



## Growing genetic regulatory networks from seed genes

Ronaldo F. Hashimoto<sup>1,2</sup>, Seungchan Kim<sup>3</sup>, Ilya Shmulevich<sup>4</sup>, Wei Zhang<sup>4</sup>, Michael L. Bittner<sup>3</sup> and Edward R. Dougherty<sup>1,4,\*</sup>

<sup>1</sup>Department of Electrical Engineering, Texas A&M University, College Station, TX 77843, USA, <sup>2</sup>Departamento de Ciencia de Computacao, Universidade de Sao Paulo, Sao Paulo, Brazil 05508-090, <sup>3</sup>Translational Genomics Research Institute, Phoenix, AZ 85004, USA and <sup>4</sup>Cancer Genomics Laboratory, Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

Received on April 22, 2003; revised on November 16, 2003; accepted on November 17, 2003  
Advance Access publication February 10, 2004

### ABSTRACT

**Motivation:** A number of models have been proposed for genetic regulatory networks. In principle, a network may contain any number of genes, so long as data are available to make inferences about their relationships. Nevertheless, there are two important reasons why the size of a constructed network should be limited. Computationally and mathematically, it is more feasible to model and simulate a network with a small number of genes. In addition, it is more likely that a small set of genes maintains a specific core regulatory mechanism.

**Results:** Subnetworks are constructed in the context of a directed graph by beginning with a seed consisting of one or more genes believed to participate in a viable subnetwork. Functionalities and regulatory relationships among seed genes may be partially known or they may simply be of interest. Given the seed, we iteratively adjoin new genes in a manner that enhances subnetwork autonomy. The algorithm is applied using both the coefficient of determination and the Boolean-function influence among genes, and it is illustrated using a glioma gene-expression dataset.

**Availability:** Software for the seed-growing algorithm will be available at the website for Probabilistic Boolean Networks: <http://www2.mdanderson.org/app/ilya/PBN/PBN.htm>

**Contact:** e-dougherty@tamu.edu

### 1 INTRODUCTION

The study of genetic networks promises to uncover the mechanisms behind cellular growth, function and failure in disease. An important consideration is the selection of a suitable model class aimed at achieving specific goals of analysis, in view of available data used for inference. Our main goal is to model biological regulation.

Our starting point is the extreme abstraction of regulation features that reduces transcriptional regulatory networks to a set of Boolean functions having few effective inputs

(Kauffman, 1993). Biologists have built up a schematic view of the kinds of molecular interactions involved in transcriptional regulation. These range from the very direct interactions of transcription factors with genomic DNA (Davidson, 2001) and the core and accessory transcriptional machinery (Locker, 2001), through various chains of complex signal transduction within or between cells that ultimately impact the behavior of the highly localized transcription machinery acting on a gene and the local structure of chromatin in the region of a gene's promoter sequence (Harold, 2001). Much of the information available has been developed on the basis of what fails to happen when a certain gene product is not present or is not present in an active form in a particular type of cell. As in classical genetics, this is an excellent method of finding out what process a particular protein participates in, given a particular situation; however, the method is unwieldy for determinations of cooperative action, and usually provides information about only a small portion of a gene product's full repertoire of activity. Further, this kind of study is limited by the ability to quantitatively measure an indicator of the defect. Thus, there are mostly qualitative data about regulation, and, for mammals, this amount is minute when considered in relation to the full complement of genes. This very incomplete level of prior knowledge does not provide rich enough clues to the details of a gene's regulatory interactions to guide our construction of networks from lists of genes found to be different between sets of samples exhibiting complex phenotypic differences. For these reasons, it is useful to consider whether techniques that make very minimal assumptions about the forms of networks can be used to generate skeletal network models that operate in ways resembling what is known of biological mechanics from the application of simple rules to the discretized representations of transcriptional behavior in a variety of settings.

There is also a strong drive to develop ways of approaching vital questions regarding the role of cellular misregulation in the genesis and development of human diseases, such as

\*To whom correspondence should be addressed.

cancer, where it is not feasible to obtain closely spaced observations of the system's states as a particular disease originates and disseminates. What can conceivably be derived is an appreciation of the kinds of transitional rule sets and connection architectures that could produce the varieties of steady states present in a set of observations. As biological networks exhibit extreme reliability and extraordinary determinism in spite of the number of elements involved and the overall plasticity of biological systems, it must be expected that there are simple, powerful underlying regulatory actions that arise from these rules even at the very high level of abstraction from which we must currently work. This can be appreciated by considering the variety of genes that respond to genes such as p53 and *c-myc*, clearly acting at or near the center of the intersection of a large skein of regulatory paths (Levine, 1997; Pelengaris et al., 2002).

There is growing evidence that biological networks (metabolic and genetic) function in what might be called a multiscale manner. This is consistent with the fact that many of these networks exhibit a scale-free topology (Jeong et al., 2000; Ravasz et al., 2002). In the context of genetic networks, this would imply that genes form small groups (or clusters) within each of which the constituent genes have close interactions; some of these clusters form larger 'meta-clusters' that themselves exhibit interactions and this process may continue on several different scales. There have been conceptual (Hartwell et al., 1999) and experimental (Ideker et al., 2001) treatments of this theme in biology.

Owing to these considerations, our aim here is to discover, using gene expression measurements, relatively small subnetworks, out of a large network, whose genes interact significantly. Secondly, we might also wish to find a subnetwork whose genes are not strongly conditioned by genes outside the network. As will be seen, this latter criterion is practically problematic and may indeed be undesirable. We refer to these two criteria collectively as the principle of autonomy. We will proceed by starting with a 'seed' consisting of one or more genes believed to participate in such a subnetwork. Functionalities and regulatory relationships among these genes may be partially known or they may simply be of interest. The size of a constructed network is limited for two reasons. Computationally and mathematically, it is more feasible to model and simulate a genetic regulatory network with a small number of genes, and it is more likely that a small gene set maintains a specific core regulatory mechanism. Given the seed, we will iteratively adjoin new genes so as to enhance subnetwork autonomy.

Since our intention is algorithmic, we wish to remain in a general graph-theoretic context. Thus, we consider a genetic regulatory network to be a directed graph in which each node corresponds to a gene and each edge to a connection, which might be a multivariate dependency or a regulatory relation. This generality makes the seed-growing algorithm applicable for many models.

In this paper, we will focus on probabilistic Boolean networks (PBNs; Shmulevich et al., 2002a,b). PBNs represent an extension of Boolean networks (Kauffman, 1993) in which the functions determining state transitions can vary according to certain selection probabilities. The basic theory is extendable to multi-valued logic and to any finite quantization of gene-expression (Zhou et al., 2003). In applying the algorithm, we will make use of both the coefficient of determination (Dougherty et al., 2000) and the Boolean-function influence (Shmulevich et al., 2002a). We illustrate subnetwork construction using gene-expression data for glioma.

## 2 METHODS

Subnetwork construction depends on a criterion to measure the relation between genes, an objective function involving this criterion to decide which genes should be adjoined to the growing network and an algorithm incorporating the objective function that includes both initialization and stopping conditions.

### 2.1 Strength of connection

Modeling a genetic regulatory network as a directed graph, we denote the strength of connection from a set  $\mathbf{X}$  of genes to a gene  $Y$  by  $\sigma_{\mathbf{X}}(Y)$ . Strength will be determined by either the coefficient of determination or the influence. We define the strength from a set  $\mathbf{X}$  of genes to a target set  $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_m\}$  of genes by

$$\sigma_{\mathbf{X}}(\mathbf{Y}) = \Psi[\sigma_{\mathbf{X}}(Y_1), \sigma_{\mathbf{X}}(Y_2), \dots, \sigma_{\mathbf{X}}(Y_m)],$$

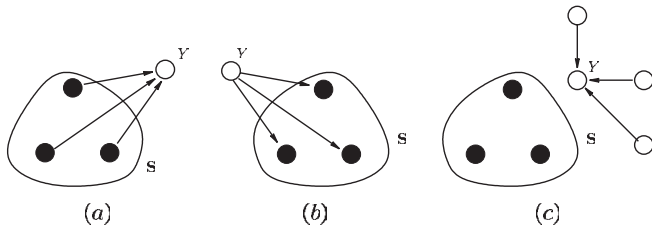
where  $\Psi$  is a function such as the sum, maximum, minimum.

The coefficient of determination measures the degree to which a set of variables improves the prediction of a target variable relative to the best prediction in the absence of any conditioning observations. Let  $Y$  be a target variable,  $\mathbf{X}$  be a set of variables and  $f$  be the function such that  $f(\mathbf{X})$  is the optimal predictor of  $Y$  relative to minimum mean-square error,  $\varepsilon[Y, f(\mathbf{X})]$ . In the binary setting,  $f$  is defined by  $f(\mathbf{x}) = 1$  if  $P(Y = 1 | \mathbf{x}) \geq 0.5$  and  $f(\mathbf{x}) = 0$  if  $P(Y = 1 | \mathbf{x}) < 0.5$ . The coefficient of determination (CoD) for  $Y$  relative to  $\mathbf{X}$  is defined by

$$\theta_{\mathbf{X}}(Y) = \frac{\varepsilon_{\bullet}(Y) - \varepsilon[Y, f(\mathbf{X})]}{\varepsilon_{\bullet}(Y)},$$

where  $\varepsilon_{\bullet}(Y)$  is the error of the best constant estimate of  $Y$  in the absence of any conditional variables. In the binary case, this estimate is the majority value of  $Y$  in the data. The CoD is between 0 and 1. When using the CoD to measure the strength of a connection, for any variable  $Y$  and set  $\mathbf{X}$  of genes, we define  $\sigma_{\mathbf{X}}(Y) = \theta_{\mathbf{X}}(Y)$ .

The influence of a variable relative to a Boolean function for which it is one among several Boolean variables is defined via the partial derivative of a Boolean function. One can define the partial derivative of a Boolean function in a number of equivalent ways (Shmulevich et al., 2002a); however, for our



**Fig. 1.** (a) Sensitivity of gene  $Y$  from the seed  $\mathbf{S}$ . (b) Effect of gene  $Y$  on the seed  $\mathbf{S}$ . (c) Sensitivity of  $Y$  from outside.

purposes here, we simply note that the partial derivative of  $f$  with respect to the variable  $x_j$  is 0 if toggling the value of variable  $x_j$  does not change the value of the function, and it is 1 otherwise. The influence of  $x_j$  on  $f$  is the expectation of the partial derivative with respect to the distribution of the variables. In the context of a PBN, there are a number of predictor functions associated with each gene, and each of these functions has associated with it a selection probability. The influence,  $I_k(x_j)$ , of gene  $x_k$  on gene  $x_j$  is the sum of the influences of gene  $x_k$  on  $x_j$  relative to the family of predictor functions for  $x_j$ , weighted by the selection probabilities for these  $x_j$ -predicting functions. The collection of these influences comprises the PBN influence matrix,  $\Gamma = [I_k(x_j)]$ . When using the influence to measure the strength of a connection, we define  $\sigma_{\mathbf{X}}(Y)$  to be the sum of the influences of the genes in  $\mathbf{X}$  on  $Y$ .

## 2.2 Growing algorithm

In accordance with the principle of autonomy, we would like to grow a subnetwork in a way that enhances a strong collective strength of connections among the genes within the subnetwork. In addition, we might also wish to limit the collective strength of the connections from outside the subnetwork. There are various ways to attach strength measures to achieve these goals. Let  $\mathbf{S}$  be a subnetwork,  $\mathbf{U}$  be the set of all genes under study and  $Y \notin \mathbf{S}$ . We let

$$\sigma_{\text{from},\mathbf{S}}(Y) = \sigma_{\mathbf{S}}(Y)$$

measure the sensitivity of  $Y$  from  $\mathbf{S}$ , i.e. the collective strength of connection from the network  $\mathbf{S}$  to the target  $Y$ . We let

$$\sigma_{\text{to},\mathbf{S}}(Y) = \sigma_{\{Y\} \cup \mathbf{S}}(\mathbf{S})$$

measure the impact of  $Y$  to  $\mathbf{S}$ , i.e. the strength of connection from the target to the network (Fig. 1). Note that the measure  $\sigma_{\text{to},\mathbf{S}}(Y)$  is defined in terms of strength of connection from  $\{Y\} \cup \mathbf{S}$ , not just  $Y$ . This is because we are concerned with the strength of  $Y$  when used in conjunction with other genes in  $\mathbf{S}$ , not just  $Y$  itself. To achieve network autonomy, if  $Y$  is a candidate member of  $\mathbf{S}$ , then  $\sigma_{\text{from},\mathbf{S}}(Y)$  and  $\sigma_{\text{to},\mathbf{S}}(Y)$  should be high.

The measures  $\sigma_{\text{from}}$  and  $\sigma_{\text{to}}$  take into account the connections between the subnetwork and target genes that might be

adjoined; one might also consider reducing the sensitivity of  $Y$  from the outside. This would mean that we want to measure the strength of connection from genes external to  $\mathbf{S}$  to  $Y$ . In this vein, we can define

$$\sigma_{\text{out},\mathbf{S}}(Y) = \max_{\mathbf{X} \subset \mathbf{U} - (\mathbf{S} \cup \{Y\}), \text{card}(\mathbf{X}) \leq m} \sigma_{\mathbf{X}}(Y),$$

where  $m$  is the maximum number of genes allowed in the strength function. Owing to the heavy computational burden posed by  $\sigma_{\text{out}}$ , its use can be impractical for the kinds of large gene sets we are considering. More importantly, its use can be counterproductive because growing the subnetwork may involve iteratively adjoining genes in an information-passing path, so that we wish to adjoin genes strongly sensitive to other genes not yet in the growing subnetwork. One way to employ  $\sigma_{\text{out}}$  is to not use it until subnetwork growth has achieved almost the desirable size. Although, we will not use  $\sigma_{\text{out}}$  in our applications, we include it in our algorithmic discussion because it represents a type of condition one might wish to employ.

To achieve network autonomy, we adjoin a gene  $\hat{Y}$  satisfying

$$\hat{Y} = \arg \max_{Y \notin \mathbf{S}} \Xi [\sigma_{\text{from},\mathbf{S}}(Y), \sigma_{\text{to},\mathbf{S}}(Y), \sigma_{\text{out},\mathbf{S}}(Y)],$$

where  $\Xi$  is a function to return a collective value for three parameters and  $\hat{Y}$  maximizes  $\Xi$ . The selection of this objective function in conjunction with the selection of a strength measure affects gene selection, and therefore the growth of the network. Examples of such functions include:

$$\Xi(x, y, z) = \alpha \cdot x + \beta \cdot y - \gamma \cdot z,$$

$$\Xi(x, y, z) = \max(\alpha \cdot x, \beta \cdot y) - \gamma \cdot z,$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are user-chosen parameters. In the applications shown in this paper, we let  $\Xi(x, y, z)$  be a linear combination with  $\gamma = 0$ .

If  $N$  is the number of genes desired in the subnetwork, then we have the following algorithm to grow the network from the seed  $\mathbf{S}$ :

```

G := S
Repeat
 $\hat{Y} = \arg \max_{Y \in \mathbf{U} - \mathbf{G}} \Xi [\sigma_{\text{from},\mathbf{G}}(Y), \sigma_{\text{to},\mathbf{G}}(Y), \sigma_{\text{out},\mathbf{G}}(Y)]$ 
G := G  $\cup$   $\{\hat{Y}\}$ 
Until  $\text{card}(\mathbf{G}) = N$ .
    
```

The stopping condition,  $\text{card}(\mathbf{G}) = N$ , can be changed to accommodate other interests besides just the number of genes in the subnetwork. For instance, suppose we are using influence and we only want to adjoin a gene if there is a gene among those outside  $\mathbf{G}$  that influences a gene in  $\mathbf{G}$  or is influenced by a gene inside of  $\mathbf{G}$ . To accomplish this end, we use the

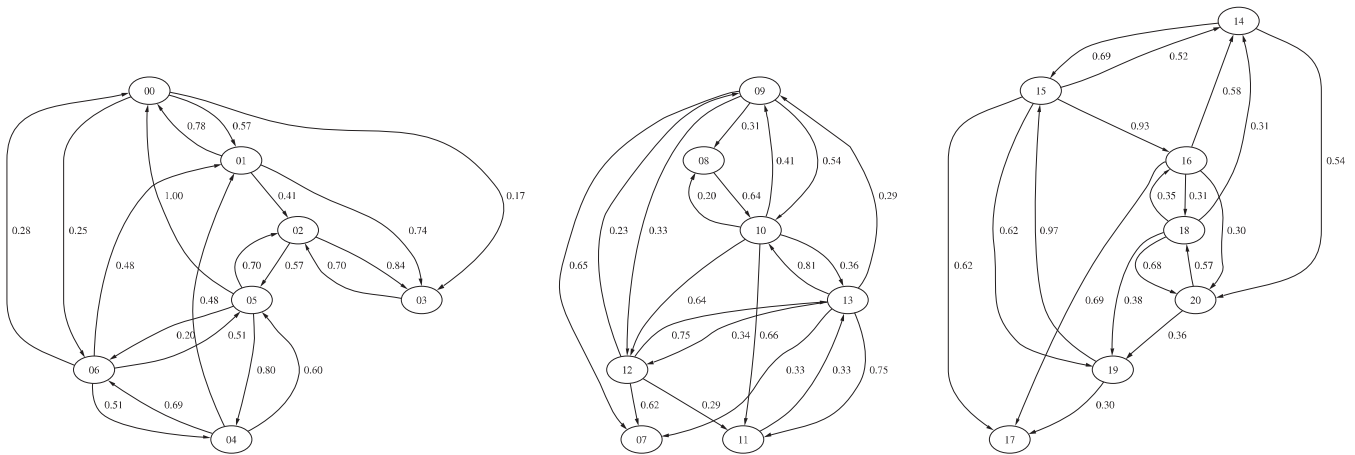


Fig. 2. Artificial influence-based three-component network.

following component-sensitive stopping condition:

$$\text{Until } \max_{Y \in U - G} \max_{X \in G} \{ \sigma_{\{X\}}(Y), \sigma_{\{Y\}}(X) \} = 0 \quad \text{or} \\ \text{card}(G) = N.$$

According to our notation,  $\sigma_{\{X\}}(Y)$  is the influence of  $X$  on  $Y$ . Now, suppose the network consists of several components relative to the influence. Relative to the preceding stopping condition, and assuming the seed is contained in a single component, the growing algorithm stops when  $G$  equals the component containing the seed or  $\text{card}(G) = N$ , whichever comes first.

To illustrate the growing algorithm based on the component-sensitive stopping condition, consider the influence-based three-component network in Figure 2. This network has resulted from an artificial PBN constructed specifically to produce three independent subnetworks. For the objective function  $\Xi(x, y) = x + y$ , if we choose  $N \geq 7$ , then the following sequences show the traces of the algorithm for various single-state seeds in the leftmost component (the seed being listed first):

- 0 → 1 → 6 → 5 → 4 → 2 → 3
- 1 → 0 → 6 → 5 → 4 → 2 → 3
- 2 → 3 → 5 → 4 → 6 → 1 → 0
- 3 → 2 → 5 → 4 → 6 → 1 → 0
- 4 → 5 → 6 → 2 → 3 → 1 → 0
- 5 → 4 → 6 → 2 → 3 → 1 → 0
- 6 → 4 → 5 → 2 → 3 → 1 → 0.

For  $N < 7$ , the algorithm stops earlier in the trace. For instance, if  $N = 4$ , the first trace becomes  $0 \rightarrow 1 \rightarrow 6 \rightarrow 5$ .

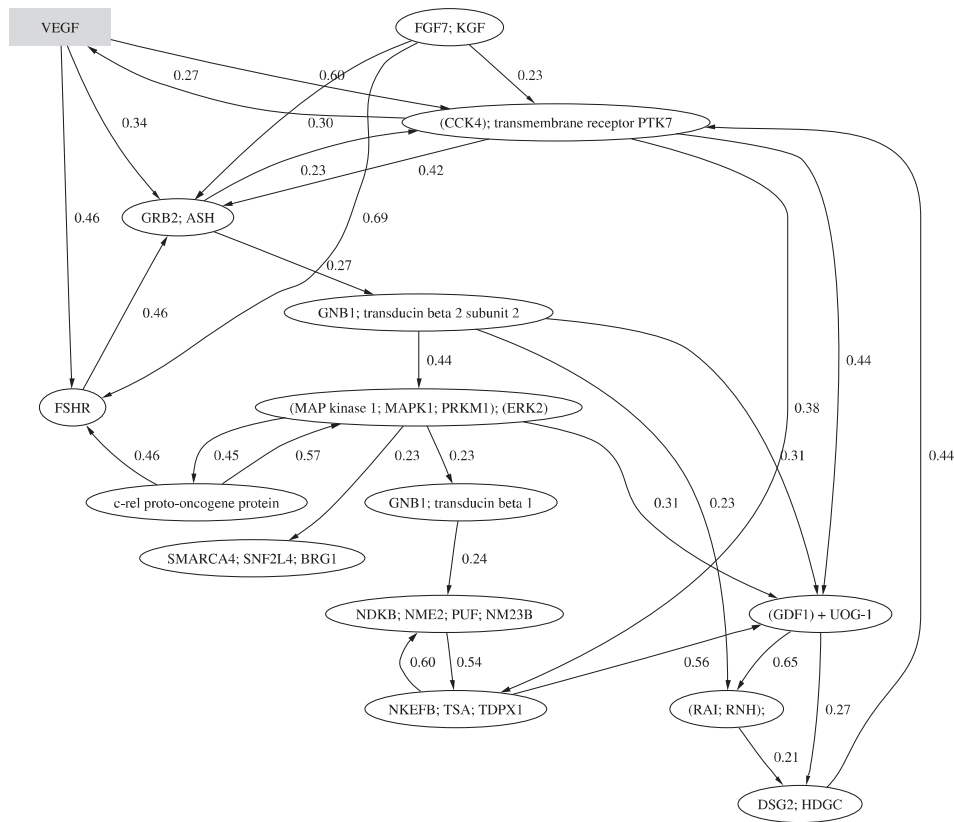
The formal algorithm structure is common to graph searching: initialization and then adjoining nodes in a loop based on maximizing an objective function until a stopping criterion is satisfied. The scientific meaning in the algorithm lies in the

structure of the graph, the initialization, the choice of objective function and the stopping criterion—the latter three being chosen by the scientist to accomplish desired ends. In fact, the formal search structure has been used in the context of genetic Bayesian networks for inferring subnetworks, where owing to the critical role of parental dependencies in Bayesian networks, the objective function is based on a measure of confidence in the Markov-neighbor (parent–child or spousal) relations within a subnetwork (Pe’er *et al.*, 2001). Since our interest is in functional relations, we have defined a flexible objective-function form whose inputs reflect either CoD or influence strength between a candidate gene and the growing network.

Given the objective function and stopping condition, the algorithm output depends on the initialization. The preceding example is encouraging because component integrity is maintained. Generally, the situation is not so clear cut. If the graph consists of a single component, or the stopping condition is not constructed to maintain component integrity, then the situation is more complex. If the seed is chosen in a part of the graph in which there is strong local connection, then the algorithm will produce a subnetwork reflecting that locally strong connectivity, and one can expect stability relative to the seed; however, should a seed be chosen so that it ‘lies between’ two regions of strong local connectivity, then the algorithm may well grow ‘toward’ one of them, and one can expect instability relative to the seed. Such behavior is not uncommon for search algorithms and points to the need for careful seed selection and/or running the algorithm with several different seeds.

### 3 APPLICATION

We have applied the growing algorithm to gene-expression data from studies of melanoma (ternary data) (Bittner *et al.*, 2000) and glioma (binary data), using both the coefficient of determination and influence, and using prior biological



**Fig. 3.** Glioma network grown from VEGF; VPF.

knowledge to choose seeds of interest. Owing to space considerations, here we consider a single glioma network in detail. Specifically, we have applied the growing algorithm to a set of transcriptome data acquired from 25 human glioma tissues of different clinical grades (Kim *et al.*, 2002). Because different genetic and molecular alterations are present in different cancer tissues, especially of different stages during cancer development, we view the 25 transcriptomes as more than one gene regulatory network under various perturbations, which may allow us to better evaluate the relationship between different genes. Although, we emphasize that the gene–gene relationship revealed in this model does not necessarily represent direct regulatory relationship in terms of gene transcription regulation and does not necessarily represent physical interactions of the protein products, this does not mean such relationships do not exist. In this regard, prior knowledge of the genes and gene products provide clues to the true meaning of at least some of the relationships in the network we build. Thus, we selected some of the best functionally characterized genes as seeds to grow the glioma gene regulatory subnetworks.

For the glioma study, the data are binary and the influence is used to define the strength of connectivities in all stages of the growth. When using influence for the growing algorithm,

one has two choices for the influence matrix. First, the influence matrix can be computed for the full network, which will include a preset number of genes. A second method is to compute the influence matrix for the genes under consideration at each stage of the growing algorithm. The second method does not take advantage of the full amount of information we have available, but it has the benefit of computing the influence only among the genes being considered for the network at any stage. It also requires increased computation time. Here, we use the full influence matrix for the full number of genes in the glioma study, which is 597. For the glioma study, the objective function we have chosen to maximize is  $\Xi(x, y) = x + y$ .

Gliomas, like other cancers, are highly angiogenic reflecting the need of cancer tissues for nutrients. To satisfy this need, the expression of the vascular endothelial growth factor (VEGF) gene is often elevated. VEGF protein is secreted outside the cells and then binds to its receptor on the endothelial cells to promote their growth. Blockage of the VEGF pathway has been an intensive research area for cancer therapeutics. The subnetwork grown from VEGF is shown in Figure 3.

Scrutiny of the VEGF network reveals some very interesting insights that are highly consistent with prior biological knowledge derived from biochemical and molecular

biology experiments. From the graph, VEGF, FGF7, FSHR and PTK7 all influence Grb2. FGF7 is a member of broblast growth factor family. FSHR is a follicle-stimulating hormone receptor. PTK7 is another tyrosine kinase receptor. The protein products of all the four genes are part of signal transduction pathways that involve surface tyrosine kinase receptors. Those receptors, when activated, recruit a number of adaptor proteins to relay the signal to downstream molecules. Grb2 is one of the most crucial adaptors that have been identified. We should note that Grb2 is a target for cancer intervention (Wei *et al.*, 2003) because of its link to multiple growth factor signal transduction pathways, including VEGF, EGF, FGF and PDGF. Thus, the gene transcript relationships among the above five genes in the VEGF subnetwork appear to reflect their known or likely functional and physical relationship in cells. Molecular studies reported in the literature have further demonstrated that activation of protein tyrosine kinase receptor-Grb-2 complex in turn activates the ras-MAP kinase-NF $\kappa$ B pathway to complete the signal relay from outside the cells to the nucleus of the cells. Although ras is not present on the VEGF network, a ras family member, GNB2, or transducin beta 2, is directly influenced by Grb2, GNB2 then influences MAP kinase 1 or ERK2, which in turn influences NF $\kappa$ B component *c-rel* (Pearson *et al.*, 2001).

We also observe some potential feedback looping relationships. For example, *c-rel* influences FSHR, which influences Grb2-GNB2-MAPK1 and then to *c-rel* itself. This may be a feedback regulation, a crucial feature of biological regulatory systems in cells to maintain homeostasis. Other feedback regulation may also exist. RAI, or rel-A (another NF $\kappa$ B component) associated inhibitor, is influenced by GNB2, which is two steps away from *c-rel*. RAI is further linked to PTK7 through GDF1, reflecting potentially another feedback regulatory mechanism. Whether those relationships are true negative feedback control mechanisms will need to be validated experimentally in the future. In this regard, the networks built from our models provide valuable theoretical guidance to experiments—and thereby constitute a hypothesis-generating paradigm.

#### 4 SELF-DETERMINACY OF SUBNETWORKS

It is useful to have criteria by which to judge the goodness of a subnetwork as a growing algorithm proceeds; indeed, such a criterion could be used as an alternative halting condition in the growing algorithm. We propose one here involving the CoD, but others are certainly possible. For any gene  $Y$ , subnetwork  $\mathbf{S}$ , and integer  $n > 0$ , define the determinacy of  $Y$  relative to  $\mathbf{S}$  at cardinality  $n$  by

$$\delta_{\mathbf{S},n}(Y) = \max_{\mathbf{X} \subset \mathbf{S}, \text{card}(\mathbf{X}) \leq n} \theta_{\mathbf{X}}(Y).$$

$\delta_{\mathbf{S},n}(Y)$  gives the maximum CoD for predicting  $Y$  via subsets of  $\mathbf{S}$  of size no greater than  $n$ . The self-determinacy of  $\mathbf{S}$  at

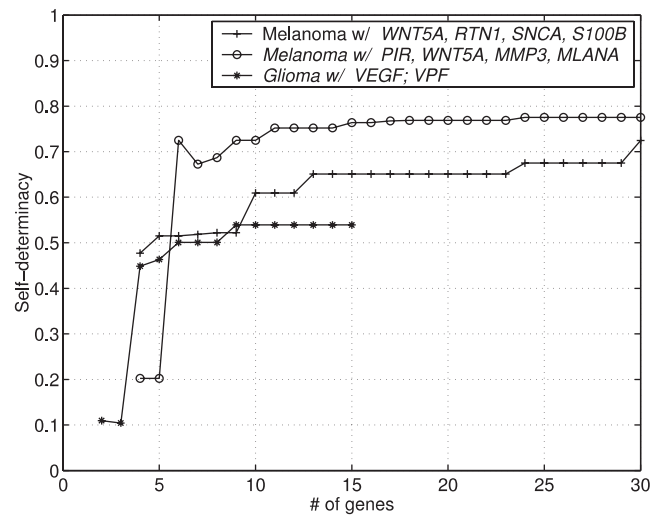


Fig. 4. Self-determinacy for each seed and growth method.

cardinality  $n$  is then defined by

$$\delta_n(\mathbf{S}) = \min_{Y \in \mathbf{S}} \delta_{\mathbf{S}-\{Y\},n}(Y).$$

This is a conservative definition, in that it measures the degree to which the subnetwork is self-determined by the minimum determinacy among all genes in the subnetwork relative to the subnetwork with the gene removed. One might also choose to average the determinacies. In an entirely different direction, one might choose to use a global measure of the degree to which the subnetwork is not determined from the outside. Finally, let us note that the self-determinacy of a network may not necessarily increase as more genes are adjoined to it, because a gene may increase determinacy of the genes that already exist in the network, but the network may not have genes necessary to determine the behavior of the added gene, thereby potentially decreasing the self-determinacy of the augmented network.

Self-determinacies for the two networks grown for melanoma and the network for glioma are shown in Figure 4. As seen from the figure, all three seeds generate the networks with high self-determinacy, especially considering that the self-determinacy is defined conservatively by taking the minimal determinacy within the network. Indeed, when using the average or median instead of the minimum for self-determinacy, they are approximately 0.2 higher (data not shown). Also, since the CoD method is more closely related to the definition of self-determinacy, it is not surprising to observe higher self-determinacy in the networks grown by the CoD method ( $\sim 0.7$ ) for melanoma, than the influence method ( $\sim 0.55$ ) for glioma. However, this does not necessarily indicate that the one network is better than the other.

## 5 CONCLUSION

This paper has proposed an algorithm for growing small genetic-regulatory subnetworks from a seed set of genes. The algorithm is based on the strength of connection between prospective genes to be added and the subnetwork at the current stage of the algorithm. It considers connections directed both in and out of the growing subnetwork. As is typical with a graph search based on an objective function, the algorithm's output will depend on the inputs and parameters of the function—more generally, the structure of the function itself. The objective function embodies the goals of the scientist. The examples have illustrated several possibilities among the goals: basing strength of connection on predictive strength or influence, weighting to give preference to a condition within the objective function and defining a stopping condition to achieve a desired end, in particular, to preserve network components or to achieve self-determinacy. These kinds of choices are natural for search algorithms, and a key intent in the present paper has been to exhibit biologically meaningful alternative forms of a basic algorithm. Ultimately, the choice of objective function and stopping condition depends on biological considerations.

## REFERENCES

- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.
- Davidson, E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.
- Dougherty, E.R., Kim, S. and Chen, Y. (2000) Coefficient of determination in nonlinear signal processing. *Signal Process.*, **80**, 2219–2235.
- Harold, F.M. (2001) *The Way of the Cell: Molecules, Organisms, and the Order of Life*. Oxford University Press, New York.
- Hartwell, L., Hopfeld, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, **402** (Suppl. 6761), C47–C52.
- Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
- Kauffman, S.A. (1993) *The Origins of Order, Self-Organization and Selection in Evolution*. Oxford University Press, New York.
- Kim, S., Dougherty, E.R., Shmulevich, I., Hess, K.R., Hamilton, S.R., Trent, J.M., Fuller, G.N. and Zhang, W. (2002) Identification of combination gene sets for glioma classification. *Mol. Cancer Ther.*, **1**, 1229–1236.
- Levine, A. (1997) p53, the cellular gatekeeper for growth and division. *Cell*, **88**, 323–331.
- Locker, J. (2001) *Transcription Factors*. BIOS. Academic Press, San Diego.
- Pearson, G., English, J.M., White, M.A. and Cobb, M.H. (2001) ERK5 and ERK2 cooperate to regulate NF-kappaB and cell transformation. *J. Biol. Chem.*, **276**, 7927–7931.
- Pe'er, D., Regev, A., Elidan, G. and Friedman, N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17** (Suppl. 1), S215–S224.
- Pelengaris, S., Khan, M. and Evan, G. (2002) *c-myc*: more than just a matter of life and death. *Nat. Rev. Cancer*, **2**, 764–776.
- Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.
- Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (2002a) Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.
- Shmulevich, I., Dougherty, E.R. and Zhang, W. (2002b) Gene perturbation and intervention in probabilistic boolean networks. *Bioinformatics*, **18**, 1319–1331.
- Wei, C.Q., Gao, Y., Lee, K., Guo, R., Li, B., Zhang, M., Yang, D. and Burke, T.R. (2003) Macrocyclization in the design of grb2 SH2 domain-binding ligands exhibiting high potency in whole-cell systems. *J. Med. Chem.*, **46**, 244–254.
- Zhou, X., Wang, X. and Dougherty, E.R. (2003) Construction of genomic networks using mutual-information clustering and reversible-jump markov-chain-monte-carlo predictor design. *Signal Process.*, **83**, 745–761.