

# Inferring genetic networks via nonlinear models and optimization algorithms

Chih Lee<sup>1</sup>, Chung-Ming Chen<sup>2</sup>, Grace S. Shieh<sup>1,\*</sup>

1. Institute of Statistical Science, Academia Sinica, Taipei City, Taiwan

2. Institute of Biomedical Engineering, National Taiwan University, Taipei City, Taiwan

\*email: gshieh@stat.sinica.edu.tw

## Introduction

Finding the optimal gene regulatory network with  $n$  genes and  $k$  factor genes involves the determination of  $n(n + k + 2)$  unknown parameters in model (1). It is, however, intractable in practice because the computation time increases exponentially with  $n$  and the chance of getting trapped in a local minimum could be high.

$$\frac{\Delta g_i(t+1)}{\Delta t} = \alpha_i \tanh\left(\sum_{j=1}^n w_{ji} g_j(t) + \sum_{k=1}^K w_{ki}^i F_k(t) - \beta_i\right) + \varepsilon_i \quad (1)$$

To circumvent these difficulties, we proposed GASA, a combination of genetic algorithm (GA) and simulated annealing (SA). GASA searches through the structure space of partially-connected networks using GA, while the fitness and parameters for each network are determined by SA.

## Restrictions on network structures

A gene is rarely regulated by all other genes. In fact, a network is usually much sparser than a fully linked one. We therefore set the maximal number of incoming links of each gene to  $max\_l$ , reducing the number of combinations for each gene to  $\binom{n+k}{max\_l}$ . One may further impose a power law, which states that  $P(K=k) \sim k^{-\gamma}$ , where  $1 < \gamma < 3$ .

## Genetic Algorithm

A network structure or chromosome specifies the number of incoming links for each gene. Given a chromosome, we enumerate all the combinations of incoming links for each gene and retain the one with the least  $SSE(g_i)$ . The fitness of a chromosome is computed as negative AIC or BIC.

$$AIC = \sum_i (SSE(g_i) / Var(g_i)) + 2(2n + L)$$

$$BIC = \sum_i (SSE(g_i) / Var(g_i)) + 2 \ln(T)(2n + L)$$

where  $SSE(g_i) = \sum_t (g_i(t) - \hat{g}_i(t))^2$ ,  $L$  is the number of links of the network, and  $Var(g_i)$  is sample variance of  $g_i$  across  $T$  time points.

In each iteration, GA performs crossover and mutation and, at the end, selects half of the chromosomes with the highest fitness from the pool of parent chromosomes and the other half from the pool of child chromosomes. The GA evolves through generations until the best fitness scores of generations converge.

### **Simulated Annealing**

By allowing the system to move into a state of a higher energy, SA inherently can escape from local minimums. We adopted a stochastic gradient descent algorithm (SGD) to obtain the parameters of a given network. Our SGD is basically a modified SA with the steepest gradient (SG) as the search direction. What is important is that when the SG direction successfully lowers the energy, there's still chance that SGD diverts randomly from this direction. Moreover, after a SA process has converged, SGD perturbs the parameter values and starts a new SA process until no further improvement can be made. All these measures are designed to avoid being trapped in local minimums.

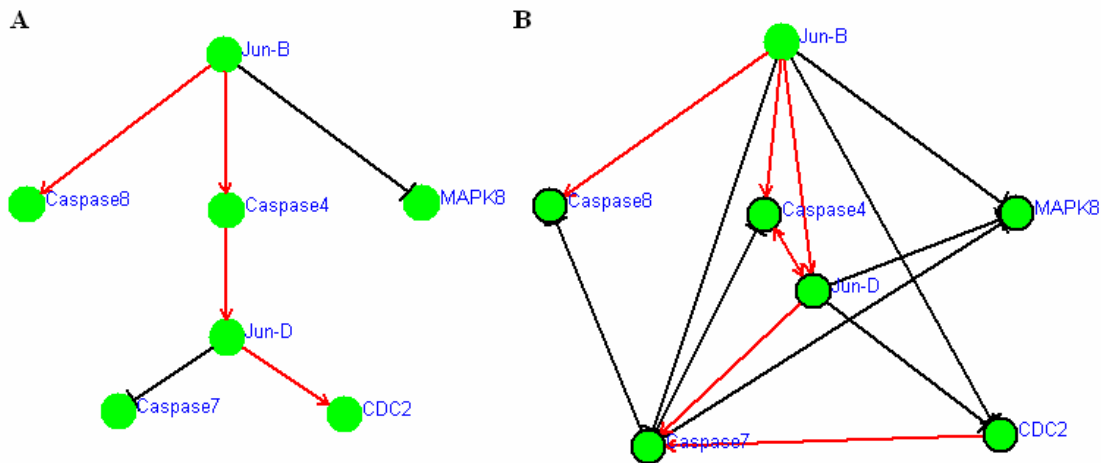
### **Experimental results and preliminary conclusions**

Table 1 and Figure 1 show the performance of GASA on the 44-replicate real data set. The True Positive Rates (TPRs) were computed assuming that the 7-gene network proposed by Beal *et al.* (2005) is the true network. True Negative Rate (TNR) and False Positive Rate (FPR) are not available because not all negative links are verified. Using AIC as the model selection criterion (MSC) and without power law, GASA recovered all the links. All the functions, activation or repression, of the links agree except for the link from *Jun-D* to *CASP7*. However, this link is only supported by 2 / 10 networks in Beal *et al.* (2005) and there's currently no literature evidence showing that *CASP4* activates the expression of *Jun-D* and *Jun-D* in turn represses the expression of *CASP7*. Additionally, the interaction between *Jun-D* and *MAPK8* and self-loops of Caspase 4, 7 and 8 predicted by GASA were found in literature (Boldin *et al.*, 1996; Earnshaw *et al.*, 1999; Riedl *et al.*, 2001; Yazgan and Pfarr, 2002).

Table 1. Performance of GASA applied to data from experiments by Rangel et al. (2004).

Space MSC	With power law			Without power law		
	TPR	TNR	FPR	TPR	TNR	FPR
AIC	0.83	N/A	N/A	1.0	N/A	N/A
BIC	0.83	N/A	N/A	0.83	N/A	N/A

Figure 1. (A) The network proposed by Beal et al. (2005). (B) The network predicted by GASA using AIC as MSC and without the power law restriction. Back circles around genes represent repressive self-loops.



## References

- Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C. and Wild, D.L. (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21, 349-356.
- Boldin, M.P., Goncharov, T.M., Goltsev, Y.V., Wallach, D. (1996). "Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1- and TNF receptor-induced cell death". *Cell* **85**, 803-815.
- Earnshaw, W.C., Martins, L.M. and Kaufmann, S.H. (1999). "Mammalian caspases: structure, activation, substrates, and functions during apoptosis", *Annu. Rev. Biochem.* **68**, 383-424.
- Riedl, S.J., Fuentes-Prior, P., Renatus, M., Kairies, N., Krapp, S., Huber, R., Salvesen, G.S. and Bode, W. (2001). "Structural basis for the activation of human procaspase-7", *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14790-14795.
- Yazgan, O. and Pfarr, C.M. (2002). "Regulation of two JunD isoforms by Jun N-terminal kinases", *J. Biol. Chem.* **277**, 29710-29718.