Contributions to transcriptomic data analysis and gene regulation network inference
ISSSB’2011 - NII Shonan meeting - November 13-17 2011

Céline Rouveirol
Overview

Inferring regulation networks from transcriptomic static data
   Motivations and goals
   Methods
   Evaluation

Finding dense regions in binary contexts
   Motivation and goals
   Methods
   Experiments and results
Regulation network inference

Building a regulation graph for a biological process

(A) Unité
Facteur de transcription
Gène cible et site de fixation

(B) Motifs
SIM
MIM
FFL

(C) Modules

(D) Réseau de régulation génique
adapté de Current Opinion in Structural Biology

Contributions to transcriptomic data analysis and gene regulation network inference
Transcriptional regulation

A transcription factor is:

- a protein, that binds to specific sequences of DNA adjacent to the genes that they regulate
- controls the flow (activates and/or respress) this gene’s transcription
Cahier des charges

- learning cooperative regulation relations from gene expression only
- no time series data available
- without any a priori assumption concerning the gene expression distribution
- local approach (one network / gene)
LICORN

LICORN (Elati et al., Bioinformatics 2007) follows these 3 steps:

1. Build a set of candidate co-regulators (predicate invention) for all genes;
2. Build a set of candidate regulation networks for each target gene $g$;
3. Select the best candidate network(s) for each gene $g$, and assign a significance score to this (these) network(s).
Co-regulators computation

- Goal: find all formulas of a language \( L \) that satisfy a constraint \( q \) on a dataset \( r \), \( Th(r, L, q) \).

- \( r \) : discrete matrix \( r \) of \( m \) observations described with \( n \) attributes \( A = g_1, g_2, \ldots, g_n \) (\( n \gg m \))
- \( L \) : language describing itemsets on \( A (2^A) \).
- \( q \) : constraint of interest, e.g. frequency of a pattern \( p \) in \( r \): \( p \) is frequent if \( \text{freq}(p) \geq \text{minsupp} \)

- Extension of Apriori (Agrawal et al., 1994) for computing frequent/closed itemsets from discrete data
Cooperative regulation model

- potentially, several cooperative activators/repressors
- AND-aggregation for activators/repressors + deterministic function for computing target gene state given the aggregated states of its activators/repressors.

### Cooperative Regulation Model Diagram

**Regulators** $\text{Reg}(g)$
- Activator set $\text{A}(g)$
  - $a_1, a_2, \ldots, a_n$
- Repressor set $\text{R}(g)$
  - $r_1, r_2, \ldots, r_m$

**Function** $E_{\text{AND}}(\text{A}(g))$

**Function** $E_{\text{AND}}(\text{R}(g))$

**Regulation program** $RP(\text{Reg}(g))$

<table>
<thead>
<tr>
<th>$E_{\text{AND}}(\text{A}(g))$</th>
<th>$E_{\text{AND}}(\text{R}(g))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Target gene $g$**
Generating candidate co-regulators for a target gene

Let $C$ be a co-regulator and $g$ be a target gene. $S_x(C)$ and $S_y(g)$ denote their support for the values $x, y \in \{-1, 1\}$.

**Definition (Overlap constraint)**

$C$ in state $x$ *co-varies* with $g$ in state $y$, denoted $\text{cov}(S_x(C), S_y(g))$ if and only if $\frac{|S_y(g) \cap S_x(C)|}{|S_y(g)|} \geq \text{minoverlap}$, a user-defined minimum overlap threshold.

Best-first search for the $k$-best co-activators and repressors of $g$. 
Assessment of candidate networks

- Rank candidate networks \(((A, I) \text{ pairs})\) wrt a local score (MAE)
- Select (n-)best network(s)
- Associate a statistical signficance to those networks: non-paramteric approach, permutation-based (Benjamini et al., 2001).

![Diagram of assessment process](image)
10-CV evaluation

- ○ majority vote
- △ Minreg
- + LICORN
- ⨄ LICORN (FDR < 0.05)

Mean Absolute Error (MAE)

Fold index

Contributions to transcriptomic data analysis and gene regulation network inference
From local to global patterns

(Birmelé et al. BMC 2008)
On-going work - Perspectives

- Combine local networks to build a global regulation graph (ILP, frequent graph mining, ...)
- Integrate other information sources (promoter sequence, genomic alterations, miRNA, proteins, epigenetic, ...)
- More powerful evaluation for networks: select networks that are supported by some domain model
Context

- A bioinformatics task from gene expression dataset:
  - Mining co-expressed genes (Sets of genes that are jointly expressed) → discretisation + extraction of frequent/closed/maximal itemsets (e.g. Apriori [Bor02]).
The problem

- **Effect of noise**
  - Shattering relevant itemsets into a small irrelevant itemsets
    → explosion in the number of resulting itemsets.

- **Aim and intuition**
  - Mine efficiently a small number of maximal regions of 1, potentially overlapping, and verifying density and minimal support constraints.
  - by combining data mining methods with graph algorithms.
Related work

- **Complete approaches**
  - Methods based on the level-wise principle [Man04, Bes05, Bes06, Liu06, Che06]
  - Handle anti-monotone constraints to prune the space search
  - Quasi-biclique methods [Uno08].

  - Large number of itemsets extracted.
  - Very expensive in execution time for dense data.

- **Non-complete approaches**
  - Bi-clustering methods [Pre06]
    - Difficulty in the choice of parameters.
  - Heuristic methods [Mou11]
    - Still too many results and redundancy.
**HANCIM**: Hybrid Approach for Noisy Contexts Itemset Mining

- Consists of two main steps:
  - Identification of a seed pattern $s_i$
  - Construction of a dense region $(O, A)$ such that $s_i \subseteq A$
- Extracts the maximal regions $M = (A, O)$ such that:
  - Density: $\text{density}(M) \geq \delta$
  - Minimal support: $\frac{|O|}{|O_{\text{context}}|} \geq \sigma$. 
Seed patterns and the adaptative support

- Use all maximal frequent patterns of D as seeds [Mou11]
  - A high redundancy in the obtained results.
  - Quite expensive especially for dense contexts.

- Seed patterns should:
  - be small enough to be easy to compute
  - favour the extraction of diverse seeds (small overlap) to avoid redundancy in the resulting regions
Seed patterns and the adaptative support (ctd.)

- $O$ set of observations, $A$ set of attributes, two contexts on $A$ and $O : D$ and $D_{cs}$ (updated after each new seed is extracted)

```
while ∃ seed pattern $s_i \subseteq D_{cs}$ do
  Compute the region $(O', A') \subseteq D$ such as $(s_i \subseteq A')$ and $(|O'| \geq \sigma)$ and $(\text{density}((O', A')) \geq \delta)$
  $Updated\_D_{cs}$: $D_{cs}$ where all elements of $(A', O')$ are set to zero
  Support = $\frac{\text{Support} \times \text{density}(Updated\_D_{cs})}{\text{density}(D_{cs})}$; $D_{cs} = updated\_D_{cs}$
end
```
Searching for dense regions

- Searching for a maximal dense region including a seed pattern $s_i$
- Based on graph algorithms: maximal flow/minimal cut.
Searching for dense regions

Searching for the maximal dense region including a seed pattern $s_i$

- Construct the augmented and weighted bipartite graph corresponding to $s_i$
- Compute a minimal st-cut: push-relabel [Che97] → a dense subgraph $G_0 = (O_0, s_i)$ where the observations $O_0$ are strongly linked to the attributes $s_i$
Searching for dense regions

- Construct the augmented and weighted bipartite graph corresponding to $O_0$
- Compute a minimal st-cut
  $\Rightarrow$ a dense subgraph $G_1=(O_0, A_1)$ where $s_i \in A_1$ and each attribute of $A_1$ has a density greater than $\delta$
Experiments on gene expression datasets

- Detection of co-expression relationships between genes from the **Gasch dataset** [Gas00]
  - Expression measures of 2993 genes over 173 observations.
  - Discretization model described by [Pre06].
  - Parameters: $\sigma=20\%$ and $\delta=80\%$

- Running time:
  - *Bimax*: calculation stopped after 1 week
  - *HANCIM*: results obtained after 12 minutes.

- Comparison with 100 biclusters published by Bimax [Pre06].
Experiments on real data

- Calculate the enrichment of extracted biclusters in Gene Ontology terms (GO)[Che98].

- The 100-top biclusters extracted:
  - *Bimax*: have p-values ranging between $3e^{-2}$ and $3e^{-4}$.
  - *HANCIM*: have p-values less than $e^{-5}$.

- The best annotated bicluster:
  - *Bimax*: has a p-value equal to $3e^{-4}$.
  - *HANCIM*: has a p-value equal to $e^{-38}$.

<table>
<thead>
<tr>
<th></th>
<th>&lt; $e^{-2}$</th>
<th>&lt; $e^{-3}$</th>
<th>&lt; $e^{-4}$</th>
<th>&lt; $e^{-5}$</th>
<th>&lt; $e^{-10}$</th>
<th>&lt; $e^{-20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HANCIM</td>
<td>94%</td>
<td>48%</td>
<td>28%</td>
<td>18%</td>
<td>7%</td>
<td>4%</td>
</tr>
<tr>
<td>BiMax</td>
<td>34%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Conclusion & Perspectives

- A new approach based on max. flow/min. cut algorithms for mining patterns in noisy contexts.

- The results are very promising regarding:
  - quality and size of the extracted patterns
  - reasonable running time
  - annotation quality of results

- Perspectives:
  - Adapt weight one of the bipartite graph to bias search towards regions that take domain knowledge into account
  - Links with ’noisy closure’
Collaborators

LIPN, Univ. Paris Nord
I. Chebil
L. Létocart
K. Mouhoubi

Univ. Evry
M. Elati
P. Neuvial
R. Nicolle

Institut Curie
E. Barillot
F. Radvanyi

IGM, Orsay
M. Bolotin