Using Bayesian Networks to Analyse Expression Data

Authors

Nir Friendman
Michal Linial
Iftach Nachman
Dana Pe'er

Introduction

- A new framework for discovering interactions between genes
- Based on multiple expression measurements
- Using a Bayesian network to represent statistical dependencies

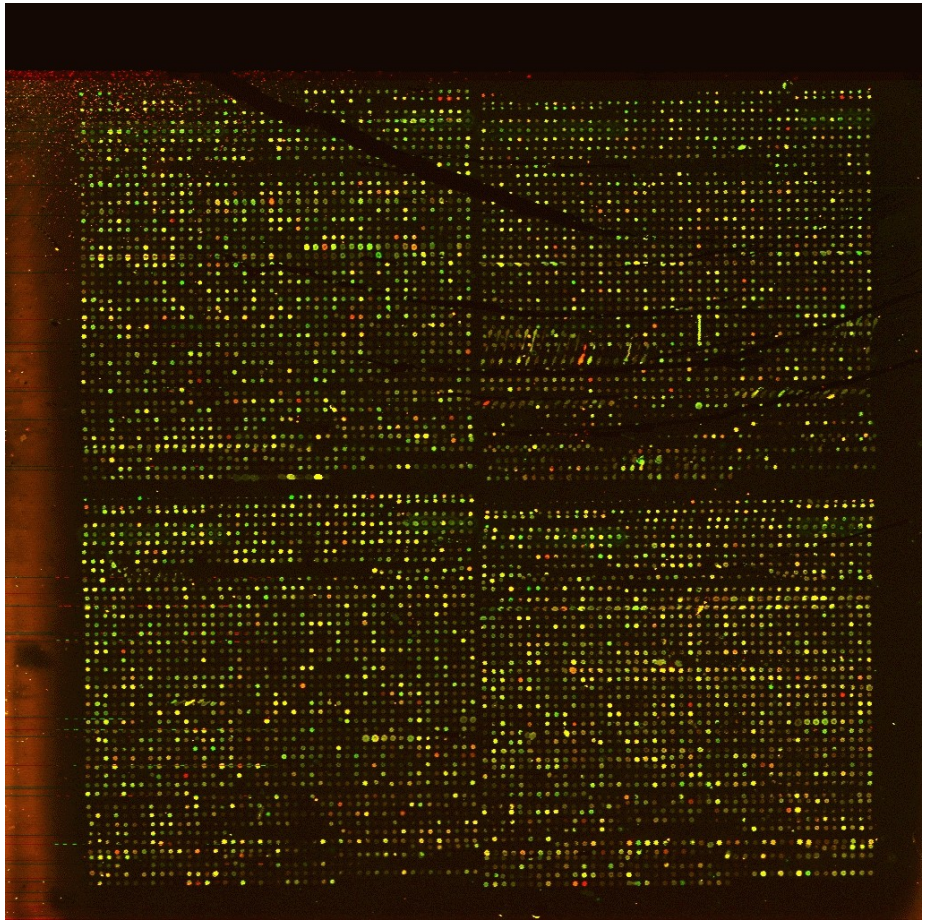
Table of Content

- Biological Background
- Bayesian Networks
- Application of Bayesian Networks to study expression data
- Robustness evaluation of this method
- Results
- Conclusion and further improvements

Biological Background

- Genes expression is responsible for cell activity.
- Protein synthesis is regulated by many mechanisms at its different levels

- Molecular Biology : Understand the regulation of protein synthesis
-Technical breakthroughs lead to development of DNA microarrays



A Machine Learning Challenge

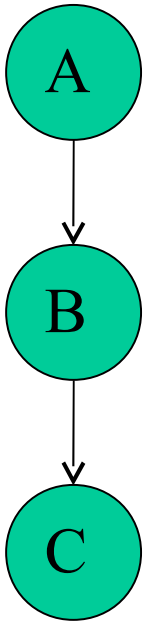
- Analysing microarrays samples to extract biological interactions :
    - Discover co-regulated genes
    - Reveal the structure of the transcriptional regulation system

- Previous attempts using clustering algorithms

Bayesian Networks

- A graphical representation of a probability distribution
- Represent the dependence structure between multiple interactive quantities

- In this basic example :
  - -P(A,B,C)=P(A)P(B|A)P(C|B)
  - - Conditional Independence :
    - -P(A|B,C)=P(A|B)

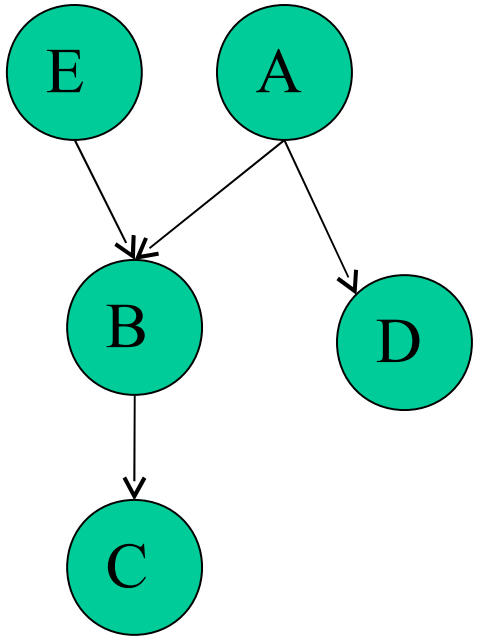


Bayesian Networks : Advantages

- Compact & intuitive representation
- Captures causal relationships
- Efficient model learning
- Deals with noisy data
- Integration of prior knowledge
- Effective inference for experiment planning

An example in context

- P(A,B,C,D,E)=P(A)P(B|A,E)P(C|B)P(D|A)P(E)
- I(A;E)
- I(B;D | A;E)
- I(C;A,D,E | B)
- I(D;B,C,E | A)
- I(E;A, D)



Equivalence classes of Bayesian Networks

- Two different graphs can imply the same set of independencies
    - Example : X→ Y & X← Y
- Two graphs G & G' are equivalent if Ind(G)=Ind(G')
- An equivalence class of network can be uniquely represented by a Partially Directed Acyclic Graph (PDAG)

Learning Bayesian Networks

- Optimization problem
- Given a training set D, find the network B = <G,Θ> that best matches D
- Evaluation of the networks is done using the Bayesian scoring metric
- Score(G:D) = log P(G|D) = log P(D|G)+ log P(G) + C
- Marginal likelihood : $P(D|G) = \int P(D|G,\Theta)P(\Theta|G)d\Theta$
- If G & G' are structure equivalent, they will have the same score

Learning Causal Patterns

- A Bayesian network models dependencies
- We need to model a flow of causality : a causal network
- Its representation is similar to a Bayesian network
- Causal networks not only models the distribution of the observations but also the effects of interventions
- We can learn an equivalence class from the data, and infer some causal directions from the PDAG

Applying Bayesian Networks to Expression Data

- The expression level of each gene is modelled as random variables
- These can include a variety of attributes such as experimental conditions
- Issues
    - -Massive number of variables
    - -Small number of samples
    - -Sparse network (only a small number of genes directly affect one another)

Representing Partial Models

- Not enough data to determine which model is the « right » one
- Pool of reasonable models should be considered
- Extract common features and focus on them
    - Two kind of features
        - Markov relations (is Y in the Markov blanket of X?)
        - Order relations (is X an ancestor of Y)

Statistical Confidence in Features

- We want to estimate a measure of confidence in the features of the learned networks
- An effective and simple approach : Bootstrap method

- For $i = 1 \ldots m$
  - Re-sample with replacement, N instances from D. Denote by D_i the resulting data set.
  - Apply the learning procedure on D_i to induce a network structure $\hat{G}_i$.
- For each feature f of interest calculate confidence$(f) = \frac{1}{m} \sum_{i=1}^{m} f(\hat{G}_i)$, where $f(G)$ is 1 if f is a feature in G, and 0 otherwise.

Sparse Candidate Algorithm

- An optimization problem in the space of directed acyclic graphs
- Complexity of the problem : super-exponential in the number of variables
- The Sparse Candidate Algorithm focuses on small regions of the search space
    - For each gene, we can identify a relatively small number of candidate parents
    - Search space restricted to the networks where the candidate parents are parents of the gene

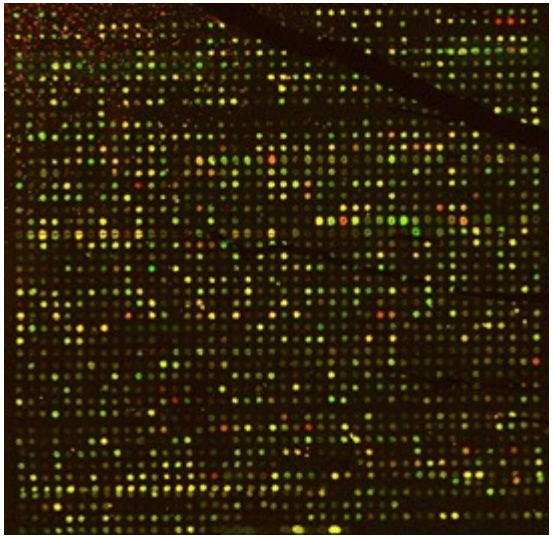
Data preprocessing

- Need to define local probability models
- Gene expression values discretized into three categories : -1, 0, and 1
    - loss of information
    - but reasonably unbiased approach compared to other alternatives such as semiparametric density models

The experimentation

- 800 genes
- 76 gene expression measurements
    - 6 time series
    - this temporal dimension was introduced as an additional variable in the network

- Bayesian networks learned using the Sparse Candidate algorithm with a 200-fold bootstrap (m)

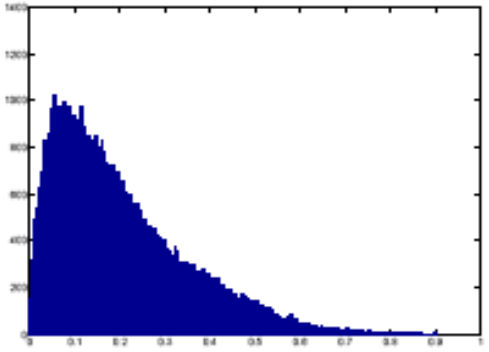


Robustness Evaluation

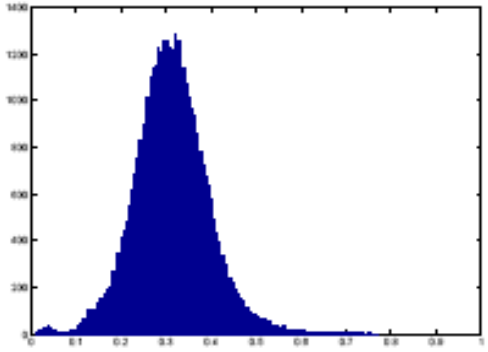
- Analysis of the statistical significance and robustness of the procedure
- Tests done on a smaller gene data set (250)
- Method : Randomize the gene order (random data set)
    - genes independent of each others : hence no expectation to find « real » features

Histograms of the number of Features at different Confidence levels

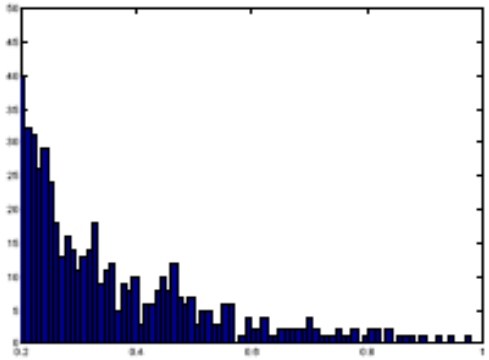
Order
Original set



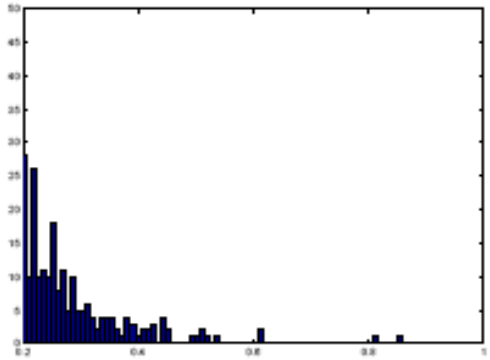
Order
Random set

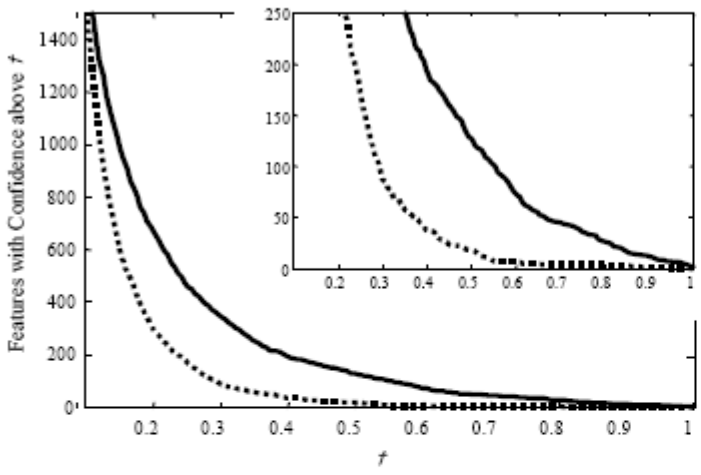
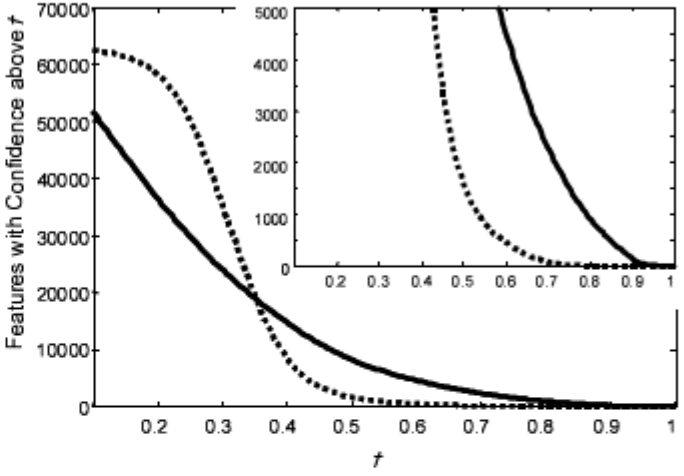
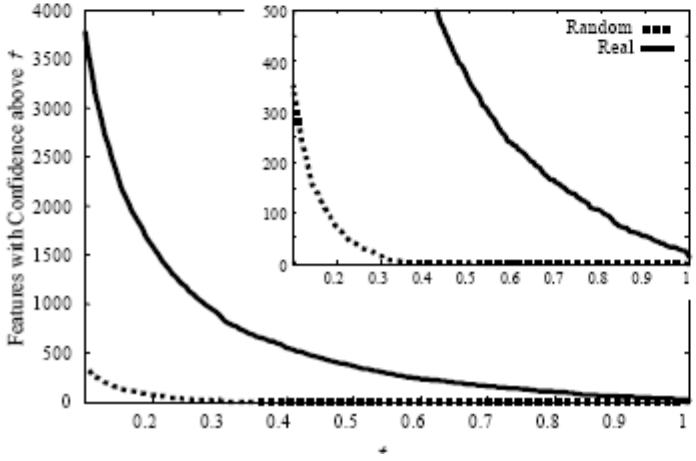
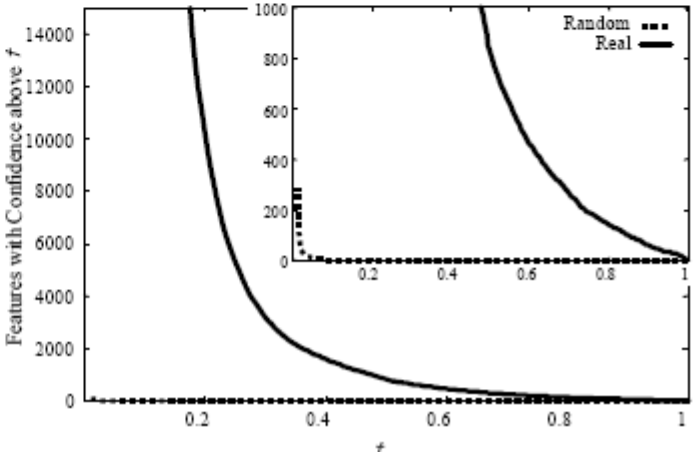


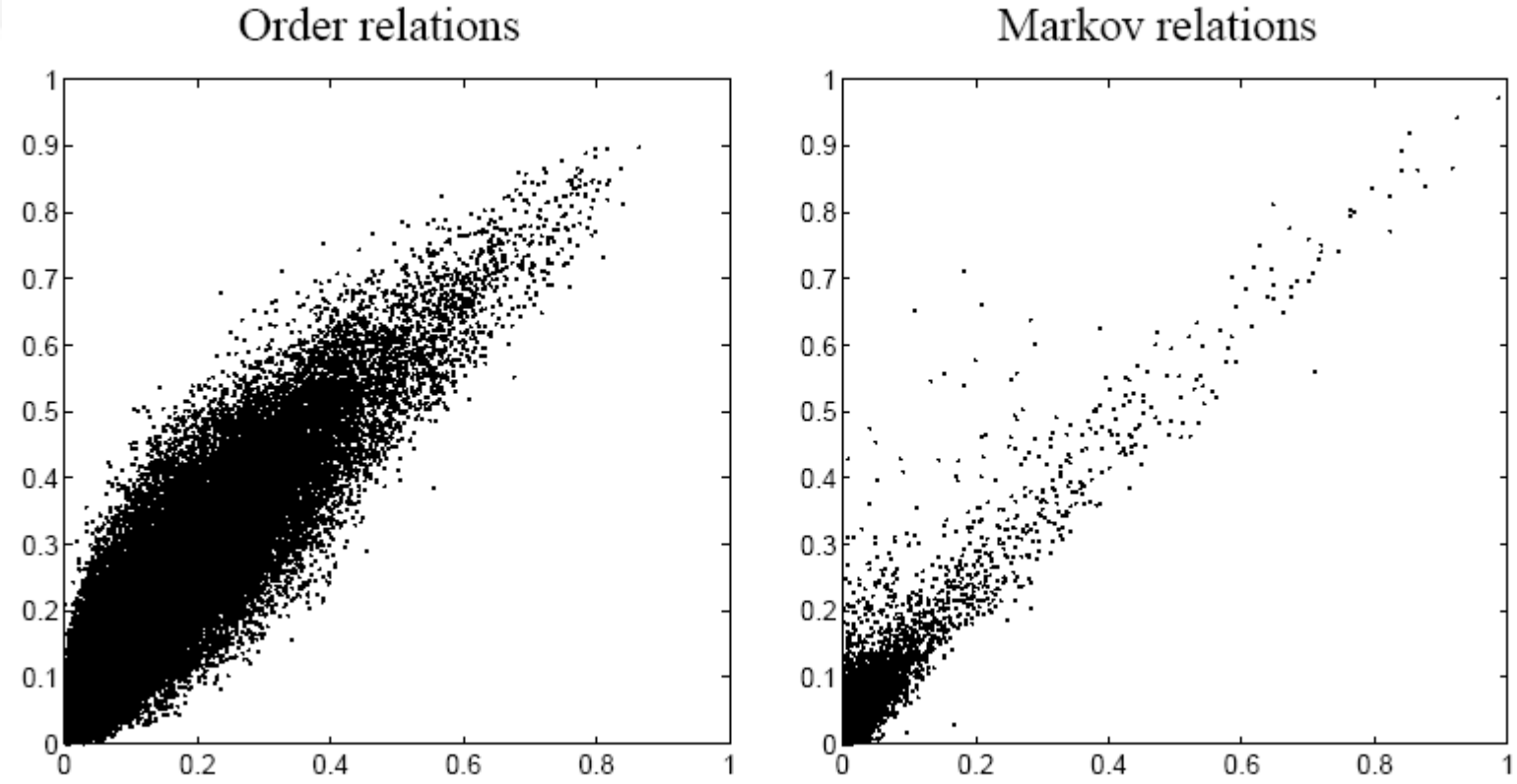
Markov
Original set



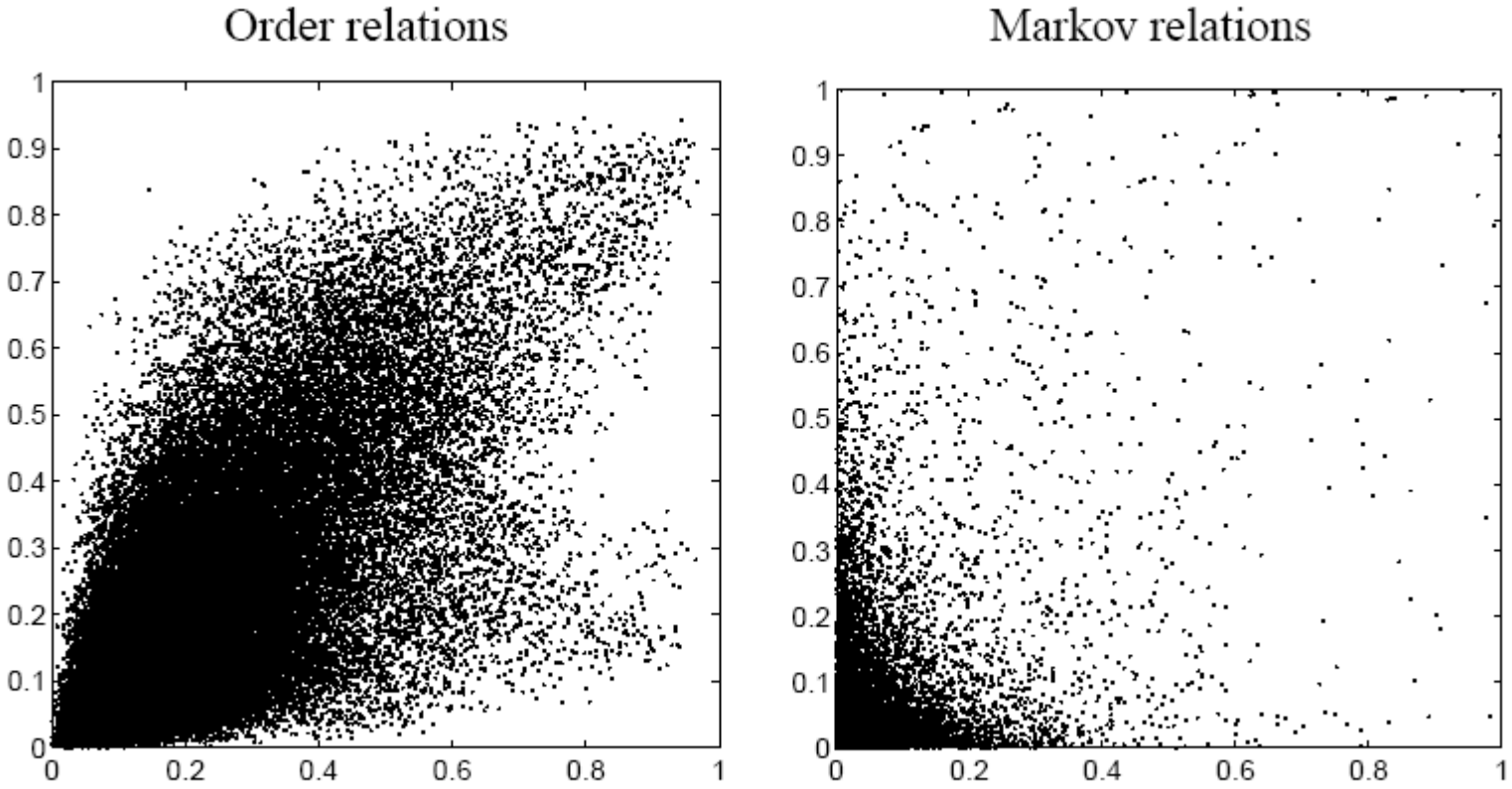
Markov
Random set



	Markov	Order
Multinomial		
Linear-Gaussian		



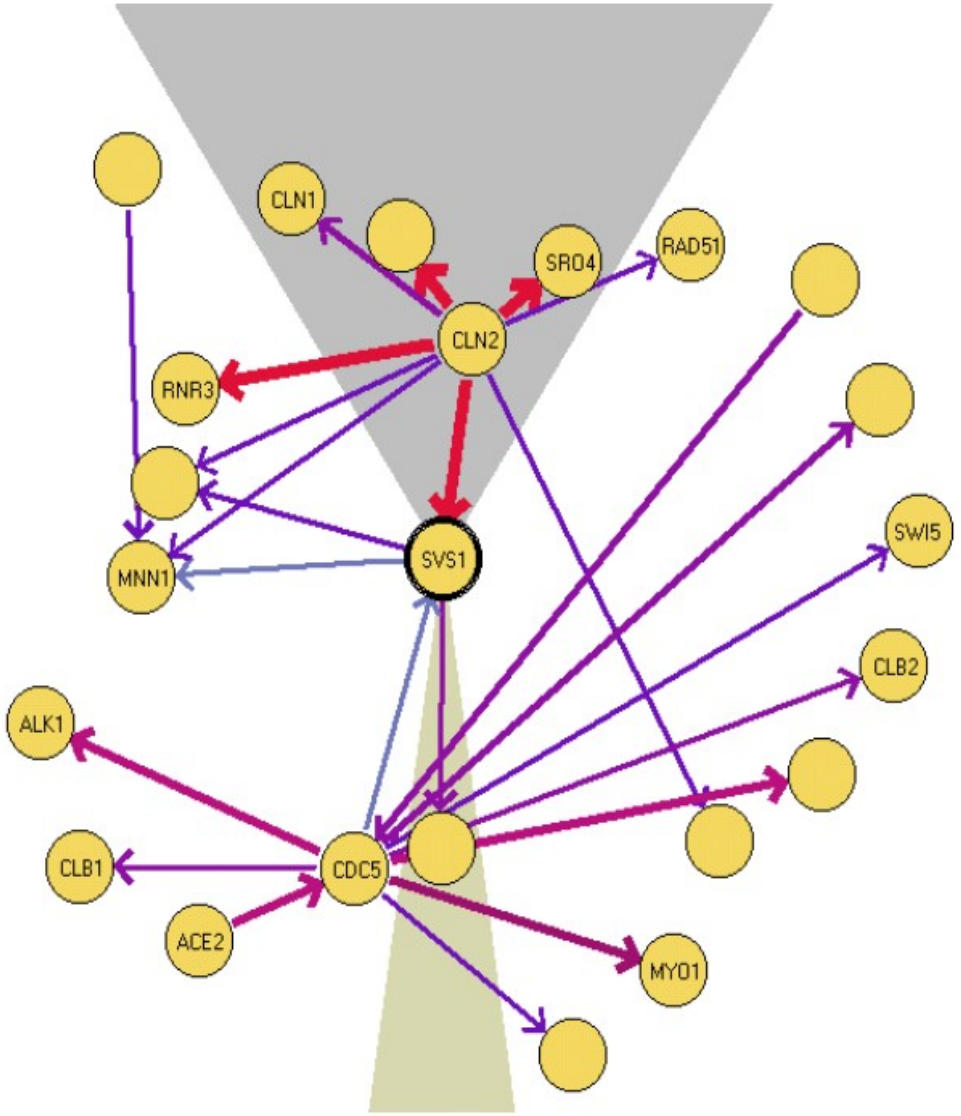
Comparison of confidence levels obtained in two datasets differing in the number of genes, on the multinomial experiment. Each relation is shown as a point, with the x-coordinate being its confidence in the the 250 genes data set and the y-coordinate the confidence in the 800 genes data set. The left figure shows order relation features, and the right figure shows Markov relation features.



Comparison of of confidence levels between the multinomial experiment and the linear-Gaussian experiment. Each relation is shown as a point, with the x-coordinate being its confidence in the multinomial experiment, and the y-coordinate its confidence in the linear-Gaussian experiment. The left figure shows order relation features, and the right figure shows Markov relation features.

Results

An example of the graphical display of Markov features. This graph shows a "local map" for the gene SVS1. The width (and color) of edges corresponds to the computed confidence level. An edge is directed if there is a sufficiently high confidence in the order between the genes connected by the edge.



Results : Order relations

- Dominant genes : with high confidence order relations
  - Could be genes involved in cell-cycle process

Gene/ORF	Dominance Score	# of descendent genes $> .8$	$> .7$	notes
YLR183C	551	609	708	Contains forkheaded assosiated domain, thus possibly nuclear
MCD1	550	599	710	Mitotic chromosome determinant, null mutant is inviable
CLN2	497	495	654	Role in cell cycle START, null mutant exhibits G1 arrest
SRO4	463	405	639	Involved in cellular polarization during budding
RFA2	456	429	617	Involved in nucleotide excision repair, null mutant is inviable
YOL007C	444	367	624	
GAS1	433	382	586	Glycophospholipid surface protein, Null mutant is slow growing
YOX1	400	243	556	Homeodomain protein that binds leu-tRNA gene
YLR013W	398	309	531	
POL30	376	173	520	Required for DNA replication and repair, Null mutant is inviable
RSR1	352	140	461	GTP-binding protein of the ras family involved in bud site selection
CLN1	324	74	404	Role in cell cycle START, null mutant exhibits G1 arrest
YBR089W	298	29	333	
MSH6	284	7	325	Required for mismatch repair in mitosis and meiosis

Results : Markov relations

- Most pairs are functionaly related
- Plus : Make biological sense

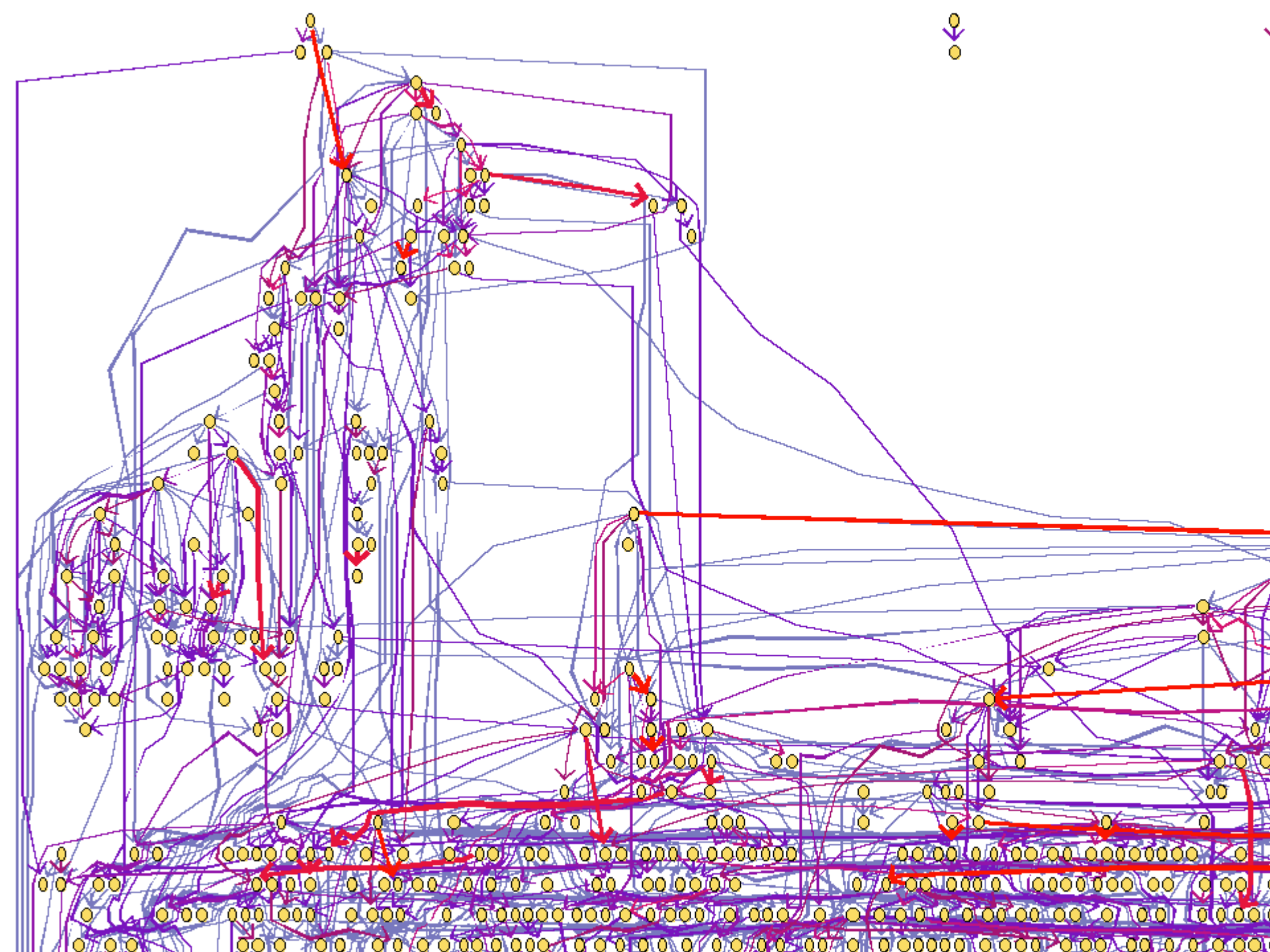
Confidence	Gene 1	Gene 2	notes
1.0	YKL163W-PIR3	YKL164C-PIR1	Close locality on chromosome
0.985	PRY2	YKR012C	No homolog found
0.985	MCD1	MSH6	Both bind to DNA during mitosis
0.98	PHO11	PHO12	Both nearly identical acid phosphatases
0.975	HHT1	HTB1	Both are Histones
0.97	HTB2	HTA1	Both are Histones
0.94	YNL057W	YNL058C	Close locality on chromosome
0.94	YHR143W	CTS1	Homolog to EGT2 cell wall control, both do cytokinesis
0.92	YOR263C	YOR264W	Close locality on chromosome
0.91	YGR086	SIC1	
0.9	FAR1	ASH1	Both part of a mating type switch, **expression uncorelated**
0.89	CLN2	SVS1	Function of SVS1 unknown, possible regulation mediated through SWI6
0.88	YDR033W	NCE2	Homolog to transmembrame proteins, suggesting both involved in protein secretion
0.86	STE2	MFA2	A mating factor and receptor
0.85	HHF1	HHF2	Both are Histones
0.85	MET10	ECM17	Both are sulfite reductases
0.85	CDC9	RAD27	Both participate in Okazaki fragment processing

Conclusion

- New approach for analysing gene expression using bayesian networks
- Make use of the Sparse Search Algorithm and bootstrap method
- Different from earlier clustering approaches : try to learn the structure of the process/data
- It fits well the stochastic nature of biological processes and noisy experiments
- Study of the statistical robustness
- Interesting biological findings without any prior of biological knowledge / constraints in the model

Improvements

- Combine clustering and Bayesian networks approaches
- Improve testing methods to estimate the confidence level
- Incorporate biological knowledge as prior knowledge
- Improve the heuristic search
- Incorporate the temporal dimension of the data (Dynamic Bayesian Networks)



BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 11 2005, pages 2739–2747
doi:10.1093/bioinformatics/bti406

Systems biology

Modularized learning of genetic interaction networks from biological annotations and mRNA expression data

Phil Hyoun Lee[1] and Doheon Lee[2,*]

[1]School of Computing, Queen's University, Canada and [2]Department of BioSystems, KAIST, Korea

Received on November 12, 2004; revised on March 2, 2005; accepted on March 22, 2005

Advance Access publication March 29, 2005

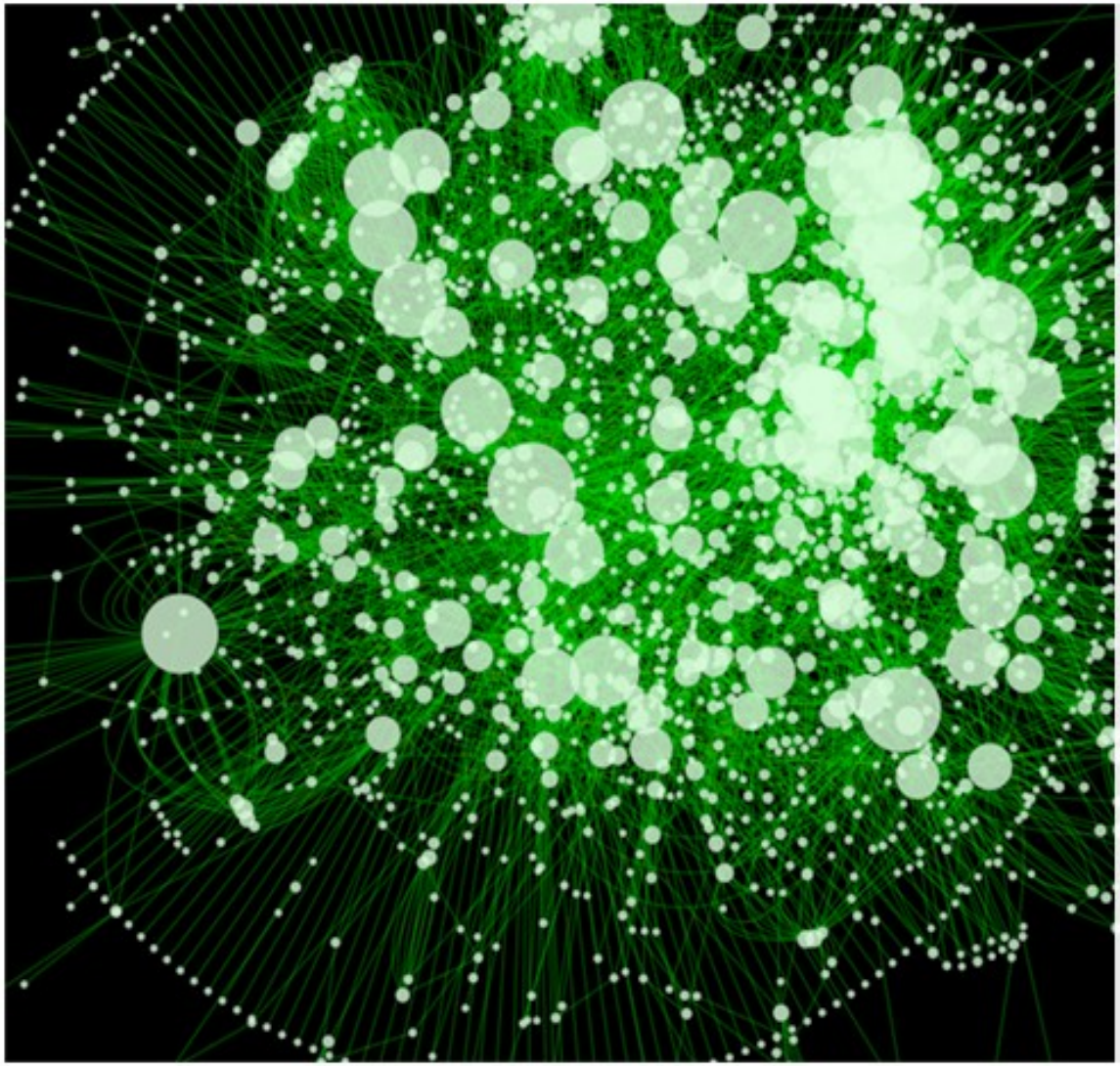
ABSTRACT

Motivation: Inferring the genetic interaction mechanism using Bayesian networks has recently drawn increasing attention due to its well-established theoretical foundation and statistical robustness. However, the relative insufficiency of experiments with respect to the number of genes leads to many false positive inferences.

Results: We propose a novel method to infer genetic networks by alleviating the shortage of available mRNA expression data with prior knowledge. We call the proposed method 'modularized network learning' (MONET). Firstly, the proposed method divides a whole gene set to overlapped modules considering biological annotations and expression data together. Secondly, it infers a Bayesian network for each module, and integrates the learned subnetworks to a global network. An algorithm that measures a similarity between genes based on hierarchy, specificity and multiplicity of biological annotations is presented. The proposed method draws a global picture of inter-module relationships as well as a detailed look of intra-module interactions. We applied the proposed method to analyze *Saccharomyces cerevisiae* stress data, and found several hypotheses to suggest putative functions of unclassified genes. We also com-

among nodes (Neapolitan, 2004). However, it is hard or nearly impossible to secure such sufficient amounts of expression profiles when hundreds or thousands of genes are considered. This shortage of observation data leads to many false positive edges; a significant portion of inferred relationships is not consistent with known biological knowledge. To alleviate this problem, several techniques incorporating statistical biases and prior biological knowledge have been proposed.

Friedman *et al*. (2000) have introduced two statistical techniques, sparse candidates (Friedman *et al*., 1999) and model averaging. The former restricts the maximum number of affecting genes for each target gene so that the search space is reduced. The latter generates multiple networks from different initial conditions, and extracts commonly inferred edges. Other groups have incorporated prior biological knowledge to refine network structures. Hartemink *et al*. (2002) have applied the chromatin immuno-precipitation (CHIP) assay and Tamada *et al*. (2003) incorporated promoter sequence motif information as prior knowledge. They both assumed that relationships between transcription factor genes and their target genes should be supported by other biological clues. Recently, modulariz-



Cytoscape

An Open Source Platform for Network Analysis and Visualization