Introduction

Causality and Expression Data

modENCODE

State of the Art

Inference

Validation

Conclusions

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

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

- 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).

Conditional Mutual Information (CMI)

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

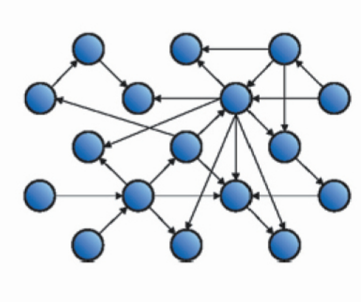
$$I(X_i; X_j|X_k) = I((X_i, X_k); X_j) - I(X_k; X_j)$$

- 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

- $gene \rightarrow RNA \rightarrow protein$
- 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_{r_i} is the realization of gene X_i at measurement s_r
- output $\hat{T}$: list of triplets $(tf, weight, tg)$ of length $\#tf \times \#tg$

DATA	X_1	X_2	...	X_n
s 1	0.1	0.9	...	0.5
...	...	...	...	...
s m	0.2	0.3	...	0.8

$\Rightarrow$

tf	w	tg
X_1	0.1	X_2
...	...	...
...	...	...
$X_{\#tf}$	0.9	$X_{\#tg}$

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 is a *cause* of X_j, denoted by $X_i \rightarrow X_j$, if there exists a value $x_i \in \mathcal{X}_i$ such that setting $X_i = x_i$ 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.

$$X_i \leftrightarrow X_j \Rightarrow I(X_i; X_j) > 0$$

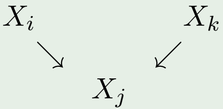
where $X_i \leftrightarrow X_j$ denote an *undirected causal link*, i.e., $X_i \rightarrow X_j$ or/and $X_i \leftarrow X_j$.

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]*)



X_i	1	1	0	0
X_k	1	0	1	0
$X_j = X_i \oplus X_k$	0	1	1	0

Indirect links

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

$$\left.\begin{array}{l} X_i \rightarrow X_k \rightarrow X_j \\ X_i \leftarrow X_k \rightarrow X_j \end{array}\right\} \Rightarrow I(X_i; X_j | X_k) = 0$$

Direct causality (2)

Introduction
Causality and Expression Data
modENCODE
State of the Art
Inference
Validation
Conclusions

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]*

$$X_i^{MRMR} = \arg\max_{X_i \in X_{-K}} \{ I(X_i;X_j) - \frac{1}{|K|} \sum_{X_k \in X_K} I(X_i;X_k) \}$$

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

State-of-the-art

Method	RBN	ARACNe	Lasso	MRNET
Speed/Size	-	+	+	+
indirect arcs	+	-	+	+
non-linearity	+	+	-	+

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

- Model Organism Encyclopedia Of DNA Elements (modENCODE) : the most comprehensive collections of functional datasets for a single organism: D.melanogaster [Celniker et al., Nature, 2009] (and C.elegans)
- 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?

Outline

1 Introduction

2 Causality and Expression Data

3 modENCODE

4 State of the Art

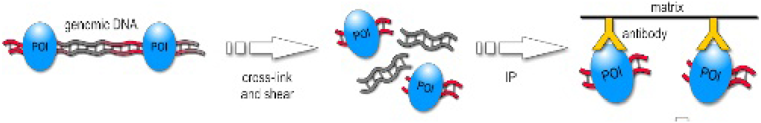
5 Inference

6 Validation

7 Conclusions

ChIP-binding based network

Binding experiments for 76 tfs (full genome)



cond.	tf	chrom.	peakStart	peakEnd	intensity
t1	CG1674	chr2L	1	5954	0.9
...	...	...	...	...	...

→ threshold on intensity

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

Gene annotation file from flybase.org

name	chrom	txStart	txEnd	cdsStart	cdsEnd
CG1678	chr4	251355	266500	252579	266389
...	...	...	...	...	...

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

ChIP-binding based network (2)

For all tf-tg pairs, an edge weight is

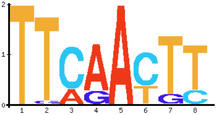
- 0 if no binding evidence at 500 bp near txStart
- 0.1 if no data for a tf
- 1 if binding

	tf	w	tg
	X_1	0.1	X_2
→	X_i	0	X_k
	...	...	...
	$X_{\#tf}$	1	$X_{\#tg}$

Binding motif-based network

From flybase.org

- DNA sequence
- 139 known tf binding motifs



→search (GREP) binding motif in the genome.
Problem: to many non-functional binding motifs

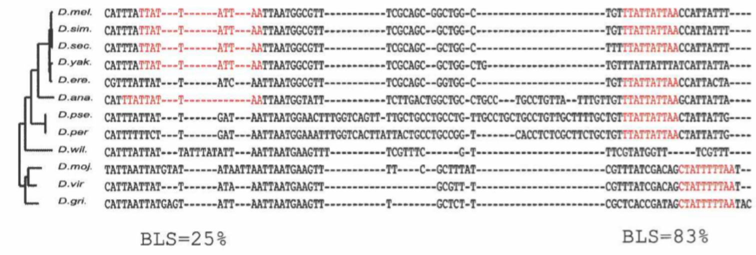
- gene annotation file

name	chrom	txStart	txEnd	cdsStart	cdsEnd
CG1674	chr4	251355	266500	252579	266389
...	...	...	...	...	...

→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]

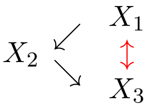

D.mel. CATTTATTAT---T------ATT---AATTAATGGCGTT-----------TCGCAGC-GGCTGG-C----------------------TGTTTATTATTAACCATTATTT----
D.sim. CATTTATTAT---T------ATT---AATTAATGGCGTT-----------TCGCAGC--GCTGG-C----------------------TGTTTATTATTAACCATTATTT----
D.sec. CATTTATTAT---T------ATT---AATTAATGGCGTT-----------TCGCAGC--GCTGG-C----------------------TTTTTATTATTAACCATTATTT----
D.yak. CATTTATTAT---T------ATT---AATTAATGGCGTT-----------TCGCAGC--GCTGG-CTG--------------------TGTTTATTATTTATCATTATTA----
D.ere. CGTTTATTAT---T------ATC---AATTAATGGCGTT-----------TCGCAGC--GGTGG-C----------------------TGTTTATTATTAACCATTACTA----
D.ana. CATTTATTAT---T------------AATTAATGGTATT-----------TCTTGACTGGCTGC-CTGCC---TGCCTGTTA--TTTGTTGTTTATTATTAAGCATTATTA----
D.pse. CATTTATTAT---T------GAT---AATTAATGGAACTTTGGTCAGTT-TTGCTGCCTGCCTG-TTGCCTGCTGCCTGTTGCTTTTGCTGTTTATTATTAACTATTATTG----
D.per CATTTTTTCT---T------GAT---AATTAATGGAAATTTGGTCACTTATTACTGCCTGCCGG-T-------CACCTCTCGCTTCTGCTGTTTATTATTAACTATTATTG----
D.wil. CATTTATTAT---TATTTATATT---AATTAATGAAGTTT----------TCGTTTC------G-T----------------------TTCGTATGGTT-----TCGTTT----
D.moj. TATTAATTATGTAT------ATAATTAATTAATGAAGTT-----------TT----C--GCTTTAT----------------------CGTTTATCGACAGCTATTTTTAAT--
D.vir CATTAATTAT---T------ATA---AATTAATGAAGTT------------------GCGTT-T----------------------CGTTTATCGACAGCTATTTTTAAT--
D.gri. CATTAATTATGAGT------ATT---AATTAATGAAGTT-----------T--------GCTCT-T----------------------CGCTCACCGATAGCTATTTTTAATAC
BLS=25%
BLS=83%

	tf	w	tg
	X_1	0.1	X_2
$\rightarrow$	X_i	0	X_k
	...	...	...
	$X_{\#tf}$	0.83	$X_{\#tg}$

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

   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]

Outline

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

Principle

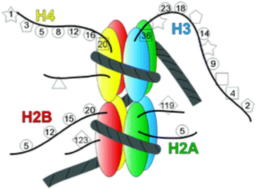
- Networks from sequence and/or tf binding
  - pro: physical connections (directed)
  - issue: elimination of non functional bindings
- Networks from expression and/or chromatin data
  - pro: functional connections (but undirected)
  - issue: elimination of indirect interactions

$$\begin{tikzcd}
 & G_2 \arrow[dl] \arrow[dd, leftrightarrow, red] \\
G_1 \arrow[dr] & \\
 & G_3
\end{tikzcd}$$

→ combine physical and functional networks to extract direct functional interactions

Chromatin regulation with histone modification

Chromatin can compact the genome up to 40000 times



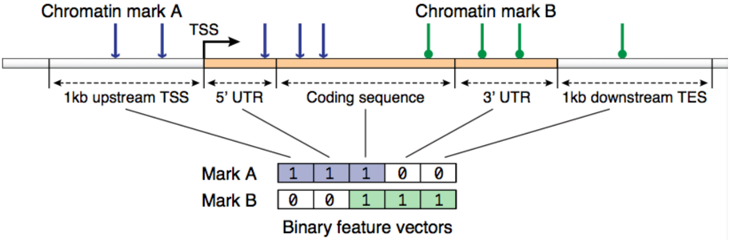
- 5 families: H1, H2A, H2B, H3, H4
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification: 4 modifications: me1, me2, me3, ac

→ 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).

Co-chromatin network

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.



Mark A	1	1	1	0	0
Mark B	0	0	1	1	1

Functional networks

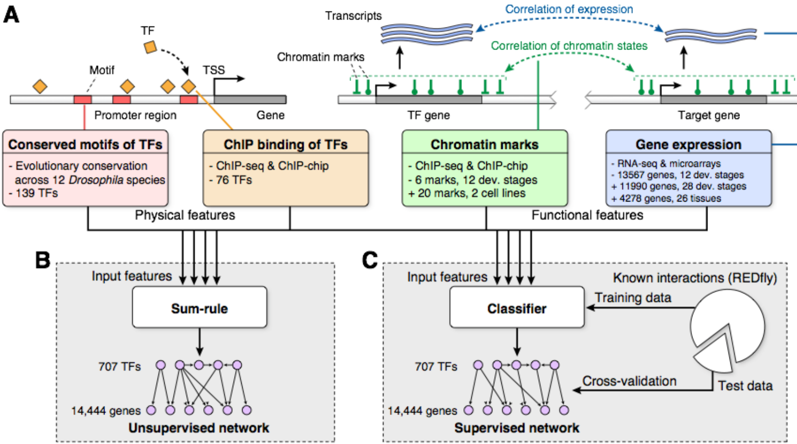
gene	M	A	R	K	1	M	A	R	K	2	...
tf	1	1	0	0	0	0	1	1	1	0	...
tg	1	0	0	0	0	0	1	1	1	1	...

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

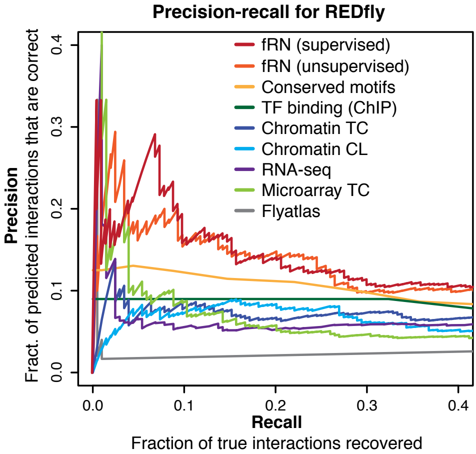
4 State of the Art

5 Inference

6 Validation

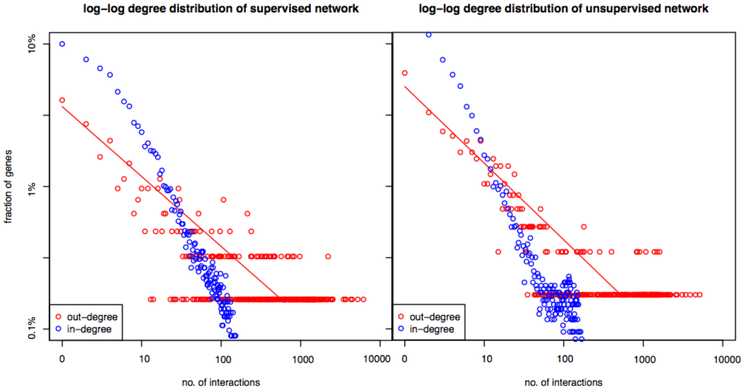
7 Conclusions

REDfly PR-Curves



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

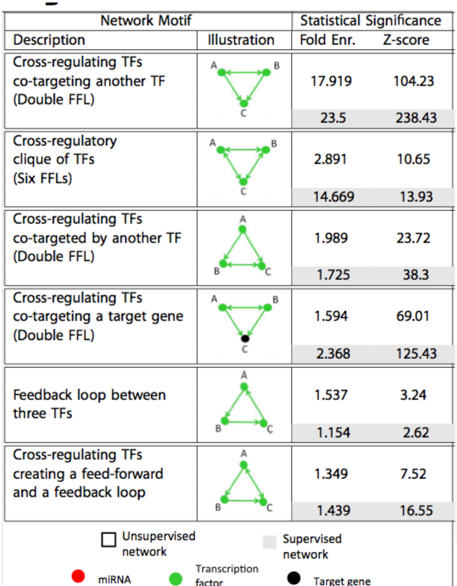
Structural properties: degree distributions



Similar to E.coli and S.Cerevisae known network topology

Most frequent three-nodes patterns

Introduction
Causality and Expression Data
modENCODE
State of the Art
Inference
Validation
Conclusions

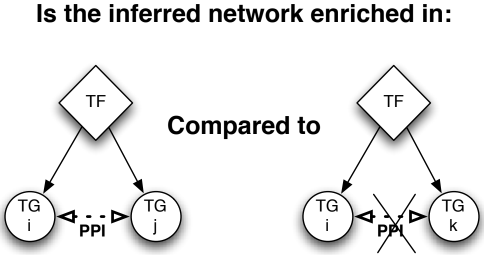


Network Motif		Statistical Significance	
Description	Illustration	Fold Enr.	Z-score
Cross-regulating TFs co-targeting another TF (Double FFL)	A B C	17.919 23.5	104.23 238.43
Cross-regulatory clique of TFs (Six FFLs)	A B C	2.891 14.669	10.65 13.93
Cross-regulating TFs co-targeted by another TF (Double FFL)	A B C	1.989 1.725	23.72 38.3
Cross-regulating TFs co-targeting a target gene (Double FFL)	A B C	1.594 2.368	69.01 125.43
Feedback loop between three TFs	A B C	1.537 1.154	3.24 2.62
Cross-regulating TFs creating a feed-forward and a feedback loop	A B C	1.349 1.439	7.52 16.55

□ Unsupervised network ▨ Supervised network

● miRNA ● Transcription factor ● Target gene

Biological Insights on co-targeted genes



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

network	PPI	GO	RNAseq
motif	1.39	1.06	1.08
ChIP	1.24	1.23	1.46
unsupervised	1.53	1.44	3.07
supervised	1.58	1.55	3.62

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 ?