Learning “graph-mer” motifs that predict gene expression trajectories in development

Xuejing Li\textsuperscript{1}, Chris H. Wiggins\textsuperscript{2}, Valerie Reinke\textsuperscript{3}, Christina Leslie\textsuperscript{4}\textsuperscript{*}

\textsuperscript{1}Department of Physics, Columbia University
\textsuperscript{2}Department of Applied Physics and Applied Mathematics, Columbia University
\textsuperscript{3}Department of Genetics, Yale University
\textsuperscript{4}Computational Biology Program, Memorial Sloan-Kettering Cancer Center
\textsuperscript{*}Corresponding author: cleslie@cbio.mskcc.org
Regulatory networks in development

Motivating problem

- Reinke lab: time series gene expression for *C. elegans* wild-type development
- Problem: decipher regulatory networks governing germline- and sex-regulated gene expression

![Diagram of germline development](figure-from-Valerie-Reinke)

**Figure:** Diagram of germline development (figure from Valerie Reinke)
Deciphering *cis* regulatory logic

**Promoter sequence and mRNA expression**

- DNA motifs: binding sites for transcription factors
- *cis* regulatory logic determines mRNA profile over time course, to first approximation (also: chromatin structure, mRNA processing, microRNAs)

![Diagram of promoter sequence and mRNA expression](image)

- Typical cluster-first approach:
  - Cluster gene expression profiles
  - Search for over-represented DNA motifs in promoter sequences of cluster (*de novo* motif discovery or enrichment of known motifs)
From sequence to expression profile

Alternative formulation

- Time series data: highly correlated profiles, few distinct clusters
- Motivation: REDUCE (Bussemaker et al, 2001), correlates motif counts with log fold change in single experiment
- Regression problem: map from promoter sequence ($k$-mer counts) to full expression trajectory

Goals

- Learn motifs *de novo* from promoters by building from $k$-mers
- No clustering of expression profiles
- Flow of information from sequence to expression

Xuejing Li$^1$, Chris H. Wiggins$^2$, Valerie Reinke$^3$, Christina Leslie$^4^*$
Motivation Algorithm Results

Regression problem

Regression Problem

- Predict gene expression levels ($Y$) from sequence information ($X$)

![Diagram](image)

- Ordinary Least Squares Regression (OLS) not applicable: number of motifs $>>$ number of time points

Xuejing Li$^1$, Chris H. Wiggins$^2$, Valerie Reinke$^3$, Christina Leslie$^4$*
Regression problem

Partial Least Squares Regression: Dimension reduction + OLS regression

- Dimension Reduction
  - Iteratively learn latent factors $t_i = Xw_i$, $w_i = \arg\max_w \text{cov}(t_i, Y)$
  - Unlike PCA, latent factors covary with gene expression

- OLS (or ridge) regression from $T = [t_1, \ldots, t_K]$ to $Y$
Regularized PLS regression

- Sparsity constraint, many $k$-mers get zero weight

\[ ||w||_1 = \sum_{p=1}^{M} |w^p| \leq b_1 \]  

(1)

- Graph Laplacian constraint, edge $p \sim q$ if close in Hamming distance

\[ w^T L w = \sum_{p \sim q} |w^p - w^q|^2 \leq b_2 \]  

(2)
Interpretation of weight vectors

From motif to expression pattern

- $c_i = Y^T t_i$ can be interpreted as an expression pattern
- Highly weighted $k$-mers in $w_i$ relevant for this expression pattern

Motif weight vector $w$  
Experiment weight vector $c$
Graph-regularized PLS

Algorithm

Loop over latent factors: For $i = 1, \ldots, K$

(1) Learn weight vectors and latent factors:

$$w = \arg \max_w (w^T S^T S w), \ S = Y^T X,$$

subject to

(i) $w^T w = 1$

(ii) $\sum_{p=1}^{M} |w^p| \leq b_1$

(iii) $\sum_{p \sim q} |w^p - w^q|^2 \leq b_2$

Compute latent factor: $t = Xw$

Normalize: $t = t / \sqrt{t^T t}, \ w = w / \sqrt{t^T t}$

$c = Y^T t, \ u = Yc$

(2) Deflate $S$:

$v = X^T t$

if $i > 1$ then

$v = v - V(V^T v)$

$v = v / \sqrt{v^T v}$

$S = S - v(v^T S)$

Store $w, t, c, u$ and $v$ into column $i$ of $W, T, C, U$ and $V$

end
**C. elegans wild-type germline development**

Gene expression matrix from time series data:

- Gene expression levels at multiple times points, averaged over replicates
- About 9000 genes, 12 time points

Motif matrix from sequence data:

- Scan promoter sequences for 6mers and 7mers
- Filter $k$-mers based on expected counts

Gene sets from mutant expression data:

- Sperm genes: high expression during spermatogenesis
- Oocyte genes: high expression during oogenesis
Standard and regularized PLS

10 fold C.V. on held-out genes

- Standard PLS achieves lowest test $\chi^2$ after the first 4 factors
- Standard PLS with genes permuted in $X$ overfits badly
- Regularized PLS more resistant to overfitting
**Regularized PLS**

- Sperm/Oocyte gene set: largest $\chi^2$ reduction from 2nd/1st factor respectively

![Graph showing normalized squared mean error on test data for regularized PLS](image)
Analysis of gene expression

Sperm gene set

- $c_2$ significantly correlated with sperm gene expression profiles; consistent with the largest $\chi^2$ reduction

Figure: Sperm gene expression profiles and $c_2$
Analysis of gene expression

Oocyte gene set

- $c_1$ significantly correlated with oocyte gene expression profiles; consistent with the largest $\chi^2$ reduction

Figure: Oocyte gene expression profiles and $c_1$
Extracting PSSMs from graph-mers

**Workflow**

- Network graph of top 50 $k$-mers ranked by $w$ (Cytoscape)
- Identify highly interconnected regions as $k$-mer clusters (Mcode)
- Generate position-specific scoring matrices (PSSMs) by hierarchical sequence agglomeration (HSA)

**Hierarchical sequence agglomeration**

Generate position-specific scoring matrix (PSSM) for $k$-mer cluster

- $k$-mers are initial PSSMs
- Iteratively merge similar PSSMs based on JS entropy measure

$$d(p, q) = \min_{\text{offsets}} [\pi_p D_{KL}(p|\pi_p p + \pi_q q) + \pi_q D_{KL}(q|\pi_p p + \pi_q q)]$$  \hspace{1cm} (3)

where $\pi_p \propto N_p$, $\pi_q \propto N_q$, $\pi_p + \pi_q = 1$, $N_{p,q}$ number of target genes
Analysis of motifs

Graph-mer: Network graph of top 50 $k$-mers ranked by $w_2$

- Sperm gene motifs: ELT-1 motif GATAAA and bHLH motif ACGTG
Analysis of motifs

Strong spatial bias of motif ACGTG

- ACGTG biased towards coding region in sperm genes

Figure: Distribution of distance to coding region
Analysis of motifs

Graph-mer: Network graph of top 50 \( k \)-mers ranked by \( w_1 \)

- CG-rich \( k \)-mers
Analysis of motifs

Conservation of CG-rich $k$-mers between *C. elegans* and *C. briggsae*

- Motif conservation score (MCS) correlated with $w_1$

**Figure:** $w_1$ versus MCS
Graph-mers for *C. elegans* development

- Graph-regularized PLS learns motifs by predicting gene expression trajectories
- No clustering of gene expression profiles
- Regularization
  - Sparsity constraint: short list of *k*-mers
  - Graph Laplacian constraint: graph-mer, *k*-mer cluster as motif pattern
- Biological Validation
  - Recover motifs from literature
  - Novel CG-rich motifs for oocyte genes
  - Additional evidence: positional and conservation information