Upstream/Downstream Relation Detection of Signaling Molecules using Microarray Data

* Ozgun Babur¹

¹Center for Bioinformatics, Computer Engineering Department, Bilkent University, Ankara 06800, Turkey

ABSTRACT

Motivation: Construction of molecular networks is one of the main challenges in biology. Most of the work in this area is done by manual curation of the data in the literature. Microarray data give information on the expression levels of genes in a certain state of the cell. When compared two microarray experiments, the differential expression of genes is highly affected by the underlying network structure. This fact brings a great motivation for using microarray data to re-construct molecular networks, even there is no demonstrated success of this approach yet.

Results: We developed a new method for detection of the upstream/downstream relation of molecule pairs that may be used in automated or semi-automated construction of molecular networks. We compare microarray experiments and calculate statistics of coordinated and independent behavior of gene pairs. This statistics is later used for estimating possible upstream or downstream relation of new gene pairs in the molecular network.

We demonstrated this method using an annotated set of gene pairs, which is compiled from CSNDB database, and 355 microarray experiments that use the array platform Affymetrix HuGeneFL, which is available in NCBIs Gene Expression Omnibus.

Contact: ozgun@cs.bilkent.edu.tr

Keywords: cellular pathways, microarray data analysis, pathway re-construction

INTRODUCTION

Molecular Networks

One of the best ways to capture data on cellular processes is by molecular networks. This network then can act as a blueprint for simulations and other analysis methods, enabling us to understand and predict the behavior of a cell much better. There are several commonly known databases of cellular network data, each of which is using a different ontology (Ogata et al., 1999; Takai-Igarashi and Kaminuma, 1999; Karp et al., 2002; WIT, 2001; BRITE, 2001; Wingender and Chen, 2001). For instance KEGG (Ogata et al., 1999) database aims to model metabolic pathways (the pathways that are related to utilize metabolites in the cell) and its ontology is enzymatic reaction centric (Figure 1).



Fig. 1. Part of the pyruvate pathway in KEGG. Rectangular nodes represent a specific enzymatic activity which may be possessed by several enzymes, small circles represent the metabolites and other small chemicals, edges represent substrate and product relations between reactions and chemicals.

On the other side, signaling pathway databases such as CSNDB (Cellular Signaling Networks Database) (Takai-Igarashi and Kaminuma, 1999) tries to model signaling events in the cell and its ontology is signaling centric, which identifies actors and signals clearly (Figure 2).

There are two types of edges in CSNDB ontology. *Cellular_Signaling* edges represent protein-protein interactions, while *Gene_Expression* edges represent transcriptional regulation, i.e. one gene product activates expression of another.

Microarray Technology

Microarray is a widely used high-throughput molecular biology technique, which provides researchers with cellwide expression profiles. It is based on detection of RNA molecule concentration in the cell for thousands of genes

^{*}To whom correspondence should be addressed.

[©] Bilkent University Press 2005



Fig. 2. p38MAPK Pathway in CSNDB Database. Circles represent a protein product of certain genes. Activation relations between proteins is shown with edges, also with edge labels redundantly. In this figure BPI -> LSP means "at least one protein state of BPI makes LSP to transform from its inactive state to active state".

with an experiment. Complementary DNA sequences are attached to a glass surface, which each point belogs to a different gene, and the surface (chip) is hybridized with the extracted RNA molecules of the cell being tested. Then the amount of hybridized RNA is measured with lazer scanning, constructing a microarray image whose intensity of the dots correspond to RNA concentrations (Figure 3).



Fig. 3. An image of two channel microarray experiment. The colors and intensity of the dots correspond to differential expression of genes when two type of cells are compared. Red: upregulated, Green: downregulated, Yellow: no change. Single channel experiment images are similar but instead of three colors they are in grayscale, where the intensity of the dot corresponds to the concentration of an RNA molecule in a single type of cell.

There may be one or two signal channels in microarray

data according to its type. When it is two channel, two different cell types are hybridized with a single chip and concentrations are read by measuring the intensities in two different wavelengths, usually red is used for one channel and green is for other channel. Then the colors are mixed according to intensities on the spots, therefore a red spot means relatively higher expression in the first cell type, while yellow spot means equal level of expressions. When one channel microarray data is used, we can compare two experiments to figure out the differential expression.

Pathway Re-construction

Expression patterns in the microarray data is highly affected by the underlying cellular network. However deriving systemic information from expression profiles remains a big challenge. There are a significant number of studies that try to re-construct the molecular network information using the gene expression data. D'haseleer *et al* classifies and reviews these methods, whose common part is that they use some correlation measure between genes and estimate an undirectional link between genes using this correlation (D'haeseleer et al., 2000).

None of the methods enumerated above targets to find *directional* links *from* a molecule *to* another. Fuente *et al* propose an experimenting system for detection of directional and quantitative affects of genes to each other (de la Fuente et al., 2002). They suggest to perform a series of microarray experiments where in each of them just a single gene is less-expressed. So, when compared with the control experiment it would be possible to detect effects of the single gene to others. They suggest to repeat these experiments for each gene, which will collectively lead to a regulatory strength matrix that would be used for network re-construction. However, there are no such experiments performed yet.

In this paper, we propose a new method for inferring the upstream/downstream relation between pairs of genes. We utilize a large number of microarray experiments to calculate co- and independent behavior of the genes, and use this data to classify a given pair. We demonstrate this method using a set of training pairs of genes compiled using CSNDB database, and 355 microarray experiments all of which use the array platform Affymetrix HuGeneFL.

SYSTEM AND METHODS

Types of relations in the cellular network

In the cellular network, molecules affect each other through paths of interactions. When we investigate a pair of molecules A and B, we identify three types of interesting relations (Figure 4).

• Upstream: Molecule A affects molecule B through a path of interactions (Figure 4(a)).

- Downstream: Molecule A is affected by molecule B through a path of interactions (Figure 4(b)).
- Co-affected: Molecule A and molecule B do not have any upstream or downstream relation but they are affected by the activity of another molecule (Figure 4(c)).



Fig. 4. Three interesting types of relations between molecules A and B. (a) Molecule A is at the upstream of molecule B, (b) molecule A is at the downstream of B, (c) molecules A and B do not affect each other but are at the downstream of a common affector. In this figure, edges represent an existing pathway between molecules, not just a direct interaction as was in Figure 2.

In each of these relations we expect to observe a correlation in the differential expression patterns of molecules A and B, i.e. when the upstream molecule changed its expression (up or down) the downstream molecule may be affected and change its expression in the same way. In the third case this correlation is expected because of being at the downstream of the same molecule. Other kinds of relations, e.g. being on the same cycle, are more complex to analyze and out of the scope of this paper.

Correlated and independent behaviors

In this paper we question whether upstream or downstream molecule behaves *more independent* than the other throughout microarray experiments. To rationalize this, consider the first case (Figure 4(c)). In a broader picture we will have other molecules that will affect the relation of molecules A and B (Figure 5). If we imagine two other upstream molecules of B, i.e. X and Y intercepting the A-B path, we see that A may affect B only when some certain conditions are satisfied and the path from A to B has the ability to carry the signal. Molecule A may behave independently when the A-B path is not carrying the signal, and B may behave independently when the A-B path is working but B is affected from other molecules like X or Y.

Another factor that affects observed correlation between A and B is the length of the A-B path. When this path is short, then A and B would correlate more because there would be less number of interceptors, which may decrease



Fig. 5. Two actors that affects the regulation of molecule A on molecule B. In this scenario, the path from A to B works only when the molecules X and Y are also present. Otherwise A cannot affect B.

the significance of independent behaviors. On the other side, when the A-B path is too long the correlation would decrease because of increased number of interceptors and independent behaviors would become more frequent. However, again because of the increased number of interceptors there would be more noise in the statistics, which may obscure the detection of any deviation from random behavior. Therefore we expect an optimum length of the path to give us some useful statistics. The optimum length is related to the observed correlation of molecules, so this correlation may be used to check if it is meaningful to make an estimation.

Taking the statistics

We need a large number of microarray data to calculate statistics, preferably all on the same platform to reduce errors because of systematic differences between arrays. Consider we have N experiments and M genes, and let all data is represented by the matrix X where X(n,m) is the value of the gene m in experiment n where $1 \le n \le N$ and $1 \le m \le M$. Values in matrix X are 0 (molecule absent), 1 (molecule present) and 2 (could not decide).

We are given the molecule pair A-B whose gene indexes are n_a and n_b respectively. Then we check for every pair of microarray experiments m_i and m_j to count the correlated and independent behavior of the gene pair. When we consider the gene pair n_a and n_b , and the experiment pair m_i and m_j one of the 5 types of situations is detected according to four corresponding values in matrix X:

- 1. $X(m_i, n_a)$ or $X(m_i, n_b)$ or $X(m_j, n_a)$ or $X(m_j, n_b)$ is 2 (undecided), so data is not useful [†].
- 2. $X(m_i, n_a) \neq X(m_j, n_a)$ and $X(m_i, n_b) \neq X(m_j, n_b)$ and $X(m_i, n_a) = X(m_i, n_b)$, i.e. there is positive correlation.
- 3. $X(m_i, n_a) \neq X(m_j, n_a)$ and $X(m_i, n_b) \neq X(m_j, n_b)$ and $X(m_i, n_a) \neq X(m_i, n_b)$, i.e. there is negative correlation.

 $^{^\}dagger \text{All}$ other situations require that those four values are either 0 or 1

- 4. $X(m_i, n_a) \neq X(m_j, n_a)$ and $X(m_i, n_b) = X(m_j, n_b)$, i.e. A behaves independent of B.
- 5. $X(m_i, n_a) = X(m_j, n_a)$ and $X(m_i, n_b) \neq X(m_j, n_b)$, i.e. B behaves independent of A.

Count of bad data (case 1) is simply discarded and others are converted to percentages among last four counts. Percentage of positive and negative correlation (cases 2 and 3) is used for deciding whether to make an estimation. We expect one of them to be in a certain range, while the other is just noise. Percentage of independent behaviors (cases 4 and 5) is used for deciding if A is at the upstream of B or vice versa.

RESULTS

We used the interaction data in CSNDB to collect pairs of genes whose upstream/downstream relation is known and have a SwissProt protein ID (a cross-reference to the most popular protein database) that will enable us to map the molecules to microarray data. We figured out all unambiguous upstream relations between molecules (say A and B), i.e. $A \rightarrow B$ is observed but $B \rightarrow A$ is not observed. We prepared a training set of pairs by choosing these unambiguous upstream relations where the path contains at least one *Gene_Expression* interaction, ensuring involvement of gene expression events through the path. The set of upstream pairs become downstream pairs when we switch the order, i.e. $B \leftarrow A$ is a downstream relation if and only if $A \rightarrow B$ is upstream. There happens 1130 upstream pairs that match this criteria in CSNDB.

However, obtaining instances for the third type of relation (co-affected) is not easy as the first two types. This is because of the missing interactions in the database. To get a co-expressed sample we need to guarantee that the pair have no upstream or downstream relation, which is impossible because CSNDB and other molecular network databases lack a great amount of data. For this reason we restrict ourselves to discriminate between only the first two cases.

Microarray data is obtained from NCBIs Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). We selected the platform GPL80, an Affymetrix gene array for human (HuGeneFL) and found 355 experiments that use this platform (from a collection of 15 published datasets).

The statistics calculation is done by considering all $\binom{355}{2}$ pairs of experiments for each training upstream

pair. We do not need to re-calculate statistics for downstream relations because it is just the same when we switch the counts of independent behavior. We sorted the pairs according to their positive correlation percentage across experiments and added an additional column (winner) indicating the most independent gene of the pair (1 indicates the first gene behaves more independent than second, and 0 indicates the vice versa) (Table 1).

gono 1	anno 2	noss oom	nog oonn	ind first	ind second	winner
gene 1	gene 2	poss_corr	neg_corr	mu_m st	mu_seconu	winner
P07339	P29597	0.616	0.010	0.159	0.213	0
P15248	P16860	0.567	0.001	0.146	0.285	0
P46108	P35354	0.454	0.029	0.230	0.285	0
P29597	P10145	0.419	0.032	0.250	0.297	0
P04637	P19838	0.406	0.029	0.345	0.219	1
P29597	P35354	0.402	0.018	0.275	0.302	0
P04637	P99999	0.400	0.017	0.393	0.188	1
P07339	P35354	0.364	0.015	0.262	0.357	0
P07339	P99999	0.359	0.034	0.385	0.221	1
P07339	P19838	0.357	0.041	0.337	0.263	1
P01042	P16860	0.351	0.002	0.563	0.083	1
P07339	P06396	0.349	0.046	0.313	0.291	1
P07339	Q92934	0.345	0.047	0.321	0.285	1
P07339	P10145	0.341	0.042	0.256	0.360	0
P04637	P35354	0.338	0.037	0.288	0.335	0
÷	:	÷	:		:	:
•	:	:	:	:	:	

Table 1. Statistics of gene pairs in descending correlation. gene 1: SwissProt ID of first gene, gene 2: SwissProt ID of second gene, poss_corr: percentage of positive correlation of genes, neg_corr: percentage of nagative correlation of genes, ind_first: percentage of first gene behaving independent, ind_second: percentage of second gene behaving independent, winner: is 1 if first gene is more independent than second, 0 otherwise. This table has 1130 rows which only top 15 are shown here.

In this table we are interested in the unequal distribution of values (1 and 0) in the winner column. In order to determine the probability of dominance of upstream gene independence (first gene being more independent in an upstream sample), we determine the k gene pairs whose positive correlation is nearest to the target correlation. The average of the *winner* value is the target probability approximated by this training set. Similarly we can calculate the same probability for a range of correlation limits, just filtering the gene pairs according to their correlation and calculate average *winner* value.

We plot the dominant independence probability of the first genes across the positive correlation values, where the probability is calculated using the k- nearest neighbor gene pairs [‡] for two different k values (Figure 6). In both of the plots it is possible to differentiate between upstream and downstream genes. With small k values we observe higher fluctuation on the dominant independence probability with the changing correlation, it gets stabilized with increasing k. The worth of points in these plots increase as the deviate from 0.5, which could be used to estimate the status of new gene pairs. In both of the plots

 $^{^{\}ddagger}$ In this k- nearest neighbor calculation, the distance function uses just the positive correlation column, so k- nearest neighbors appear in consequent order on the sorted table.



Fig. 6. The plots of dominant independence probability of the first genes across the positive correlation values calculated with 20 nearest neighbors and 50 nearest neighbors. Each pair in the training sets contribute to a point on the plots.

we observe that in the correlation region from 0.09 to 0.20 there is a consistent deviation of probability. There are 290 gene pairs (more than 1/4 of the whole set) whose positive correlation falls in this region and we calculate 0.69 dominant independence probability for the region.

DISCUSSION

In this paper we definitely showed that the upstream / downstream relation between gene pairs affects the coand independent behavior of genes across microarray experiments. As we expected, some useful statistics appear in the area of certain correlation levels between gene pairs. We observe that microarray experiments do not reveal the differences between upstream and downstream molecules when they are highly correlated or when they have very little correlation. But in about 1/4 of the cases, when the correlation value between genes fall in a certain range, we can expect a dominant independence of upstream genes with 0.69 probability.

This work would be more valuable if we could also have some *co-affected* pairs in our training set. Here what we achieve is to obtain some probability to distinguish between upstream and downstream genes *given that they have an upstream or downstream relation*.

REFERENCES

- BRITE (2001). Biomolecular Relations in Information Transmission and Expression. http://www.genome.ad.jp/brite/.
- de la Fuente, A., P. Brazhnik, and P. Mendes (2002). Linking the genes: inferring quantitative gene networks from microarray data. *TRENDS in Genetics* 18(8), 395–398.
- D'haeseleer, P., S. Liang, and R. Somogyi (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16(8), 707–726.
- Karp, P., M. Riley, M. Saier, I. Paulsen, J. Collado-Vides, S. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro (2002). The EcoCyc database. *Nucleic Acids Research* 30(1), 56–58.

- Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleid Acids Research* 27, 29–34. http://www.genome.ad.jp/kegg/.
- Takai-Igarashi, T. and T. Kaminuma (1999). A pathway finding system for the cell signaling networks database. *In Silico Biology 1*, 129–146.
- Wingender, E. and X. Chen (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Research* 29(1), 281– 283.
- WIT (2001). What Is There? Interactive Metabolic Reconstruction on the Web. http://wit.mcs.anl.gov.