

Predictive Modeling for Transcription Regulation

Jun Liu
 Department of Statistics
 Harvard University
 Email: jliu@stat.harvard.edu
 Http://www.fas.harvard.edu/~junliu

6/20/2007

1

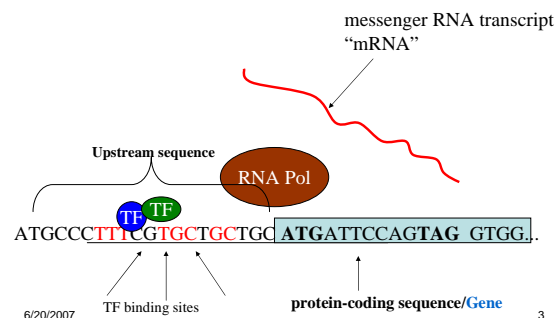
Outline

- ◆ 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

6/20/2007

2

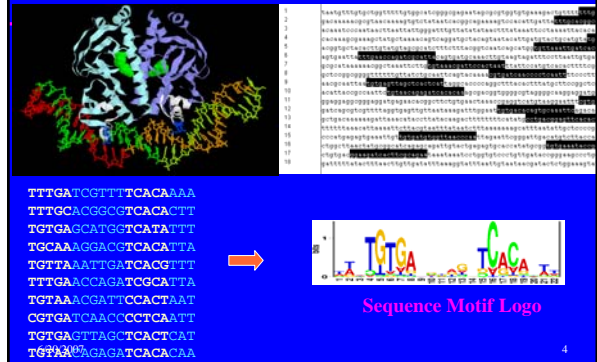
Transcription



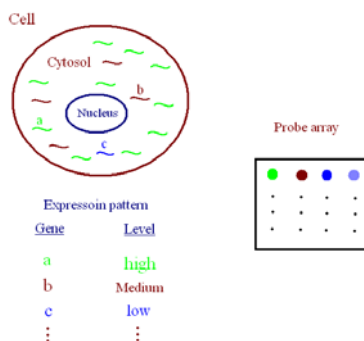
6/20/2007

3

An example of TF binding motif (pattern)



Principle of gene expression microarray:



6/20

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

6/20/2007

6

Expression Information

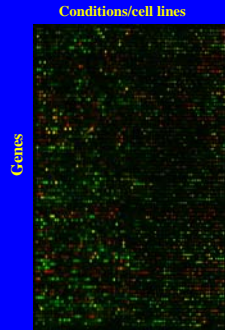
Microarray data tell us which genes are "up" or "down"

"co-expression"



"Co-regulation"

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



6/20/2007

Task: an abstraction

◆ Single-array experiment

| | Sequence Information | Expression/Enrichment |
|--------|------------------------|-----------------------|
| Gene 1 | ... CACTAAGATAAGCGA | Y_1 |
| Gene 2 | ... CAACATAGAAAACAGAAG | Y_2 |
| ⋮ | ⋮ | ⋮ |
| Gene G | ... CTGGATCTGTCCATAA | Y_G |

"Features"

Responses

-800 bps upstream of the gene

6/20/2007

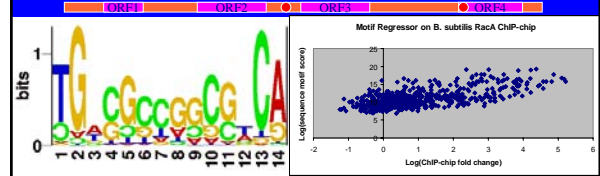
8

| | A | B | C | D | E | F | G | H | I |
|----|-------|--------|----------|----------------|-------------|------|------|------|------|
| 1 | Score | Score | Location | Site | Closest ORF | ChIP | Near | GC3 | GC6 |
| 2 | 1 | 668982 | 2320606 | TGAGCGGGGCTCA | ypwA | 0.98 | 0 | 53.3 | 46.6 |
| 3 | 2 | 464994 | 3804572 | TGGCGCGGGCTCA | ywlJ | 20.7 | 1 | 60 | 50 |
| 4 | 3 | 408056 | 3862909 | TGTGCGCGGCGACA | ywlL ywlK | 6.62 | 1 | 70 | 65 |
| 5 | 4 | 344237 | 3847549 | TGTGCGCAGGCTCA | ywhG | 18.6 | 1 | 53.3 | 55 |
| 6 | 5 | 218459 | 237967 | TGTGCGCGGCGTCA | ybfB ybfE | 1.13 | 0 | 63.3 | 56.6 |
| 7 | 6 | 208948 | 1318599 | TGAGCTGGGCTCA | yjqB | 1.02 | 0 | 56.6 | 61.6 |
| 8 | 7 | 204045 | 110339 | TGAGCGCAGGCTCG | yacM yacN | 7.17 | 1 | 63.3 | 58.3 |
| 9 | 8 | 198744 | 1783376 | TGTGCGCGGGCACA | pksC | 0.68 | 0 | 63.3 | 46.6 |
| 10 | 9 | 195596 | 3935462 | TGAGCGCGGCGACA | ywbD | 10.8 | 1 | 70 | 58.3 |
| 11 | 10 | 134515 | 3993264 | CGAGGCGGGGACG | yxjM yxjL | 7.13 | 1 | 66.6 | 56.6 |
| 12 | 11 | 107573 | 132262 | TGTGCGTGGGGTCG | fus | 1.51 | 1 | 56.6 | 48.3 |
| 13 | 12 | 99427 | 233800 | TGAGCTGCGGACA | glpO glpT | 4.99 | 0 | 53.3 | 50 |
| 14 | 13 | 90094 | 3888091 | TGTCTCGGGCTCA | spsD | 34.4 | 1 | 73.3 | 58.3 |
| 15 | 14 | 81938 | 3889333 | TGAGCGCGGCTCA | spsC | 29.4 | 1 | 63.3 | 55 |
| 16 | 15 | 75897 | 66091 | TGTGCGCGGCTTA | yabM | 14.8 | 1 | 53.3 | 58.3 |
| 17 | 16 | 70435 | 3866446 | TGAGCGCAGGCTCA | ywhH | 2.78 | 1 | 53.3 | 53.3 |
| 18 | 17 | 68274 | 4210531 | CGGCGCGGCGGACA | gidA thdF | 11.8 | 1 | 66.6 | 56.6 |
| 19 | 18 | 65196 | 3847072 | TGCGCGCGGCTCA | 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 | TGAGCGCTGGCTCG | dnaB | 0.93 | 0 | 56.6 | 38.3 |
| 22 | 21 | 60080 | 407673 | TGAGCGCGGCGACA | erfA ywA | 1.47 | 0 | 53.3 | 46.6 |

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?

TCGCGATGGCTGCACCTC**ATCG**TATGCCCTACGACCTC **YER042W**
 CACATCGCAT**CTCATCG**ACCAGTTCCTCATCGGACGGC **YPL250C**
 GCCTCG**CTCATCG**TGGTACAGTTCAAACCTGACTAAA **YPL054**
 TCTCGTTAGGACCAT**CTCATCG**ACCACATCGAGAGCG **YPL223C**
 CGTAGCC**CTCATCG**GATCTTGTTCGAGAATTGCCTAT **YAL012W**

A binding site

← Upstream sequence → Gene Starts

6/20/2007

11

In reality ...

- ◆ 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

Aligned TF binding sites

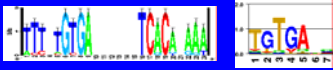
6/20/2007

12

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 |

A motif

```

TTGTGTCGTTTCACAAAA
TTGCACGGGATCACACTT
TGTGAGCATGATCATAAT
TSCAAAGGAGATCACATTA
TSTTAAATTGATCAGCTTT
TTTGAACCAATCGCATTA
TSTAACCAATCCACTAAT
CTGTATCAATCCCTCAAT
TGTGAGTTAGATCAGTCAT
TSTAACAGAGATCACACAA
    
```

6/20/2007

13

References for motif finding algorithms

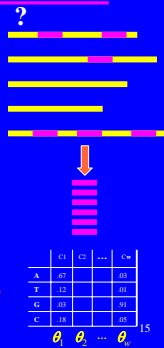
- ◆ BioProspector (<http://a1.med.nyu.edu/~yulin/BioProspector>)
- ◆ MDscan (<http://a1.med.nyu.edu/~yulin/MDscan>)
- ◆ Motif Regressor (<http://www.techtransfer.lacvax.edu/Software/MotifRegressor>)
- ◆ MEME (<http://meme.sdsc.edu/meme/meme.html>)
- ◆ Gibbs Motif Sampler (<http://hpcweb.wadsworth.org/gibbs/gibbs.html>)
- ◆ AlignACE
- ◆ CONSENSUS
- ◆ ANN-Spec
- ◆ YMF
- ◆ MotifSampler

6/20/2007

14

The Motif Sampler

- ◆ Initialized by *random starting* positions $a_1^{(0)}, a_2^{(0)}, \dots, a_k^{(0)}$
- ◆ 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



6/20/2007

15

Back to reality

- ◆ We talked about genome expression data
- ◆ ... and clustering analysis using these data
- ◆ We asked how co-expression and co-regulation agree with each other
- ◆ ... then explained some methods for discovering significant “words/motifs/features” in upstream sequences of co-regulated genes
- ◆ **Are we now able to discover real motifs just based on genomic data (expression+sequence)?**
- ◆ **How well does the sequence predict expression?**

6/20/2007

16

Motif Regressor

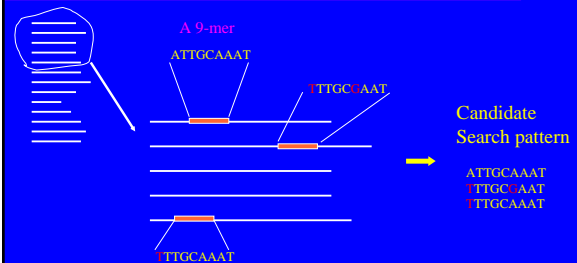
(Conlon et al 2003, *Proc Natl Acad Sci*)

- ◆ 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

6/20/2007

17

MDscan – “feature” extraction?



6/20/2007

18

Consequently:

- ◆ Many candidate motif patterns (N-w)

```

ATTGCAAAAT   GCCACCGT   TTACTAA   GCAAA
TTGC  AAT   -CCACCGT   TT  CTAA   GCAAA
TTGCAAAAT   -CCAC  GT   TTA  TAA   GCAAA
              GCCAC  G  T   T  ACTAA
              ...           ...

AGGGGC
GGGGC
AGGGG
AGGGG
    
```

6/20/2007

19

Evaluation of candidate motifs

- ◆ Entropy type

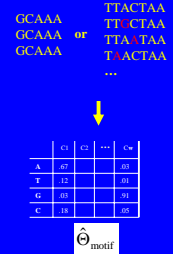
$$\text{Score}(A) = |A| \left[I_{\text{seq}}(\hat{\theta}_{\text{motif}} | \Theta_{\text{background}}) + \log \frac{P_A}{1 - P_A} \right]$$

- ◆ Other scoring functions?

$$\text{BP}(A) = \log(|A|) \times I_{\text{seq}}(\hat{\theta}_{\text{motif}} | \Theta_{\text{background}})$$

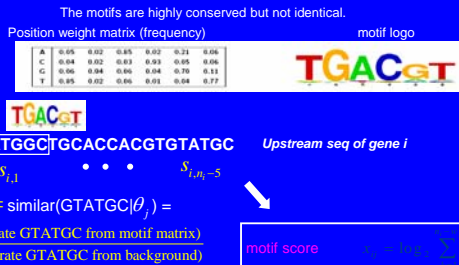
$$\text{MAP}(A) = \log(P(A | \text{Sequences}))$$

- ◆ Keep the top 30-50 candidates for further enhancement



6/20/2007

Motifs



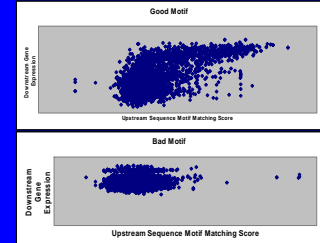
6/20/2007

21

Motif Regressor (Conlon et al. 2003, PNAS)

- ◆ For each TF:

| | Upstream Seq Motif 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

6/20/2007

22

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

6/20/2007

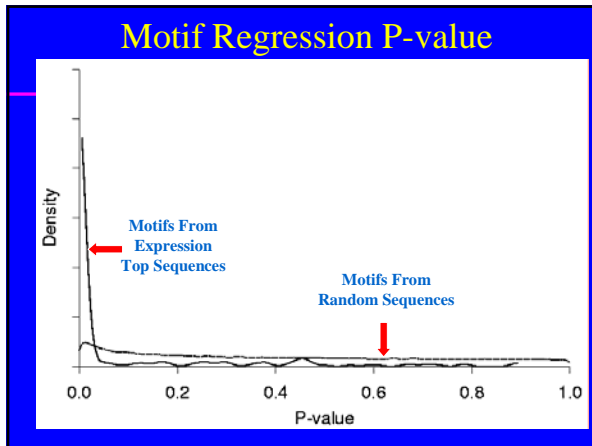
23

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

6/20/2007

24



- ### 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

| | Sequence Information | Expression |
|--------|--|------------|
| Gene 1 | ... CACTAAGATAAGCGA | y_1 |
| Gene 2 | ... TGGGTGCAC ATATG ATATGC | y_2 |
| | ... CAACATAGAAAACAGAAG | y_3 |
| | ...TCG ATCAT ATATG ACC | y_4 |
| Gene n | ... CTGGATCTGTCCATAA | y_n |

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

A simple derivation

$$R^2(T, \beta) = \text{corr}^2(T(Y), X\beta)$$

$$= \text{corr}^2(T(Y), E(X\beta|Y)) \frac{\text{var}(E(X\beta|Y))}{\text{var}(X\beta)}$$

$$T(Y) = E(X\beta|Y)$$

$$\max_{\beta} \left(\max_{T, \beta} R^2(T, \beta) \right) = \max_{\beta} \left(\text{var}(E(X\beta|Y)) \right) = \max_{\beta} \left(\beta^T \text{var}(E(X|y)) \beta \right)$$

$$\max_{T, \beta} R^2(T, \beta) = \lambda_1$$

λ_1 is the maximum eigenvalue of $\text{var}(E(X|y))$

Residual sum of square

- ◆ Define

$$RSS(\beta) = E \left(\sum_{i=1}^n (T(y_i) - X_i \beta)^2 \right)$$

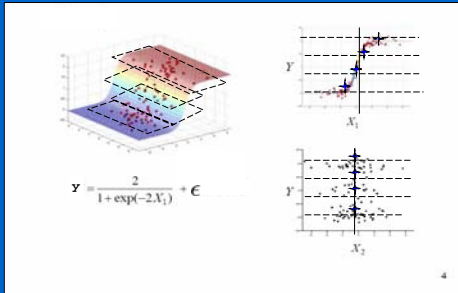
$$= nE \left((E(X\beta|y_i) - X_i \beta)^2 \right)$$

$$= nE \left(\beta^T (E(X|y_i) - X_i)^T (E(X|y_i) - X_i) \beta \right)$$

$$= n \left(\beta^T \text{var}(X) \beta - \beta^T \text{var}(E(X|y)) \beta \right)$$

$$= n(1 - \lambda)$$

Sliced inverse regression for estimating λ (SIR, Li 1991)

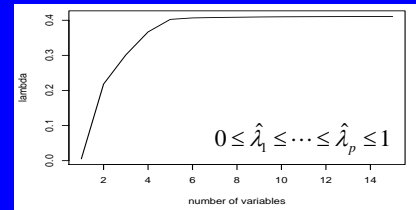


6/20/2007

31

λ increases as more variables included

We need to penalize the model complexity $\hat{\lambda}_k = \max_{\beta, S_k} \beta^T \text{var}(E(X; X \in S_k | y)) \beta$



6/20/2007

32

QPBC

$$\frac{\log(n) + k \log(n)}{n}$$

Example: $p=3$:

$S_{11}=\{x1\}$ $S_{21}=\{x1,x2\}$

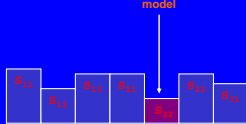
$S_{12}=\{x2\}$ $S_{22}=\{x1,x3\}$

$S_{13}=\{x3\}$ $S_{23}=\{x2,x3\}$

$S_{31}=\{x1,x2,x3\}$

$$\text{QPBC} = n \log(1 + \frac{b_j}{n}) + k \log(n)$$

$$b_j = \max_{\beta} - \log \text{var}(E(X_j | y))$$



Lemma: QPBC is consistent under the Gaussian and monotonicity assumption

6/20/2007

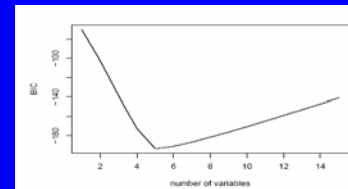
33

Empirical performance 15 candidate variables

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

| | 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 |



6/20/2007

34

Stepwise variable selection

Normalize each x_j to have mean 0, variance one

Step 1

Add one more variable to the current subset and check the significance of adding the new variable.

Forward addition

Delete the insignificant variable from the selected subset of variables.

Backward deletion

Stop if no more variable can be added in or excluded from the model.

Stopping

6/20/2007

35

Criterion for variable in and out

Recall $RSS = n \sum_{i=1}^n \hat{\epsilon}_i^2$

$$RSS(S_{k+1} | S_k) = RSS(S_k) + RSS(S_{k+1}) = n(\hat{\beta}_{S_{k+1}}^T \hat{\beta}_{S_k})$$

$$n(\hat{\lambda}_{k+1} - \hat{\lambda}_k) \xrightarrow{D} (1 - \lambda_k) \chi_1^2$$

(A theorem conjectured by W Zhong and proved by Tingting Zhang)

Variable in if $n(\hat{\lambda}_{k+1} - \hat{\lambda}_k) > (1 - \lambda_k) \alpha_{\chi_1^2(p_{in})}$

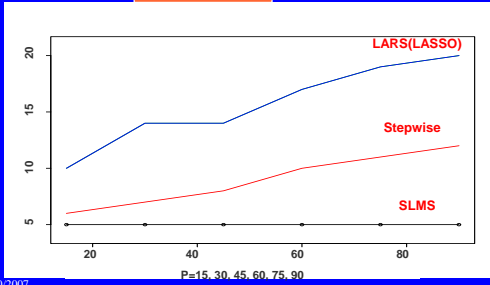
Variable out if $n(\hat{\lambda}_{k+1} - \hat{\lambda}_k) < (1 - \lambda_k) \alpha_{\chi_1^2(p_{out})}$

6/20/2007

36

Empirical performance: linear model with p increasing

Robustness Linear link function



6/20/2007

37

Motif identification for heat shock

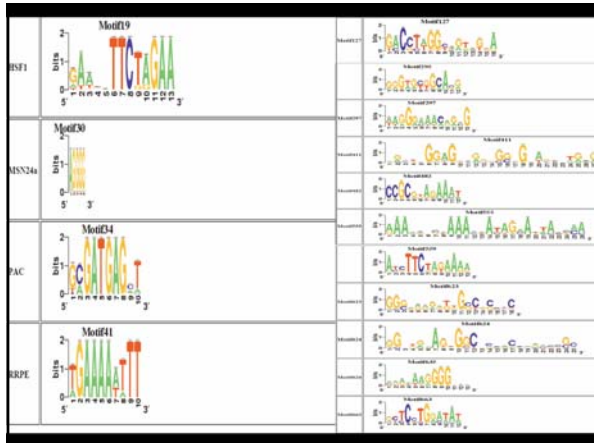
| Response (y) | | Regressor (X) | | | |
|--------------|------|--------------------|-----|----------------------|--|
| Gene name | | Motif ₁ | ... | Motif ₆₆₆ | |
| YAL001 | 5.20 | 3.5 | ... | 2.7 | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| YBL0610 | 6.31 | 4.1 | ... | 6.4 | |

- ◆ 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

6/20/2007

38



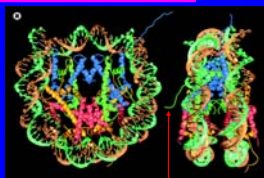
Comparison with LARS and stepwise

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

6/20/2007

40

Histone modification (Yuan et al. 2006)



Luger et al.
Nature, (1997)

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

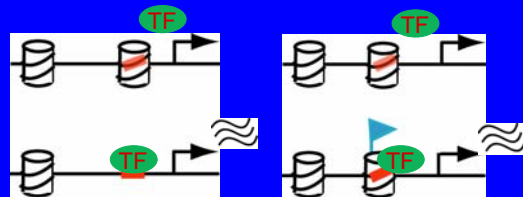
- Acetyl
- Ubiquityl
- Methyl
- Phosphoryl



6/20/2007

Regulatory role of chromatin

- ◆ Nucleosome positioning
- ◆ Histone modification



6/20/2007

42

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?

6/20/2007

Yuan et al. in preparation 43

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)

6/20/2007

44

A simple regression approach

- ◆ A regression model:

$$y_i = \alpha + \sum_j \beta_j A_{ij} + f(MS_{i_1}, \dots, MS_{i_k}) + \varepsilon_i$$

y_i : expression; A_{ij} : acetylation; S_j : 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 $f(MS_{i_1}, \dots, MS_{i_k})$.
 - RSIR selected 104 motifs from MDscan results
 - RSIR selected 69 motifs from the 666 motifs of B&T

6/20/2007

- ◆ We found that $f(\dots)$ is approximately linear

45

Estimating sequence dependent regulation

Linear regression model with TFBMs

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

S_{ij} motif score

linear $f(S)$

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

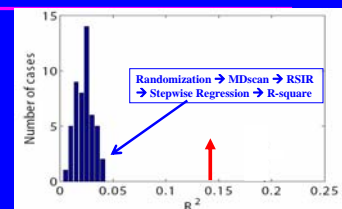
| | adjusted R-square | | |
|--------------|-------------------|-----------------------|--------------------------|
| | acetyl alone | acetyl + Mdsan motifs | acetyl + AlignACE motifs |
| No H3 or H4 | 0 | 0.2094 | 0.2223 |
| H3 (K9, K14) | 0.2443 | 0.354 | 0.3546 |
| H4 | 0.1342 | 0.2979 | 0.3128 |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

6/20/2007

47

Over-fitting issue

- Randomization Test of Motif Effect



- Cross-validation MSE

| | Training | Predict |
|-----|----------|---------|
| IGR | 1.50 | 1.52 |
| ORF | 1.48 | 1.50 |

6/20/2007

48

Acknowledgement

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

- **NIH,NSF**