ECS 234: Combining 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)

4. Refining Models of Regulation (Filkov and Shah, ’08)
1. Motivation:
   a) Co-expressed Genes Share Function

Eisen et al. (1998), PNAS 95
b) Cis-elements determine expression

- 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
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

  independent

  synergistic
Computational Methodologies

- Clustering $\rightarrow$ Co-expression
- Motif finding $\rightarrow$ Co-regulation
- Clustering $+$ motif finding $\rightarrow$ 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)

“Over-cluster” Genes Based on Their Expression Profiles

Identify “Meaningful Clusters” Through Functional Enrichment

Feed the Genes in Each “Meaningful Cluster” Through a Motif Finder

Significant Motifs Found are Responsible for the Observed Expression Patterns
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 upstream 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

- Replication & DNA synthesis
- Organization of centrosome
- Budding and cell polarity

Clusters: MCB, SCB, M14a, M14b, ECB
Non-Periodic Clusters
Cis-Element Distance Distribution

- **a** cluster 2
- **b** cluster 14
- **c** cluster 7
- **d** cluster 1
- **e** cluster 30
- **f** cluster 3
- **g** cluster 8
- **h** cluster 4

Each cluster shows a distribution of distance from ATG (b.p.) with a bar chart indicating the number of sites at different distances.
(b) Cis-Element Discovery Without Clustering

*REDUCE*, Bussemaker et al., ’01

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

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

*Ex.*

Gene i’s cis-region

- \( F_1 = +0.5 \)
- \( F_2 = +0.3 \)
- \( N_{i1} = N_{i2} = 1 \) (or weight matrix score)

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

Observable sequence elements

Similar expression pattern

B. Mathematical Structure of Network

Sequence Features \(x_1, x_2, \ldots, x_n\) → Expression Patterns \(e_i\)

\{0,1\} \(x_1\)

\{0,1\} \(x_2\)

\{0,1\} \(x_n\)

\(P(e_i | x_1, x_2, \ldots, x_n)\)

C. PAC: RRPE:

Graph showing frequency vs. correlation

D. Graph showing frequency vs. correlation
Parameters of Cis-element Interaction

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

Distance Constraints Between Cis-elements
A

- Frequency
- Correlation

B

- Actual expression of test set genes
- Predicted expression pattern
Criticism of the Beer-Tavazoie method

• BT overstated the accuracy (73%): overfitted 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, ’08)

- 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
• Filkov and Shah, Simple Model of the Modular Structure of Transcriptional Regulation in Yeast, J Comp Bio 2008