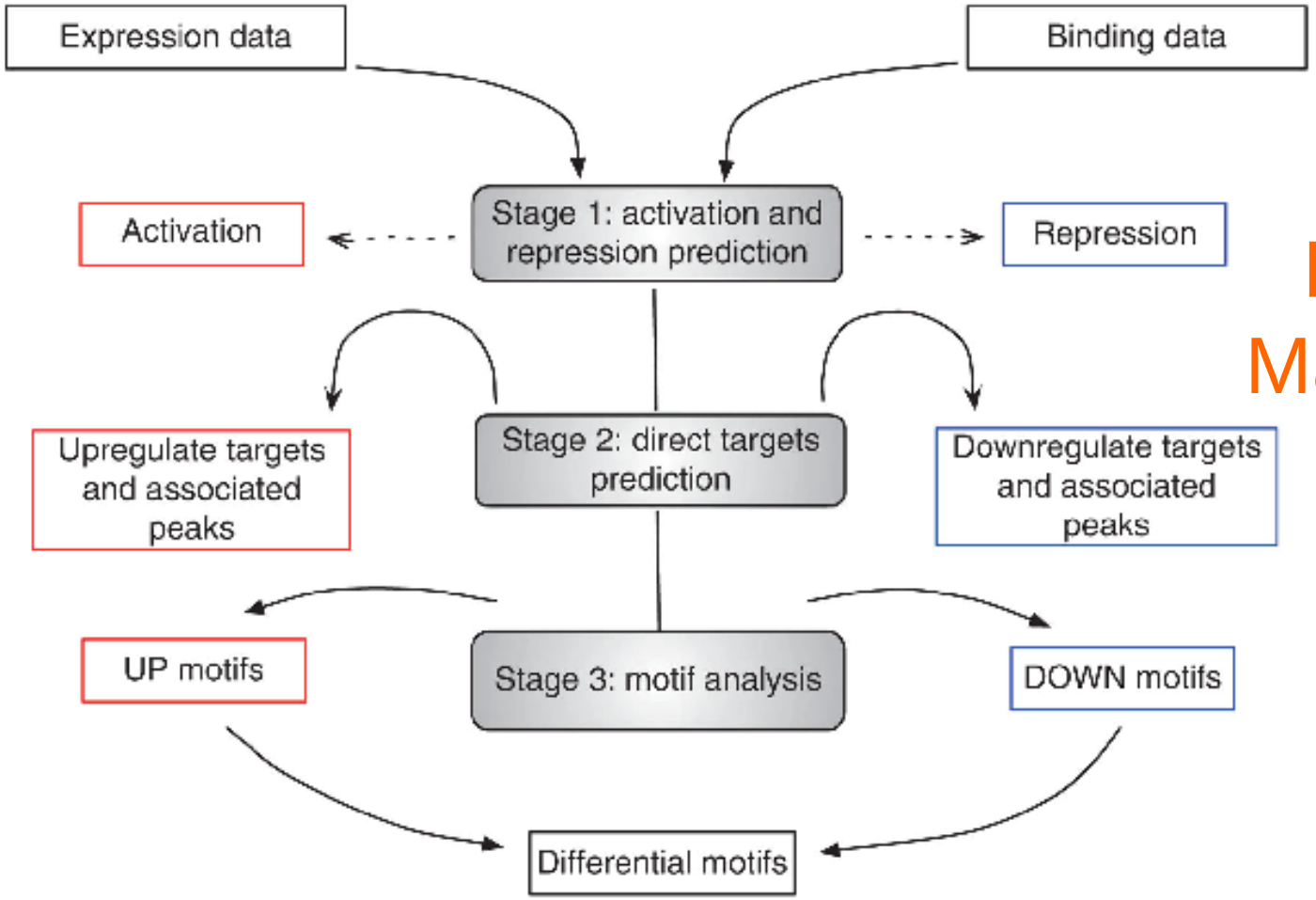


Figure 1.
BETA workflow. Stage 1 analyzes the differential expression and ChIP-seq binding data to predict whether a factor generally activates or represses gene expression. Stage 2 predicts direct target genes by their upregulation or downregulation. Stage 3 conducts motif analysis to identify putative collaborating factors that contribute to upregulation (UP) or downregulation (DOWN).

Wang et al. *Nature Protocol,* 8:2502-2515, 2013
Tang et al. Cancer Research, 71:6940-6947, 2011

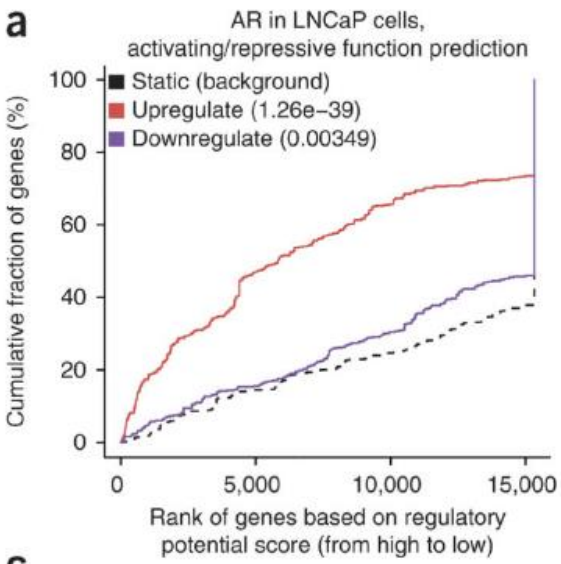
- **For Stage 1, i.e. direct gene-target prediction, BETA ranks genes on the basis of both regulatory potential of factor binding and differential expression upon factor binding, and then it calculates the rank product of the two to predict direct targets**

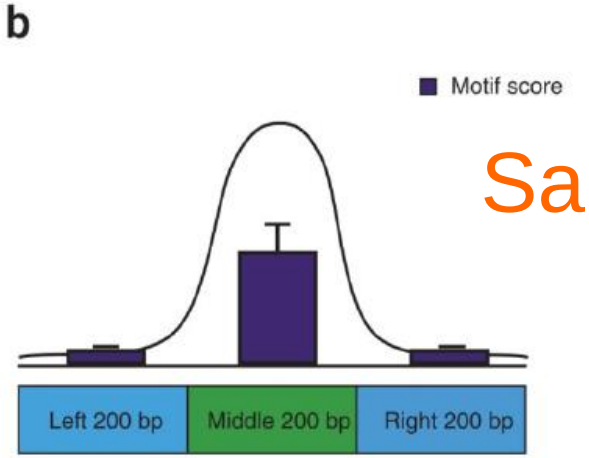
- **The regulatory potential is calculated as**

$$S_g = \sum_{i=1}^{k} e^{-(0.5+4\Delta_i)}$$

- **All binding sites (*k*) near the transcription start site of the gene (*g*) within a user specified range (100 kb as default) are considered. Δ is the exact distance between a binding site and the TSS proportional to 100 kb (Δ = 0.1 means the exact distance = 10 kb)**

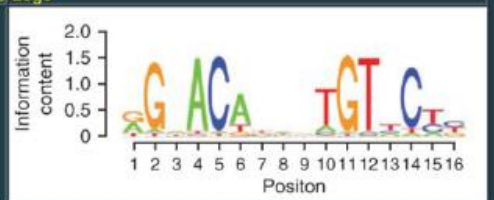
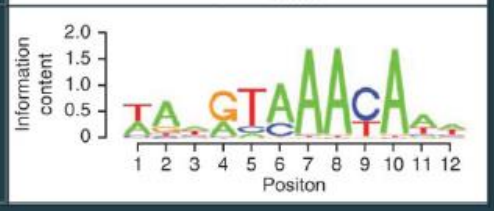
BETA: Sample result





c

PART1: UP TARGET GENES

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
NR3C1	Hormone-nuclear Receptor Family	Homo sapiens	9.67e-16	8.03	
PGR	Hormone-nuclear Receptor Family				
NR3C2	Hormone-nuclear Receptor Family				
AR	Hormone-nuclear Receptor Family				
FOXC2	Forkhead Domain Family	Homo sapiens	2.86e-09	5.86	
FOXC1	Forkhead Domain Family				
FOXA2	Forkhead Domain Family				
FOXB1	Forkhead Domain Family				
FOXA1	Forkhead Domain Family				
FOXJ1	Forkhead Domain Family				

BETA output of activating/ repressive function prediction and motif analysis of AR. (a) BETA activating/repressive function prediction of the AR data set from the LNCaP prostate cancer cell line. The red and the purple lines represent the upregulated and downregulated genes, respectively. The dashed line indicates the non-differentially expressed genes as background. Genes are cumulated by the rank on the basis of the regulatory potential score from high to low. P values that represent the significance of the UP or DOWN group distributions are compared with the NON group by the Kolmogorov-Smirnov test. (b) Motif scan algorithm. Motif scores in each binding peak are compared among three regions. The middle region consists of 200 bp centered on the peak summit; the left and right regions comprise 200 bp in either direction of the middle region. The significance of motif summit enrichment is measured by the P value from a one-tailed t test. (c) Screenshot of binding motif analysis on UP target regions of AR. Similar motifs are grouped together, and the motif logo of the most significant factor in the group is provided in the last column. The motif symbol, DNA-binding domain and species are shown in the first three columns; the t score and the P value from the t test are shown in the middle two columns.

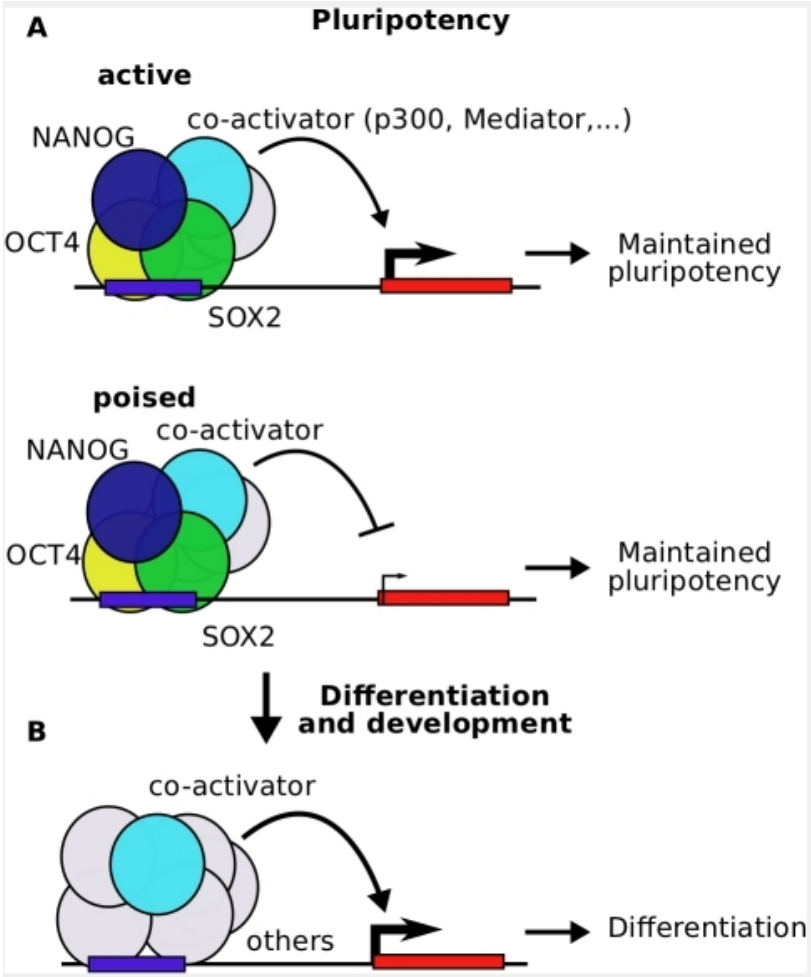
What happens when BETA is used w/o gene expression data?

Some expts a student (Iana Pyrogova) did using OCT4/SOC2/Nanog data…

OCT4/SOX2/NANOG

(A) Enhancers are bound by OCT4, SOX2 and NANOG together with p300 in embryonic stem cells. These enhancers maintain pluripotency by activating gene expression in ES cells (top) or poisoning expression for activation after differentiation (bottom)

(B) After differentiation of the cell, the same enhancers are bound by p300 in developmental tissues together with other transcription factors. The target gene is expressed



Göke et al., *PLoS Comput. Biol., 7(12):e1002304, 2011.*

Oct4 mm10

Input

- **TF name: OCT4**
- **Cell: ES**
- **Expt: ChIP-seq**

BETA pamaters

- **Genome: mm10**
- **Distance: 100kbp**

Predict target genes

Predicted target genes by BETA

#Chr	TSS	TTS	RefseqID	Score	Strand	GeneSymbol
chr10	60002804	60099990	NM_026937	1.167	+	Ascc1
chr15	58094046	58135082	NM_027435	1.158	-	Atad2
chr10	59987908	60003112	NM_025514	1.105	-	Anapc16
chr8	70539674	70592858	NM_007924	1.092	+	Ell
chr2	31572650	31617526	NM_001033389	1.064	+	Fubp3
chr2	31572650	31617526	NM_001290548	1.064	+	Fubp3

Compare top 500 of these two lists

- Perturbation expression data (from W.Sikora 2013)
  - Absolute expr fold change in response to TF perturbation

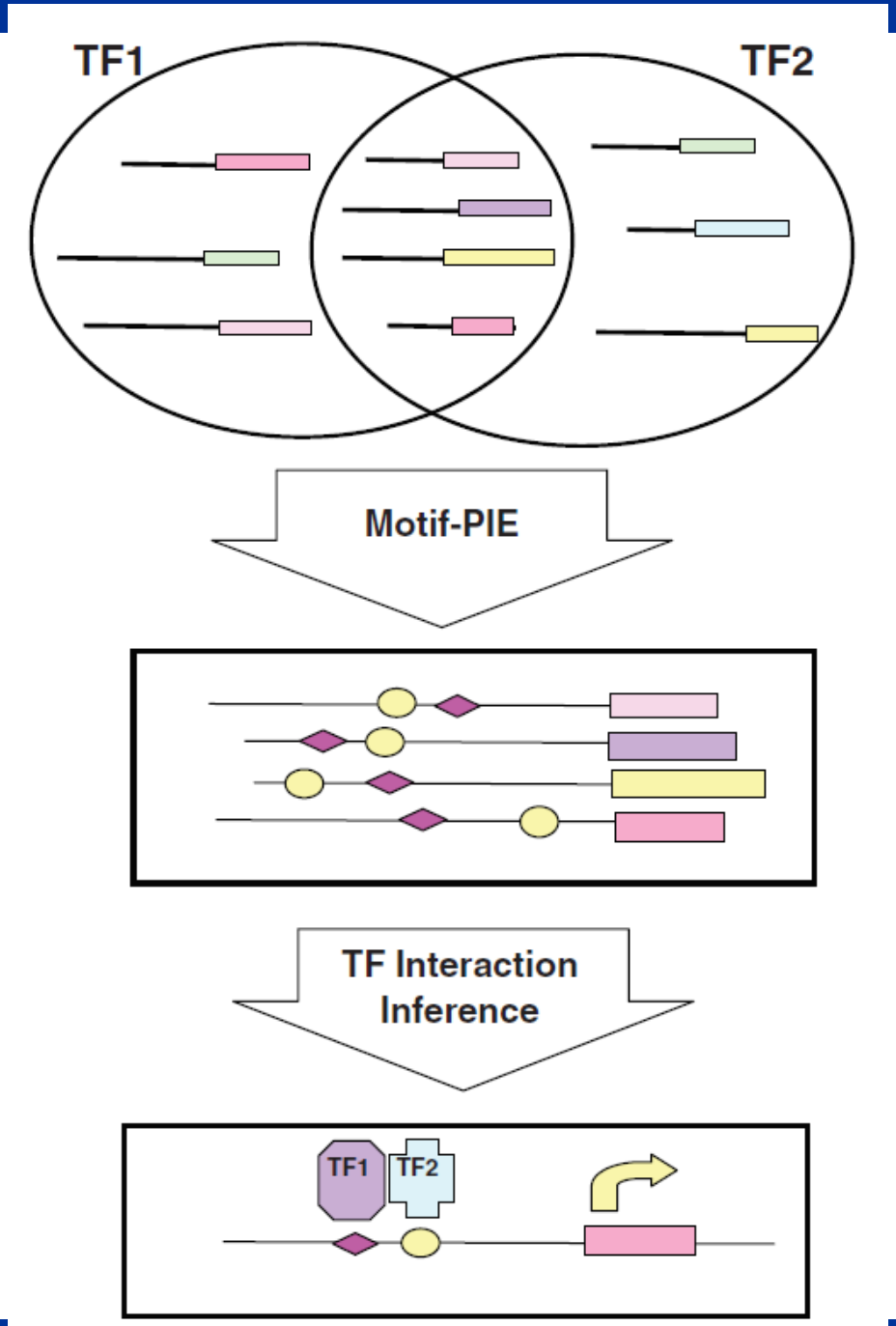
EnsemblGeneID	ExpressionFoldChange
ENSMUSG00000031179	8.868013732
ENSMUSG00000006200	8.385737772
ENSMUSG00000068048	7.99106154
ENSMUSG00000061082	7.980707822
ENSMUSG00000049382	7.736291716
ENSMUSG00000032085	7.599653246
ENSMUSG00000023039	7.40733782
ENSMUSG00000020911	7.407259106
ENSMUSG00000057615	7.26657795

overlap = 25

Out of top 500 BETA predictions based on OCT4 binding data only, a mere 5% (=25) target genes are confirmed by gene expression data

TF-TF interactions

NUS
National University
of Singapore



Basic ideas for identifying co-operative TFs

- **Check regulatory region of their common target genes for**
  - Binding-site co-occurrence enrichment
  - Relatively fixed binding distance between their binding sites

Yu et al., NAR, 34(3):917-927, 2006

Genome Res., 14(2):201-208, 2004

A Motif Co-Occurrence Approach for Genome-Wide Prediction of Transcription-Factor-Binding Sites in *Escherichia coli*

Martha L. Bulyk,[1,2,3,4] Abigail M. McGuire,[1,2,3] Nobuhisa Masuda,[2] and George M. Church[1,2,5]

Author information ▶ Article notes ▶ Copyright and License information ▶

Abstract

Go to:

Various computational approaches have been developed for predicting *cis*-regulatory DNA elements in prokaryotic genomes. We describe a novel method for predicting transcription-factor-binding sites in *Escherichia coli*. Our method takes advantage of the principle that transcription factors frequently coregulate gene expression, but without requiring prior knowledge of which groups of genes are coregulated. Using position weight matrices for 49 known transcription factors, we examined spacings between pairs of matrix hits. These pairs were assigned probabilities according to the overrepresentation of their separation distance. The functions of many open reading frames (ORFs) downstream from predicted binding sites are unknown, and may correspond to novel regulon members. For five predictions, knockouts with mutated replacements of the predicted binding sites were created in *E. coli* MG1655. Quantitative real-time PCR (RT-PCR) indicates that for each of the knockouts, at least one gene immediately downstream exhibits a statistically significant change in mRNA expression. This approach may be useful in analyzing binding sites in a variety of organisms.

1998). The matrix pairs were ranked according to either their most significant single spacing between 0 and 500 bp (e.g., exactly 3 bp) or their most significant spacing bin (McGuire 2000). Eight different spacing bins were examined (the bins including separation distances 0–30 bp, 30–60 bp, 60–90 bp, 0–100 bp, 100–200 bp, 200–300 bp, 300–400 bp, and 0–450 bp).

The rankings were based on the probability of obtaining the observed number of hits for the most overrepresented bin or spacing, given the number expected by chance for that particular bin or spacing. This number expected by chance was determined in the following manner:

$$E(x) = N_a \cdot N_b \cdot \pi(x - c), \tag{1}$$

where N_a and N_b are the number of hits in the genome using search matrices a and b, c is a correction factor to account for the lengths of the search matrices, and $\pi(x)$ is the probability that two randomly chosen noncoding base pairs are separated by a distance x. $\pi(x)$ was computed by tabulating the actual frequen-

Some technical details

Similarly, the probabilities of obtaining the observed number of hits within the eight different spacing bins was calculated:

$$P_{bin} = 1 - \sum_{s=0}^{obs(bin)-1} \binom{N_a \cdot N_b}{s} \cdot \Pi^s \cdot (1 - \Pi)^{N_a \cdot N_b - s}, \tag{3}$$

$$\Pi = \sum_{x=0}^{binsize} \pi(x - c), \tag{4}$$

$$obs(bin) = \sum_{x=0}^{binsize} obs(x), \tag{5}$$

where $obs(bin)$ is the observed number of hits in that spacing bin.

Time for Exercise #1

- **Motif co-occurrence is a useful approach for identifying TF-TF interactions. This approach has the advantage of not needing experimental data**
- **Discuss how you can make better prediction when some experimental data is available, and what type(s) of experimental data you should look for**

Must read

- [MEME] Bailey & Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning Journal*, 21:51-83, 1995

- [BETA] Tang et al. A comprehensive view of nuclear receptor cistromes. *Cancer Research*, 71:6940-6947, 2011

- Bulyk et al. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. *Genome Research*, 14:201-208, 2004

Good to read

- Bailey et al. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34:W369-W373, 2006
- Wang et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nature Protocol,* 8:2502-2515, 2013
- Geertz & Maerkl. Experimental strategies for studying transcription factor-DNA binding specificities. *Briefings in Functional Genomics*, 9:362-373, 2011
- Shin et al. Computational methodology for ChIP-seq analysis. *Quantitative Biology*, 1:54-70, 2013
- Jankowski et al. TACO: A general-purpose tool for predicting cell-type–specific transcription factor dimers. *BMC Genomics*, 15:208, 2014