Combining Large-Scale Gene Expression and Promoter Sequence Data

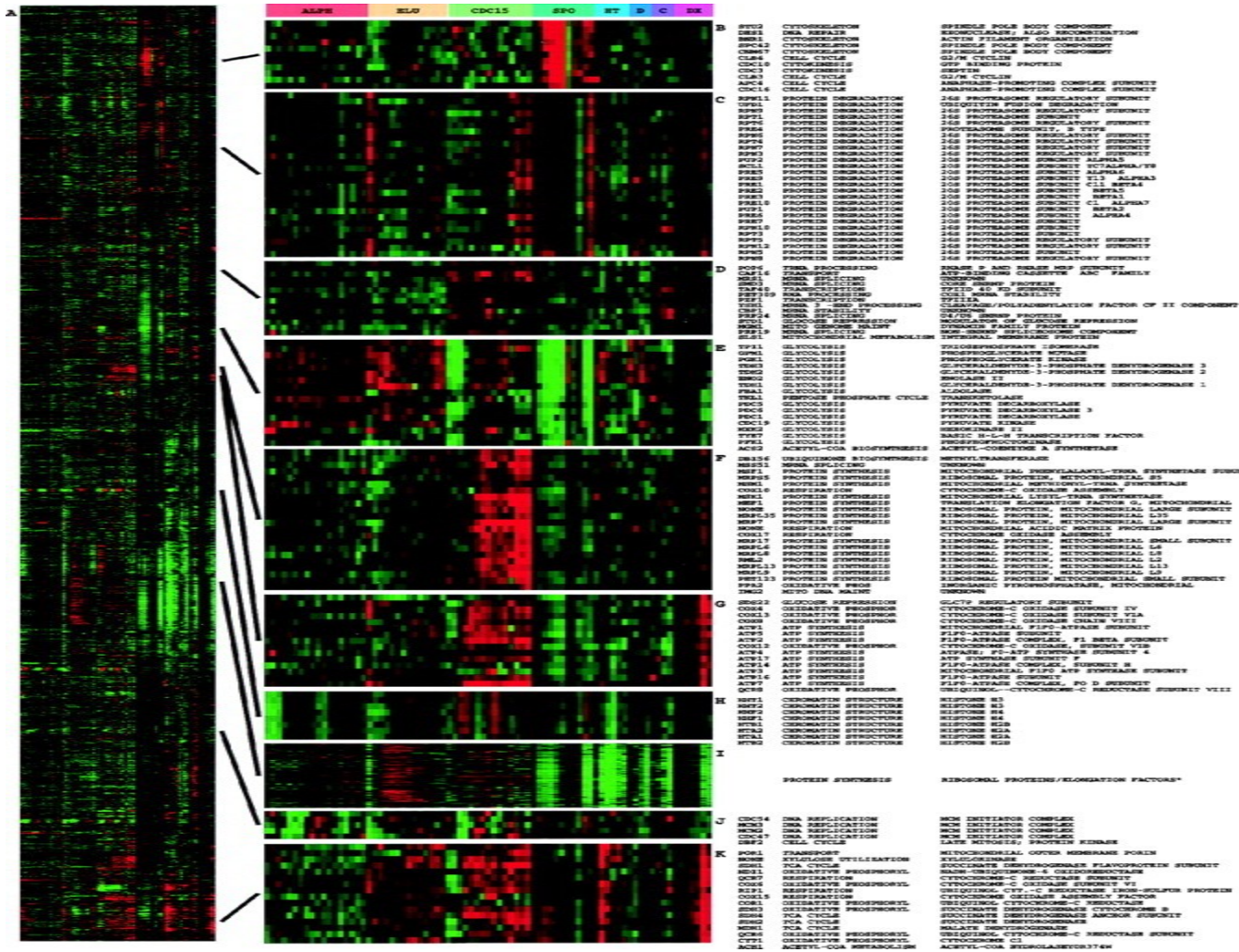
Outline

1. Motivation
   - Functionally related genes cluster together
   - genes sharing cis-elements cluster together
   - transcriptional regulation is modular
2. Models and Methods
   - Model the cis-elements as functioning exclusively or independently
   - A lot of data available: reason on genomic scale
   - Gene co-expression + motif finding = more than either by itself
3. Practical Approaches
   - clustering gene expression followed by motif finding (Tavazoie et al., ‘99, Beer and Tavazoie, ‘04)
   - finding motifs correlated with gene expression (Bussemaker et al., ‘01, Bussemaker et al., ’03)

- Refining Models of Regulation (Filkov and Shah, ’04)

1. Motivation

Co-expression and Function

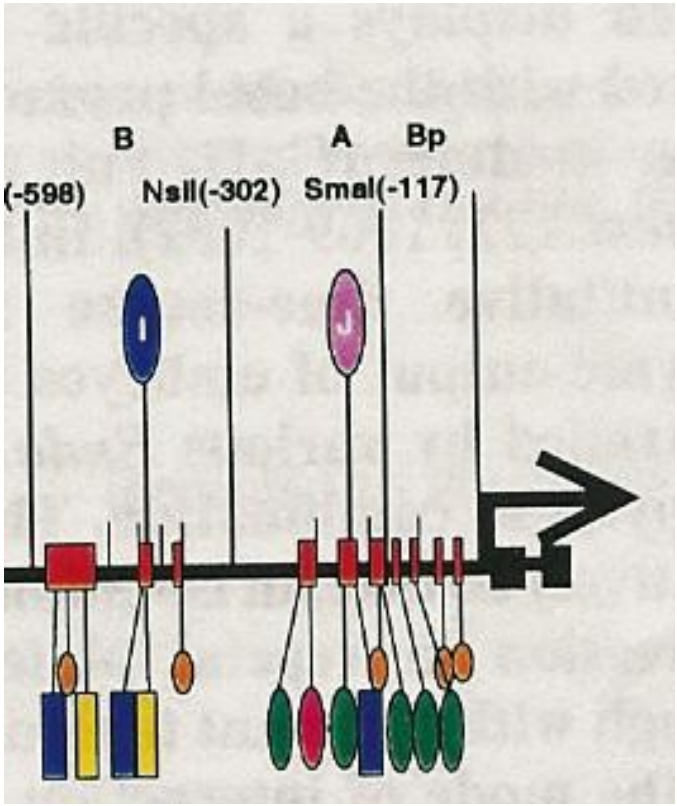


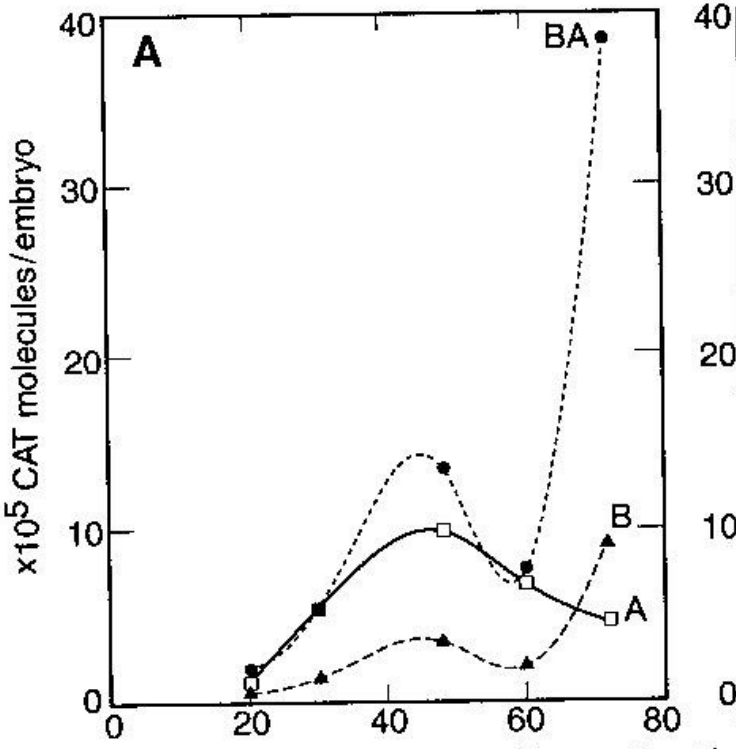
Eisen et al. (1998), PNAS 95

Co-expression and cis-elements

- Genes that have the same TFs bound to their upstream region have been shown to have the same pattern of expression (Tavazoie and Church, ‘98)
- TFs are sequence specific: **cis-trans equivalence**
- Thus, the same cis-elements will result in the same expression

Transcriptional Regulation is Modular



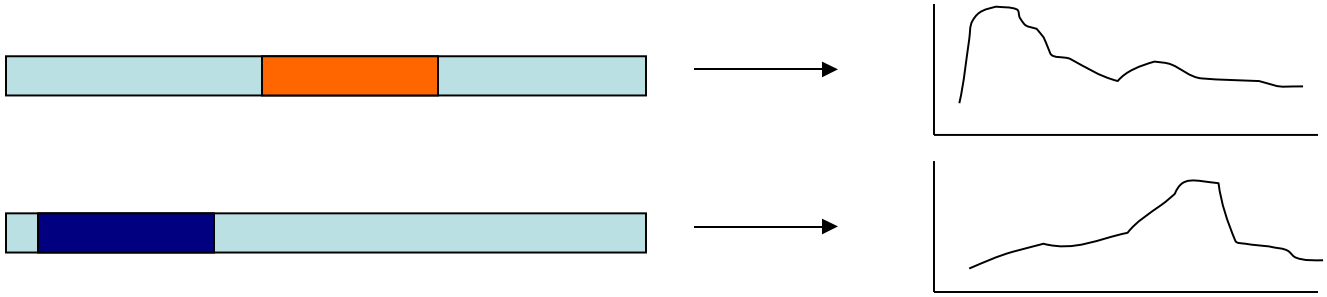


The modularity of Endo16's cis-region and its effect on the gene's expression (Davidson et al., 1995, 2001)

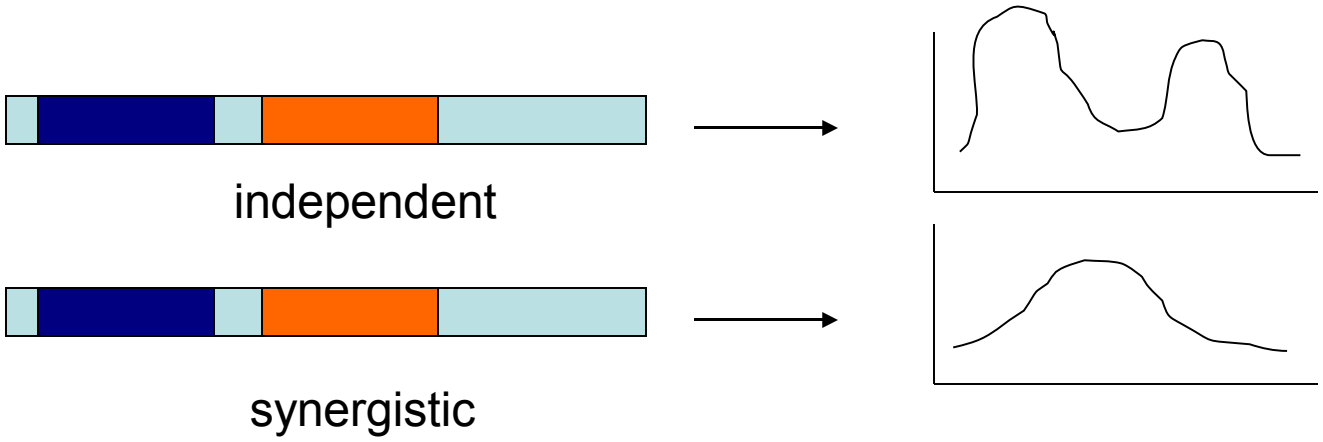
2. Models and Methods

Computational Models

- The same cis-elements recur in many upstream regions
- Cis-elements function exclusively



- Cis-elements function combinatorially



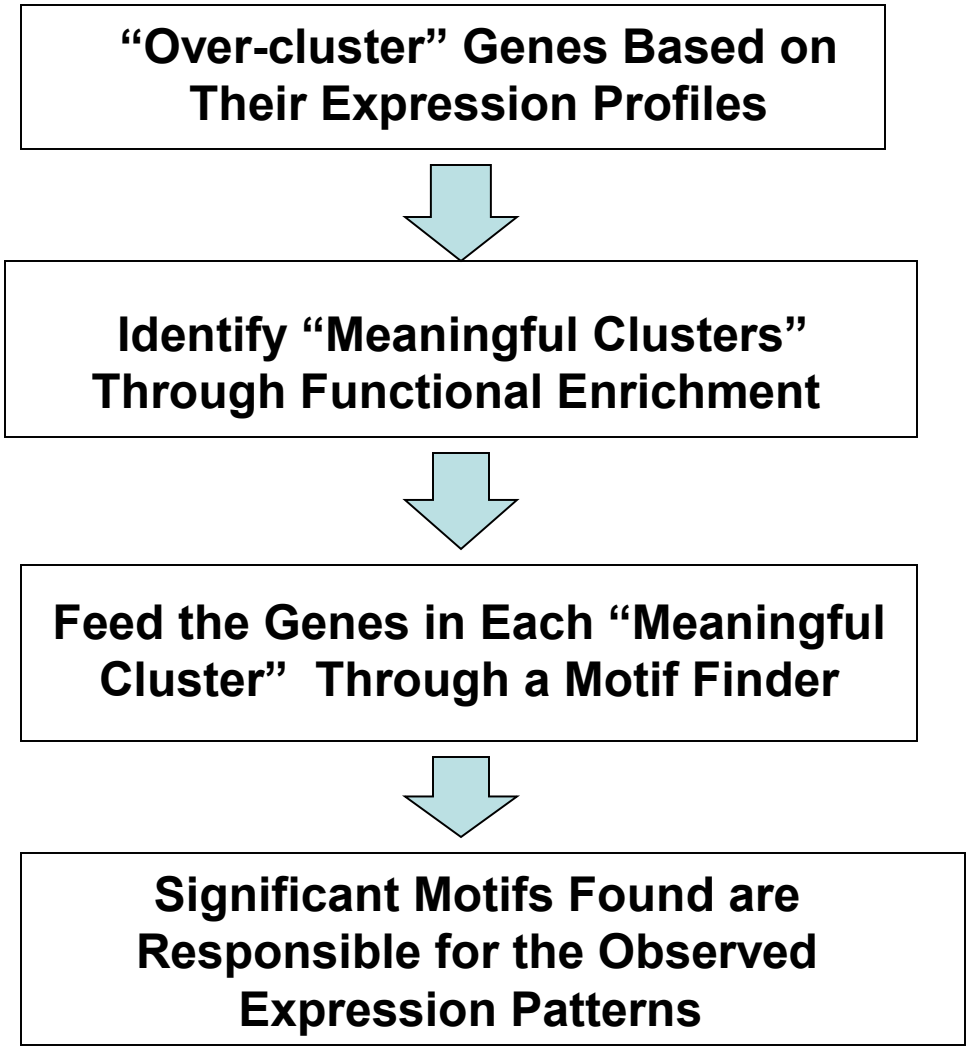
Computational Methodologies

- Clustering → Co-expression
- Motif finding → Co-regulation
- Clustering + motif finding → discovering clusters of co-regulated genes and the responsible cis-elements
- How to execute?

3. Practical Approaches

(a) Cis-Element Discovery With Clustering

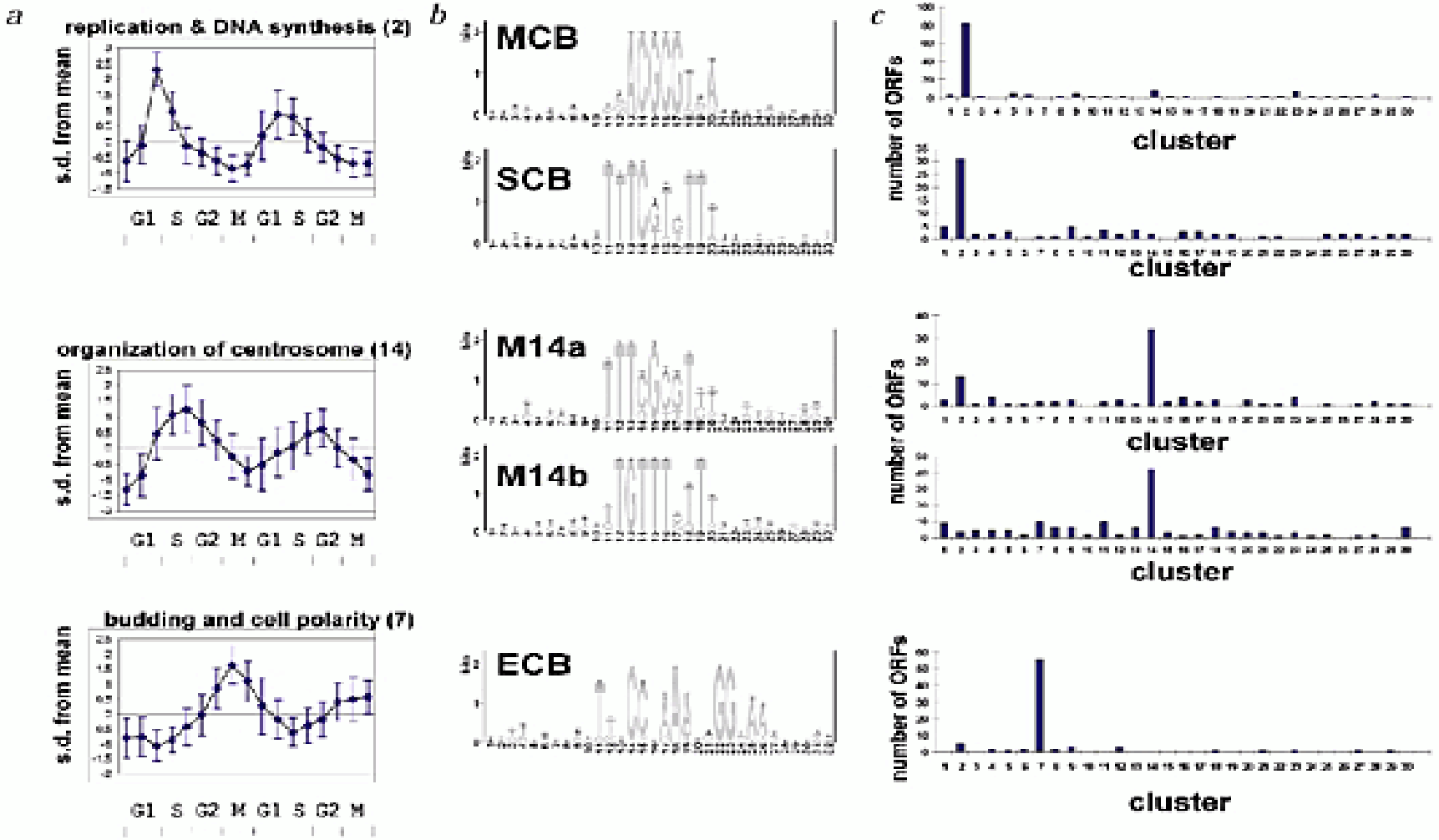
(Tavazoie et al., '99)



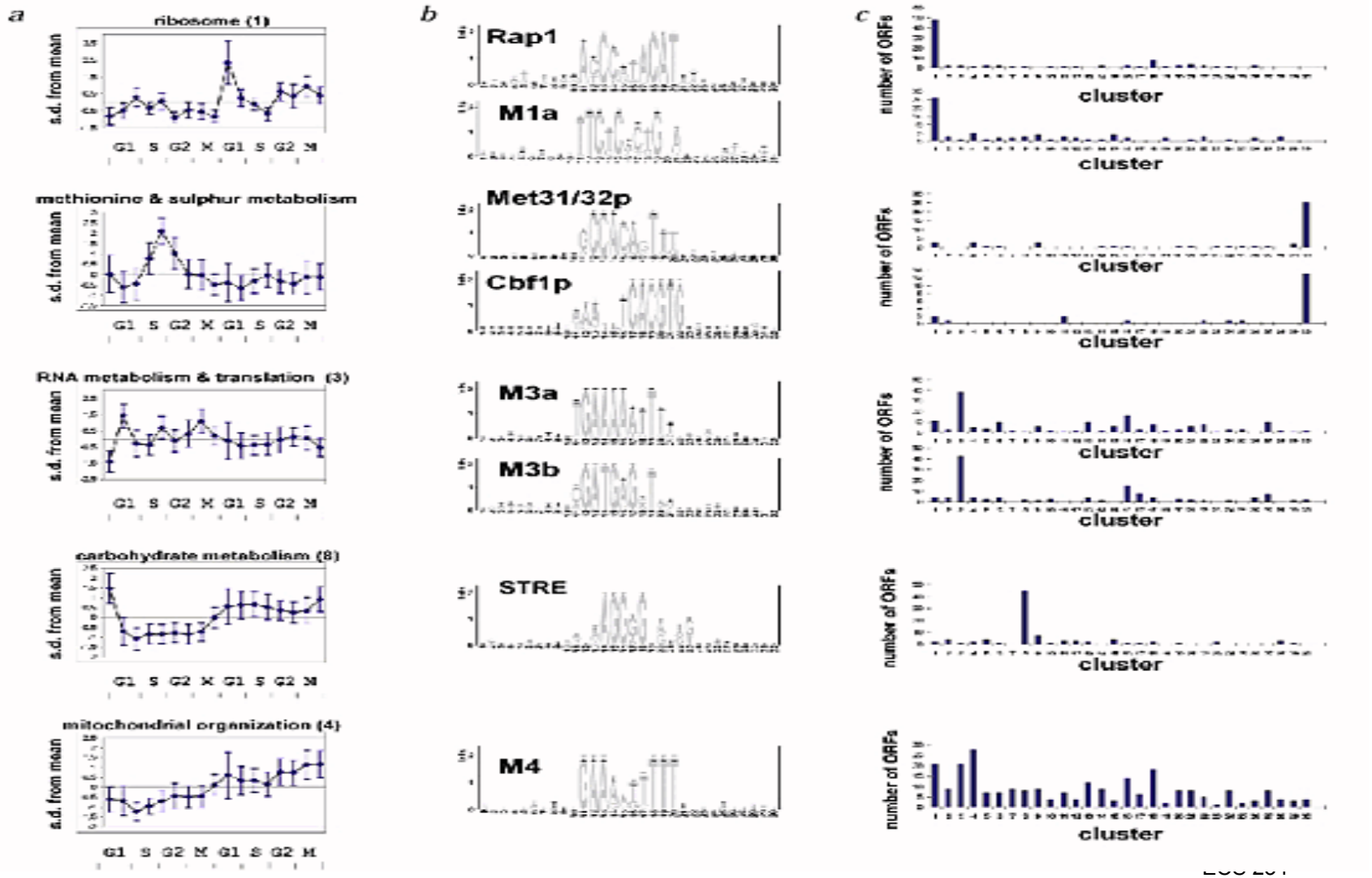
Methods Used

- Expression Data Preprocessing
- Clustering
  - K-means
  - over-cluster
- Cluster Validation
  - MIPS functional annotation
  - Hypergeometric distr. (based on Fisher's exact test)
  - p-value for enrichment reported
- Motif Finding
  - AlignACE
  - 600 bp ustream of ORFs
  - Motifs Significance
    - MAP score (AlignACE)
    - if found in at least two of three groups of ORFs (group 1: 50 "top" ORFs, groups 2 and 3: ½ of next 50 ORFs each)

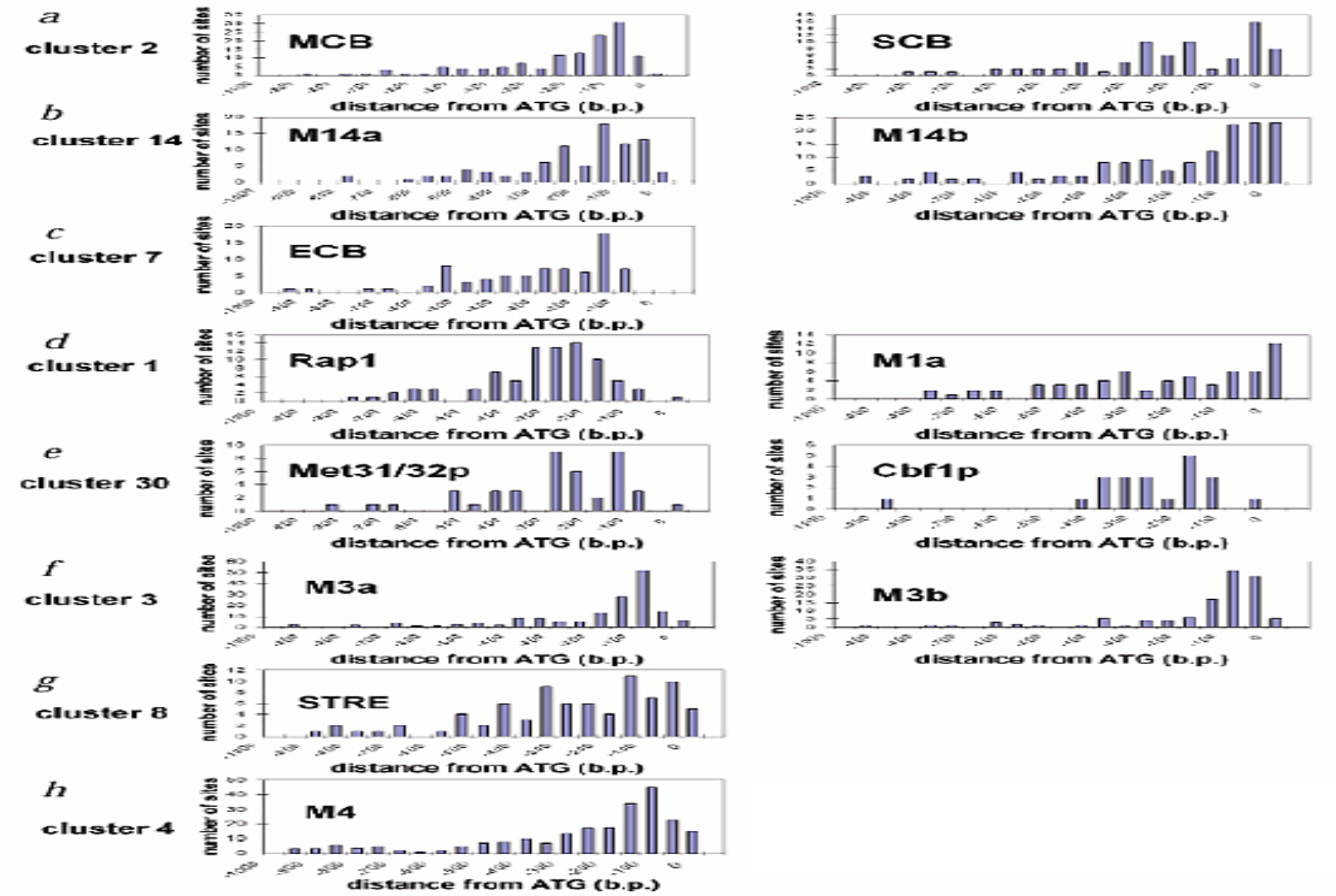
Periodic Clusters



Non-Periodic Clusters



Cis-Element Distance Distribution



(b) Cis-Element Discovery Without Clustering

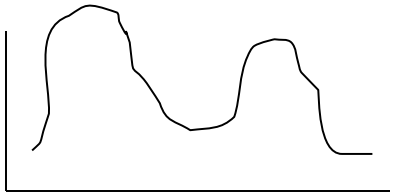
(**REDUCE,** Bussemaker et al., '01)

Model: Upstream motifs contribute additively to the overall expression of the gene

$$Expr_i = C + \sum_{\substack{\text{all motifs,} \\ \text{j, ingene i}}} F_j N_{ij}$$

Ex.

Gene i's cis-region

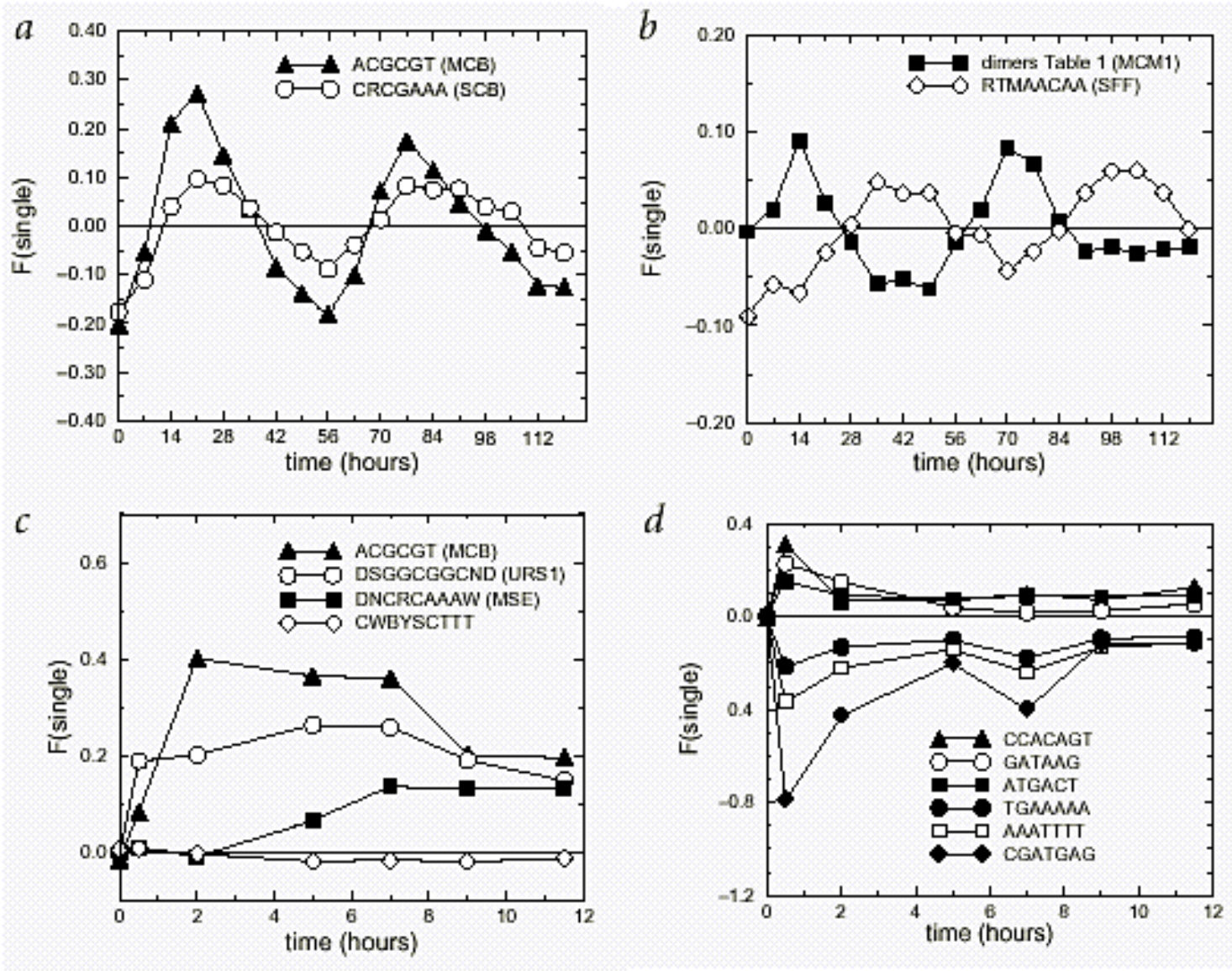


F_1=+0.5 F_2=+0.3

N_{i1}=N_{i2}=1 (or weight matrix score)

$Expr_i$

(Least squares fit used to find F and C, iterative algorithm used to get the fewest motifs for the best fit)


a
ACGCGT (MCB)
CRCGAAA (SCB)
F(single)
time (hours)
b
dimers Table 1 (MCM1)
RTMAACAA (SFF)
F(single)
time (hours)
c
ACGCGT (MCB)
DSGGCGGCND (URS1)
DNCRCAAAW (MSE)
CWBYSCTTT
F(single)
time (hours)
d
CCACAGT
GATAAG
ATGACT
TGAAAAA
AAATTTT
CGATGAG
F(single)
time (hours)

c) General Combinatorial Cis-element Interaction (Beer and Tavazoie '04)

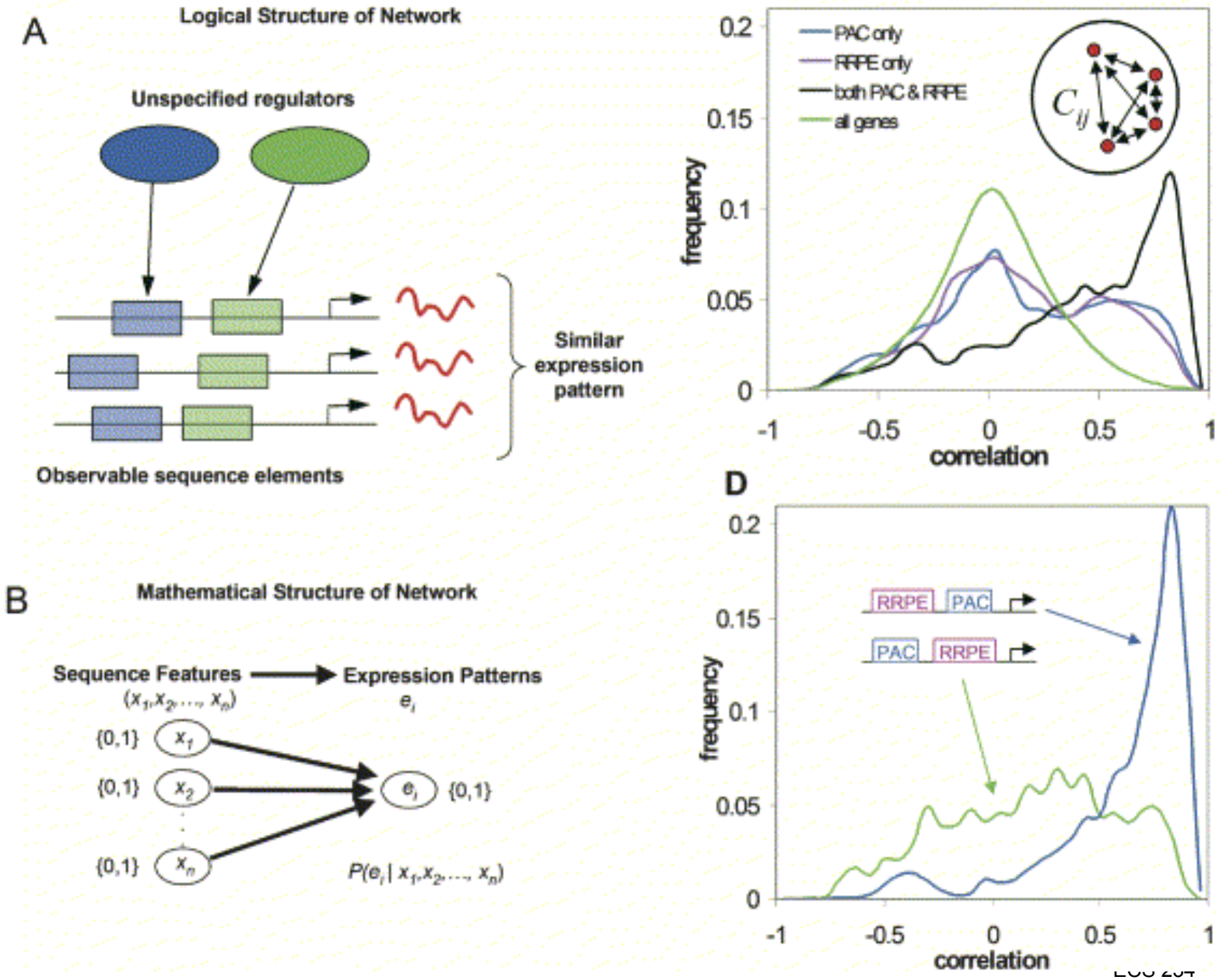
Approach:

(1) Clustering

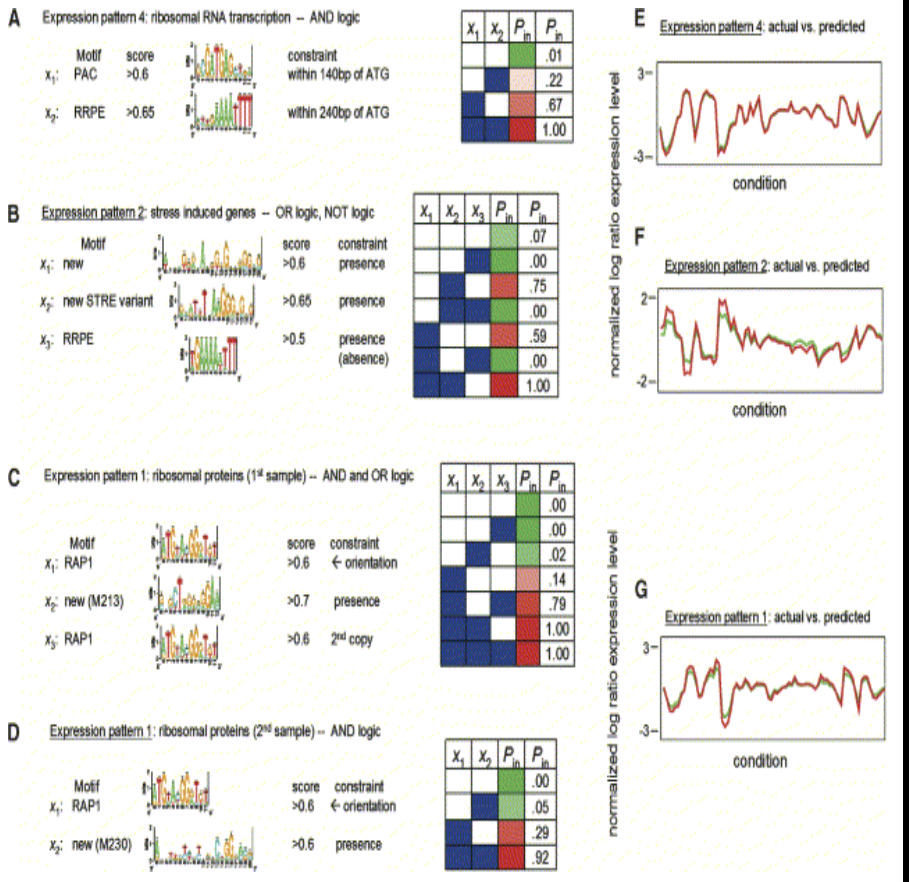
(2) Motif finding

(3) Motif Interaction Discovery Using a Bayesian Network Approach

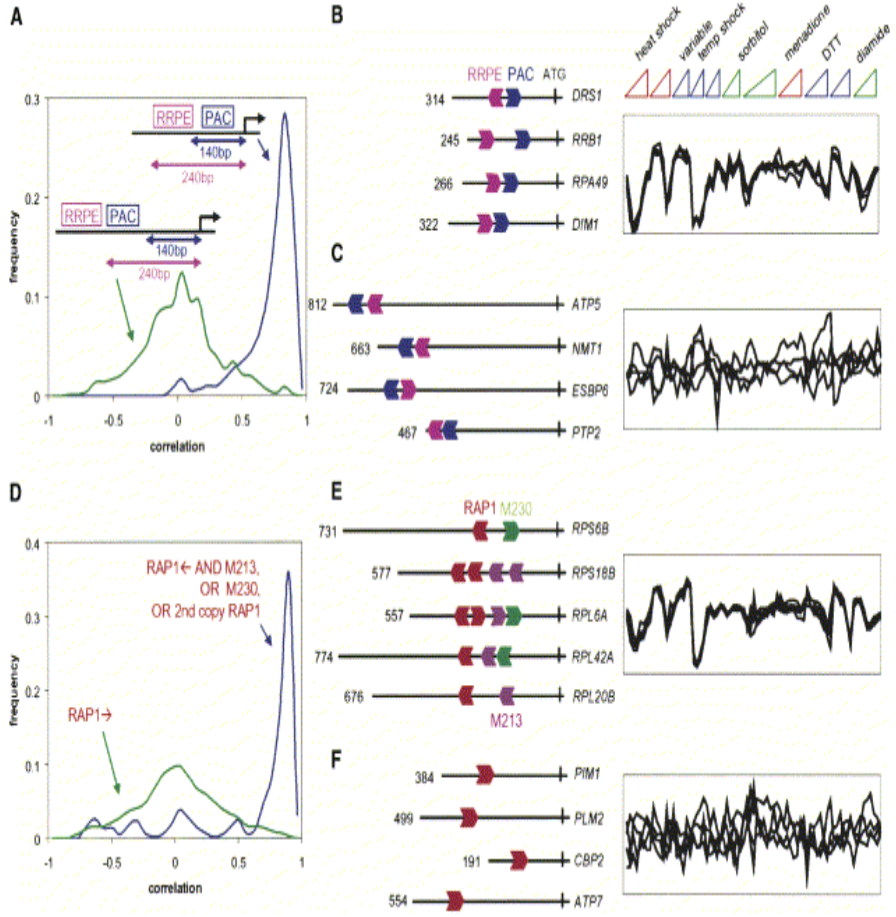
Goal: Predicting Gene Expression From the Promoter Sequences


A
Logical Structure of Network
Unspecified regulators
Similar expression pattern
Observable sequence elements
B
Mathematical Structure of Network
Sequence Features
Expression Patterns
$(x_1, x_2, \ldots, x_n)$
e_i
{0,1}
x_1
x_2
x_n
e_i
{0,1}
$P(e_i \mid x_1, x_2, \ldots, x_n)$
C
PAC:
RRPE:
PAC only
RRPE only
both PAC & RRPE
all genes
C_{ij}
frequency
correlation
D
RRPE
PAC
PAC
RRPE
frequency
correlation

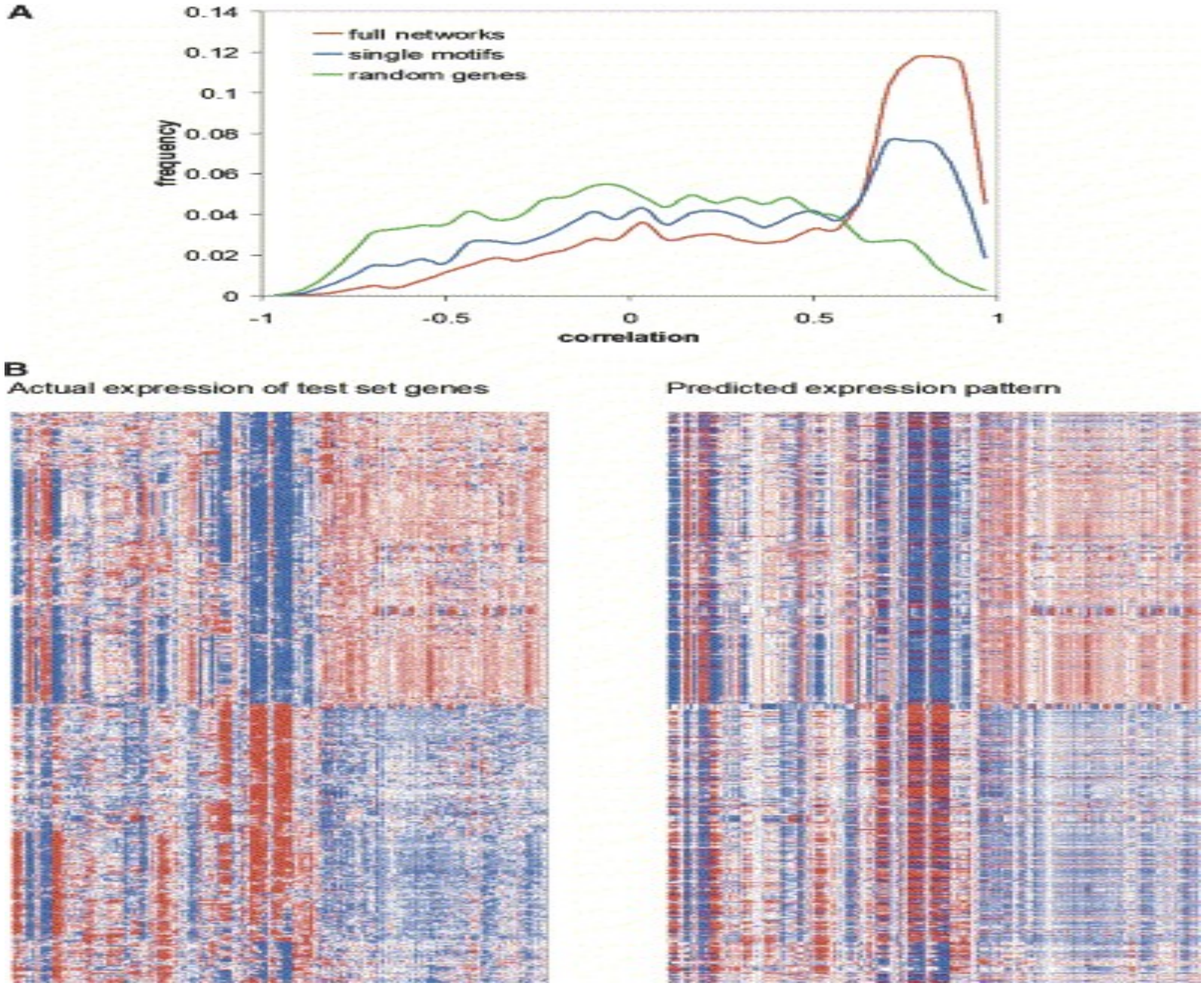
Parameters of Cis-element Interaction



Probability of Cis-element Contribution (red=high, green=low)



Distance Constraints Between Cis-elements


A
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
frequency
full networks
single motifs
random genes
-1
-0.5
0
0.5
1
correlation
B
Actual expression of test set genes
Predicted expression pattern

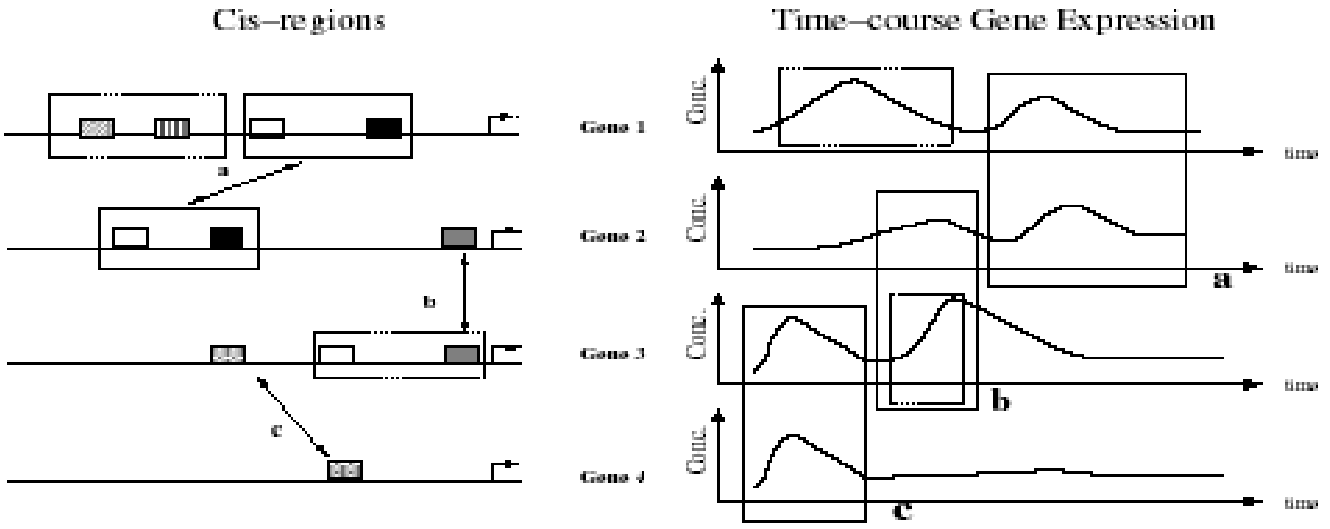
Criticism of the Beer-Tavazoie method

Yuan et al (2007) PLoS Comput Biol 3(11)

- BT overstated the accuracy (73%): over-fitted the data by training and testing on the same set; if correctly done drops to 61%.
- Simpler predictors do better! Eg naïve Bayes classifiers
- Position and orientation of TFBS is circumstantial: without them the prediction is better

4. Refining Models of Regulation

(Filkov and Shah, ’04)



- Cis-modules responsible for gene expression “events”
- Cis-modules are recurrent in genomes and between genomes
- Gene hierarchies formed from overlapping cis-modules
- Hierarchies are a refinement of the linear additive model

References

- Tavazoie et al., Systematic determination of genetic network architecture
- Beer and Tavazoie, Predicting Gene Expression from Sequence
- Bussemaker et al., (2001) Regulatory element detection using correlation with expression, Nat. Genetics v. 27, 167-171
- Roven and Bussemaker, REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic Acids Res. 2003 Jul 1;31(13):3487-90.
- Filkov and Shah, Regulation Hierarchies from Expression Data, RECOMB Satellite Workshop on Regulatory Genomics, 2004, also J Comp Bio 2008
- Yuan Y,et al (2007) Predicting Gene Expression from Sequence: A Reexamination. PLoS Comput Biol 3(11)