The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 109–118

Transcriptional Regulatory Network Discovery via Information Mining Approach in Rett Syndromes Study

Xiaorong Yang[1,2] Xianwen Ren[2,3] Xiaobo Zhou[2]

[1]School of Statistics & Mathematics, Zhejiang Gongshang University, Hangzhou, 310018, China.
[2]Center for Biotechnology & Informatics, The Methodist Hospital Research Institute
Weill Medical College, Cornell University, Houston, TX 77030, USA.
[3]Academy of Mathematics & Systems Science, Chinese Academy of Science, Beijing, 100190, China.

Abstract Methyl-CpG binding protein 2 (MeCP2) was identified as both activator and repressor in Rett syndrome. Numerous genes in the hypothalamus expressed differentially under the regulation of MeCP2. Experimental results indicated that only a small fraction of genes were directly bound by MeCP2. The transcriptional regulatory networks from the source MeCP2 to its downstream genes were built to reveal the mechanism of molecular processes and signal transduction. Information mining approaches, including database search, literature mining, interaction prediction, and computational inference were integrated to maximize the discovery due to the limitation of prior knowledge of MeCP2. Hub cofactors, whose binding sites were enriched in the MeCP2 activated and repressed groups, comparing with that of the whole genome, were finally identified by using novel linear programming algorithm. The network topology analysis results proved the hub cofactors were important for network structure.

Keywords Rett Syndromes; MeCP2; Transcription Factor; PPI

1 Introduction

Rett syndrome (RTT) is a neurodevelopmental disorder which is caused by mutation in the gene encoding the transcriptional repressor methyl-CpG binding protein 2 (MeCP2). A recent study indicated that MeCP2 regulates the expression of a wide range of genes in the hypothalamus and that it can function as both an activator and a repressor of transcription (see [1]). Gene expression profiles analysis in MECP2-Tg (overexpress the gene MECP2 under the control) and Mecp2-null (knock out the MECP2 under the control) generated the list of MeCP2 activated and repressed targets. However, not all the activated or repressed genes are direct target of MeCP2, and they will be regulated by other cofactors bound to MeCP2. The discovery of transcriptional regulatory network from MeCP2 to its downstream genes advances our better understanding of mechanisms of molecular processes and signal transduction, and is of particular importance in heuristic research of the disease from a nosogenetic perspective.

In the past fewer years, numerous approaches were contributed to the transcriptional regulatory network inference from high-throughput profiles. The output of high dimensional data, like gene expressions, can be regarded as the downstream products of some specific regulatory signals driven through an interacting network. With the growth of vast databases, multiple method integrated methodologies sprung up and potentially provided system level discussion regarding the underlying mechanism. Protein-protein interaction and protein-gene interaction were involved in the network construction due to the consensus that genes in the same pathway or having the similar functions are linked with high possibility in the network; see [1] and [2] for instance.

A major challenge posed in the construction of transcriptional regulatory network is to vigorously maximize the information of the hidden dynamics of the regulatory signals. Although MeCP2 was recognized as a key contributor to neurological disease, limited information is included in the existing databases in regards to transcriptional factors, like TRANSFAC and Genomatix. Therefore, network construction approaches, which strongly depend on prior knowledge of the network elements or need well known network structure, will not take effect. In the study of Chahrour ([3]), more than 2000 genes were identified as being activated by MeCP2 and over 300 genes were repressed MeCP2 targets. Finally, the authors performed quantitative real-time reverse transcription polymerase chain reaction to validate parts of the gene expression changes, and only fewer genes were directly bound by MeCP2. The regulatory signals, like how the rest genes were regulated by MeCP2, and if there exist some key cofactors in regulatory network, attract us for further analysis to explore the regulatory mechanisms.

As mentioned above, the known information regarding MeCP2 is a little. To maximally discover the regulatory mechanisms of MeCP2 and its target genes, an information mining approach is employed in this article. The methodology is an integration of database search, literature mining, interaction prediction, and computational inference. Given the genes that validated as MeCP2 activated targets or repressed targets, candidate cofactors of the transcriptional regulatory network were listed according to the transcription factor binding sites search algorithm. Protein-protein interactions and protein-DNA interactions derived simultaneously from database, literatures and computational predictions were involved in the network building. A linear programming based algorithm was developed to heuristically infer the pathways from MeCP2 to its downstream genes.

The whole flowchart (Fig.1) of the network construction filled in the gap between MeCP2 and its target genes, and as much information as possible was used to reveal the latent dynamics of the regulatory signals. Especially, the incorporation of the direction information, i.e. protein-DNA interactions, well interpreted that how MeCP2 finally regulates the target genes. The linear programming based method used here provided a choice to search the optimal pathways that transmit the maximal information from the source MeCP2 to the target genes. The transcriptional regulatory networks for both the activated MeCP2 targets and the repressed MeCP2 targets exhibit hierarchical structures, which shed light on the selection of the potentially some hub transcription factors in the whole network. The hub transcription factors were inferred based on the results starting from the validated genes, and then enrichment analyses were performed by comparing the binding sites frequencies between the whole bunches of target genes identified by microarrays expression analysis and the whole genome. The significantly lower p-values of the proportion tests indicated that the hub transcription factors were indeed enriched. Most of the cofactors found in the network are brain development related or neuronal function related, and it may guide the therapeutic strategies for clinical studies.

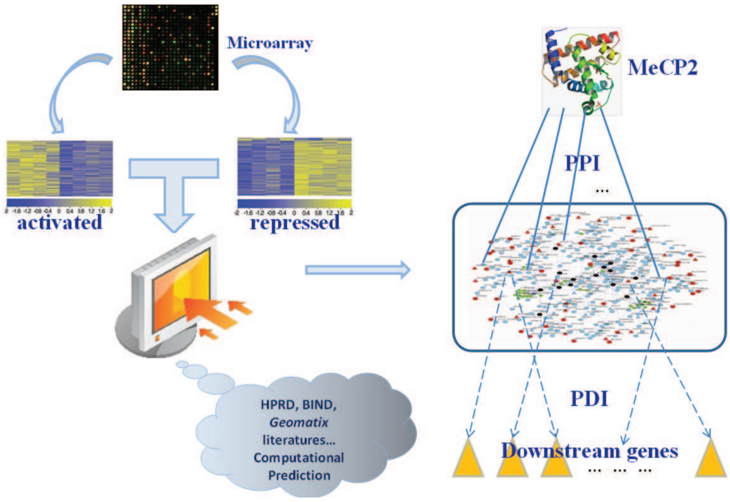

Figure 1: The flowchart of transcriptional regulatory network discovery.

2 Data-set & Primary Results

2.1 The data-set

Microarray analysis using hypothalamic RNA from four MeCP2-Null males, four MeCP2-Tg males, and their respective wild-type (WT) littermates at 6 weeks of age using the Affymetrix Mouse Exon 1.0 ST microarray (see [4] and [5]), were performed. Given that MeCP2 functions as a transcriptional repressor in vitro, 2184 genes were activated by MeCP2 and 377 were repressed. Among the genes with alterations in the expression, 66 genes went to the further validation; of these, 46 genes were activated MeCP2 targets and 20 were repressed targets ($P < 0.05$). ([6]) For the further analysis, the human homolog genes to the selected mouse genes were used since there are very few protein-protein interactions regarding mouse data. The numbers of mouse genes and the corresponding human homolog genes were listed in Table 1.

Table 1: Numbers of mouse genes and the corresponding human homolog genes.

	Mouse Gene	Human Homolog Gene
Activated (All/Validated)	2184/46	1832/42
Repressed (All/Validated)	377/20	284/20
The Whole Genome	17213	14488

2.2 Primary Results

2.2.1 The Generation of Candidate Transcription Factors

The analyses of the two groups of genes, MeCP2 activated and repressed targets, followed the same methodologies. The upstream promoter regions of genes that activated by MeCP2 were significantly enriched in CpG islands ([3]), which guided us to search potential binding sites based on the upstream sequence with length of 5-kb. Transcription factors binding sites were extracted from the database Genomatix MatInd, which employs an alignment algorithm based on the method described by Cartharius et al. ([7]), and creates the nucleotide distribution matrix by counting the bases at each position of the alignment. We restricted the core similarity (the highest conserved position of the matrix) at 1.0, and the matrix similarity was chosen as optimal (thresholds that minimize false positives for each individual matrix are supplied with its library). All the single matrixes were finally mapped to their corresponding transcription factors.

2.2.2 Protein-protein Interactions and Protein-DNA Interactions

Once we generated the list of candidate TFs, the interactions among them were forced to uncover the potential relationships between MeCP2 and other TFs. Interactions from the major human protein-protein interaction database HPRD ([8]) and the database of Genomatix MatBase were collected. For protein-DNA level, interactions were extracted from the database BIND ([9]). However, the direct interactions between MeCP2 and the candidate TFs were very few; only two TFs interact with MeCP2. Directly use the information from database probably narrow the discovery of the true mechanisms of MeCP2; therefore, prediction approach was integrated in the construction.

A lot of works have been devoted to the PPI prediction in the network studies. Most existing methods relied on the information about protein homology or interaction marks of the protein partners, which could not take effects for MeCP2 related prediction since current databases only covered a small fraction of MeCP2 related knowledge. This motivated us to employ a method that only amino acid sequences were requested because it is virtually axiomatic that "sequence specifies structure" might be sufficient to estimate the interacting propensity between two proteins for a specific biological function ([10] and [11]).

The PPI prediction approach used here followed the work of Shen ([11]). In [11], a machine learning method based on a support vector machine (SVM) combined with a kernel function and a conjoint triad feature abstract was developed for the prediction of PPIs based only on the primary sequences of proteins. The web-based tool Sequence-based Protein Partners Search (SPPS) for rapidly predicting potential partners for a given protein sequence is available at [12]. The accuracy of human PPIs prediction can be reached at 83.9% as claimed in their website.

The algorithm returns an interacting probability of each protein pair. To reduce the overfitting problem, we only remained the interacting pairs with probability larger than 0.95. Finally, hundreds of predicted PPIs were used in the further analysis.

2.2.3 Transcriptional Regulatory Network Construction by Linear Programming

The goal of our study is to construct a network which reveals the regulatory mechanism of MeCP2 through its downstream genes. Here we developed a novel computational approach based on linear programming algorithm to build the network. We search the optimal pathways from the source (MeCP2) to its targets (activated genes or repressed genes). This algorithm heuristically guided the identification of biologically meaningful pathways. Each link between two nodes was assigned a weight according to expression data or the protein-DNA interaction frequencies (details were exhibited in METHOD part). Figure 2 and Figure 3 are the two transcriptional regulatory networks of MeCP2 activated group and repressed group, respectively.

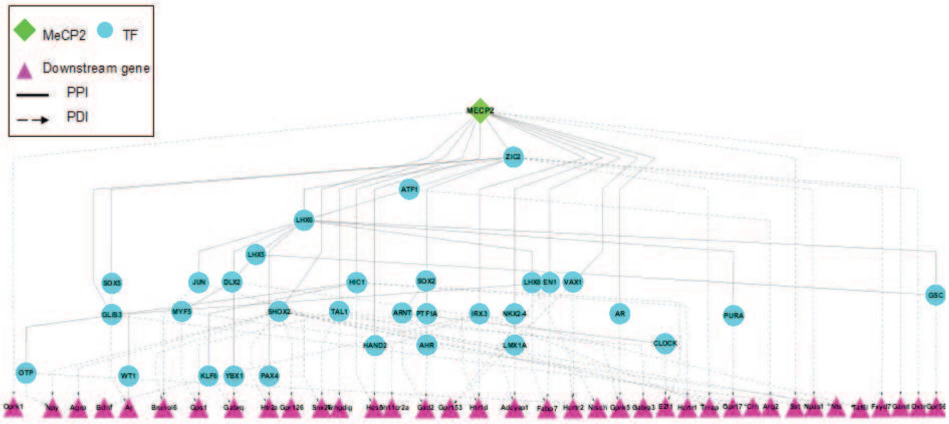

Figure 2: Transcriptional regulatory network for MeCP2 activated targets.

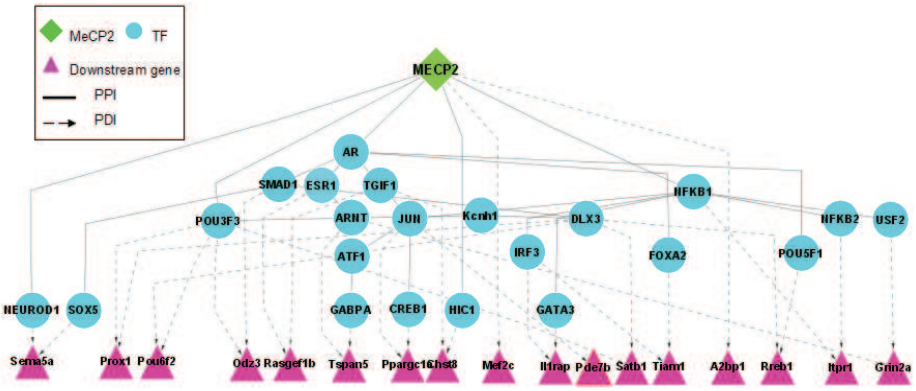

Figure 3: Transcriptional regulatory network for MeCP2 repressed targets.

The above two figures hierarchically displayed how MeCP2 regulates the downstream genes. Some potential cofactors were identified by using our algorithm, and their names and functions were listed in Table 2. More interesting, we found that in the activated group, ZIC2 was recognized as a hub cofactor; while in the repressed group, AR plays the same role. Notice that in [3], the authors mentioned over 2000 genes were activated by MeCP2, and about 300 genes were repressed by MeCP2. We then analyzed the upstream sequences of those genes to check if the binding sites of the hub cofactors were enriched. The frequency of binding sites of each extensive group was generated by Genomatix MatInd. Two group proportion tests were performed based on the following null hypothesis H_0 against alternative hypothesis H_1,

H_0: Comparing with the whole genome, the binding sites are enriched in the activated (repressed) group.

H_1: Comparing with the whole genome, the binding sites are not enriched in the activated (repressed) group.

The comparison results were shown in Table 3. The significantly low p-values the binding sites of hub cofactors were indeed enriched in the activated or repressed group, respectively.

Table 2: Celluar functions of the cofactors in both activated and repressed group.

Cofactors in Activated Group	Function
ZIC2	Brain Development
LHX6	Brain Development
LHX5	Cerebellum Development
IRX3	Regulation of Neuron Differentiation
DLX2	Brain Development
DEUROD1	Neurogenesis
Cofactors in Repressed Group	Function
AR	Androgen Receptor
DLX3	Endocrine System Development
ERS1	Neuroprotection
JUN	Positive Regulation of Smooth Muscle Cell Proliferation
SMAD1	Positive Regulation of Osteoblast Differentiation
POU3F3	Nervous System Development
SOX5	Cartilage Development

Table 3: Binding sites enrichment analysis of the two hub cofactors.

	ZIC2	AR
Extensive Activated	92%	63%
Extensive Repressed	73%	80%
The Whole Genome	79%	69%
p-value	3.36e-40	3.59e-5%

3 Methods

To infer the pathway from MeCP2 to its downstream genes, we develop a novel algorithm based on linear programming algorithm to search the optimal paths in the network when given the source s (MeCP2) and targets t (downstream genes). The major contribution is that the information flow from the source to the target was considered. We built the model to set the balance condition that can ensure that the amount of the input flow equals the amount of the output flow at each node, except the source node and the target node.

3.1 The Linear Programming

For each pair of nodes (i, j) in the network, we define the four variables O_{ij}, O_{ji}, I_{ij} and I_{ji} as following

O_{ij}, the output flow at i from i to j,

O_{ji}, the output flow at j from j to i,

I_{ij}, the input flow at j from i to j,

I_{ji}, the input flow at i from j to i.

Let $E(i,j)$ denote the edge between nodes i and j. The search approach aims at solving the linear programming model

$$\min \sum_j -I_{jt} \tag{1}$$

subject to

$$\sum_k O_{sk} = I_0 \quad \text{for the output flow at } s \tag{2}$$

$$\sum_i I_{is} = 0 \quad \text{for the input flow at } s \tag{3}$$

$$\sum_j O_{tj} = 0 \quad \text{for the output flow at } t \tag{4}$$

$$\sum_i I_{ij} \geq \sum_k O_{jk} \quad \text{for each node } j \text{ except } s \text{ and } t \tag{5}$$

$$O_{ij} \geq 0 \quad \text{for each edge } E(i,j) \tag{6}$$

$$I_{ij} \geq 0 \quad \text{for each edge } E(i,j) \tag{7}$$

$$I_{ij} = D_{ij}O_{ij} \quad \text{for each edge } E(i,j) \tag{8}$$

$$O_{ij} \leq C_{ij} \quad \text{for each edge } E(i,j) \tag{9}$$

$$O_{ij} = 0 \quad \text{for each e } (i,j) \in T \tag{10}$$

where I_0 is the total amount of output flow at the source node. C_{ij} represents the capacities on the edge $E(i,j)$. D_{ij} denotes the dissipation index on the edge $E(i,j)$ and T refers to the set of interactions with directions (protein-DNA interactions). Formulation (1) illustrates that the objective is to maximize the received input flow at the target node. Equations (2) and (3) determine that the source node only sends information. And equation (4) ascertains that the target node does not send out any information flow. Equation (5) shows that the amount of input flow has to be larger than or equal to the amount of output flow at each internal node (nodes except the source and target). Inequations (6) and (7) require the flow to be nonnegative. Equation (8) defines that the flow from i to j is dissipated, in which part of the output flow O_{ij} at i was converted into the input flow I_{ij} at j according to the dissipation index D_{ij}, defined as the absolute value of Pearson correlation coefficient of gene i and j based on gene expression data. Inequation (9) confines the output flow on each edge cannot exceed the capacity limit of that edge. The linear programming model (1)-(9) does not consider the direction of the interactions. So it is based on an undirected network, which can be constructed from the large-scale protein-protein interactions and gene expression data. However, in biological systems, most interactions have orientations. Thus, constraints that confine the directions of flows on interactions should be added. The formulation is given by equation (10). Equation (10) restricts the flow only along the direction of the edge. The reverse flow should be zero.

3.2 The Search Algorithm

The linear programming model (1)-(10) infers the pathways given the source, the target and the whole network $G(V,E,D,C,T)$, where V is the set of proteins and DNAs, E is the set of protein-protein interactions and protein-DNA interactions, D defines the dissipation indices on each edge, C defines the capacities of each edge, and T defines the orientations of the interactions. V, E and T can be easily constructed from the large-scale protein-protein and protein-DNA interactions. D is defined by the absolute value of correlation coefficients determined by using the expression values of genes ([13]-[16]). C cannot be assigned that easily, because now there is no sufficient experimental information available. We design a stochastic searching algorithm in this study to bypass the assignment problem of C in practice. The algorithm is described as follows:

(i) For $k=1$, set C_1 large enough for each edge (e.g. input for each edge), solve the linear programming model (1)-(10) with parameters $G(V,E,D,C,T)$ and get the solution X_1. X_1 is a simple path from the source to the target.

(ii) For $k=i$ (i>1), randomly select one of the edges of $X_i - 1$ and denote the selected edge as p. Let $C_i = C_i - 1$, set the capacity of p as zero and update C_i. Solve the linear programming model (1)-(10) with parameters $G(V,E,D,C,T)$ and get the solution X_i.

(iii) Repeat (ii) until k reaches the allowable times K.

(iv) $X_1, \cdots, X_k$ are all simple paths. Assemble $X_1, \cdots, X_k$ will get a subnetwork connecting the source and the target. Set the subnetwork as the last solution to the original problem defined by (1)-(10) in which C is unknown.

The idea behind the algorithm is to search the optimal path at first, then to search the suboptimal paths after blocking the optimal path, and repeat this procedure. Saturation is simulated through blocking the available paths. This algorithm is likely to identify the optimal path from the source to the target. The difference lies in the simulations of saturation through blocking.

Due to the stochastic nature of the algorithm, it will run several times, e.g. N times, and then half of the solutions with the lower objective values are selected as the candidate pathways from the source to the target.

There are totally three parameters in this searching algorithm. The first parameter is I_0, which represents the amount of information flow the source sends out. The result is independent of the value of I_0, as long as it is positive. In this study we set I_0 to be 1. The second parameter is K, the number of zeros in C, which measures the complexity of the inferred pathways. The larger K, the more complicated is the predicted pathway. A pathway predicted with a smaller K is more significant. The pathway predicted with lager K is more complete and includes the pathway predicted with the smaller . The third parameter is N, which represents the number of repetitions to counteract the random effect in the stochastic search. N is positively related to the number of edges of the predicted pathway.

4 Discussion and Conclusion

To better understand the importance of the hub cofactors and depict the properties of a network, the network topologies were considered in this section. The network diameter defined as the average minimum distance between pairs of nodes. Once remove the hub cofactors in the two regulatory networks, respectively, the network diameters increase (see Table 4).

Table 4: Binding sites enrichment analysis of the two hub cofactors.

	ZIC2	AR
With	6	7
Without	11	10

In this article, transcriptional regulatory network revealing the mechanism of the key contributor MeCP2 was considered in RTT study. In view of the limitation study of MeCP2, we used information mining method to maximize the potential protein-protein interaction and protein-DNA interaction. Database search, literature mining, and computational prediction were adopted, and finally we developed a novel algorithm to search the optimal pathways from the source MeCP2 to its downstream genes. In both MeCP2 activated and repressed groups, two hub cofactors, whose binding sites were enriched in the extensive groups, were finally identified. The network topology analysis indicated that the hub cofactors play the important roles because the network diameters will significantly increase once they were removed.

Acknowledges

The authors would like to thank Dr. Maria Chahrour in Baylor College of Medicine and the colleagues of Bioinformatics Center, The Methodist Hospital Research Institute for their discussion and valuable suggestion in the research. This work was partially supported by the National Natural Science Foundation of China(No.10926036), the Scientific Research Foundation of Zhejiang Gongshang University (No.1020KH110043 & No.1020XJ200973) and the TMHRI Scholarship (Xiaobo Zhou).

References

[1] Haverty PM, Weng Z, Hansen U. CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic. Acids. Res.* 2004: W213-216.

[2] Sun J, Tuncay K, Haidar AA, Ensman L, Stanley F, et al. Transcriptional regulatory network discovery via multiple method integration: application to e.coli K12 Algorithms for Molecular Biology 2007, 2:2

[3] Chahrour M, Jung SY, Shaw C, Zhou XB, et al. MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* 2008, 30 vol. 320, no. 5880, 1224-1229.

[4] Gardina PJ, Clark TA, Shimada B, Staples MK, et al., Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 2006, 7, 325.

[5] Srinivasan K, Shiue L, Hayes JD, Centers R, et al., Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, 2005, 37(4), 345-359.

[6] `http://www.sciencemag.org/cgi/content/full/sci;320/5880/1224/DC1`

[7] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 2005, 21, 2933-2942.

[8] Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A et al: Human Protein Reference Database–2009 update. *Nucleic. Acids. Res.*, 2008:gkn892.

[9] Bader GD, Betel D, Hogue CWV. BIND: the Biomolecular Interaction Network Database. *Nucleic. Acids. Res.*, 2003, 31(1):248-250.

[10] Zaki N, Lazarova-Molnar S, El-Hajj W, Campbell P. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, 2009, 10:150.

[11] Shen J, Zhang J, Luo X, Zhu, W, et al. Predicting protein-protein interactions based only on sequences information *Proc. Natl. Acad. Sci. USA*, 2007,104(11):4337-4341.

[12] `http://www.dddc.ac.cn/spps/document.php`

[13] Scott J, Ideker T, Karp RM, and Sharan, R. Efficient algorithms for detecting signaling pathways in protein interaction networks, *Journal of Computational Biology*, 2006, 13, 133-144.

[14] Suthram S, Beyer A, Karp RM, Eldar Y, and Ideker T. eQED: an efficient method for interpreting eQTL associations using protein networks, *Mol. Syst. Biol.*, 2008, 4.

[15] Ren X, Zhou X, Wu LY and Zhang XS. An information-flow-based model with dissipation saturation and direction for active pathway inference. *BMC System Biology*, 2010, 4:72

[16] Tu Z, Wang L, Arbeitman MN, Chen T, and Sun, F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, 2006, 22, e489-496.