Transcriptional network structure assessment via the Data Processing Inequality

Enrique Hernández-Lemus,
Computational Genomics,
National Institute of Genomic Medicine
(México)

Outline

- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

Outline

- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

Introduction

Information theoretical measures have been been used successfully to **infer 2-way interactions in gene regulatory networks** (GRNs). In particular the family of measures that includes mutual information, Markov random fields and Kullback-Liebler divergences, has established itself as a sound and robust alternative for this task.

Introduction

However, due to the fact that conditional probabilities obey a **'tower property'**, a number of 'false positives' links appear, in some instances as a consequence of indirect correlations. For instance, if process A has a high value of conditional probability (say, mutual information) with process B, and process B is also highly correlated with process C, **most common algorithms would predict also a (possibly non-existent) link** between processes A and C.

Introduction

One way to assess and correct for this indirect links is by use of the **Data Processing Inequality (DPI)** which is a simple but useful theorem that states that no matter what processing you do on some data, you cannot get more information (in the sense of Shannon) out of a set of data than was there to begin with. In a sense, it provides **a bound on how much can be accomplished with signal processing**.

Introduction

As we will see DPI states simply that if genes g1 and g3 interact only through a third gene, g2; we have that I(g1; g3) <= min[I(g1; g2); I (g2; g3)]. Hence, **the least of the three conditional measures can come from indirect interactions** only so that the proposed algorithm examines each gene triplet for which all three measures are greater than a threshold value and removes the edge with the smallest value

Introduction

Hence DPI is thus useful to **quantify efficiently the dependencies among a large number of genes** because eliminates those statistical dependencies that might be of an indirect nature, such as between two genes that are separated by intermediate steps in a transcriptional cascade.

We will outline an algorithmic implementation of the DPI within the framework of GRN inference and structure assessment.

Outline

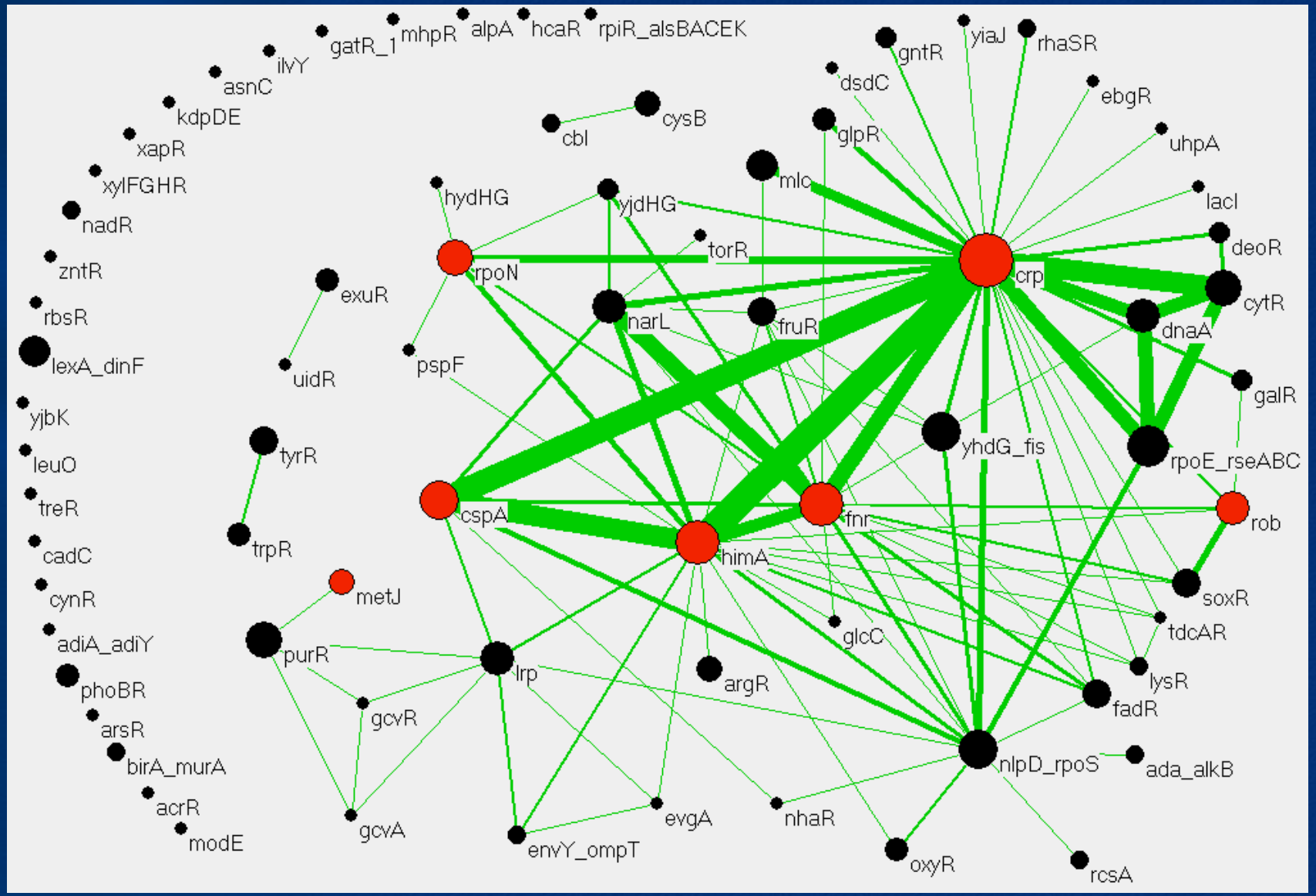
- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

Motivation

Most **common pathologies** are not caused by the mutation of a single gene, rather they are **complex diseases** that arise due to the dynamic interaction of many genes and environmental factors. To construct dynamic maps of gene interactions (i.e. GRNs) we need to understand the interplay between thousands of genes.

Motivation

One important problem in contemporary computational biology, is thus, that of reconstructing **the best possible set of regulatory interactions between genes** (a so called gene regulatory network -GRN) from **partial knowledge**, as given for example by means of gene expression analysis experiments.


gatR_1
mhpR
alpA
hcaR
rpiR_alsBACEK
ilvY
asnC
kdpDE
xapR
xylFGHR
nadR
zntR
rbsR
lexA_dinF
yjbK
leuO
treR
cadC
cynR
adiA_adiY
phoBR
arsR
birA_murA
acrR
modE
cbl
cysB
hydHG
yjdHG
rpoN
exuR
uidR
pspF
narL
torR
mlc
glpR
dsdC
gntR
yiaJ
rhaSR
ebgR
uhpA
lacI
deoR
crp
cytR
dnaA
fruR
galR
yhdG_fis
rpoE_rseABC
rob
cspA
fnr
himA
soxR
tdcAR
lysR
fadR
glcC
argR
lrp
tyrR
trpR
metJ
purR
gcvR
gcvA
envY_ompT
evgA
nhaR
oxyR
rcsA
nlpD_rpoS
ada_alkB

Outline

- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

The gene network inference problem

Several issues arise in the analysis of experimental data related to gene function:

- The nature of measurement processes generates **highly noisy signals**
- There are **far more variables involved** (number of genes and interactions among them) **than experimental samples**.
- Another source of complexity is the **highly nonlinear** character of the underlying biochemical dynamics.

The gene network inference problem

Information theory (IT) has resulted on a powerful theoretical foundation to develop algorithms and computational techniques to deal with network inference problems applied to real data. There are however goals and challenges involved in the application of IT to genomic analysis.

The applied algorithms should return **intelligible models** (i.e. they must result understandable), they must also **rely on little a priori knowledge**, deal with **thousands of variables**, detect **non-linear dependencies** and all of this starting from tens (or at most few hundreds) of **highly noisy samples.**

The gene network inference problem

There are several ways to accomplish this task, in our opinion, the best benchmarking options for the GRN inference scenario, are the use of **sequential search algorithms** (as opposed to stochastic search) and **performance measures based on IT**, since this made feature selection fast end efficient, and also provide an easy means to communicate the results to non-specialists (e.g. molecular biologists, geneticists and physicians).

Outline

- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

The joint probability distribution approach

The deconvolution of a GRN could be based on optimization of the Joint Probability Distribution of gene-gene *interactions* as given by gene expression experimental data could be implemented as follows:

$$P(\{g_i\}) = \frac{1}{Z}\, exp^{H_{gen}}$$

$$H_{gen} = [-\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j) - \sum_{i,j,k}^{N} \Phi_{i,j,k}(g_i, g_j, g_k) - \ldots]$$

The joint probability distribution approach

$$P(\{g_i\}) = \frac{1}{Z}\, exp^{H_{gen}}$$

$$H_{gen} = [-\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j) - \sum_{i,j,k}^{N} \Phi_{i,j,k}(g_i, g_j, g_k) - \ldots]$$

Here N is the number of genes, Φ_i's are *interactions* (i.e. correlation measures) and Z is a normalization factor (called a Partition function). The functional H is termed a *Hamiltonian* (in analogy with statistical physics)

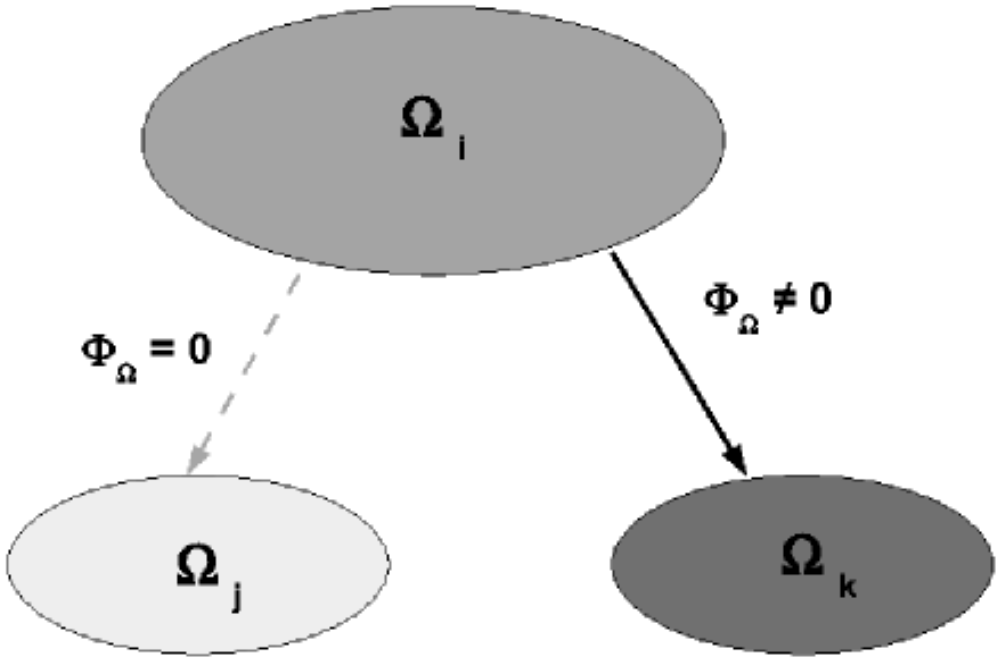


Figure 1. A set of genes Ω_i interacts with another set of genes Ω_k by means of a potential $\Phi_\Omega \neq 0$ and is non-interacting with another set of genes Ω_j since the corresponding potential functional is equal to zero.

Hernández-Lemus, E., Rangel-Escareño, C., *The role of information theory in gene regulatory network inference* en **Information Theory: New Research**, Pierre Deloumeaux, Jose D. Gorzalka (Ed.), Mathematics Research Developments Series, Nova Publishing (2012) ISBN: 978-1-62100-395-3

Outline

- Introduction
- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

We will introduce here the essential notions of IT that will be used, like entropy, mutual information and other measures. In order to do so, let X and Y denote two discrete random variables having the following features:

- Finite alphabet $\mathcal{X}$ and $\mathcal{Y}$ respectively
- Joint probability mass distribution $p(X, Y)$
- Marginal probability mass distributions $p(X)$ and $p(Y)$

Let also $\hat{X}$ and $\hat{Y}$ denote two additional discrete random variables defined on $\mathcal{X}$ and $\mathcal{Y}$ respectively, the associated probability mass distributions will be $\hat{p}(X)$ and $\hat{p}(Y)$, their joint probability mass distribution $\hat{p}(X, Y)$ and defined on $\mathcal{J}$, the joint probability sampling space; $\mathcal{J} = \mathcal{X} \times \mathcal{Y}$. For particular realizations, we have $p(x) = P(X = x)$ and $\hat{p}(y) = P(\hat{Y} = y)$.

Some useful information theoretical measures

- Entropy (Shannon-Weaver's entropy)

$$H = -K_s \sum_{\nu} p_\nu(X) \log p_\nu(X)$$

- Kullback-Leibler divergence

$$\mathcal{KL}\left[\, p(Y);\, \widetilde{p}(Y)\,\right] = \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{\widetilde{p}(y)}$$

Mutual information

- Mutual information

$$I(X,Y) = H(X) - H(X|Y)$$

- The meaning is more evident if we look at the corresponding KL representation

$$I(Y,X) = \mathcal{KL}\left[p(X,Y);\, p(X)p(Y)\right]$$

MI measures the degree of statistical dependency between two random variables. From the definition one can see that *I*(*X*, *Y*) = 0 if and only if *X* and *Y* are statistically independent.

Estimating MI between gene expression profiles under high throughput experimental setups typical of today's research in the field is a computational and theoretical challenge of considerable magnitude. One possible approximation is the use of estimators. Under a Gaussian kernel approximation, the JPD of a 2-way measurement is given as:

$$f(\vec{\mathcal{X}}) = \frac{1}{M} \sum_i \frac{\mathcal{G}\left(h^{-1}|\vec{\mathcal{X}} - \vec{\mathcal{X}}_i|\right)}{h^2}$$

$$I(\{x_i\},\{y_i\}) = \frac{1}{M} \sum_i \log \frac{f(x_i, y_i)}{f(x_i)\, f(y_i)}$$

$$I(g_i, g_j) \approx \Phi(g_i, g_j)$$

Mutual information

- Mutual information allows to distinguish different kinds of 2-way interactions (one particular case of interest is that of triplets):

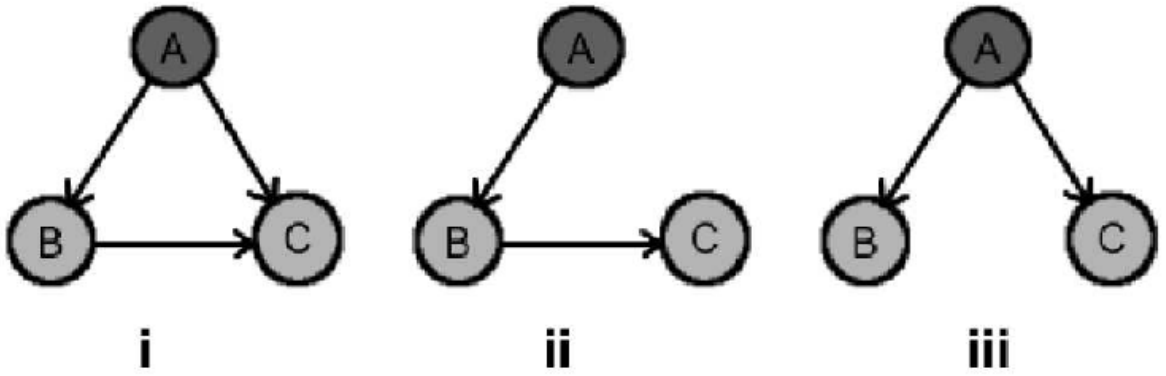


Figure 2. Panel **i** shows a bivariate interaction between gene A and genes B and C, panel **ii** shows an indirect interaction of gene A on gene C mediated by gene B, panel **iii** depicts two independent interactions between gene A and B and gene A and C.

Hernández-Lemus, E., Rangel-Escareño, C., *The role of information theory in gene regulatory network inference* en **Information Theory: New Research**, Pierre Deloumeaux, Jose D. Gorzalka (Ed.), Mathematics Research Developments Series, Nova Publishing (2012) ISBN: 978-1-62100-395-3

Direct and indirect interactions: How to tell?

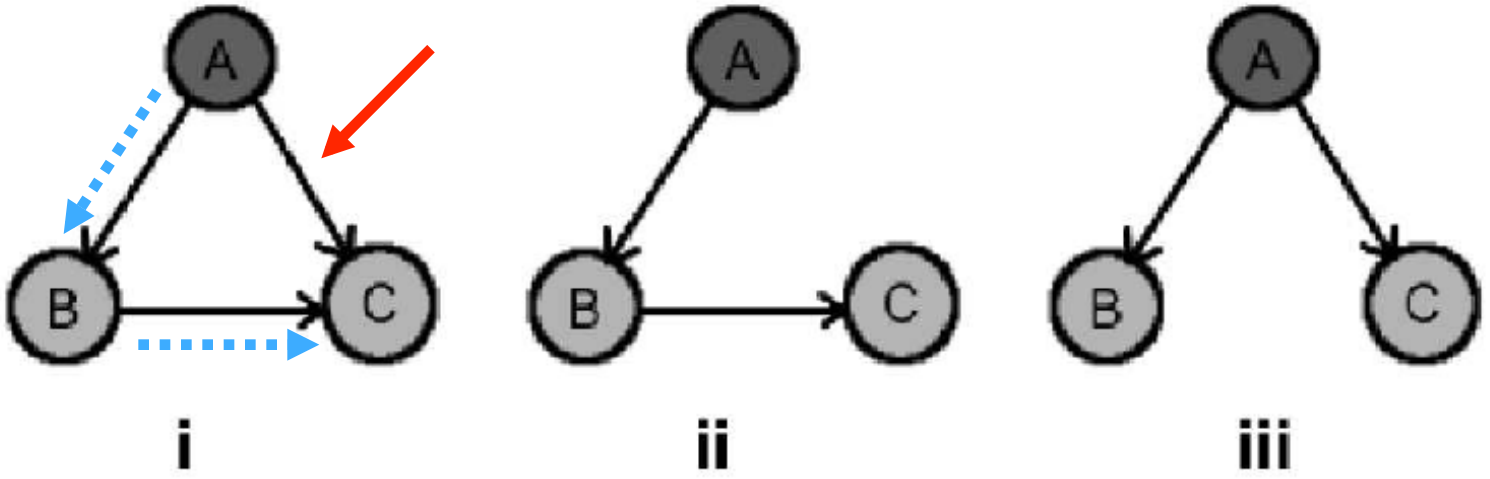


Figure 2. Panel **i** shows a bivariate interaction between gene A and genes B and C, panel **ii** shows an indirect interaction of gene A on gene C mediated by gene B, panel **iii** depicts two independent interactions between gene A and B and gene A and C.

Hernández-Lemus, E., Rangel-Escareño, C., *The role of information theory in gene regulatory network inference* en **Information Theory: New Research**, Pierre Deloumeaux, Jose D. Gorzalka (Ed.), Mathematics Research Developments Series, Nova Publishing (2012) ISBN: 978-1-62100-395-3

The Data Processing Inequality (DPI) for Markov chains

Definition: Three random variables X, Y and Z are said to form a **Markov chain** (in that order) denoted X → Y → Z if the conditional distribution of Z depends only on Y and is independent of X. That is, if we know Y, knowing X does not tell us any more about Z than if we know only Y.

If X, Y and Z form a Markow chain, then the JPD can be written:

$$P(X,Y,Z) = P(X)\,P(Y|X)\,P(Z|Y)$$

The Data Processing Inequality

Theorem: If X, Y and Z form a **Markov chain** X → Y → Z then

$$I(X;Z) \leq I(X;Y)$$

Proof By the chain rule for mutual information we can write

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned}$$

By the Markov property, since X and Z and independent given Y,

$$I(X;Z|Y) = 0.$$

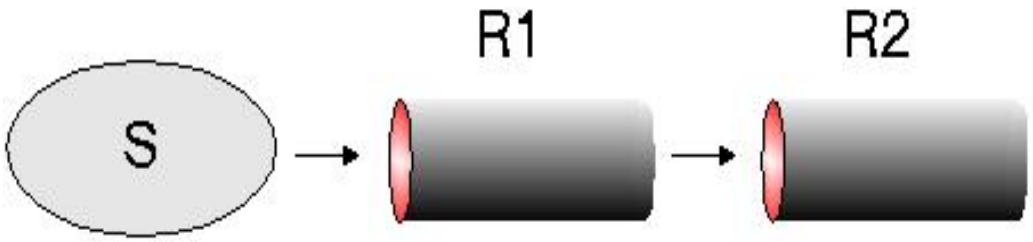
Since $I(X;Y|Z) \geq 0$ we have

$$I(X;Y) \geq I(X;Z).$$

□

Information theoretical interpretation of the DPI

'Data Processing Inequality' Theorem



$$I(S, R1) \geq I(S,R2)$$

Information can not be recovered after being degraded.

DPI as a tool to find *sufficient statistics*

The data processing inequality is the foundation for the idea of sufficient statistics. Suppose you have observations $x_1, x_2, x_3, \ldots, x_n$ of a random variable X distributed according to $f_\theta(x)$. A statistic $T(X)$ extracts some of the information in your observed sample: $X \rightarrow T(X)$. By the data processing inequality, $I(\theta, X) \geq I(\theta, T(X))$. If equality obtains, T is a *sufficient statistic* for θ. In other words, a sufficient statistics for some distribution $f_\theta(x)$ extracts *all* of the information within your data $x_1, x_2, x_3, \ldots, x_n$ about the value of θ.

DPI is thus useful to infer **Minimum Networks**, i.e. the smaller GRNs that captures μ-almost-all information content of the correlation structure of the actual biological network

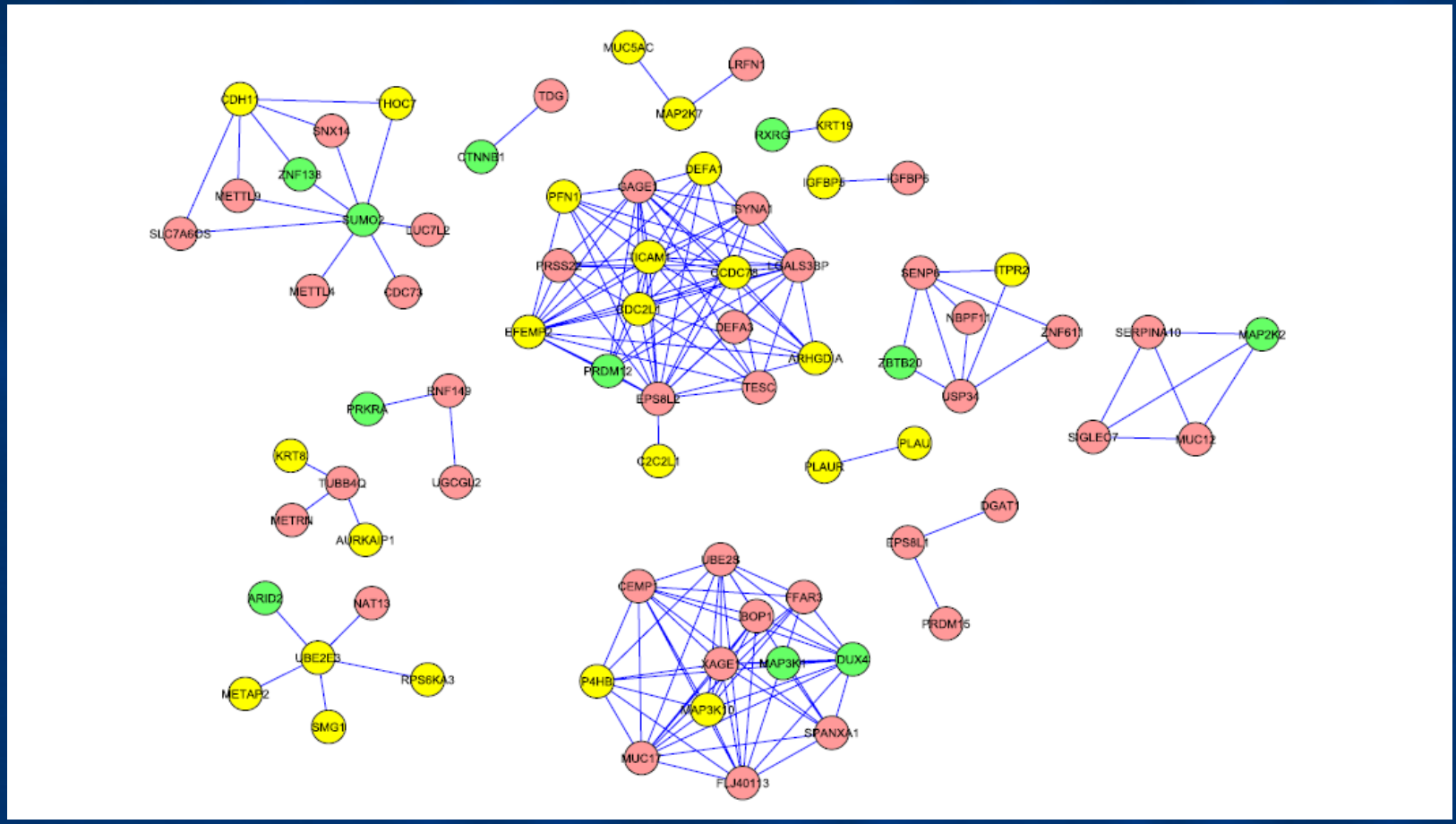
Outline

- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

Aplications I

Minimal networks

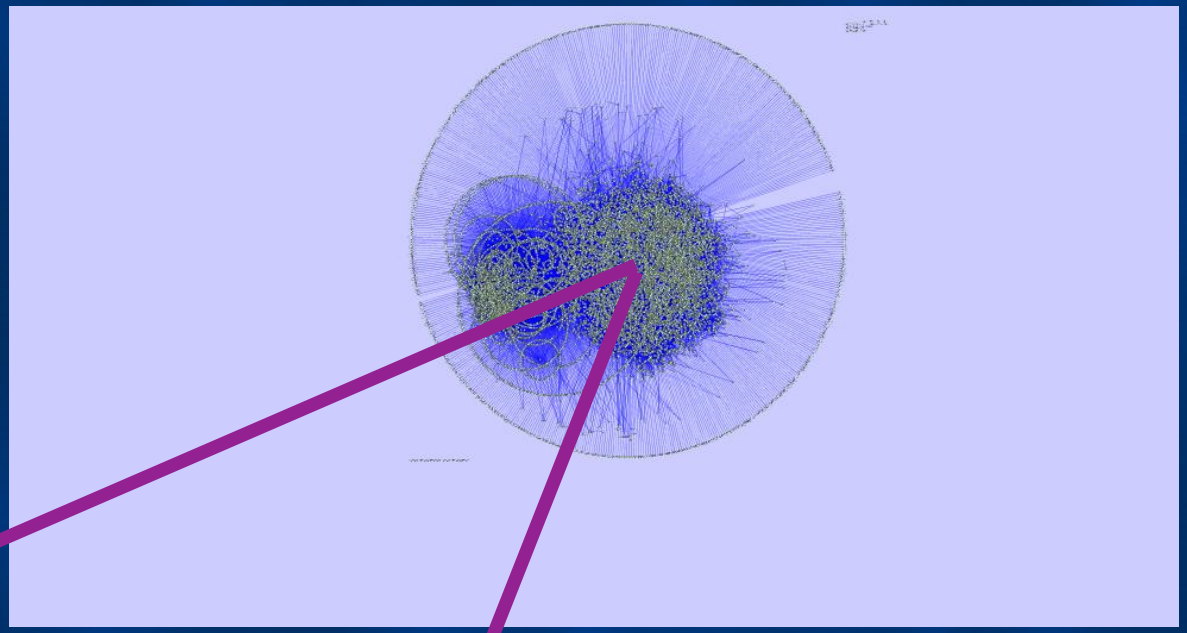
A minimal GRN for Papillary Thyroid Cancer went form 285 to 170 gene interactions

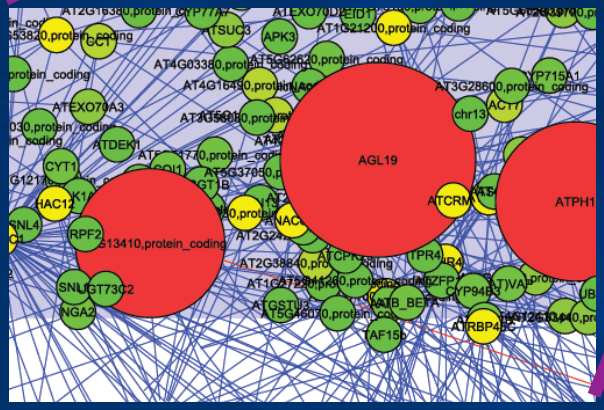


E. Hernández-Lemus et al., Physica A 388 (2009) 5057-5069

Aplications II

Non-DPI lookup for master regulators

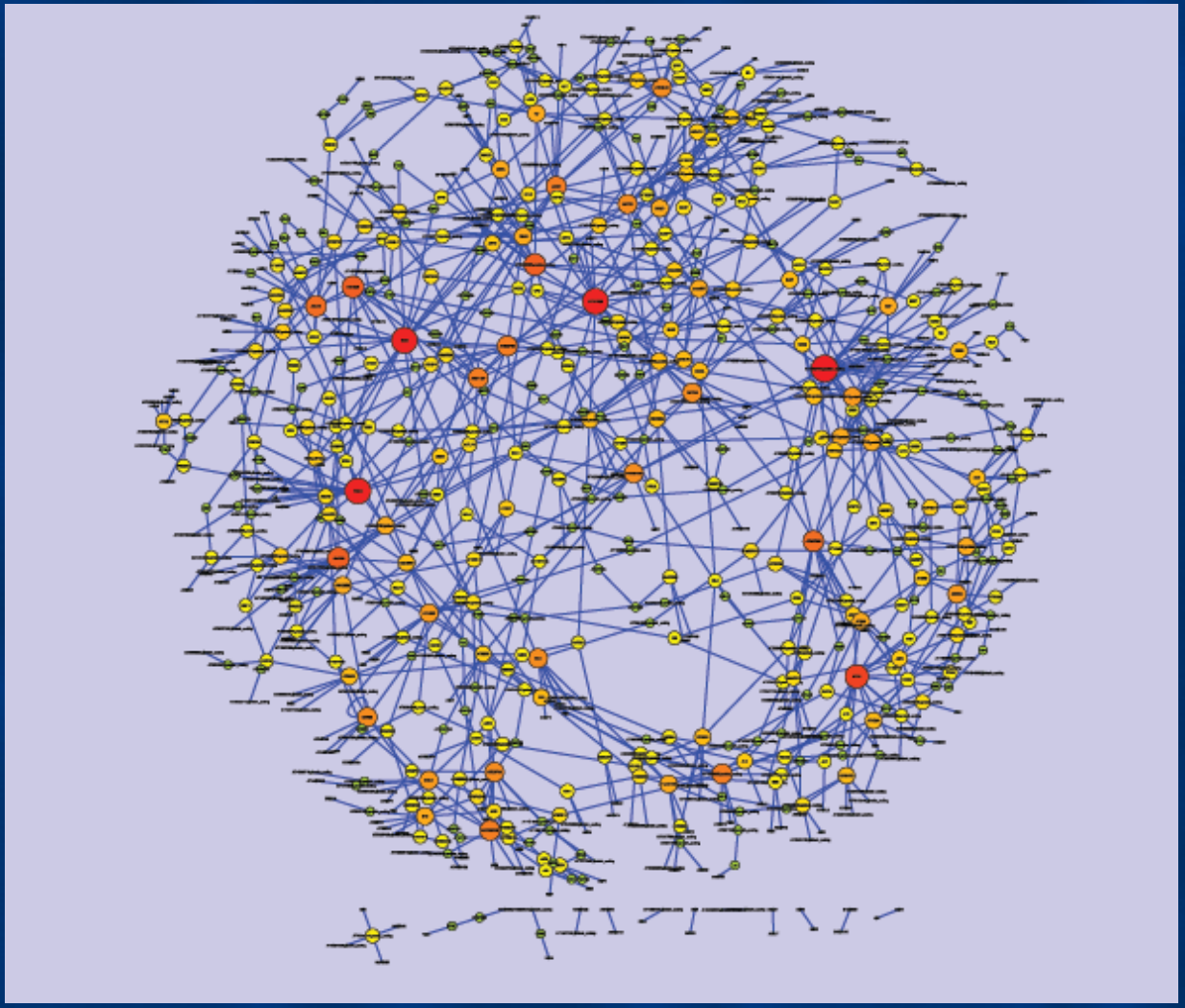




Chavéz Montes, R, et al. (in preparation)

Aplications III

DPI lookup for transcription factor interactions



Chavéz Montes, R, et al. (in preparation)

Network statistics (aplications II & III)

Network features for Transcription factors and Master regulators		
Topological features	**TFs only**	**TFs vs All genes**
Number of nodes	806	14269
Number of links	1076	37276
Number of connected components	13	18
Network diameter	15	12
Network centralization	0.02	0.053
Number of shortest paths	601456	202393380
Percentage of shortest paths	92	99
Characteristic path length	6.523	5.334
Average number of neighbors	2.665	5.224
Network heterogeneity	0.905	3.076
Degree distribution power law ($Y = aX^b$)		
a	765.54	4040.6
b	-2.136	-1.637
Correlation	0.912	0.998
R-Squared	0.925	0.911
Topological coefficients power law ($Y = aX^b$)		
a	1.001	0.813
b	-0.98	-0.742
Correlation	0.991	0.981
R-Squared	0.992	0.929
Shortest path length distribution type	Normal	Bimodal
Singular node observations	Several hubs	AGL92 is by far the strongest hub

Outline

- Motivation
- The gene network inference problem
- The joint probability distribution approach (Guilt by association)
- Information theoretical measures and the data processing inequality (DPI)
- Applications
- Conclusions and perspectives

Conclusions

The use of the DPI allows Network structure assessment in at least 4 ways:

- Identifying and removing indirect transcriptional interactions that may be false positives of the network inference.
- Identifying and highlighting indirect interactions to search for master network regulators.
- Searching for sufficient statistics inference, i.e. minimal networks with potential (although still unreachable) use in the clinic
- If used in a “hard” way (DPI-e = 0) it allows to extract the DAG-structure out of an IT-inferred network to compare with Bayesian-inferred versions

Perspectives

Development of efficient computational methods to calculate DPIs in large networks: Current methods are O[N(N+2)] without bootstrapping and O[N^3] with bootstrapping.

Implementation of DPI methods in other types of networks: biological, technological and social

Extension of the DPI theorem to other IT-measures and to Non-markovian scenarios