On Network Inference and Validation Methods

Patrick E. Meyer

Bioinformatics and Systems Biology (BioSys)
PhytoSystems, Université de Liège (ULg, Belgium)

November 2014 (NETBIO)
Our BioSys Lab

Our unit:
Bioinformatics and Systems Biology (Biosys)
Université de Liège, Belgium

Team biased towards large networks, machine learning and algae...

Collaborating with three PhD students:

- Ngoc Pham (From Vietnam)
  Expression-Based Transcriptional Networks

- Eoin Marron (From Ireland)
  Chlamydomonas reinhardtii data-mining

- Pau Bellot (From Spain, co-tutelle with UPC)
  Meta-network inference
Outline

1. Introduction
2. Causality and Expression Data
3. modENCODE
4. State of the Art
5. Inference
6. Validation
7. Conclusions
Outline

1 Introduction

2 Causality and Expression Data

3 modENCODE

4 State of the Art

5 Inference

6 Validation

7 Conclusions
Notation

- $X = (X_1, X_2, ..., X_n)$: the set of $n$ variables
- $X_k \in X$: one variable of the set
- $X_K \subset X$: a subset of variables
- $X_{-k} = X \setminus X_k$: set of variables without $X_k$
- $X_{-K}$: the set $X$ without the subset of variables $X_K$
- $X_{i,j} = \{X_i, X_j\}$: two variables of the set $X$
- $X_{-(i,j)}$: set of variables $X$ without $X_i$ and $X_j$
Mutual Information (MI)

Definition ([Thomas and Cover])

Let $X_i$ and $X_j$ be two (discrete) random variables, the mutual information between $X_i$ and $X_j$ is

$$I(X_i; X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log \left( \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right)$$

- Mutual information is a divergence between the joint and the product distribution.
- $I(X_i; X_j)$ is maximal if $X_i$ or $X_j$ is perfectly predictable from the other.
- $I(X_i; X_j) = 0$ if $X_i$ or $X_j$ are independent (unpredictable).
Definition ([Thomas and Cover])

Let $X_i$, $X_j$ and $X_k$ be three random variables, the conditional mutual information between two random variables $X_i$ and $X_j$ knowing $X_k$ is

$$I(X_i; X_j|X_k) = I((X_i, X_k); X_j) - I(X_k; X_j)$$

- It measures the gain of information on $X_j$ (or $X_i$) due to the other variable $X_i$ (or $X_j$), when $X_k$ is given.
- $I(X_i; X_j|X_k) \geq 0$ with equality iff $X_i$ and $X_j$ are conditionally independent given $X_k$. 
Transcriptional Network

- $\text{gene} \rightarrow RNA \rightarrow protein$

- some protein (tf) can modify RNA production of target genes (tg)

$\Rightarrow$ Each cell has an encoded network (circuit) in DNA.

- Each node is a gene.
- An arc connects a regulator gene (tf) to a regulated one (tg).
Problem Formalization

- inputs $X$: $m \times n$ matrix, where $x_{ri}$ is the realization of gene $X_i$ at measurement $s_r$
- output $\hat{T}$: list of triplets $(tf, weight, tg)$ of length $\#tf \times \#tg$

<table>
<thead>
<tr>
<th>DATA</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>...</th>
<th>$X_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s 1</td>
<td>0.1</td>
<td>0.9</td>
<td>...</td>
<td>0.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>s m</td>
<td>0.2</td>
<td>0.3</td>
<td>...</td>
<td>0.8</td>
</tr>
</tbody>
</table>

$\Rightarrow$

<table>
<thead>
<tr>
<th>$tf$</th>
<th>$w$</th>
<th>$tg$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.1</td>
<td>$X_2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$X_{#tf}$</td>
<td>0.9</td>
<td>$X_{#tg}$</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Causality and Expression Data
3. modENCODE
4. State of the Art
5. Inference
6. Validation
7. Conclusions
Cause

**Definition (Cause [Neapolitan, 2003])**

\[ X_i \text{ is a cause of } X_j, \text{ denoted by } X_i \rightarrow X_j, \text{ if there exists a value } x_i \in X_i \text{ such that setting } X_i = x_i \text{ leads to a change in the probability distribution of } X_j. \]

In other words: causality creates a (bivariate) dependency between a cause and its effect.

\[ X_i \leftrightarrow X_j \Rightarrow I(X_i; X_j) > 0 \]

where \( X_i \leftrightarrow X_j \) denote an *undirected causal link*, i.e., \( X_i \rightarrow X_j \) or/and \( X_i \leftarrow X_j \).
Assumption

\[ X_j \leftrightarrow X_i \Rightarrow I(X_i; X_j) > 0 \]

This bivariate dependency is true in most cases but not always: cancellation of two causal pathways, the XOR.

Example (XOR problem [Neapolitan 2003])

\[
\begin{array}{c|cccc}
  & X_i & X_k & X_j \\
\hline
X_i & 1 & 1 & 0 & 0 \\
X_k & 1 & 0 & 1 & 0 \\
X_j = X_i \oplus X_k & 0 & 1 & 1 & 0 \\
\end{array}
\]
Indirect links

- In most cases, $X_j \leftrightarrow X_i \implies I(X_i; X_j) > 0$

- Unfortunately, reverse is not true:
  There are three cases of indirect interaction with three variables:

  1. $X_j \rightarrow X_k \rightarrow X_i$
  2. $X_j \leftarrow X_k \rightarrow X_i$
  3. $X_j \rightarrow X_k \leftarrow X_i$

Two of them typically lead to high $I(X_j; X_i)$
### Direct Causality

**Definition (Direct cause [Neapolitan, 2003])**

\( X_i \) is a direct cause of \( X_j \) if \( X_i \) is a cause of \( X_j \) and there is no other variable \( X_k \) such that once we know the value of \( X_k \), a manipulation of \( X_i \) no longer changes the probability distribution of \( X_j \).

It means:

two dependent variables are no longer dependent once given the direct cause.

\[
\begin{align*}
X_i &\rightarrow X_k \rightarrow X_j \\
X_i &\leftarrow X_k \rightarrow X_j \\
\Rightarrow & \quad I(X_i; X_j | X_k) = 0
\end{align*}
\]
Direct causality (2)

Equivalently: if there are no set of variables that cancel the dependency between two variables, then one of these variables is a direct cause of the other. More formally:

\[ \forall X_K \subseteq X_{-(i,j)} : I(X_i; X_j|X_K) > 0 \Rightarrow X_i \leftrightarrow X_j \]

Implicit assumption of *causal sufficiency*, that is all the variables that cause at least two effects (two variables in the dataset) should also be present in the dataset:

\[ \forall (X_i, X_j) \in X : \exists X_k, X_i \leftarrow X_k \rightarrow X_j \Rightarrow X_k \in X_{-(i,j)} \]
MRNET

**Network Inference** Based on Variable selection min-redundancy-max-relevance (mRMR) \[ [Meyer et al., 2007] \]

\[
X_i^{M R M R} = \arg \max_{X_i \in X_{-K}} \left\{ I(X_i; X_j) - \frac{1}{|K|} \sum_{X_k \in X_K} I(X_i; X_k) \right\}
\]

*Bivariate approx. of \(I(X_i; X_j | X_K) \rightarrow adapted to expression data*

**State-of-the-art**

<table>
<thead>
<tr>
<th>Method</th>
<th>RBN</th>
<th>ARACNe</th>
<th>Lasso</th>
<th>MRNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed/Size</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>indirect arcs</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>non-linearity</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

**Package:** Bioconductor (5000+ downloads/year/since 2008)
Outline

1. Introduction
2. Causality and Expression Data
3. modENCODE
4. State of the Art
5. Inference
6. Validation
7. Conclusions
modENCODE project

- 4 years of work from 50+ different institutions
- Kellis lab (CSAIL MIT + BROAD Institute) coordinating the integrative analysis to gain insights into the regulatory circuitry that controls gene expression in response to changing environments. [The modENCODE Consortium et al. Science 2010, genome Research 2012]
Problem

Drosophila melanogaster data:

- Publicly available data:
  - list of $>700$ known tf
  - $>14k$ genes
  - 12 Drosophila genomes
  - 139 known tf binding motifs
  - GO functional terms database
  - $>1000$ Protein-Protein Interactions
  - REDfly data
  - 2 "big" microarray datasets (Flyatlas + GSE6186)

- modENCODE data:
  - 2 RNAseq datasets
  - 2 histone modifications datasets
  - 76 tf-binding experiments (ChIP full genome)

→ Transcriptional network?
Introduction

Causality and Expression Data

modENCODE

State of the Art

Inference

Validation

Conclusions
ChIP-binding based network

Binding experiments for 76 tfs (full genome)

<table>
<thead>
<tr>
<th>cond.</th>
<th>tf</th>
<th>chrom.</th>
<th>peakStart</th>
<th>peakEnd</th>
<th>intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>CG1674</td>
<td>chr2L</td>
<td>1</td>
<td>5954</td>
<td>0.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

→ threshold on intensity
but lots of non-functional binding (not intensity dependent)

Gene annotation file from flybase.org

<table>
<thead>
<tr>
<th>name</th>
<th>chrom</th>
<th>txStart</th>
<th>txEnd</th>
<th>cdsStart</th>
<th>cdsEnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG1678</td>
<td>chr4</td>
<td>251355</td>
<td>266500</td>
<td>252579</td>
<td>266389</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

→ There is a link if binding near (+ - 500bp) of txStart
ChIP-binding based network (2)

For all tf-tg pairs, an edge weight is

- 0 if no binding evidence at 500 bp near txStart
- 0.1 if no data for a tf
- 1 if binding

<table>
<thead>
<tr>
<th></th>
<th>tf</th>
<th>w</th>
<th>tg</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.1</td>
<td>X2</td>
<td></td>
</tr>
<tr>
<td>Xi</td>
<td>0</td>
<td>Xk</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>X_#tf</td>
<td>1</td>
<td>X_#tg</td>
<td></td>
</tr>
</tbody>
</table>
Binding motif-based network

From flybase.org

- DNA sequence
- 139 known tf binding motifs

→ search (GREP) binding motif in the genome.

**Problem:** too many non-functional binding motifs

- gene annotation file

<table>
<thead>
<tr>
<th>name</th>
<th>chrom</th>
<th>txStart</th>
<th>txEnd</th>
<th>cdsStart</th>
<th>cdsEnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG1674</td>
<td>chr4</td>
<td>251355</td>
<td>266500</td>
<td>252579</td>
<td>266389</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

→ There is a link if tf motif near (+ - 500bp) of txStart
Binding motif-based network (2)

Use 12 Drosophila genomes with Branch Length Score (BLS) confidence [Kheradpour et al., gen.res., 2007]
Expression based Networks

Two steps:

1. Co-expression network: compute MI/correlation for all couples of genes
   but false positive trends because of indirect links
   Assume $X_1$ influence $X_3$ through $X_2$

   \[
   \begin{align*}
   X_1 & \leftrightarrow X_2 \\
   X_2 & \leftrightarrow X_3 \\
   \end{align*}
   \]

   Then $I(X_1; X_2)$ and $I(X_2; X_3)$ will be high
   but also $I(X_1; X_2)$, hence it adds a false link between $X_1$ and $X_3$.

2. Use an indirect-arc elimination algorithm on the correlation/MIM matrix.
   - ARACNE [Margolin et al, BMC Bioinfo, 2006]
   - MRNET [Meyer et al., BMC Bioinfo., 2008]
Networks from sequence and/or tf binding
- pro: physical connections (directed)
- issue: elimination of non functional bindings

Networks from expression and/or chromatin data
- pro: functional connections (but undirected)
- issue: elimination of indirect interactions

$G_1 \leftrightarrow G_2 \leftrightarrow \downarrow \uparrow \leftrightarrow G_3$

→ combine physical and functional networks to extract direct functional interactions
Chromatin can compact the genome up to 40000 times

- 5 families: H1, H2A, H2B, H3, H4
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification: 4 modifications: me1, me2, me3, ac

→ H3K4me1 denotes the monomethylation of the 4th residue (a lysine) from the start of the H3 protein.

51 distinct chromatin states suggests distinct biological roles (Ernst et al. Nature 2010).
We have two datasets of measurements (ChIP)

- **Ts**: H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K27ac, H3K9ac
- **Ct**: H3K4me2, H4K16ac, H3K36me1, H3K36me3, H3K79me1, H3K79me2, H3K23ac, H3K18ac, H4K12ac, H4K5ac, H2BK5ac, H4K8ac.
Functional networks

<table>
<thead>
<tr>
<th>gene</th>
<th>M</th>
<th>A</th>
<th>R</th>
<th>K</th>
<th>1</th>
<th>M</th>
<th>A</th>
<th>R</th>
<th>K</th>
<th>2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>tf</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>tg</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

Squared Spearman correlation between

- tf and tg chromatin profiles (2 datasets) → 2 co-chromatin networks
- tf and tg expression profiles (3 datasets) → 3 co-expression networks
- 1 expression dataset kept for validation → 5 functional networks inferred + 2 physical networks inferred (ChIP and motif)
Consensus Networks
Supervised Network

Method: supervised logistic regression

- Weight $w_{ij}$ from tf $i$ to tg $j$, $w_{ij}^{output} = \frac{1}{1+e^{-m}}$
  
  $m = \alpha_0 + \alpha_{motif}w_{ij}^{motif} + \alpha_{ChIP}w_{ij}^{ChIP} +$
  $\alpha_{chromtc}w_{ij}^{chromtc} + \alpha_{chromcl}w_{ij}^{chromcl} +$
  $\alpha_{RNAseqtc}w_{ij}^{RNAseqtc} + \alpha_{arraytc}w_{ij}^{arraytc} + \alpha_{flyatlas}w_{ij}^{flyatlas}$

- 10 fold cross-validation

- positive set: random sampling (with replacement) of 2k interactions of the 233 REDfly interactions

- negative set: random sampling of 2k interactions out of the 7k non-REDfly interactions

- fitting using iterative reweighted least squares

- final network: 318k edges (0.6 confidence)
Outline

1. Introduction
2. Causality and Expression Data
3. modENCODE
4. State of the Art
5. Inference
6. Validation
7. Conclusions
REDfly PR-Curves

Precision-recall for REDfly

Logistic regression weights: $\alpha_{motif,chromtc} = 2$, $\alpha_{ChIP,chromcl,RNAseq} = 1$, $\alpha_{array,flyatlas} = 0.4$
Structural properties: degree distributions

Similar to E.coli and S.Cerevisae known network topology
## Most frequent three-nodes patterns

<table>
<thead>
<tr>
<th>Network Motif</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td><strong>Illustration</strong></td>
</tr>
<tr>
<td>Cross-regulating TFs co-targeting another TF (Double FFL)</td>
<td><img src="image1" alt="Illustration" /></td>
</tr>
<tr>
<td>Cross-regulatory clique of TFs (Six FFLs)</td>
<td><img src="image2" alt="Illustration" /></td>
</tr>
<tr>
<td>Cross-regulating TFs co-targeted by another TF (Double FFL)</td>
<td><img src="image3" alt="Illustration" /></td>
</tr>
<tr>
<td>Cross-regulating TFs co-targeting a target gene (Double FFL)</td>
<td><img src="image4" alt="Illustration" /></td>
</tr>
<tr>
<td>Feedback loop between three TFs</td>
<td><img src="image5" alt="Illustration" /></td>
</tr>
<tr>
<td>Cross-regulating TFs creating a feed-forward and a feedback loop</td>
<td><img src="image6" alt="Illustration" /></td>
</tr>
</tbody>
</table>

- **Unsupervised network**
- **Supervised network**
- miRNA
- Transcription factor
- Target gene
Biological Insights on co-targeted genes

Is the inferred network enriched in:

Compared to

1. protein-protein interactions (PPI)
2. co-expressed in developmental cycle (RNAseq)
3. similar function profiles (GO terms)
## Results

**Fold enrichment of co-targeted genes**

<table>
<thead>
<tr>
<th></th>
<th>PPI</th>
<th>GO</th>
<th>RNAseq</th>
</tr>
</thead>
<tbody>
<tr>
<td>motif</td>
<td>1.39</td>
<td>1.06</td>
<td>1.08</td>
</tr>
<tr>
<td>ChIP</td>
<td>1.24</td>
<td>1.23</td>
<td>1.46</td>
</tr>
<tr>
<td>unsupervised</td>
<td>1.53</td>
<td>1.44</td>
<td>3.07</td>
</tr>
<tr>
<td>supervised</td>
<td>1.58</td>
<td>1.55</td>
<td>3.62</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Causality and Expression Data
3. modENCODE
4. State of the Art
5. Inference
6. Validation
7. Conclusions
Results

Our integrative networks outperform feature-specific networks
- PR-Curves on REDfly
- Enrichment of co-targeted genes on PPI, expression and GO terms

Our integrative networks fit known topological properties observed in E.coli and S.cerevisae
- In-degree and out-degree
- Most frequent three-nodes patterns
http://homepage.meyerp.com

Thank you!

Questions?