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

- <u>Clustering</u> \rightarrow Co-expression
- <u>Motif finding</u> \rightarrow Co-regulation
- <u>Clustering + motif finding</u> → 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



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 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



ECS 234

Non-Periodic Clusters

ь



61 5 62 М 61 8 62 М

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

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

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

c) General Combinatorial Ciselement 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

Parameters of Cis-element Interaction

Contribution (red=high, green=low)

ECS 234

Between Cis-elements

| and the second sec | the second | |
|--|--|--|
| | A REAL PROPERTY OF THE PROPERT | - |
| The second se | and the second se | 1.000 |
| the second se | A REAL PROPERTY OF THE REAL PR | C -100 |
| THE REPORT OF A DECK | A REAL PROPERTY OF A REAL PROPER | 1.000.00 |
| and an all the second sec | the second s | 1000 |
| | and the second se | - |
| | the second | |
| the second se | | |
| The second se | the second se | C |
| and the second sec | a gas a second of the second s | 1 |
| A DESCRIPTION OF A DESC | | - 200 |
| | the second se | 1.000 |
| The second se | A REAL PROPERTY OF A REAL PROPER | |
| | | |
| and the second se | the second se | |
| the second se | and the second sec | the summer of |
| | the second | - |
| | The second se | 2 ACT |
| | | - |
| and an and the second se | No. of the second | 1.000 |
| | the second | 1.000 |
| and the second se | and the second se | - |
| The second se | and the second of the second sec | - |
| | the second se | - Contra - C |
| | and the second se | |
| | the second se | |
| the second se | the second se | 1000 |
| and the second of the second sec | the second | and so the second |
| | the second | - |
| THE REPORT OF A DECK | A REAL PROPERTY OF A REAL PROPER | 2000 |
| | the second of the second | - Second |
| and the second se | and the second s | the local little |
| THE REPORT OF A REAL PROPERTY AND A REAL PROPE | | The state of the s |
| the second se | the second is the local day in the second second | of the local division of the local divisiono |
| A REAL PROPERTY OF A REAL PROPER | A REAL PROPERTY OF THE PARTY OF | 1.00 |
| | the second se | |
| and the second se | and the second of the second se | - |
| The second se | the second second the second s | - |
| | Contraction 2, 12 (2012) 2012 (2012) 2018 | 0.000 |
| | A REAL PROPERTY OF A REAL PROPER | 2-22.2 |
| the second se | | 1.000 |
| | | 100 |
| The second se | a part and a contract to the second second second second | |
| | | 2.000 |
| | | |
| | | - E I |
| | | C - 44 - 2 |
| | | |
| STATES IN THE REAL PROPERTY AND A DESCRIPTION OF A DESCRI | THE REPORT OF COLUMN TO A DESCRIPTION OF THE | 10 mar 1 |
| BEREITE CONTRACTOR | And the second s | 100 |
| Participant and the second sec | The subject employees (| |
| Contraction of the second | and the second second second | |
| | A CONTRACTOR | - |
| | an manage | 1 |
| | an maria | THE |
| | | C THE |
| | B M THE | - Harris |
| | | - Harrison |
| | | - Harrison |
| | | Harmer |
| | | - Harrison |
| | | - Hardward |
| | | Manual Property |
| | | South States of the |
| | | - Harrison and and |
| | | - Hardward and and |
| | | - Harrison and a state |
| | | - Harrister Barrist |
| | | - HAR AND AND A STATE |
| | | " Harris and the second of the |
| | | All Report Manual Property in |
| | | " Hardson and a state of the |
| | | "Harris and the same of the |
| | | "Har have been and the second |
| | | Billy a line has a surrow of |
| | | · Harnan Barren Barr |
| | | Barris and Street Proping in |
| | | BHE BUT THINK & BUT THE BUT |
| | | And the state of the second se |
| | | AMAGE IN COMPANY AND INCOMPANY |
| | | Party and the state of the second second |
| | | Providence of the second |
| | | Particular particular and the |
| | | Providence of the party of the party of the |
| | | Party and a survey of the second second |
| | | - Harrison Street Lines and |
| | | · Hardware in the state of the second |
| | | 「「「「「「」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」 |
| | | 「「「「「「」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」」 |
| | | 「日本の日本の日本の日本の日本の日本」 |
| | | 一下の一下の一下の一下の一下の一下の一下の一下の一下の一下の一下の一下の一下の一 |
| | | 一一日日の一日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日 |
| | | |
| | | |

Criticism of the Beer-Tavazoie method

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

- 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, '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)