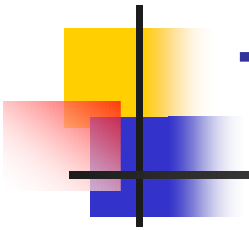
Dynamic Algorithm for Inferring Qualitative Models of Gene Regulatory Networks

Zheng Yun and Kwoh Chee Keong

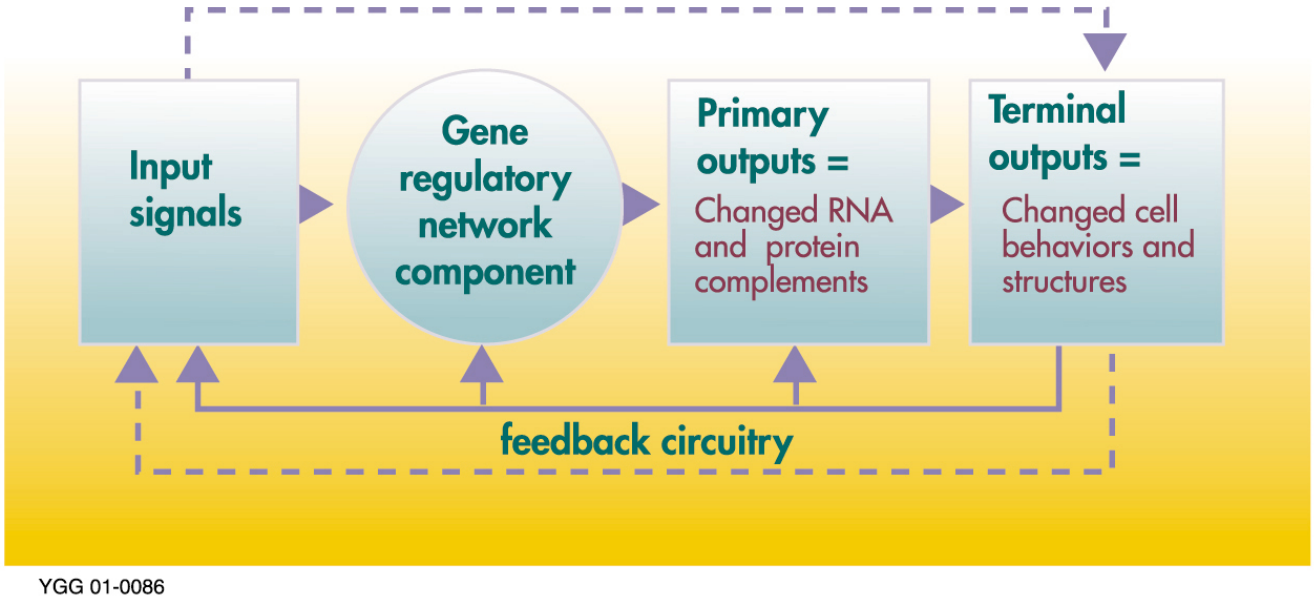
Bioinformatics Research Center
School of Computer Engineering
Nanyang Technological University
Singapore 639798
Tel:+65-67906613, +65-67906057
Fax:+65-63162780, +65-67906559
Email: pg04325488@ntu.edu.sg, asckkwoh@ntu.edu.sg



Topics

- Gene Regulatory Networks (GRNs)
- Reverse engineering methods for reconstructing GRNs
- Information theory foundation for our method
- The DFL algorithm
- Experimental results
- Conclusion
- Summary

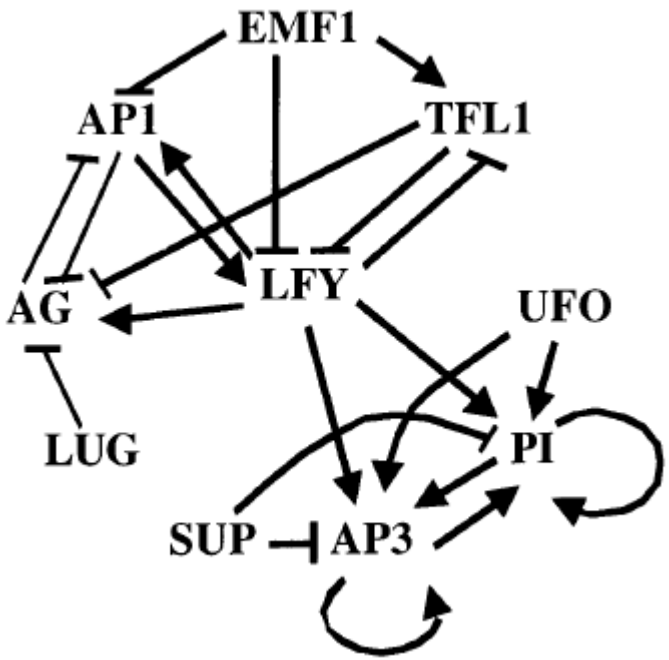
GRNs macro-view



Gene regulatory networks, Courtesy of Genomes to Life Program of U.S. Department of Energy, http://www.doegenomestolife.org.

Input signals are from both intra-cellular and inter-cellular sources. The upper dashed arrow is the signaling responses, which may act directly on cell behaviors and structures. The solid and the dashed arrows at the bottom are direct and indirect feedbacks respectively.

GRN for *Arabidopsis thaliana* flower morphogenesis



Generalized logical model of flower morphogenesis,

Courtesy of Mendoza *et al.* 1999

Experimental data supporting the GRN of *Arabidopsis thaliana* flower morphogenesis

Interactions	Main evidence	Main references
AG —| *AP1*	*AP1* mRNA accumulates uniformly in *ag*-1 mutant flowers	Gustafson-Brown *et al.* (1994)
AP1 —| *AG*	Sepals are replaced by carpels, and petals by stamens in *ap1* mutants. *AG* mRNA found in all the flower primordium of *ap1-1* plants	Bowman *et al.* (1993) Weigel and Meyerowitz (1993)
AP1 —> *LFY*	Reduction of *LFY* mRNA in *ap1 cal* double mutants (independent pathways)	Weigel and Nilsson (1995) Bowman *et al.* (1993) Kempin *et al.* (1995)
AP3/PI —> *AP3/PI*	*AP3* and *PI* mRNA levels are not maintained in *ap3*-3, *pi*-1 or double mutants. Co-immunoprecipitation of AP3 and PI proteins	Goto and Meyerowitz (1994) Jack *et al.* (1992)
EMF1 —| *AP1,LFY*	Inferred by morphological evidence that *EMF1* inhibits the flowering promoting genes	Mendoza and Alvarez-Buylla (1998)
EMF1 —> *TFL1*	Inferred by morphological evidence that *EMF1* activates the late late-flowering genes	Mendoza and Alvarez-Buylla (1998)
LFY —> *AG*	Early expression of *AG* is abnormally low in *lfy*-6 flowers	Weigel and Meyerowitz (1993)
LFY —> *AP1*	*AP1* mRNA delayed in *lfy* mutants. Earlier *AP1* promoter induction in plants overexpressing *LFY*	Weigel and Nilsson (1995) Parcy *et al.* (1998)
LFY —> *AP3*	Amount and domain of *AP3* expression reduced in *lfy*-6 mutants	Weigel and Meyerowitz (1993)
LFY —> *PI*	Amount and domain of *PI* expression reduced in *lfy*-6 mutants	Weigel and Meyerowitz (1993)
LFY —| *TFL1*	Plants overexpressing *LFY* are very similar to *tfl1* mutants	Weigel and Nilsson (1995)
LUG —| *AG*	Ectopic expression of *AG* in *lug*-1 mutants	Liu and Meyerowitz (1995)
SUP —| *AP3*	Ectopic expression of *AP3* in *sup*-1 mutants	Sakai *et al.* (1995)
SUP —| *PI*	Contrary to wild type, *PI* expression is not reduced in the center of *sup*-1 flowers	Goto and Meyerowitz (1994)
TFL1 —| *AG*	Inferred from morphological evidence. Double mutants *ap1-1 ap2-2* have a disrupted C activity, which is rescued with the addition of *tfl1* mutation	Mendoza and Alvarez-Buylla (1998)
TFL1 —| *LFY*	Precocious appearance of floral buds expressing *LFY* in *tfl1*-2 plants	Weigel *et al.* (1992)
UFO —> *AP3*	*AP3* protein and messenger levels reduced in *ufo*-2 plants	Levin and Meyerowitz (1995)
UFO —> *PI*	*PI* mRNA reduced in early stages of flower development in *ufo*-2 plants	Levin and Meyerowitz (1995)

Table Courtesy of Mendoza *et al.* 1999

Reverse engineering?

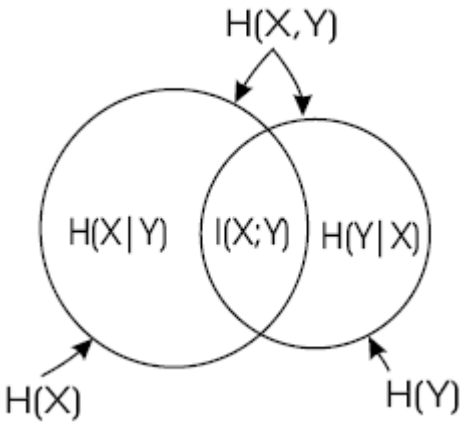
- By mapping the output of each gene to the inputs of other genes, it is possible to reverse engineer developmental circuits and even whole networks.

  Meredith L. Howard and Eric H. Davidson.*Development* 271:109–118. 2004

- If the expression of gene *A* is regulated by proteins *B* and *C*, then *A*'s expression level is a function of the joint activity levels of *B* and *C*.

  Nir Friedman.*SCIENCE* 303:799-805.2004

Information theory foundation of our approach



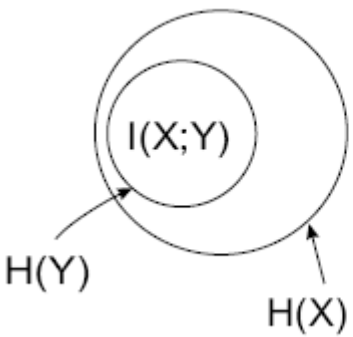
$$H(X) = -\sum_{x} P(X = x) log P(X = x)$$

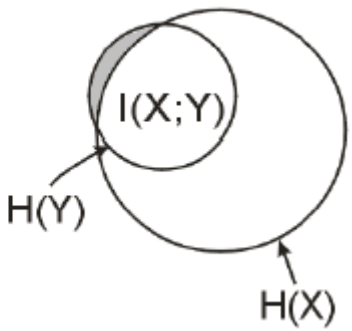
$$H(Y|\mathbf{X}) = -\sum_{\mathbf{x}} \sum_{y} p(\mathbf{x}, y) log p(y|\mathbf{x})$$

$$I(\mathbf{X}; Y) = H(Y) - H(Y|\mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X}|Y)$$

Theorem 2.1 *If the mutual information between* $\mathbf{X}$ *and* Y *is equal to the entropy of* Y*, i.e.,* $I(\mathbf{X}; Y) = H(Y)$*, then* Y *is a function of* $\mathbf{X}$*.*

Definition 5.1 *If* $H(Y) - I(\mathbf{X}; Y) \leq \epsilon \times H(Y)$*, then* $Y = f_{\epsilon}(\mathbf{X})$ *where* ϵ *is a significant factor.*



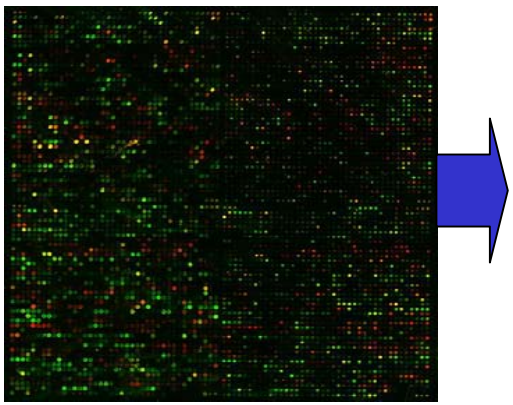


Models under consideration

$$X_i(t+1) = f_i(X_{i1}(t), \ldots, X_{ik}(t)),$$

- Boolean networks: (Liang, Fuhrman. & Somogyi 1998), (Akutsu, Miyano & Kuhara 1999), (Wuensche 1998), etc.
- Generalized Logical Formalism (GLF): (Sanchez & Thieffy 2001), (Thomas & d'Ari 1990), (Thomas, Thieffry & Kaufman 1995), etc.
- Partial Linear Differential Equations (PLDE): (Mendoza, et al. 1999.), (Mestl et al. 1995), (de Jong et al. 2002), etc.

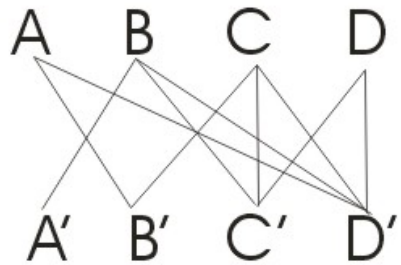
Reverse engineering approach



Gene expression data

A	B	C	D	→	A’	B’	C’	D’
0	0	0	0		0	0	0	0
0	0	0	1		0	0	0	0
0	0	1	0		0	1	0	0
0	0	1	1		0	1	1	1
0	1	0	0	→	1	0	0	0
0	1	0	1		1	0	1	0
0	1	1	0		1	1	1	0
0	1	1	1		1	1	1	1
1	0	0	1		0	1	0	0
1	0	1	0		0	1	0	0
1	0	1	1		0	1	1	1
1	1	0	0		1	1	0	1

State-transition pairs



GRNs

How much data

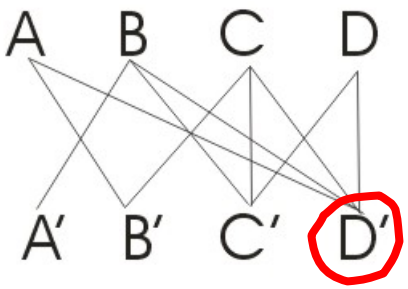
Theorem 3.1 (Akutsu 1998) *$\Omega(2^k + klog_2n)$ transition pairs are necessary in the worst case to identify the Boolean network of maximum indegree $\leq k$.*

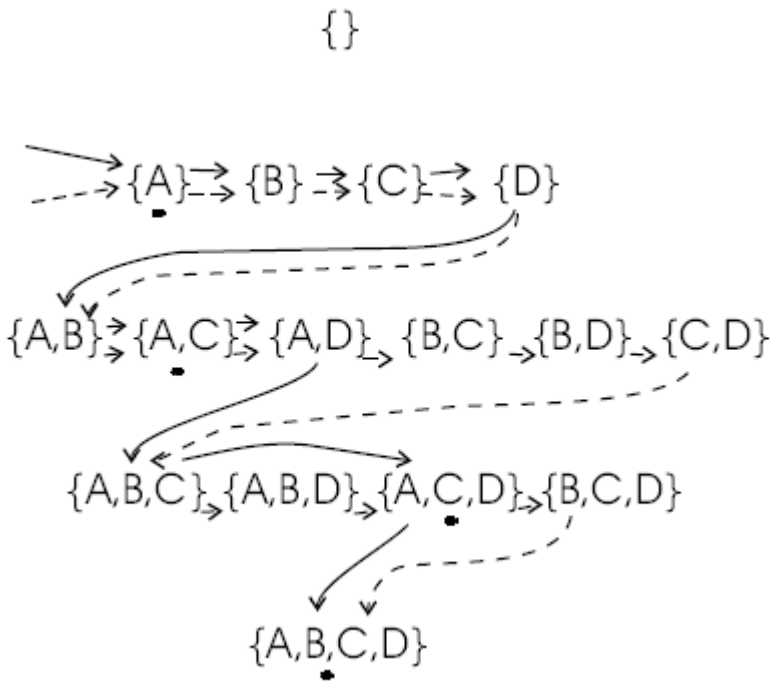
We generalize Theorem by Akutsu 1998 to meet the multilevel datasets.

Theorem 3.2 *$\Omega(b^k + klog_bn)$ transition pairs are necessary in the worst case to identify the qualitative GRN models of maximum indegree $\leq k$ and the maximum number of discrete level for variables $\leq b$.*

$$N = c \times (b^k + klog_bn).$$

The DFL algorithm





Search methods:

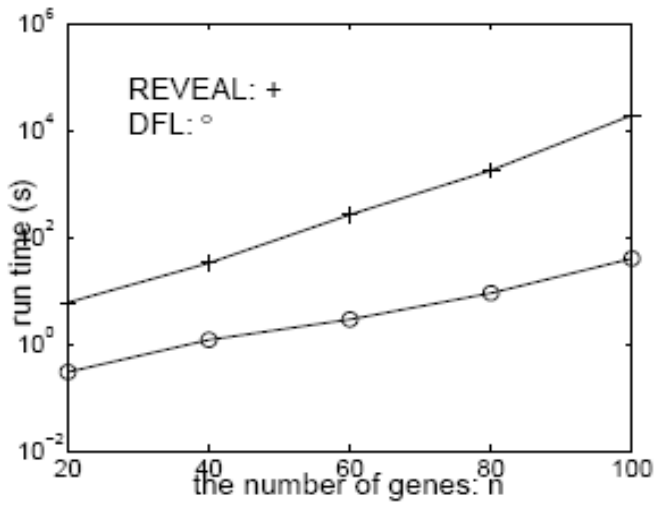
REVEAL(Liang *et al.* 1998): dashed line

$O((b^k + klog_b n)n^{k+1})$

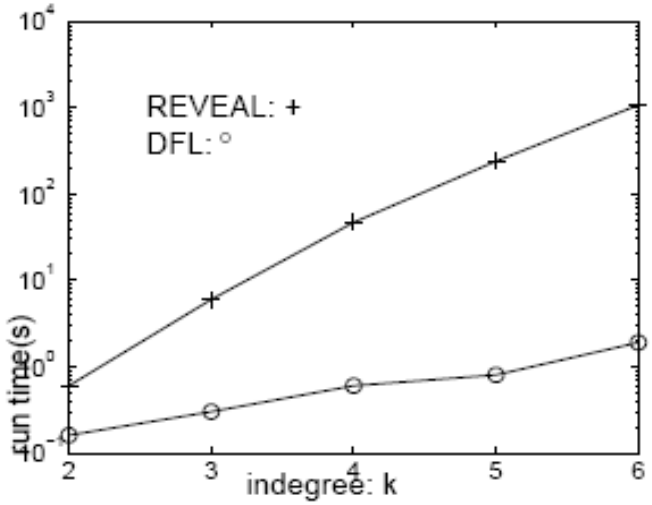
DFL: solid line

$O((kb^k + k^2 log_b n)n^2)$

Experiments for time complexity



(a)



(b)

k = 3, c = 3

n = 20, c = 3

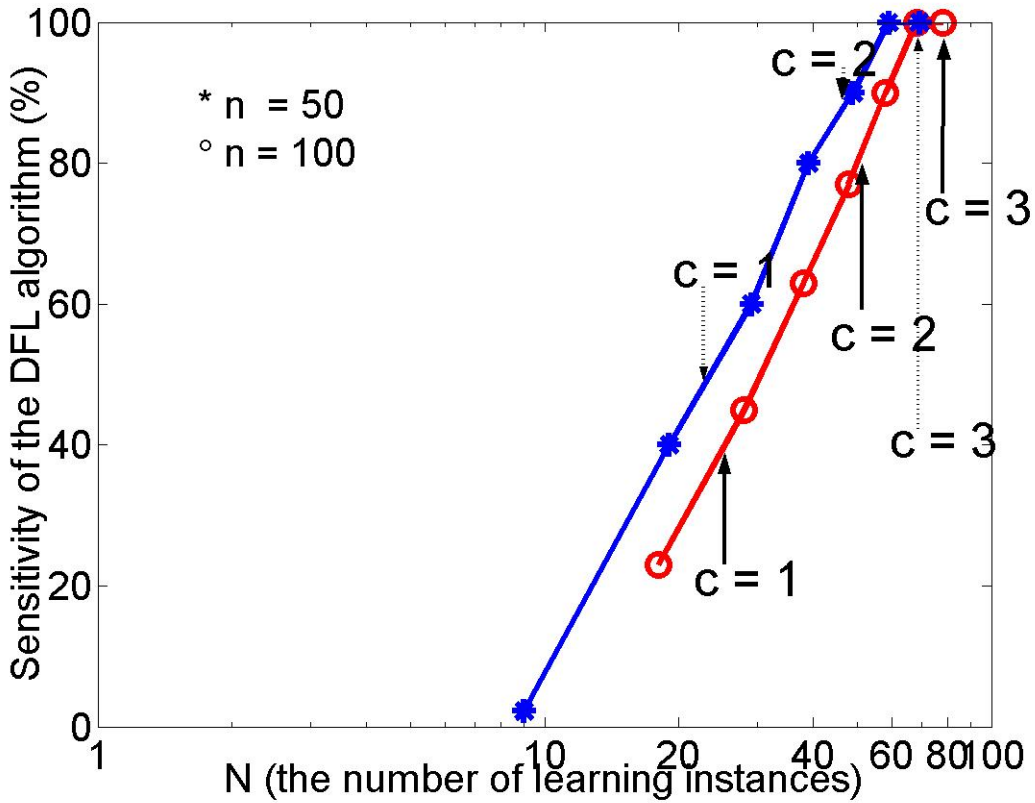
Sensitivity

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

$$\text{Sensitivity}_{\text{wrt positives}} = \frac{\text{No. of correct positive predictions}}{\text{No. of positives}}$$

$$= \frac{TP}{TP + FN}$$

Experiments for sensitivity

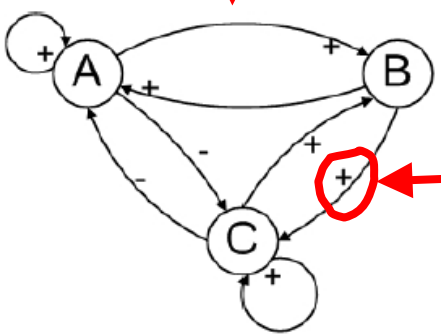


Reconstruct GLF Model

a b c	A B C
0 0 0	0 0 0
0 0 1	0 0 1
0 0 2	0 1 1
0 1 0	1 0 1
0 1 1	0 0 2
0 1 2	0 1 2

A simple example of GLF from (Thieffry and Thomas 1998).

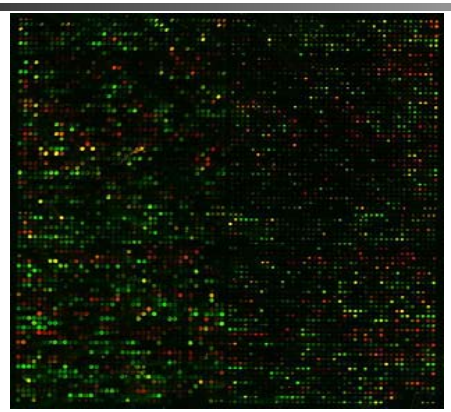
DFL



	A'	B'	C'
A	0.2	0.6	-0.7
B	0.3	0	0.3
C	-0.7	0.6	0.2

Table 4. The correlation coefficient matrix of the GLF example in Figure 6.

Experiments on yeast cycle-cycle gene expression profiles



Cell-cycle expression profiles, from Cho et al. 1998, cover approximately two full cell cycles.

DFL

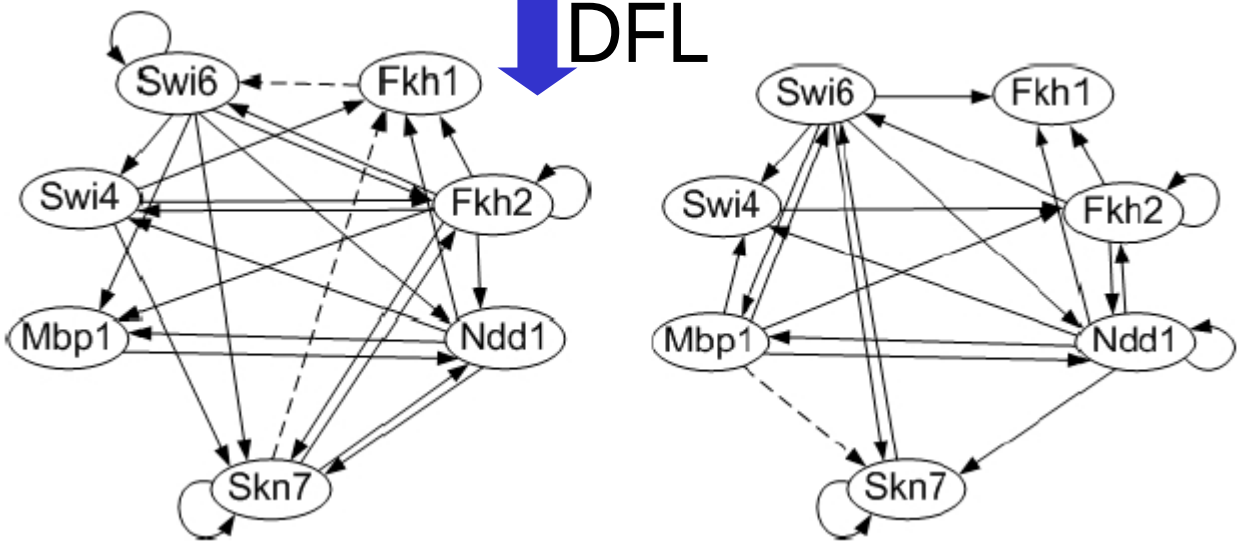


Figure 7. The learned GRN model. (a) The number of discrete levels for gene expression value is 3 and the indegree of the GRN is set to 5. (b) Idem, where the base for gene expression value is 4. The regulators are represented by ovals. The directed edge from Gene A to Gene B means that Gene A is a regulator of Gene B. The solid edges represent regulatory relations that have been verified by other approaches. The dashed edges represent regulatory relations that have not been verified.

Literature Evidences

Gene	Regulator (Protein)							Evidence
	M1	S4	S6	F1	F2	N1	S7	
MBP1	*3	*	*34		*34	*34		[19], [26]
SWI4	*34	*3	*34		*3	*34	*	[19], [26]
SWI6	* 4	*	*34	3	*34	*	* 4	[19], [26]
FKH1	* 4	*3	*4	*	*34	*34	3	[19], [26]
FKH2	* 4	*34	*3	*3	*34	* 4	*3	[19], [26]
NDD1	*34	*	*34	*	*34	* 4	*3	[19], [26]
SKN7	34	*3	*34		*3	*34	*34	[19]

"*" means regulatory relations. For example, "*" in the first cell of first line means that Mbp1 gives MBP1 gene autoregulation [19]. "3" and "4" represent the regulatory relations found with the DFL algorithm when the bases for expression values are 3 and 4 respectively. M1, S4, S6, F1, F2, N1 and S7 are Mbp1, Swi4, Swi6, Fkh1, Fkh2, Ndd1 and Skn7 respectively.

Table 5. The literature evidences for the GRN model in Figure 7 and Figure 8.

The ε-Function method

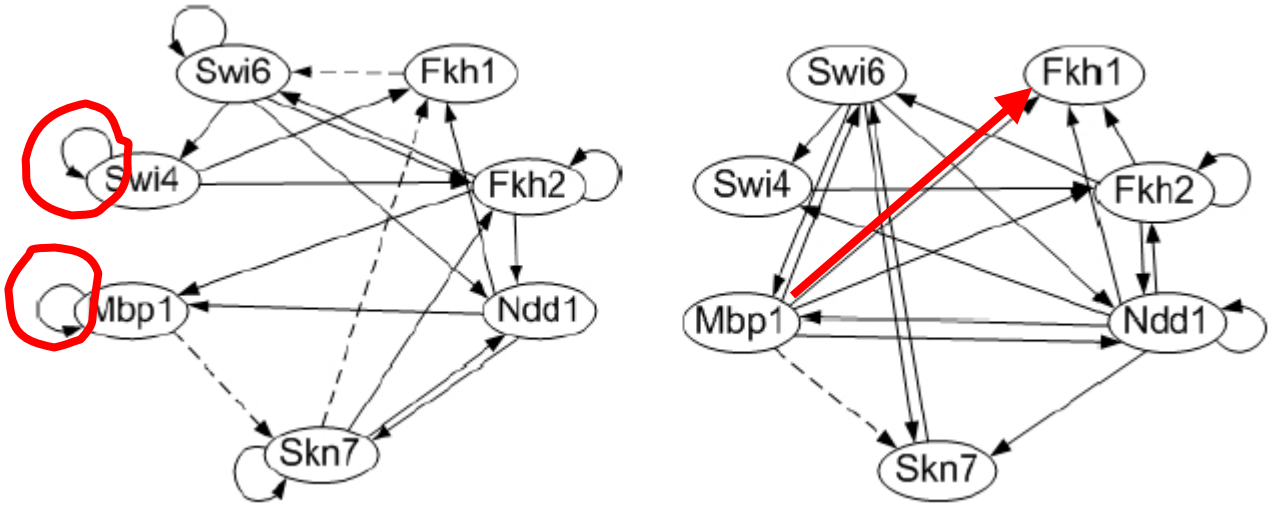


Figure 8. The learned GRN model for yeast cell cycle with the ϵ function method. (a) The base for gene expression value is 3, the indegree of the GRN is 5, and the ϵ is 0.2. (b) The base for gene expression value is 4, the indegree of the GRN is 5, and the ϵ is 0.15. The legends are the same as those of Figure 7.

A comparison with K2 for learning Bayesian networks

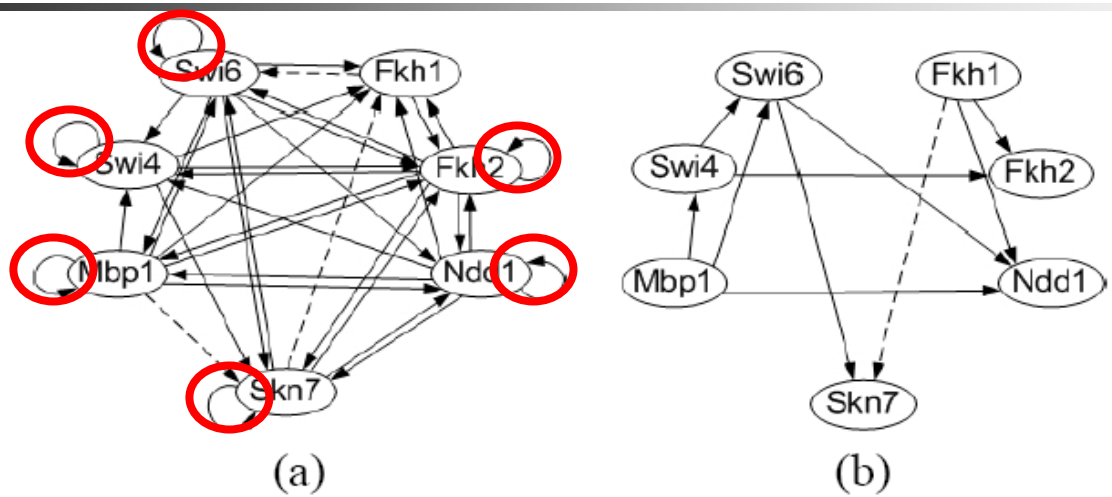


Figure 9. The combined GRN models. (a) Combined model of Figure 7 and Figure 8. (b) Combined Bayesian network structure learned with the K2 algorithm where the base for expression value is set to 3 and 4 respectively. The legends are the same as those of those of Figure 7.

Accuracy and precision

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{No. of predictions}}$$

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

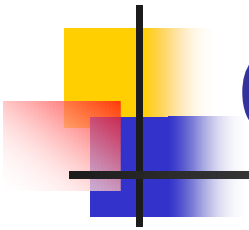
$$\text{Precision}_{\text{wrt positives}} = \frac{\text{No. of correct positive predictions}}{\text{No. of positives predictions}}$$

$$= \frac{TP}{TP + FP}$$

The comparison of prediction performances

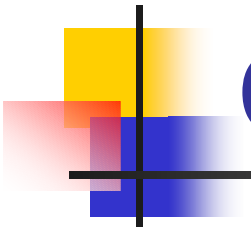
	Accur.	Sensi.	Preci.
DFL ($b = 3$)	65	67	90
DFL ($b = 4$)	63	60	96
DFL (Combined)	80	83	92
K2 ($b = 3$)	27	17	88
K2 ($b = 4$)	22	12	83
K2 (Combined)	33	24	91

Table 6. The accuracy, sensitivity and precision of the DFL algorithm and the K2 algorithm.



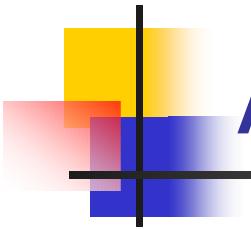
Conclusion

- The DFL algorithm is more efficient than current algorithms for reconstructing qualitative models of GRNs without loss of prediction performances.
- The ε-Function method is a good supplement to the DFL algorithm.
- The DFL algorithm identifies biologically meaningful GRN models from a limit gene expression profile.



Questions and suggestions

Thanks for your interests!



Acknowledgements

We appreciate Prasanna R Kolatkar and Ng See-Kiong for their reviews on an early version of this paper.

We also thank the two anonymous reviews of the paper.