Computational Analyses for Transcription Regulation

Jun Liu

Department of Statistics

Harvard University

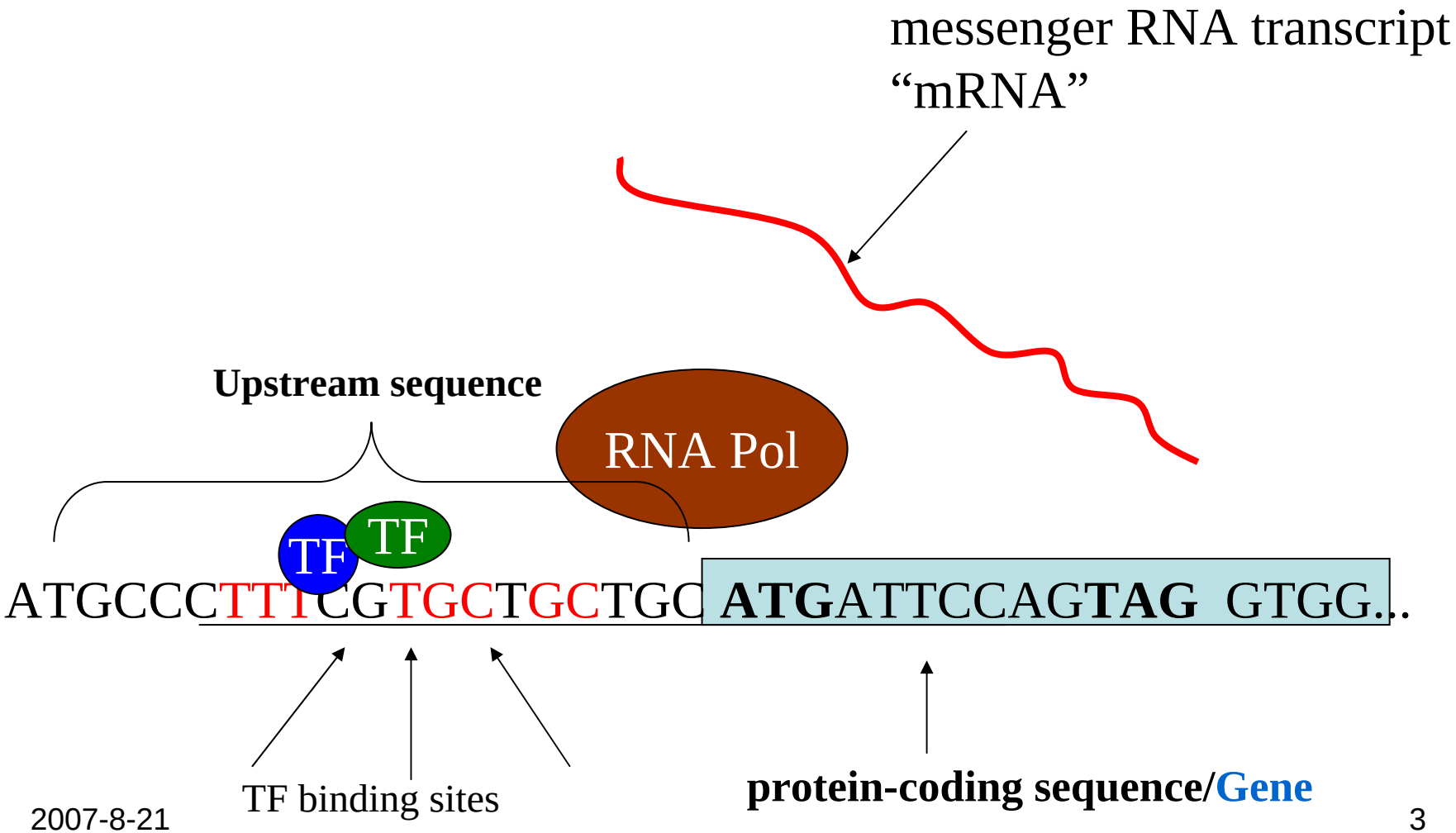
Email: jliu@stat.Harvard.edu

Http://www.fas.Harvard.edu/~junliu

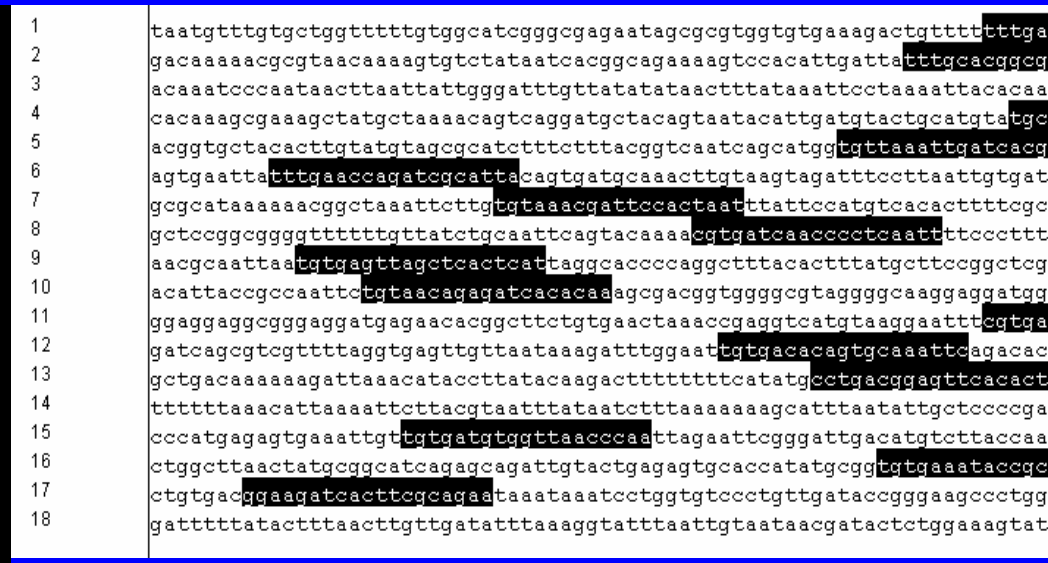
Outline

- General background of transcription regulation and gene expression
- Relating gene expression profiles to gene upstream/promoter sequence information
  - Sequence feature “extraction” – i.e., motif pattern discovery and filtering
- Regression and dimension reduction
- Summary

Transcription



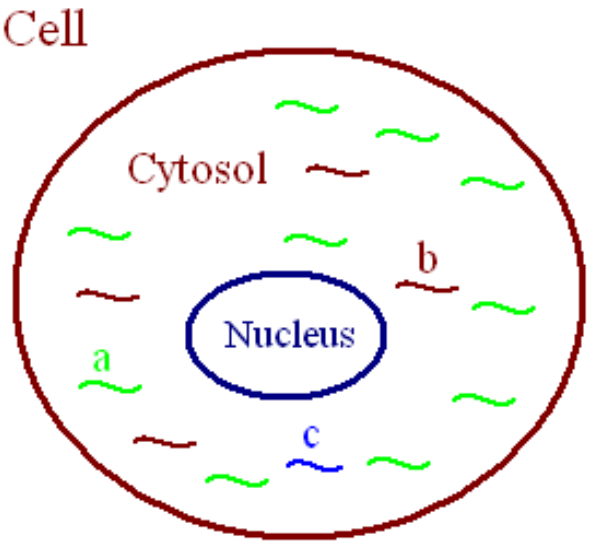
An example of TF binding motif (pattern)



TTTGATCGTTT**TCACA**AAA
TTTGCACGGCG**TCACA**CTT
TGTGAGCATGG**TCATA**TTT
TGCAAAGGACG**TCACA**TTA
TGTTAAATTGA**TCACG**TTT
TTTGAACCAGA**TCGCA**TTA
TGTAAACGATT**CCACT**AAT
CGTGATCAACC**CCTCA**ATT
TGTGAGTTAGC**TCACT**CAT
TGTAACAGAGA**TCACA**CAA

Sequence Motif Logo

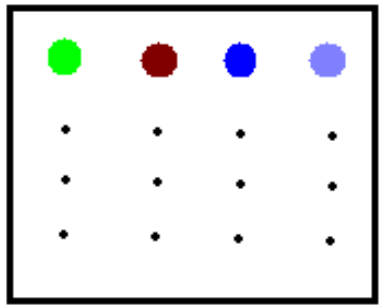
Principle of gene expression microarray:



Expressoin pattern

Gene	Level
a	high
b	Medium
c	low
⋮	⋮

Probe array



Microarray experiments

- mRNA expression array chips
  - cDNA microarrays
  - Affymatrix oligonucleotide arrays
  - Overall: giving a snapshot of the cell
- Chromatin Immuno-Precipitation + Chip
  - Study binding locations of a specific protein

Expression Information

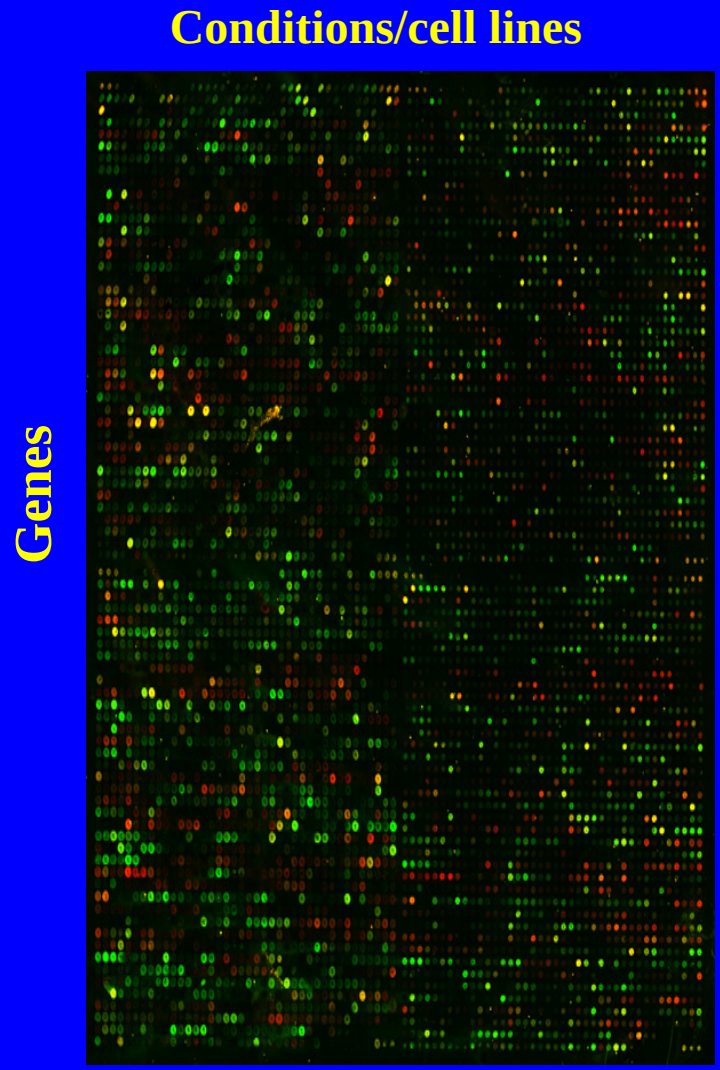
Microarray data tell us which genes are “up” or “down”

“co-expression”

?

“Co-regulation”

ChIP-chip data tell us which genes are under control of a TF



Task 1: single-array analysis

- Single-array experiment

	Sequence Information	Expression/Enrichment
Gene 1	… CACTAAGATAAGCGA	Y_1
Gene 2	… CAACATAGAAAACAGAAG	Y_2
⋮	⋮	⋮
Gene G	… CTTGGATCTGTCCATAA	Y_G

"Features" **Responses**

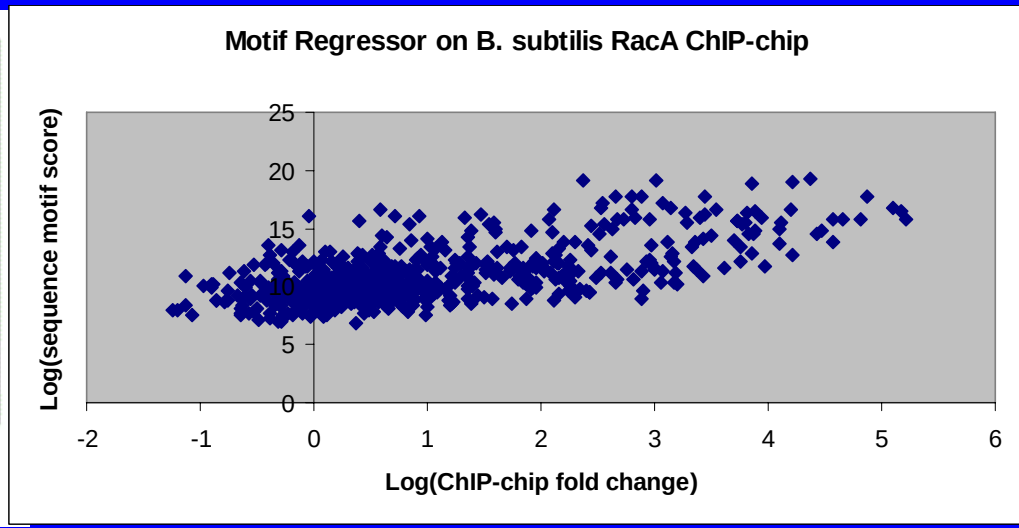
~800 bps upstream of the gene

	A	B	C	D	E	F	G	H	I
1	Scor	Score	Locatio	Site	Closest O	ChIP	Near	GC3	GC6
2	1	668982	2320606	TGACGCCGGGCTCA	ypwA	0.98	0	53.3	46.6
3	2	464994	3804572	TGGCGCCGGCCTCA	ywjI	20.7	1	60	50
4	3	408056	3862909	TGTCGCCGGCGACA	ywfL ywfK	6.62	1	70	65
5	4	344237	3847549	TGTCGCCAGCGTCA	ywhG	18.6	1	53.3	55
6	5	218459	237967	TGTGCCCGGCGTCA	ybfB ybfE	1.13	0	63.3	56.6
7	6	208948	1318599	TGAGCCTGGCGTCA	yjqB	1.02	0	56.6	61.6
8	7	204045	110339	TGACGCCAGCCTCG	yacM yacN	7.17	1	63.3	58.3
9	8	198744	1783376	TGTCGCCGGGCACA	pksC	0.68	0	63.3	46.6
10	9	195596	3935462	TGACGCCGCCGACA	ywbD	10.8	1	70	58.3
11	10	134515	3993264	CGAGGCCGGCGACG	yxjM yxjL	7.13	1	66.6	56.6
12	11	107573	132262	TGTCGGTGGGGTCG	fus	1.51	1	56.6	48.3
13	12	99427	233800	TGACGCTGCCGACA	glpQ glpT	4.99	0	53.3	50
14	13	90094	3888091	TGTCTCCGGGCTCA	spsD	34.4	1	73.3	58.3
15	14	81938	3889333	TGACGGCCGCCTCA	spsC	29.4	1	63.3	55
16	15	75897	66091	TGTCGCCGGCCTTA	yabM	14.8	1	53.3	58.3
17	16	70435	3866446	TGAGGCGAGCGTCA	ywfH	2.78	1	53.3	53.3
18	17	68274	4210531	CGGGGCCGCCGACA	gidA thdF	11.8	1	66.6	56.6
19	18	65196	3847072	TGCCGCCGGCGTCA	ywhH ywhG	13.7	1	66.6	60
20	19	63495	565099	TGAGTCCGGAGTCA	ydeF	1.2	0	43.3	36.6
21	20	63308	2963577	TGACGCCTGGCTCG	dnaB	0.93	0	56.6	38.3
22	21	60989	402603	TGACGGCGGGCACA	srfAD ycxA	1.42	0	53.3	46.6

Sheet1 Sheet2 RacAAllSiteLocGC

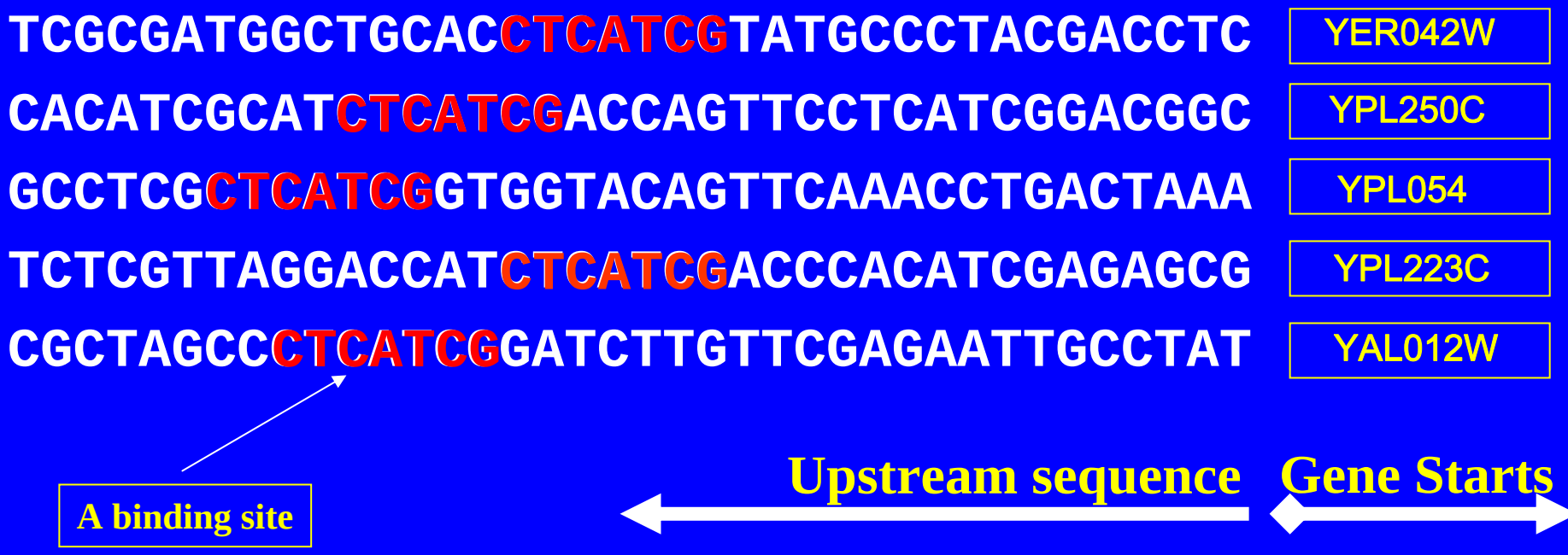
Motif Regressor on RacA ChIP-chip
(Ben-Yehuda et al. 2005, *Molecular Cell* **17**(6) 773-82)

- RacA: functional protein for sporulation, not TF
- ChIP-chip on ORF array, most targets near ORI
- Search space is too large for other algorithms
  - 49 ORFs with ChIP-chip fold change >9 (ORF plus 500 bps each side; average sequence length 2112)
  - 98 ORFs with fold change > 5 9 (average 2020)



The General Motif Finding Problem:

- Given a set of co-regulated genes, can we find "enriched" patterns in their upstreams?



Richard Bayes (1596-1675), a great-grandfather of Thomas Bayes, was a successful cutler in Sheffield. In 1643 Richard Bayes served in the rotating position of Master of the Company of Cutlers of Hallamshire. Richard was sufficiently well off that he sent one of his sons, Samuel Bayes (1635-1681) to Trinity College Cambridge during the Commonwealth period; Samuel obtained his degree in 1656. Another son, Joshua Bayes (1638-1703) followed in his father's footsteps in the cutlery industry, also serving as Master of the Company in 1679. Evidence of Joshua Bayes's wealth comes from the size of his house, the fact that he employed a servant and the size of the taxes that he paid. His influence may be taken from his activities in the town government. Following the 1662 Act of Uniformity, Samuel Bayes was ejected from his parish, eventually living in Manchester (Matthews, 1934). Joshua Bayes was closely involved in the erection of one Nonconformist chapel in Sheffield and had two sons-in-law involved in another Sheffield chapel. The second son of Joshua Bayes (1638-1703) was another Joshua Bayes (1671-1746). In 1686 the younger Joshua Bayes entered a dissenting academy where he studied philosophy and divinity. Joshua Bayes and his wife Anne née Carpenter were married some time, probably within days, after their marriage license was issued on October 23, 1700. Joshua and Anne Bayes had seven children and Thomas was the eldest.

Richard Bayes (1596-1675), a great-grandfather of Thomas Bayes, was a successful cutler in Sheffield. In 1643 Richard Bayes served in the rotating position of Master of the Company of Cutlers of Hallamshire. Richard was sufficiently well off that he sent one of his sons, Samuel Bayes (1635-1681) to Trinity College Cambridge during the Commonwealth period; Samuel obtained his degree in 1656. Another son, Joshua Bayes (1638-1703) followed in his father's footsteps in the cutlery industry, also serving as Master of the Company in 1679. Evidence of Joshua Bayes's wealth comes from the size of his house, the fact that he employed a servant and the size of the taxes that he paid. His influence may be taken from his activities in the town government. Following the 1662 Act of Uniformity, Samuel Bayes was ejected from his parish, eventually living in Manchester (Matthews, 1934). Joshua Bayes was closely involved in the erection of one Nonconformist chapel in Sheffield and had two sons-in-law involved in another Sheffield chapel. The second son of Joshua Bayes (1638-1703) was another Joshua Bayes (1671-1746). In 1686 the younger Joshua Bayes entered a dissenting academy where he studied philosophy and divinity. Joshua Bayes and his wife Anne née Carpenter were married some time, probably within days, after their marriage license was issued on October 23, 1700. Joshua and Anne Bayes had seven children and Thomas was the eldest.

Richard Bayes (1596-1675), a great-grandfather of Thomas Bayes, was a successful cutler in Sheffield. In 1643 Richard Bayes served in the rotating position of Master of the Company of Cutlers of Hallamshire. Richard was sufficiently well off that he sent one of his sons, Samuel Bayes (1635-1681) to Trinity College Cambridge during the Commonwealth period; Samuel obtained his degree in 1656. Another son, Joshua Bayes (1638-1703) followed in his father's footsteps in the cutlery industry, also serving as Master of the Company in 1679. Evidence of Joshua Bayes's wealth comes from the size of his house, the fact that he employed a servant and the size of the taxes that he paid. His influence may be taken from his activities in the town government. Following the 1662 Act of Uniformity, Samuel Bayes was ejected from his parish, eventually living in Manchester (Matthews, 1934). Joshua Bayes was closely involved in the erection of one Nonconformist chapel in Sheffield and had two sons-in-law involved in another Sheffield chapel. The second son of Joshua Bayes (1638-1703) was another Joshua Bayes (1671-1746). In 1686 the younger Joshua Bayes entered a dissenting academy where he studied philosophy and divinity. Joshua Bayes and his wife Anne née Carpenter were married some time, probably within days, after their marriage license was issued on October 23, 1700. Joshua and Anne Bayes had seven children and Thomas was the eldest.

In reality …

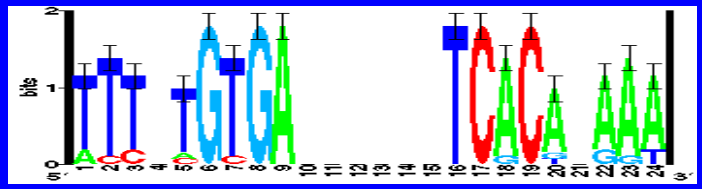
```
...TTTGATCGTTTTCACAAAA...
...TTTGCACGGCGTCACACTT...
...TGTGAGCATGGTCATATTT...
...TGCAAAGGACGTCACATTA...
...TGTTAAATTGATCACGTTT...
...TTTGAACCAGATCGCATTA...
...TGTAAACGATTCCACTAAT...
...CGTGATCAACCCCTCAATT...
...TGTGAGTTAGCTCACTCAT...
...TGTAACAGAGATCACACAA...
```

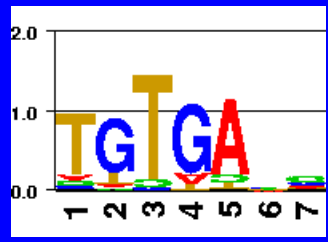
Aligned TF binding sites

- The “word” can be long
- Each “letter” of the “word” is not 100% conserved.
- Not every gene in consideration contain copies of words
- Positions of the word are very variable

Representation of a motif

- **Consensus:** TGTGA.....TCACA
- **Logo**





- **Weight matrix**

(product multinomial)

	1	2	3	…	*w*
A	.01	.01	.01		.39
C	.11	.01	.04		.01
G	.01	.55	.01		.01
T	.87	.43	.94		.59

TTTGATCGTTTTCACAAAA
TTTGCACGGCGTCACACTT
TGTGAGCATGGTCATATTT
TGCAAAGGACGTCACATTA
TGTTAAATTGATCACGTTT
TTTGAACCAGATCGCATTA
TGTAAACGATTCCACTAAT
CGTGATCAACCCCTCAATT
TGTGAGTTAGCTCACTCAT
TGTAACAGAGATCACACAA

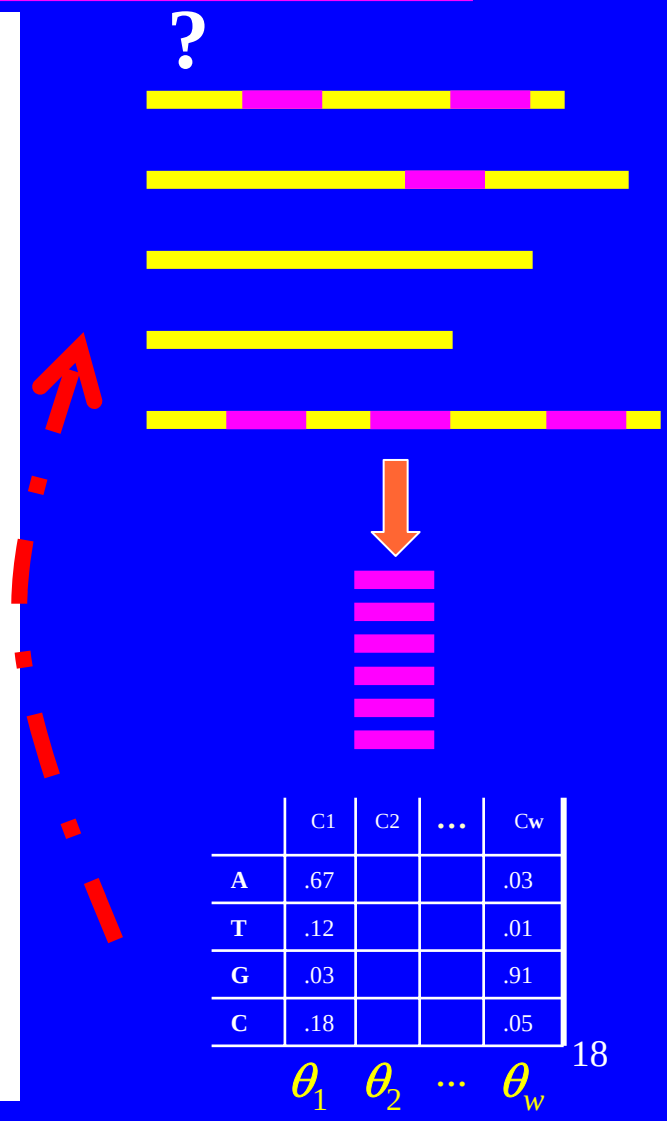
A motif

References for motif finding algorithms

- BioProspector (http://ai.stanford.edu/~xsliu/BioProspector/)
- MDscan (http://ai.stanford.edu/~xsliu/MDscan/)
- Motif Regressor (http://www.techtransfer.harvard.edu/Software/MotifRegressor)
- MEME (http://meme.sdsc.edu/meme/intro.html)
- Gibbs Motif Sampler (http://bayesweb.wadsworth.org/gibbs/gibbs.html)
- AlignACE
- CONSENSUS
- ANN-Spec
- YMF
- MotifSampler

The Motif Sampler

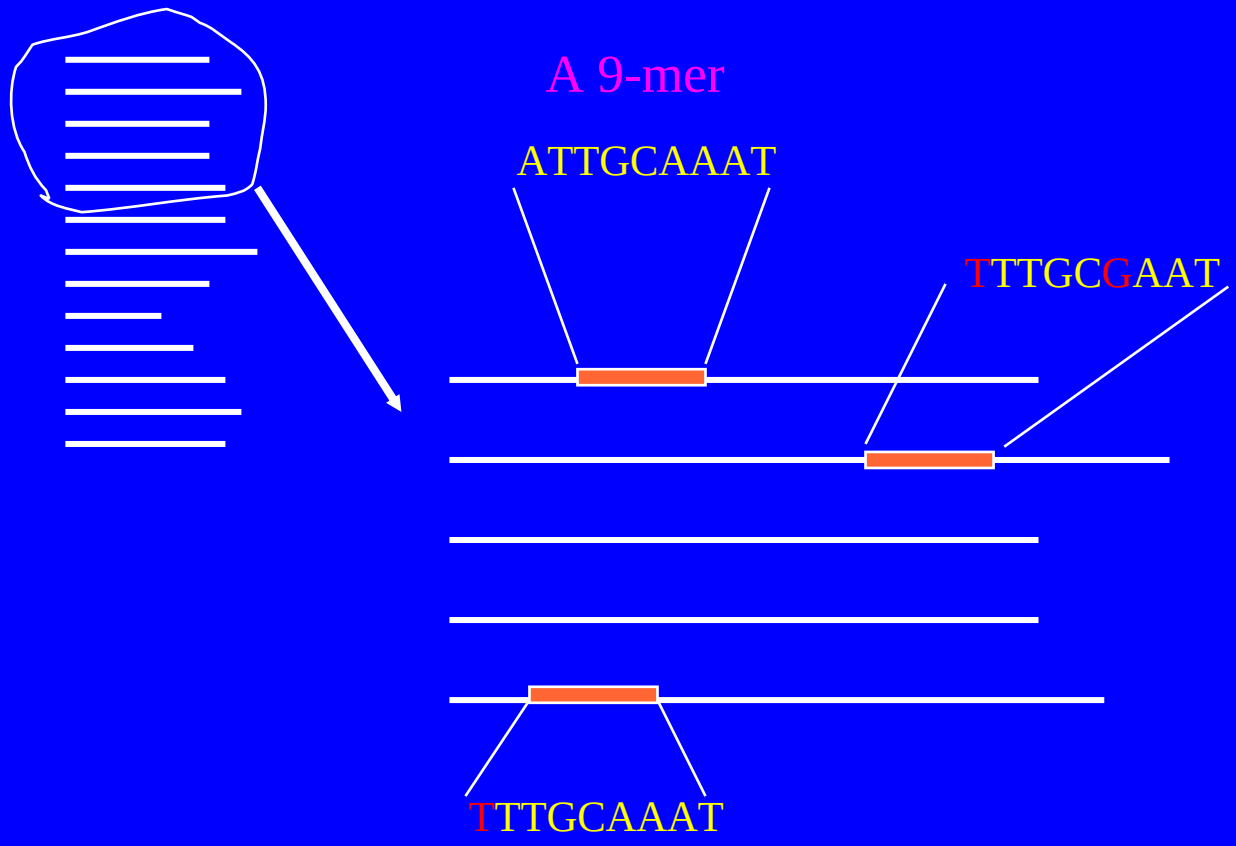
- Initialized by *random starting* positions $a_1^{(0)}, a_2^{(0)}, \ldots\ldots, a_K^{(0)}$
- Form a *weight matrix*
- Systematically go through every position in every sequence
  - Compute the *ratio* for that position being the motif start or not (signal-to-noise ratio)
  - Turn the corresponding *segment* on (as a motif site) or off according to the ratio (a Metropolis step)
- Stop when no significant changes, or some criterion met



Motif Regressor
(Conlon et al 2003, *Proc Nat'l Acad Sci*)

- Motivated by "REDUCE" (Bussemaker et al. 2001, *Nat Genet*)
- Goal: using array data to further enhance the motif discovery
  - "Reduce" and its drawback
- Idea:
  - Use a tool, MDscan, to quickly generate some sequence "features" (or "words")
  - Use regression to find "words" (*motif patterns*) that are related to the expression values

MDscan – “feature” extraction?



Candidate
Search pattern

ATTGCAAAT
TTTGCGAAT
TTTGCAAAT

Consequently:

- Many candidate motif patterns (N-w)

ATTGCAAAT
TTTGCGAAT
TTTGCAAAT

GCCACCGT
ACCACCGT
ACCACGGT
GCCACGGC
…

TTACTAA
TTGCTAA
TTAATAA
TAACTAA
…

GCAAA
GCAAA
GCAAA

AGGGGC
CGGGGC
AGGGGT
AGGGGT

……

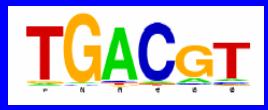
Motifs

The motifs are highly conserved but not identical.

Position weight matrix (frequency) motif logo

μ_j

A	0.05	0.02	0.85	0.02	0.21	0.06
C	0.04	0.02	0.03	0.93	0.05	0.06
G	0.06	0.94	0.06	0.04	0.70	0.11
T	0.85	0.02	0.06	0.01	0.04	0.77



GATGGCTGCACCACGTGTATGC *Upstream seq of gene i*

$S_{i,1}$ • • • S_{i,n_i-5}

S_{i,n_i-5} = similar(GTATGC|θ_j) =

$$\frac{\text{P(generate GTATGC from motif matrix)}}{\text{P (generate GTATGC from background)}}$$

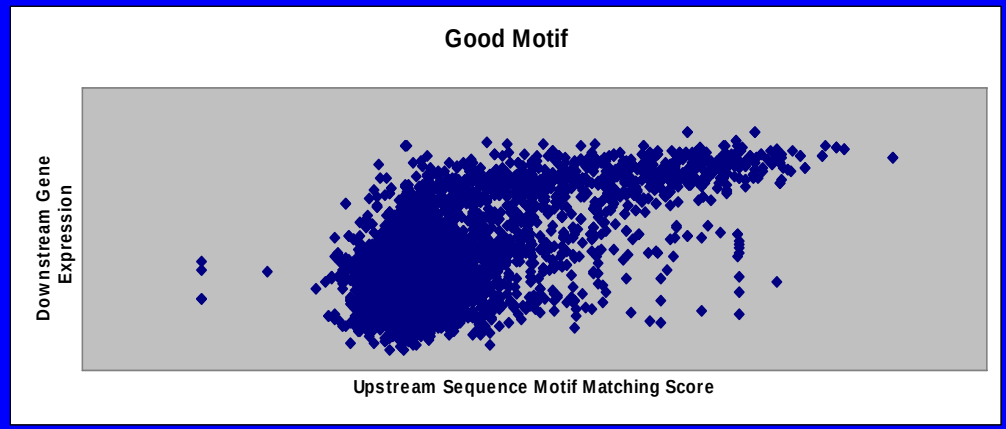
motif score $x_{ij} = \log_2 \sum_{k=1}^{n_i-w_j} s_{i,k}$

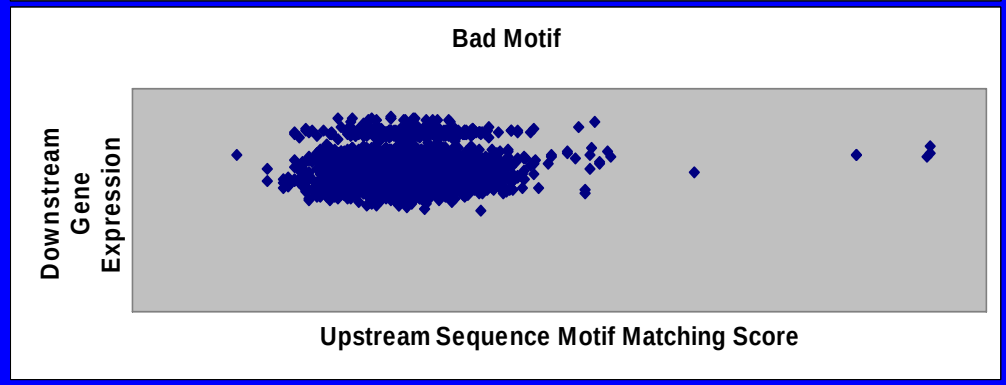
Motif Regressor (Conlon et al. 2003, PNAS)

- For each TF:

	Upstream Seq Mtf Mat	Downstream Gene Exp
Gene1	3.2	1.8
Gene2	2.8	0.3
Gene3…		

- Upstream sequence *X* motif matching score measures:
  - Number of sites
  - Strength of matching





Simple Regression MODEL

For each motif m:

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

where:

Y_g = $\log_2$ - ratio of expression for gene g

α = baseline expression

β_m = regression coefficient

S_{mg} = score for each motif m, gene g

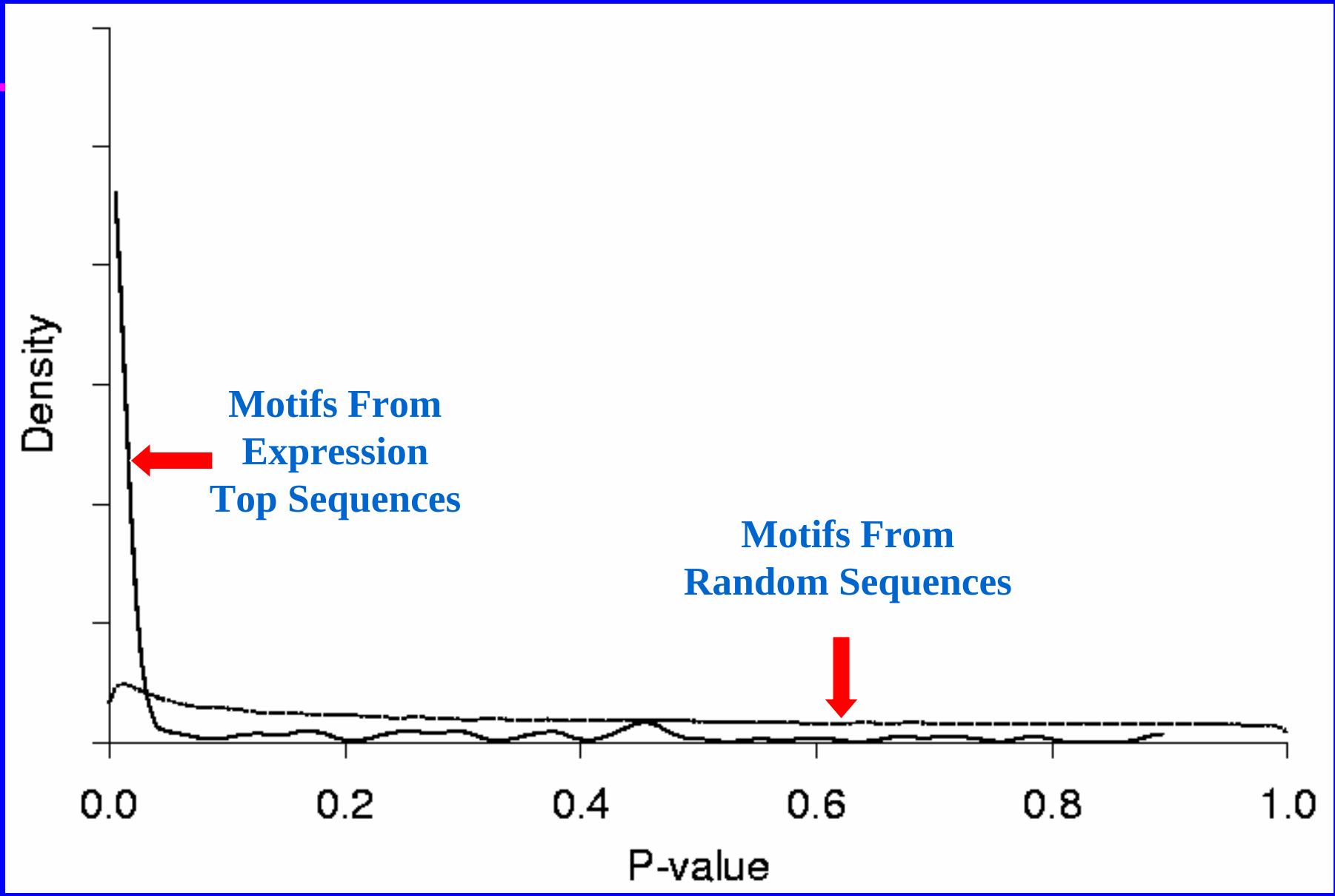
e_g = gene - specific error term

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

Validation from simulation

- Shuffle the gene names
- Find ~300 motifs in the “top” 100 genes
- Use the same regression to “re-confirm” the finding
- Results:
  - 1398 of 40,324 motifs have p-values <0.01 (3.5%)
  - For real data 235 out of 322 have p-values<0.01

Motif Regression P-value

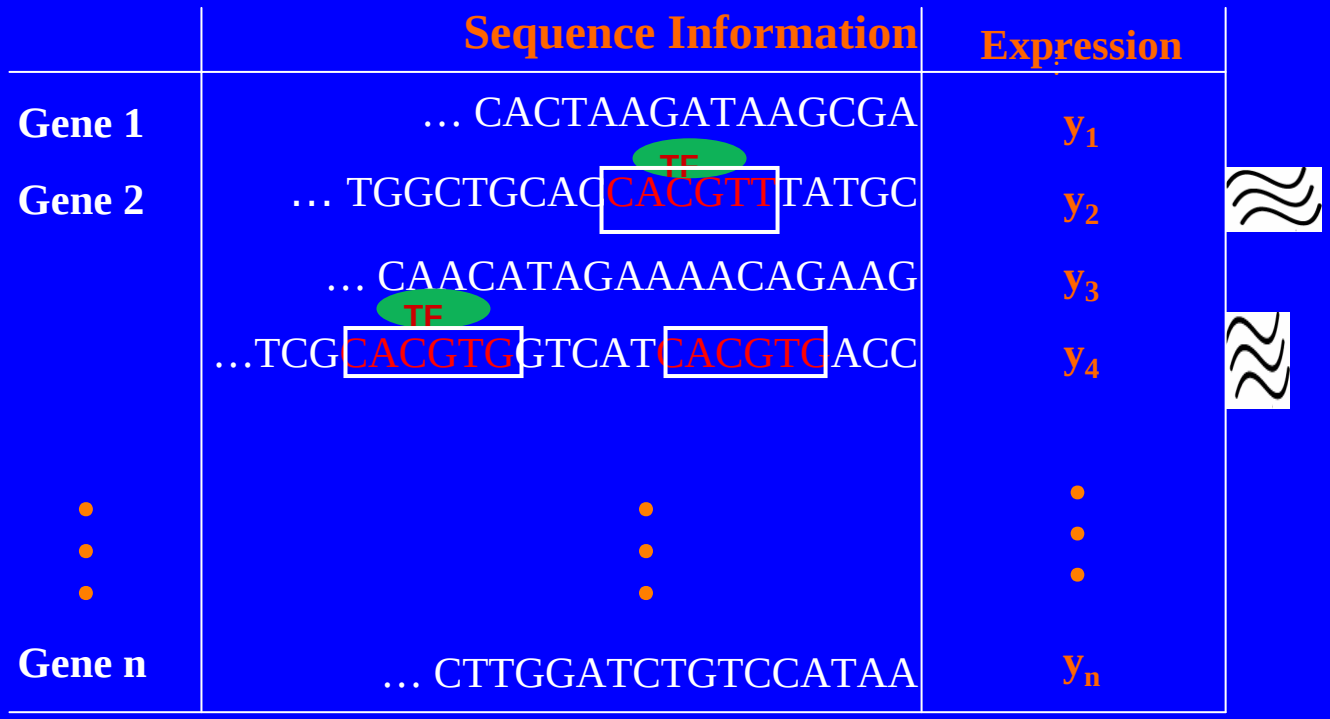


More sophisticated methods?

- Challenges:
  - Nonlinearity
  - Large number of possible variables
  - Generation of meaningful variables
- Strategies
  - Refine motif models (Zhou and Liu 2005)
  - Machine learning methods (Hong et al. 2005)
  - Slice-inverse regression (SIR)+variable selection
  - Modified Lasso et al?

Recap:

- Data structure and goals



Single-index model (Wenxuan Zhong)

- Nonlinear function of a linear combination

$$Y = f(X\beta, \varepsilon)$$

  - The link function f is unknown (continuously diff)
  - Assumptions:

$$X \sim N(0, I_p); \quad \|\beta\| = 1; \quad \varepsilon \sim N(0, \sigma^2)$$

- Connection with linear regression
  - LS: $\hat{\beta}_{LS} = \arg\max_{\beta} R^2(\beta) \equiv \arg\max_{\beta} corr^2(Y, X\beta)$
  - SI: $\arg\max_{T,\beta} R^2(T, \beta) \equiv \arg\max_{T,\beta} corr^2(T(Y), X\beta)$

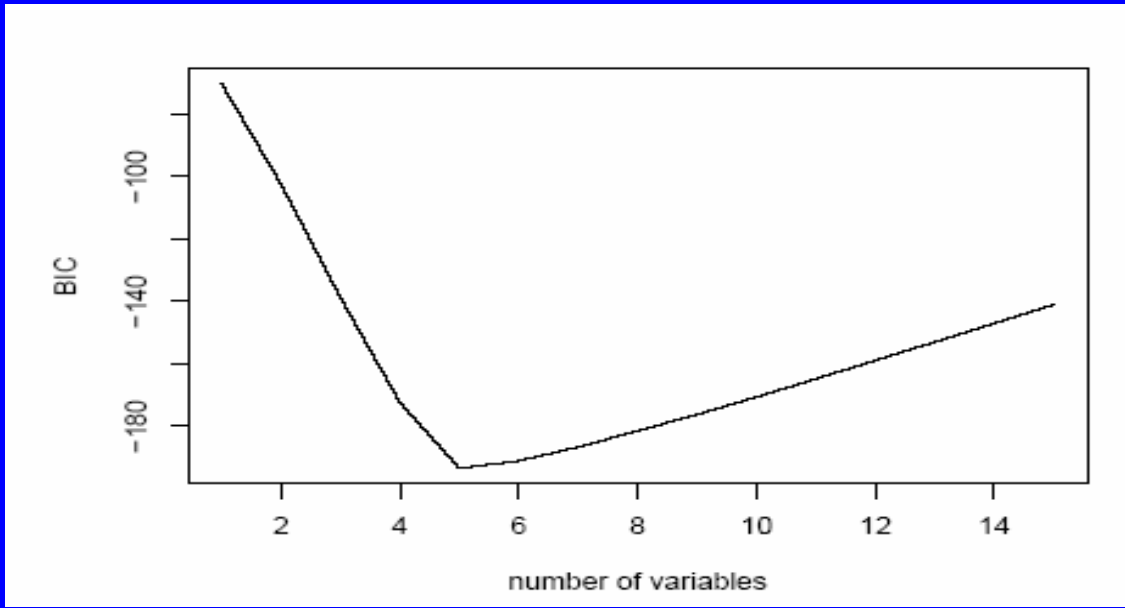
Empirical performance

15 candidate variables

$$y = \frac{x_2 + x_4 + x_6 + x_8 + x_{10}}{0.5 + (x_2 + x_4 + x_6 + x_8 + x_{10} + 1.5)^2} + \varepsilon^2$$

	Correct (2,4,6,8,10)	Wrong		
		1	2	3
Selection	0.75	**0.20**	**0.04**	**0.01**

	15 variables	5 variables
R^2	0.40	0.39



Motif identification for heat shock

Response (y)

Gene name	
YAL001	5.20
⋮	⋮
YBL0610	6.31

Regressor (X)

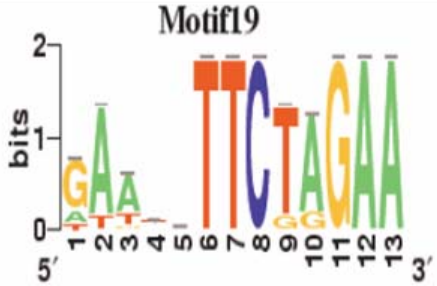
	$Motif_1$	...	$Motif_{666}$
YAL001	3.5	…	2.7
⋮	⋮	⋮	⋮
YBL0610	4.1	…	6.4

- Species: Yeast
- n=2587
- Gasch et al 2000

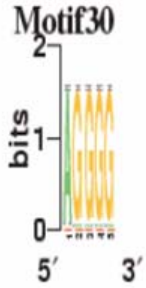
- Candidate motifs: (Beer and Tavazoie, 2004) p=666
- 51 known in literature
- 615 novel motifs

Goal: find motifs most related to the gene expression

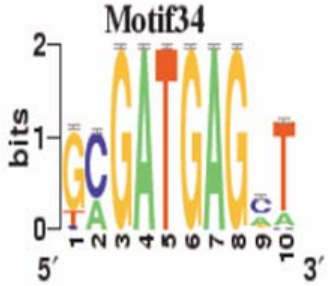
HSF1



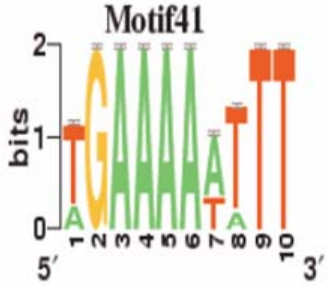
MSN24a



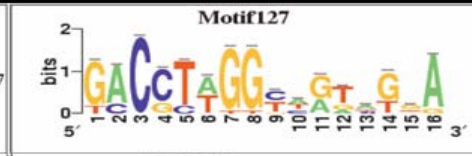
PAC

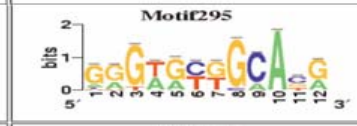


RRPE

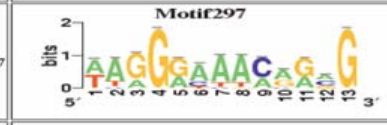


Motif127

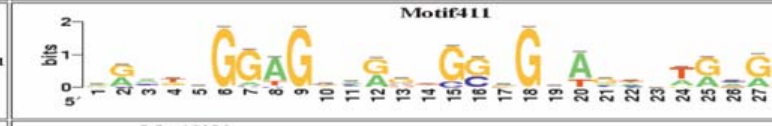




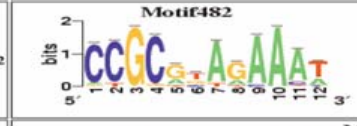
Motif297



Motif411



Motif482

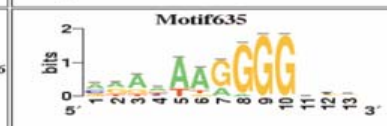


Motif555

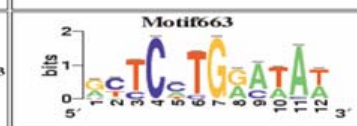
Motif623

Motif624

Motif626



Motif663

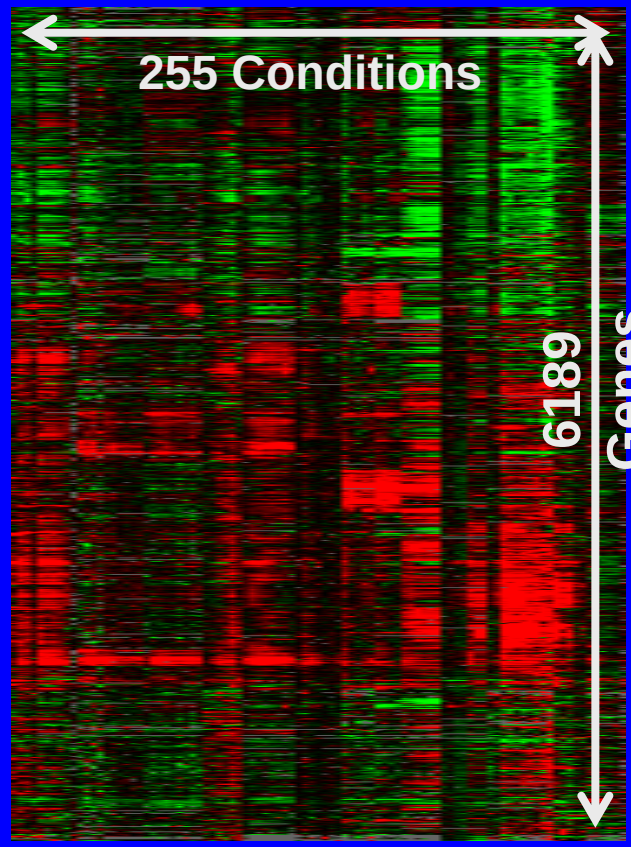


Comparison with LARS and stepwise

	Number of variables	CV (MSE)
Stepwise Regression	134	2.150144
LARS(LASSO)	122	1.062253
SLMS	15	0.7066146

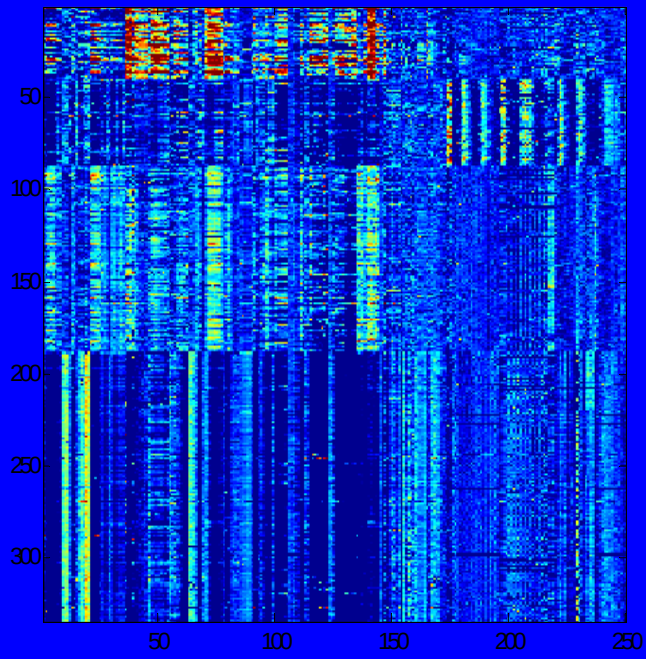
Multiple arrays – biclustering

- Example: yeast data
- Typical approaches: clustering
  - Hierarchical clustering
  - K-means
  - SOM, etc.
- Two-way clustering
  - ISA, PISA, etc.
  - More models? Plaid model



Motivation of biclustering

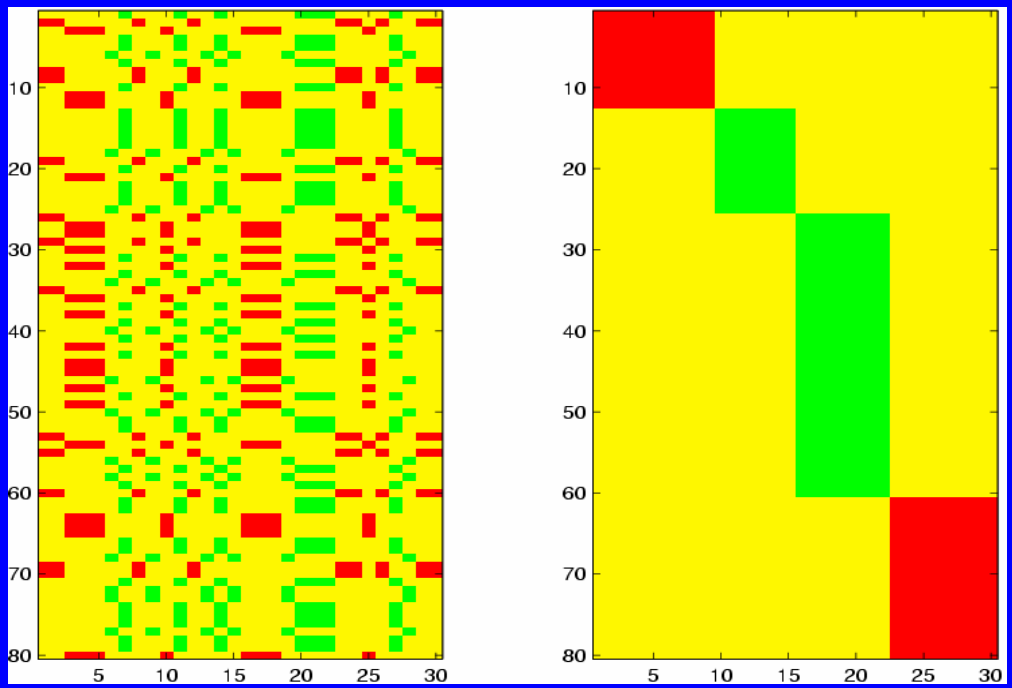
- Biclustering of genes expression data
  - Find a subset of genes and corresponding subset of conditions simultaneously.
  - Genes/conditions do not necessarily belong to a cluster .

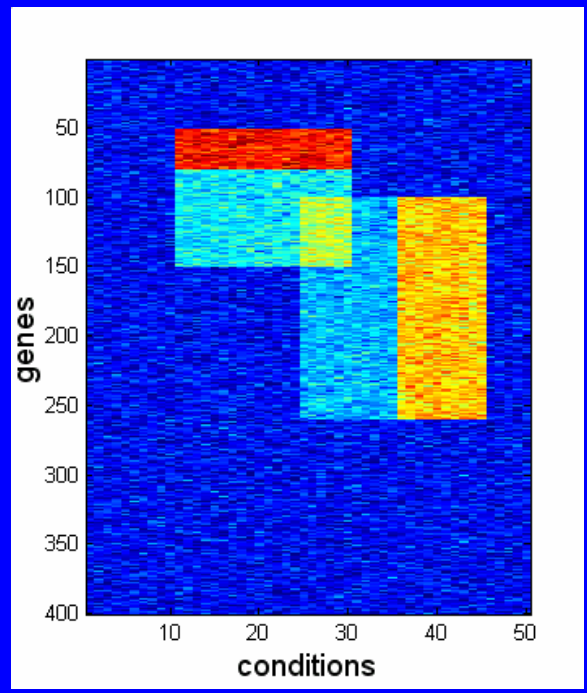


(Prepared by Jiajun Gu)

The Plaid model (Lazzeroni & Owen 2001)

- A two-way additive model with column and row selections $Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk}$
- Simultaneously selecting genes and experiments/samples





Bayesian Plaid Model

- Bayesian method

$$P(\textit{parameters} \mid Y_{obs}) \propto P(Y_{obs} \mid \text{parameters}) P(\text{parameters})$$

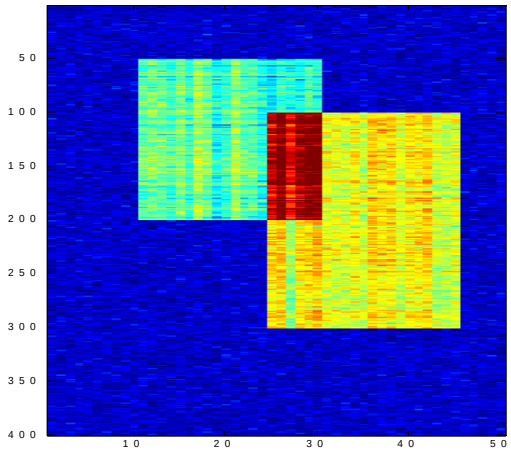
- Hierarchical Model

$$P(Y_{ij} \mid \mu_0, \mu_k, \alpha_{ik}, \beta_{jk}, \delta_{ik}, \kappa_{jk}, \tau_\varepsilon, k = 1 \ldots K) = N(Y_{ij}; \theta_{ij}, \tau_\varepsilon)$$

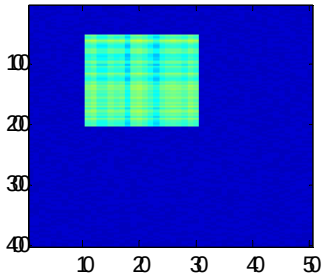
$$\text{where} \quad \theta_{ij} = \mu_0 + \sum_{k=1}^{K} \left(\mu_k + \alpha_{ik} + \beta_{jk} \right) \delta_{ik} \kappa_{jk}$$

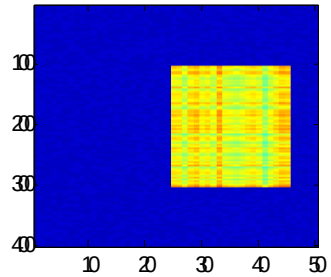
Comparisons of Biclustering results

Data

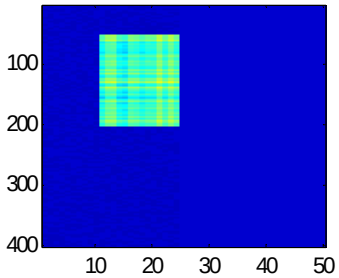


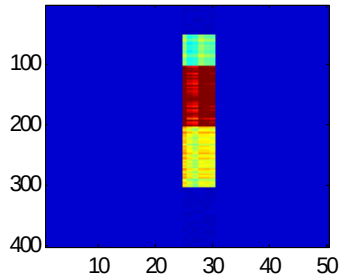
Method 1

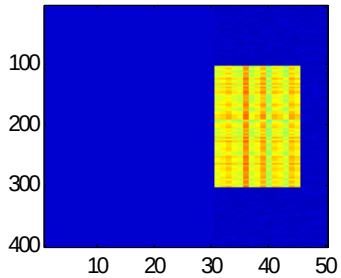




Method 2







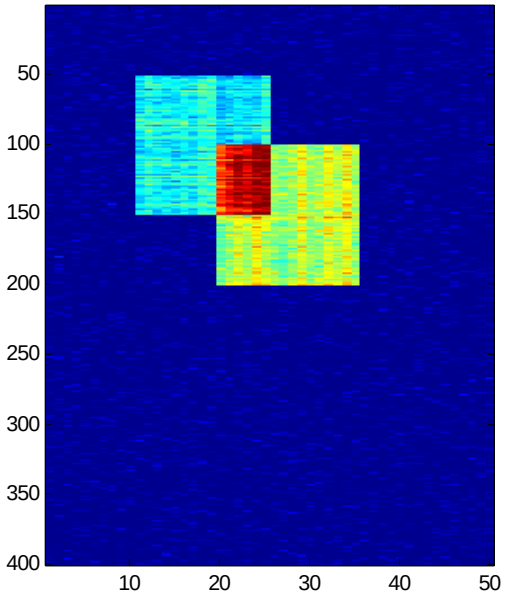
Method 3,

……

(Prepared by Jiajun Gu)

Comparison results I: Different Algorithms

Scenario II: Two clusters with additive overlapping effects



	thres	sensitivity	specificity	overlapping	# clusters
ISA	0.6, 1.2	0.53	0.90	0.08	8
ISA	0.7, 1.1	0.68	0.84	0.16	8
ISA	0.6, 1	0.84	0.84	0.12	3
Plaid		1	0.73	0.63	11
Cheng & Church		0.98	0	0	10
Bayesian Plaid		1	1	0	3

- The performance of ISA is very sensitive to thresholds.
- The performance of the Plaid model is affected by the complexity of the data structure.
- Cheng and Church's algorithm includes too many background genes.

Application of BBC to yeast data: discovery of biologically significant biclusters

- The yeast microarray data of 6108 genes and 250 conditions: yeast cell cycle (Spellman et al. 1998) and yeast environmental stress data (Gasch et al 2000) combined.
- We first normalized using 90% IQRN, and then use the Bayesian biclustering model.
- 50 biclusters were found.
- 3 types of enrichment tests for each cluster:
  - gene functional terms (MIPS): 43 significant clusters.
  - experimental condition categories: 44 significant clusters
  - TFBS on the promoter sequences (using TFBS score for 51 known yeast TFs): 27 significant clusters.

(Prepared by Jiajun Gu)

Biologically significance was shown in most biclusters

- Examples of biclusters

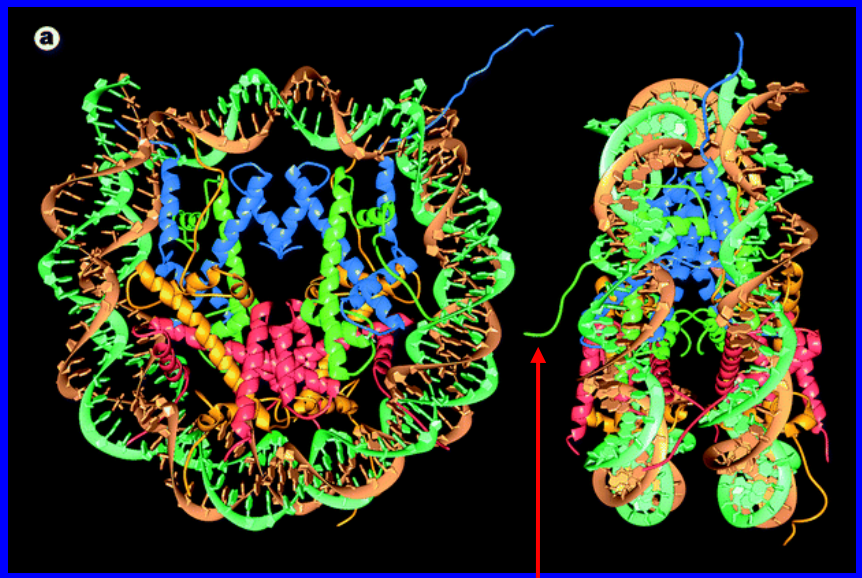
Cluster name	Size	Significant conditions (P-value)	Enriched TFBS (P-value)	Enriched gene functions (P-value)
G1/S phase	47,170	Cln3/clb2(6.4e-3), alpha factor (2.7e-4), cdc15(6.6e-3)	MBP1(5.8-11)	Cell cycle & DNA processing (4.5e-7)
G1 phase I	23,173	Alpha factor (3.1e-5), cdc15 (1.7e-5)	ACE2(1.1e-5)	Cell cycle (5.8e-3), cytokinesis (5.6e-3)
G1 phase II	130,182	Cln3&clb2(8.2e-3), alpha factor(4.8e-4), cdc28(2.4e-4)	MBP1(1.8e-30), SWI4(1.3e-12)	Cell cycle and DNA processing
Nitrogen, sulfur & selenium metabolism	66,104	Amino acid starvation (5.7e-5), nitrogen depletion (2.4e-5)	CBF1(2.1e-7), GCN4(4.9e-9), MET31(1.5e-2), MET4(3e-2)	Amino acid metabolism (1.3e-30), nitrogen, sulfur and selenium metabolism (3.4e-27)
Oxidative stress response	39,124	H2O2 (4.6e-8), Menadione(8.3e-4), diamide(3e-4)	CAD1 (4.4e-15), YAP1(7.9e-12)	Oxidative stress(2.2e-15), detoxification (5.8e-14)
Mating	27,99	Alpha factor synchronization	STE12(9.3E-9)	Pheromone response, sex specific proteins (3.5e-14), mating (3.9e-11)
Ribosome protein	147,224		RAP1(2.9e-60)	Ribosome protein (2.1e-160)

What next steps?

- Motif finding?
- Transcription module building?
- Predictive models? (Beer & Tavazoie 04).

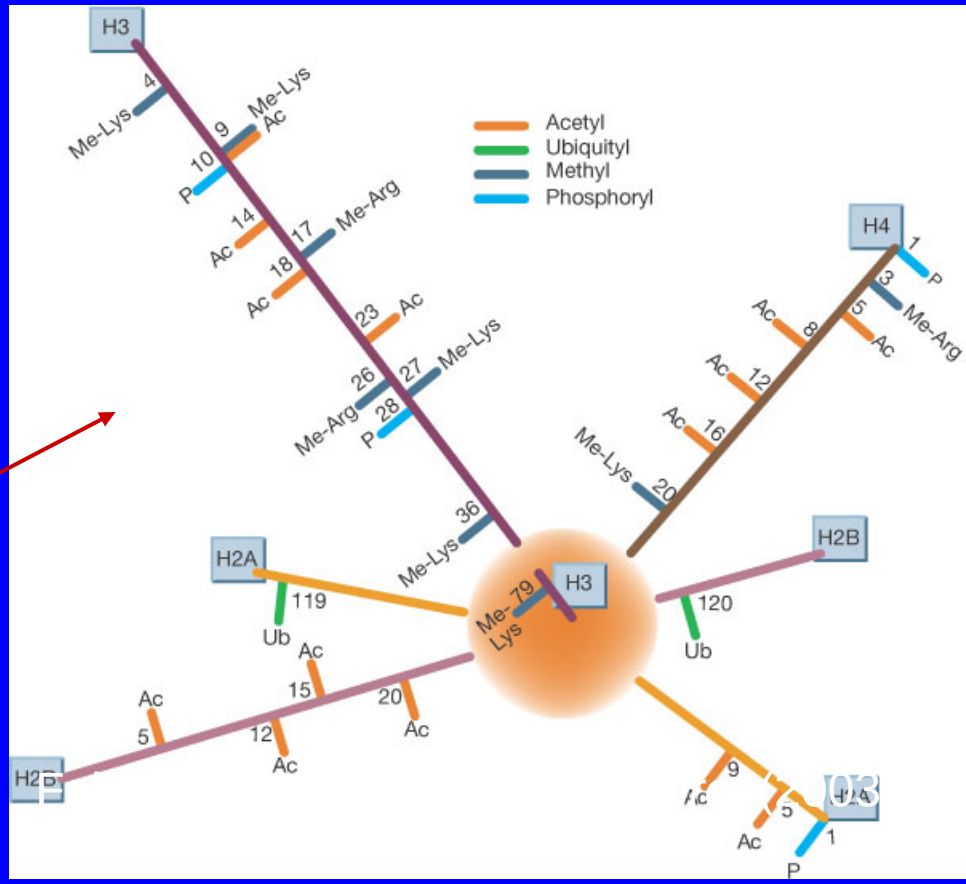
Histone modification (Yuan et al. 2006)

Acetyl
Ubiquityl
Methyl
Phosphoryl



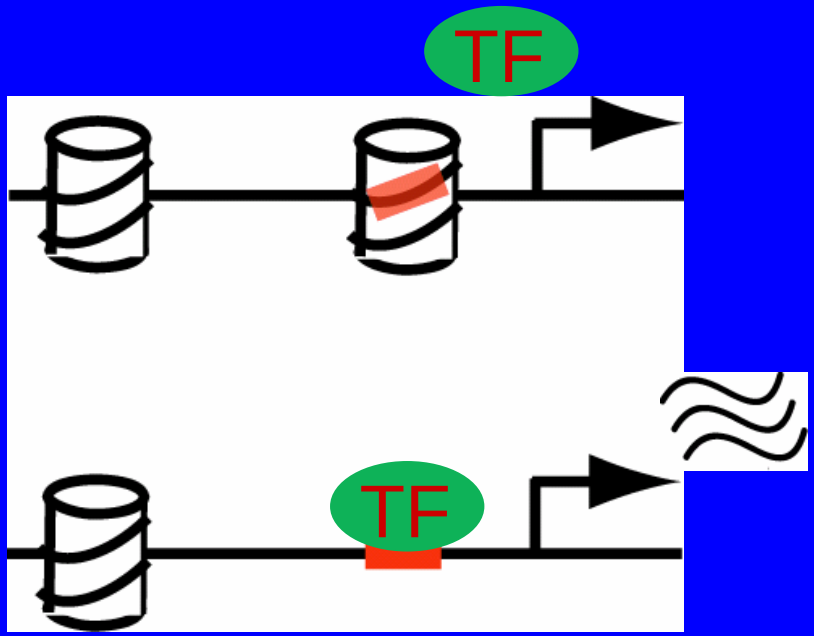
Luger *et al. Nature*, (1997)

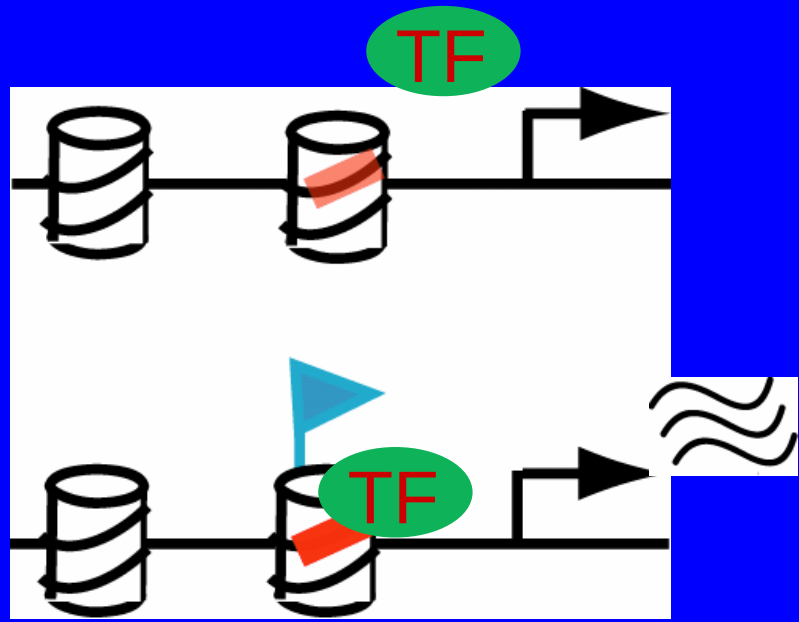
Histone tails can be covalently modified in multiple ways each at multiple sites



Regulatory role of chromatin

- Nucleosome positioning
- Histone modification





Assessing the global impact of histone acetylation on gene expression

- What are the regulatory effects of various combinations of histone acetylation sites?
- Do different acetylation sites play different regulatory roles?
- What are “real” effects of histone acetylation?

Yuan *et al.* *in preparation*

Challenges and Data Sources

- Many combinations of different histone acetylation
- Confounding effect of sequence dependent gene regulation, histone occupancy
- Multiple data sources:
  - Histone acetylation: H3K9, H3K14, H4 (Kurdistani et al. 2004; Pokholok et al. 2005)
  - Nucleosome occupancy data (Bernstein et al. 2004; Lee et al. 2004; Pokholok et al. 2005),
  - Transcriptional rate data (Bernstein et al. 2004)

A simple regression approach

- A regression model:

$$y_i = \alpha + \sum_j \beta_j A_{ij} + f(MS_{1i}, \ldots, MS_{ki}) + \varepsilon_i$$

y_i expression; A_{ij} acetylation; S_i promoter sequence

- Motif search via MDscan or AlignACE (resulting hundreds of motif patterns).
- Motif Regressor: Stepwise regression
- Dimension reduction using a modified SIR (Li 1991) method, RSIR, to estimate major "directions" of $f(MS_{1i}, \ldots, MS_{ki})$.
  - RSIR selected 104 motifs from MDscan results
  - RSIR selected 69 motifs from the 666 motifs of B&T
- We found that $f(...)$ is approximately linear

Estimating sequence dependent regulation

Linear regression model with TFBMs

$$y_i = \alpha + \sum_j \beta_j A_{ij} + \sum_j \eta_j S_{ij} + \varepsilon_i$$

linear $f(S_i)$

S_{ij} motif score

Results: R-squares

Model	Motifs only	H3 only	H4 only	H3+H4 only	H3 + Motifs	H4 + Motifs	H3+H4 Motifs	H3+H4 Mot+Occ
R-sq	0.1435	0.1808	0.0849	0.1841	0.2684	0.2093	0.2689	**0.3262**
AIC	11848	11612	11948	11601	**11369**	11605	11368	-
BIC	12486	11636	11967	11631	12020	12249	12024	-

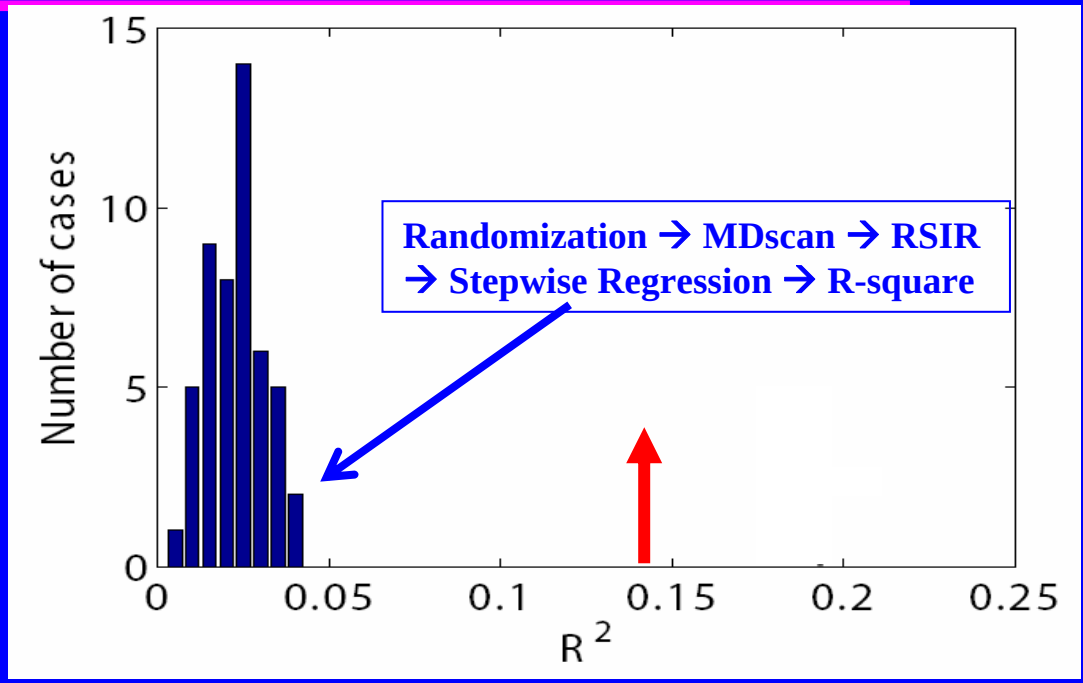
Using known motifs (from B&T 2004)

R-squares are adjusted by model degrees of freedom

	adjusted R-square		
	acetyl alone	acetyl + Mdscan motifs	acetyl + AlignACE motifs
No H3 or H4	0	0.2094	0.2223
H3 (K9, K14)	0.2443	0.354	0.3546
H4	0.1342	0.2979	0.3128
H3 and H4	0.246	0.3535	0.3543

Over-fitting issue

- Randomization Test of Motif Effect



- Cross-validation MSE

	Training	Predict
IGR	1.50	1.52
ORF	1.48	1.50

Acknowledgement

- Jiajun GU, Harvard University
- Guocheng Yuan, DFCI, Harvard Biostat
- Ping Ma, UIUC
- Michael Zhu, Purdue University
- Wenxuan Zhong, Harvard University
- Tingting Zhang, Statistics, Harvard

- **NIH,NSF**