Journal of Computer Science and Cybernetics, V.27, N.1 (2011), 23-35

COMPUTATIONAL RECONSTRUCTION OF METABOLIC NETWORKS FROM HIGH-THROUGHPUT PROFILING DATA

NGUYEN QUYNH DIEP[1] , PHAM THO HOAN[1] , HO TU BAO[2]
TRAN DANG HUNG[1] , PHAM QUOC THANG[3]

[1]*Hanoi National University of Education, 136 Xuan-Thuy, Cau-Giay, Hanoi, Vietnam*

[2]*Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

[3]*Tay-Bac University, Son-La city, Vietnam*

Tóm tắt. Hầu hết các phương pháp tính toán tái hiện mạng sinh học hiện nay mới chỉ tập trung tìm các tương tác giữa hai phân tử, trong khi đó mạng chuyển hóa lại bao gồm các phản ứng liên quan đến từ 2 đến 6 chất. Vì vậy mà các phương pháp khám phá mạng sinh học đang tồn tại không thích hợp để tái hiện các phản ứng sinh hóa có nhiều hơn hai chất tham gia.

Bài báo này giới thiệu một phương pháp tính toán tái hiện mạng các chất chuyển hóa từ dữ liệu đo nồng độ/khối lượng các chất ở các điều kiện hoặc thời điểm khác nhau. Phương pháp không chỉ phát hiện các tương tác giữa hai phân tử mà còn phát hiện được các tương tác nhiều hơn hai phân tử, đó là các tương tác ba chất, tương tác bốn chất, v.v. Trong phương pháp đề xuất, chúng tôi sử dụng độ đo thông tin phụ thuộc bậc ba để dò tìm các tương tác đa chất. Chúng tôi cung cấp một cách nhìn mới về độ đo thông tin phụ thuộc bậc ba mà thích hợp trong việc phát hiện các tương tác nhiều hơn hai biến. Hiệu năng của phương pháp đề xuất đã được đánh giá trên các dữ liệu mô phỏng các hệ chuyển hóa sinh học. Tính chính xác của phương pháp tái hiện lại mạng chuyển hóa được đánh giá ở hai mức: các tương tác hai chất và các tương tác ba chất. Kết quả tái hiện của Phương pháp đề xuất là rất triển vọng.

Abstract. All computational methods of biological network reconstruction up to now aim only to find pairwise interactions. While metabolic networks composed mainly of reactions that often consist of from 2 to 6 substrates/products, the existing computational methods may not be appropriate to reconstruct interactions of more than two variables like reactions in the metabolic networks.

In this paper, we develop a computational method for the metabolic network reconstruction that can uncover not only pairwise interactions but also interactions involving more than two substrates/products such as triple interactions, quartic interactions, etc. In the proposed method we use the ternary mutual information to capture high order interactions. The key idea is to propose a novel view on the ternary mutual information that can be appropriately used to reconstruct reactions involving more than two substrates/products. We have applied the proposed method to synthesized metabolome data; the reconstruction accuracy has been evaluated at the levels of pairwise and triple interactions. The performance of the method is promising.

Keywords: Mutual information, entropy, biological network reconstruction.

1. INTRODUCTION

Thanks to the advancement of high-throughput technologies, we can now measure simultaneously the concentrations of thousands of molecular species in a biological system, such as mRNAs [22] and metabolites [18]. These high-throughput data are snapshots of a biological system and are informative to infer what has happened in the system. The analysis of the high-throughput data to uncover underlying biological mechanisms, e.g. gene regulatory networks (see [12] for an overview) or metabolic networks [6, 20] is one of the challenges in systems biology.

Computational reconstruction of gene regulatory networks from transcriptome data has been deeply investigated by different approaches. These reverse engineering methods fall into three broad categories: (1) information theory models [24, 5, 19] with a variety of measures of pairwise mutual information between genes; (2) Bayesian and graphical networks [10, 25] that maximize a scoring function over some alternative network models to find the best model fitting the data; (3) differential and difference equations [11, 4] that explain the data by a system of mathematical equations. All the work on the gene regulatory network reconstruction until now aims to find only pairwise interactions (concerning with two genes).

Different from gene regulatory networks that mainly concern with pairwise interactions, metabolic networks are composed mainly of reactions that often consist of from 2 to 6 metabolites (substrates/products). Thus, the metabolic network reconstruction should aim to find groups of metabolites that each involves in the same reaction. Up to now, there have been efforts to reconstruct metabolic networks that use methodologies of gene regulatory network reconstruction [6, 20]. As a consequence, they can only detect pairwise interactions but not interactions of more than two metabolites.

In this work, we develop a computational method *net-reconstruct* for the metabolic network reconstruction that can uncover not only pairwise interactions but also interactions involving more than two substrates/products, for example, triple interactions, quartic interactions, etc. In this method we use the interaction mutual information [9] to capture multiple interactions. The key idea is to propose a novel view on the interaction mutual information that can be appropriately use to reconstruct reactions involving more than two substrates/products.

When applying on the synthetic perturbation data of full-random networks (all structures, kinetic laws and parameter values are randomly generated, [2]) as well as of a semi-random networks, the human red blood cell metabolism ([14, 20]), our method gave promising results of interaction subsets that are close to the validated metabolic reactions. The interaction subsets with highest mutual information found from our method often correspond to metabolic reactions in the original networks, also many original reactions have been found in the results of our software. When evaluating accuracy at the level of pairwise interactions, the results of our method agreed with those of recent research on reconstruction methods.

2. METHODS

2.1. Mutual information between two variables

Mutual information measure is more general than Pearson's correlation coefficient (PPC) to capture dependency between two variables. While PPC accounts only for linear or monotonic relationships, the mutual information takes into account all types of dependence. Given

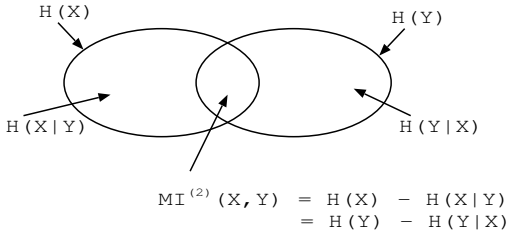

Figure 2.1. The Venn diagram for mutual information $MI^{(2)}$ of two variables

two random variables X and Y with the joint density function $f_{X,Y}$ and marginal density functions f_X, f_Y, the mutual information $MI^{(2)}$ of two variables X and Y [8] is defined as follows:

$$MI^{(2)}(X,Y) = \int\int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \tag{2.1}$$

(we use the superscript number 2 to emphasize that the mutual information here is for 2 variables)

If X and Y are independent, the mutual information $MI^{(2)}(X,Y) = 0$; if they are perfectly dependent, $MI^{(2)}(X,Y)$ approaches infinity.

The mutual information $MI^{(2)}(X,Y)$ can also be interpreted in terms of information entropy [8] as

$$\begin{aligned} MI^{(2)}(X,Y) &= H(X) + H(Y) - H(X,Y) && (2.2)\\ &= H(X) - H(X|Y) && (2.3)\\ &= H(Y) - H(Y|X) && (2.4) \end{aligned}$$

From Eq. 2.3 and Eq. 2.4 we can interpret the meaning of $MI^{(2)}(X,Y)$ as it measures the reduction the uncertainty of X due to the knowledge of Y, or vice versa [3]. The above interpretation of Shannon entropy can be visualized by the Venn diagram in Figure 2.1, where $MI^{(2)}(X,Y)$ is the intersection of two entropy circles $H(X)$ and $H(Y)$, and $H(X,Y)$ is the union of two sets $H(X)$ and $H(Y)$ [3, 13].

2.2. Mutual information for more than two variables

The mutual information $MI^{(2)}$ can detect interactions (edges) between two variables in a network. However, in most biological networks, each node (variable) may interact (link) with some others in the same or different mechanisms. Metabolic networks are an example of such networks, where each metabolite may interact with some others in different reactions. In this section, we present an extension of $MI^{(2)}$ that allows capturing the interactions of three variables.

The generalization of mutual information of three variables from that of two variables is not trivial [3, 13]. One of those generations is *i*nteraction mutual information [9] that has received much attention but with controversial interpretations, defined as follows:

$$\begin{aligned} MI^{(3)}(X,Y,Z) &= H(X) + H(Y) + H(Z) - H(X,Y)\\ &\quad -H(Y,Z) - H(X,Z) + H(X,Y,Z) && (2.5)\\ &= MI^{(2)}(X,Y) - MI^{(2)}(X,Y|Z) && (2.6) \end{aligned}$$

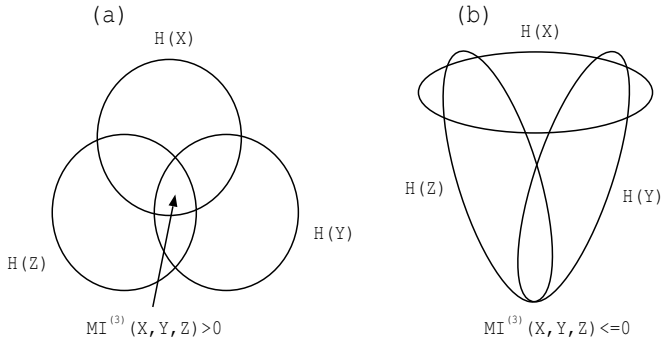

Figure 2.2. The Venn diagram for mutual information of three variables.

(Note: (1) we use again the superscript 3 to emphasize the mutual information of 3 variables; (2) some authors used the similar formulas but with the opposite sign.)

From Eq. 2.6, similarly to the interpretation of $MI^{(2)}$, the mutual information $MI^{(3)}$ common to three variables can be understood as it measures the reduction of $MI^{(2)}$ of two variables due to the knowledge of the third variable [3]. We can generalize the "reduction" interpretations of $MI^{(2)}$ (Eq. 2.3) and of $MI^{(3)}$ (Eq. 2.6) as follows. Suppose that $H(X)$ is the primary mutual information of one variable, we can view that $MI^{(2)}$ is reputedly the *s*econdary mutual information for two variables (since it is defined as the reduction of the primary mutual information after introducing a new variable), and $MI^{(3)}$ is as the ternary mutual information for three variables (also defined as the reduction of the secondary mutual information after introducing a new variable). Noting that, the extension is not only for the number of variables (from 2 to 3) but also the physical meaning or interpretation like the second order derivative and higher order derivatives of functions in calculus.

There have been many interpretations and usages of the ternary $MI^{(3)}$ ([17]). In this work, we propose a novel interpretation of $MI^{(3)}$ in order to capture interactions of more than two variables. Indeed, if three variables involve in a mechanism (such as a metabolic reaction), they are cohesive and thus the information on one variable often relates to the information on dependence of the other two. Therefore, the secondary mutual information $MI^{(2)}$ of two variables in general decreases after the introduction of the other variable, and so $MI^{(3)}$ of these three variables will be a positive number (since $MI^{(3)}$ measures the *r*eduction of $MI^{(2)}$). The higher the dependence among three variables, the higher the $MI^{(3)}$ is. This explanation agrees with the work of [13], in which the authors said that there is a redundance among three variables if mutual information of these three variables is positive, i.e. $MI^{(3)} > 0$ (in fact, in the paper they defined the mutual information for three variables by the same formulas of $MI^{(3)}$ but with the opposite sign).

Figure 2.2a illustrates the case that three variables involve in the same mechanism, and thus $MI^{(3)}$ is positive. Figure 2.2b illustrates the case that three variables do not involve in the same mechanism, but each two of them involves in the same mechanism. In this case, $MI^{(3)}$ is negative or equal to zero.

2.3. Using mutual information to detect multiple metabolite interactions

In metabolic networks, if three metabolites participate in a reaction, their concentrations often change simultaneously (cohesively) when the reaction is active. If they only appear in a unique reaction, they are completely dependent and $MI^{(3)}$ has usually a high value. If they appear in several different reactions at the same or different time, they are partially dependent and $MI^{(3)}$ has usually a lower value. Thus, we can use $MI^{(3)}$ to capture the triple metabolite interactions.

Naturally, we can think of higher order mutual information to capture the multiple interactions for more than 3 variables, such as $MI^{(4)}$ is a measurement of the reduction of $MI^{(3)}$ common to three variables after the introduction of the fourth variable. However, the reduction of $MI^{(3)}$ does not make sense in the metabolic network reconstruction.

In this method we detect such multiple interactions by recursively building them up starting from the set of triple interactions. If four metabolites X, Y, Z, T involve in the same reaction, there is a quartic interaction among them, and each of the four triples (X, Y, Z), (Y, Z, T), (X, Z, T) and (X, Y, T) corresponds to a triple interaction.

2.4. Algorithm

Given a metabolome dataset, we want to reconstruct from it multiple interactions. The proposed algorithm is similar to the Apriori in [1] to find frequent itemsets. First, we find the set L_2 of pairwise interactions between metabolites by using $MI^{(2)}$ (Subsection 2.1) with a threshold of minimal pairwise mutual information $min\_threshold_2$. Second, we join L_2 with itself to build a set of candidates of triple interactions L_3, then apply $MI^{(3)}$ (Subsection 2.2) to keep only triple interactions that $MI^{(3)}$ is greater than a minimal ternary mutual information $min\_threshold_3$ (Subsection 2.2). Third, we find the set of quartic interactions L_4 and those of higher order interactions L_5, L_6, etc. as described in Subsection 2.3. The final result is the set union $L = L_2 \cup L_3 \cup \ldots$

Table 1 describes in more detail the proposed algorithm to reconstruct metabolic networks from metabolome data. In this algorithm, we use the procedure "join" that joins a supersubset L_i with itself like the Apriori algorithm: two subsets in L_i will be joined to generate a candidate of L_{i+1} if they share i-1 variables. We used the k-nearest neighbor statistic-based method in [16] and MI-libraries provided the authors to estimate $MI^{(2)}$ and $MI^{(3)}$.

2.5. Datasets

Similarly to the gene regulatory network reconstruction, we use *in* silico generated metabolome data from random metabolic network models to evaluate the method. It is known that metabolic networks often consist of three components: stoichiometries (which contain the network structure), kinetic equations, and parameters therein. We used Matlab *RMBNToolbox* program developed by [2] to generate fully random metabolic networks with all three components randomly generated (kinetic laws are randomly selected from a library of 17 kinds of kinetic equations), then we used Matlab's built-in ordinary differential equation solver $ode15s$ to generate perturbation or time course data from the model.

Usually, $ode15s$ generates time course data points from a metabolic model until reaching the steady state, after that data is unchanged. This kind of data can be used to infer underlying networks. However, the use of these data is not always informative in all reconstruction

Table 1. Algorithm *n*et-reconstruct for inferring metabolic reactions . *The procedure $join(L_k, L_k)$ is completely the same as in the original Apriori algorithm.

Algorithm for inferring metabolic reactions	
Input:	(1) A matrix $X_{N\times T}$, measurements of N metabolites in T experiments; (2) $threshold_2, threshold_3$ (for finding L_2 and L_3)
Output:	$L = L_2 \cup L_3 \cup \ldots$ (see Subsection 2.4)

```
1   L_1 ← {all metabolites M_1, M_2, ..., M_N}
2   L_2 ← join(L_1, L_1)*
3   forall C ∈ L_2
4       if MI^(2)(C) < threshold_2
5           L_2 ← L_2 \ C
6   L_3 ← join(L_2, L_2)
7   forall C ∈ L_3
8       L_temp ← {all 2-subsets of C}
9       if ¬(L_temp ⊆ L_2) ∨ (MI^(3)(C) < threshold_3)
10          L_3 ← L_3 \ C
11  k ← 3
12  while L_k ≠ φ
13      L_{k+1} ← join(L_k, L_k)
14      forall C ∈ L_{k+1}
15          L_temp ← {all k-subsets of C}
16          if not (L_temp ⊆ L_2)
17              L_{k+1} ← L_{k+1} \ C
18  Return L = L_2 ∪ L_3 ∪ ... ∪ L_{k+1}
```

Table 2. Stoichiometry of a simple randomly generated metabolic network. An entry (i, j) of the stoichiometry matrix is 1 or -1 if metabolite Mi appears in the left or right side of reaction r_j, respectively. An entry (i, j) of the interaction matrix is 1 if there exists a reaction that both i and j participated in.

	Stoichiometry				**Interaction matrix**						
	r_1	r_2	r_3	r_4	M1	M2	M3	M4	M5	M6	M7
M1	1	0	1	0	0	0	1	0	1	0	1
M2	0	-1	0	1	0	0	1	1	0	1	1
M3	-1	0	0	-1	1	1	0	0	0	1	1
M4	0	-1	0	0	0	1	0	0	0	1	1
M5	0	0	1	0	1	0	0	0	0	0	1
M6	0	1	0	-1	0	1	1	1	0	0	1
M7	0	1	-1	1	1	1	1	1	1	1	0

methods, especially when the data contain many unchanged values. People often collect data at steady state by as many as possible perturbation experiments, and as many kinds of perturbation experiments as they can. These perturbation data are very informative for network reconstruction.

In our work, we use both kinds of data. However, with some metabolic models, the time course data contains constant values, so we use perturbation data only in those cases. The time course data is easily generated by running $ode15s$ on the metabolic model. With the perturbation data, the generation is quite complicated. We generate simulated perturbation metabolome data of metabolic networks following the ways described in [6]. The data points were collected from different runs for each variability type (biological/environment, enzymatic variabilities). We slightly modified the function WriteODEFunction in $SBMLToolbox$ [15], incorporated with random metabolic networks generation $RMBNToolbox$ [2] for different types of perturbation experiments.

In addition to the datasets prepared by ourselves, we also use available datasets of red blood cell metabolism in [20]. In the website, they provide both perturbation as well as time course data.

3. RESULTS AND DISCUSSION

3.1. Reconstruction of multiple interactions

Table 3. A part of a dataset generated from a random metabolic network model. This dataset will be input to our program.

M1	M2	M3	M4	M5	M6	M7
≈	≈	≈	≈	≈	≈	≈
2.190340	1.445580	4.871360	0.118320	3.023370	0.000293	0.294074
2.208590	1.434610	4.654120	0.116079	2.946740	0.000295	0.286802
1.994790	1.560900	4.659180	0.115525	2.943530	0.000255	0.374311
1.723110	1.694470	5.141490	0.119788	3.113190	0.000218	0.512416
1.767730	1.673290	5.654130	0.124854	3.288450	0.000227	0.488985
2.038860	1.533710	5.888540	0.127822	3.365870	0.000271	0.357428
2.896100	0.939036	5.043200	0.124665	3.064910	0.000533	0.094865
2.837500	0.987277	3.590650	0.107567	2.528740	0.000485	0.105223
2.027790	1.546790	2.987990	0.095363	2.289370	0.000246	0.355421
1.177920	1.855020	4.027860	0.106670	2.714580	0.000154	0.897059
1.266220	1.838360	4.951670	0.116731	3.057540	0.000167	0.825420
1.537640	1.765240	5.669790	0.124358	3.298910	0.000198	0.627115
1.874470	1.621190	6.101030	0.129257	3.435270	0.000245	0.434340
2.668980	1.114570	5.669880	0.128638	3.297110	0.000435	0.146444
2.953670	0.893437	4.112310	0.115540	2.706430	0.000554	0.082897
2.511160	1.234210	2.982650	0.096848	2.300660	0.000354	0.184628
≈	≈	≈	≈	≈	≈	≈

We have developed a computational method *net-reconstruct* that uncovers multiple metabolite interactions: pairwise interactions, triple interactions, quartic interactions, etc., from metabolome data. For pairing interactions, like previous network reconstruction methods,

Table 4. List of multiple interactions found on the simple metabolome data. *Those reactions have been completely reconstructed.

	Predicted interaction	MI	Matched reactions
pairwise interaction		$MI^{(2)}$	
1	{M6, M8}	0.409	r_3
2	{M1, M8}	0.327	r_3
3	{M3, M7}	0.298	r_4
4	{M1, M6}	0.217	r_3
5	{M5, M7}	0.190	r_2
6	{M2, M7}	0.135	r_2, r_4
7	{M1, M5}	0.095	
8	{M5, M8}	0.080	r_2
9	{M1, M2}	0.073	
10	{M1, M3}	0.050	r_1^*
11	{M2, M5}	0.043	r_2
triple interaction		$MI^{(3)}$	
12	{M1, M6, M8}	0.233	r_3^*
13	{M1, M2, M5}	0.088	
14	{M2, M5, M7}	0.067	r_2

net-reconstruct produces the mutual information matrix representing the confidence of pairing interactions. Moreover, our method can additionally uncover other multiple interactions that all together contribute to make a progress toward the reconstruction of complete metabolic networks.

To illustrate the advantages of the method, we carry out an experiment on the perturbation metabolome data of a randomly generated metabolic network (see Subsection 2.5). Table 2 shows stoichiometry and interaction matrix of a small randomly-generated metabolic network that consists of 7 metabolites (denoted by M1, M2, ..., M7) and 4 reactions (denoted by $r1, r2, r3, r4$). After generating the random network, we use the MATLAB $ode15s$ to generate perturbation data, which is input to our program. Table 3 shows a apart of the input data.

Applying the proposed method to the above metabolome data (100 perturbation data points), minimum thresholds $threshold_2 = 0.04$ for pairing mutual information $MI^{(2)}$ and $threshold_3 = 0.04$ for ternary mutual information $MI^{(3)}$, we obtained multiple interactions presented in Table 4. We have matched multiple interactions with the four original reactions, and the matched reactions are presented in the last column. Among eleven pairing interactions, only two are false positive. Especially, among three triple interactions with the highest $MI^{(3)}$, two of them are confirmed to be true positive. We also found that two of four original reactions (r_2 and r_3, marked with *) have been completely reconstructed by the method. The remained ones have been partially reconstructed in different multiple interactions.

We also applied the proposed method to the *i*n silico metabolome data of red blood cell metabolism (RBC) published by [20]. The RBC model consists of 39 metabolites and 44 reactions. The datasets can be downloaded at *h*ttp://menem.com/∼ilya/wiki/index.php/RBC_Metabolic_Network. Table 5 shows multiple interactions found by the method on the

Table 5. List of multiple interactions found on RBC metabolome data. *Those reactions have been completely reconstructed.

	Predicted interaction	MI	Matched reactions
pairwise interaction		$MI^{(2)}$	
1	{LAC, NAHD}	4.555303	ldh
2	{G6P, F6P}	4.554703	pgi*
3	{DHAP, GAP}	4.552503	ald,tpi*
4	{RU5P, X5P}	4.542912	xu5pe*
5	{NADPH, GSH}	4.541625	gssgr*
6	{RU5P, R5P}	4.483696	ru5pi*
7	{R5P, X5P}	4.474055	tki
8	{PG2, PEP}	4.468436	en*
9	{PG3, PG2}	4.418211	pgm*
10	{PG3, PEP}	4.326527	
11	{FDP, DHAP}	3.673954	ald
12	{FDP, GAP}	3.671634	ald
13	{AMP, IMP}	3.482159	ampda
14	{F6P, GO6P}	3.297622	
15	{G6P, GO6P}	3.297123	
16	{DPG23, MGDPG23}	2.948046	cplex(DPG23,MG)
17	{R5P, R1P}	2.865513	prm*
18	{RU5P, R1P}	2.844739	
19	{X5P, R1P}	2.841181	
20	{GL6P, KI}	2.776472	
21	{GO6P, NADPH}	2.766081	gl6pdh
22	{GO6P, GSH}	2.763634	
23	{G6P, R5P}	2.750224	
24	{F6P, R5P}	2.750112	
25	{G6P, X5P}	2.749533	
26	{F6P, X5P}	2.749280	tkii
27	{G6P, RU5P}	2.747932	
28	{F6P, RU5P}	2.747910	
triple interaction		$MI^{(3)}$	
29	{RU5P, R5P, X5P}	4.505767	(ru5pi,tki)
30	{PG3, PG2, PEP}	4.396024	(pgm,en)
31	{FDP, DHAP, GAP}	4.097788	ald*
32	{G6P, F6P, GO6P}	3.927646	
33	{GO6P, NADPH, GSH}	3.860016	(gl6pdh,gssgr)
34	{RU5P, X5P, R1P}	3.690052	
35	{G6P, F6P, R5P}	3.647796	
36	{G6P, F6P, X5P}	3.647577	
37	{G6P, F6P, RU5P}	3.646949	
38	{G6P, RU5P, X5P}	3.643090	
39	{F6P, RU5P, X5P}	3.642950	(tkii,xu5pe)
40	{RU5P, X5P, R1P}	3.634046	
41	{R5P, X5P, R1P}	3.623439	(tki,prm)
42	{G6P, RU5P, X5P}	3.588749	
43	{F6P, RU5P, X5P}	3.588471	(tkii,xu5pe)
44	{G6P, R5P, X5P}	3.579587	
45	{F6P, R5P, X5P}	3.579301	(tki,tkii)
quartic or higher interaction		Total $MI^{(2)}$	
46	{RU5P, R5P, X5P, R1P}	10.254909	
47	{G6P, F6P, RU5P, X5P}	10.052438	
48	{G6P, RU5P, R5P, X5P}	10.045316	
49	{F6P, RU5P, R5P, X5P}	10.045237	
50	{G6P, F6P, R5P, X5P}	10.041268	
51	{G6P, F6P, RU5P, R5P}	10.040485	
52	{G6P, F6P, RU5P, R5P, X5P}	13.692359	

Table 6. Accuracy (AUC) of different methods on different datasets of some metabolic networks evaluated at pairing interactions. All columns different from two columns on "ts" (time series) data are on perturbation data.

Method	Random net. with 10 metabolites			RBC (39 metabolites and 44 reactions)				*S.*cerevisiae glycolysis	
	5 react.	10 react.	20 react.	chemostat	Natural	correlated	ts	pertu. steady state	ts
Correlation	0.63	0.77	0.81	0.65	0.66	0.65	0.64	0.89	0.91
Partial Correlation	0.74	0.79	0.87	0.61	0.60	0.61	0.67	0.93	0.95
Graphical Models	0.70	0.79	0.84	0.64	0.65	0.66	0.68	0.93	0.92
Mutual Info. ($MI^{(2)}$)	0.62	0.79	0.82	0.64	0.65	0.62	0.65	0.85	0.81
Cond. MI	0.88	0.67	0.82	0.65	0.65	0.63	0.67	0.82	0.83
MI & DPI	0.56	0.71	0.73	0.59	0.60	0.61	0.63	0.79	0.75

data *R*BC_set2_0_0 (1000 data points). Both mutual information $MI^{(2)}$ and $MI^{(3)}$ tend to increase when the number of data points increased. In this experiment, we used the parameters $threshold_2 = 2.7$ and $threshold_3 = 3.5$. We found 28 pairing interactions, among them 16 are confirmed by reactions of the RBC model. Especially, in reconstructed 17 triple interactions, we confirmed 8 ones concerning with a reaction or two adjacent reactions when checking them on the list of RBC reactions. For example, {RU5P, R5P, X5P} concern with two adjacent reactions *ru*5*pi* (RU5P, R5P) and *tki* (R5P, X5P, GAP, F6P), {PG3, PG2, PEP} concern with *pgm* (PG3, PG2), *en* (PG2, PEP), and {FDP, DHAP, GAP} is completely matched with a reaction *ald* (FDP, GAP, DHAP), etc. The notation *ru*5*pi* (RU5P, R5P) describes that the reaction *ru*5*pi* consists of two metabolites RU5P and R5P.

3.2. On the use of secondary mutual information $MI^{(2)}$ in metabolic network reconstruction

It is confirmed that the mutual information-based methods ($MI^{(2)}$) can capture pairing interactions well as correlation-based methods in gene regulatory reconstruction [23]. We aim to experimentally verify this characteristics in the case of metabolic network reconstruction. In our experiments we use the implementation of some methods done by [23], in which there are two broad categories: correlation-based and mutual information-based methods. The correlation-based methods include Correlation, Partial Correlation, and Graphical models. The mutual information-based methods include Mutual Information (Mutual Info.), Conditional Mutual Information (Cond. MI), and Mutual Information and Data Processing Inequality (MI_DPI). Their detail description can be found in [23].

We evaluate the accuracy of these methods on 3 kinds of metabolome datasets. First, we generate random networks with 10 metabolites and different network complexities (5 reactions, 10 reactions and 20 reactions, see Section 2.5). For each network topology, we run all methods 10 times with the same parameters and the final prediction results are averaged. Second, we use previously generated RBC datasets [20, 7] to evaluate these six methods. The last metabolome dataset is randomly generated from the *S.*cerevisiae glycolysis model [21]. From Table 6, we can observe that the mutual information-based methods generally achieved comparable accuracy. Different from reconstruction of gene regulatory networks where the accuracy on time course data often lower then that on perturbation data, we can see in these experiment results their accuracies are not considerably different.

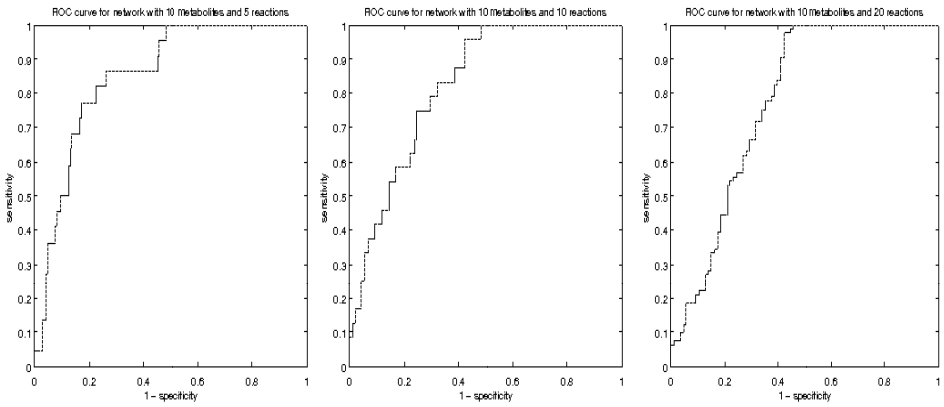

Figure 3.3. ROC of the reconstruction of triple interactions using $MI^{(3)}$ in metabolic networks with different complexities.

3.3. On the use of ternary mutual information $MI^{(3)}$ in metabolic network reconstruction

Different from existing reconstruction methods, our method can capture triple interactions by using ternary mutual information $MI^{(3)}$. As can be seen in Table 4 and Table 5, the ternary mutual information $MI^{(3)}$ allows us to detect many triple interactions that matched the real ones in the considered metabolic networks. We aim to experimentally verify that it can function well for metabolic networks with different complexities. In the experiments, we consider three random networks having 10 metabolites, but the first one has 5 reactions, the second has 10 reactions and the third has 20 reactions. Figure 3.3 shows the receiver operating characteristic of these three networks. The area under the curve of the three networks are 0.85, 0.82 and 0.77, respectively. As we can see, the reconstruction accuracy decreases when the complexity of networks increases. Although the third network is much more complicated than the others, the proposed method's performance is not significantly different.

From the found triple interactions, we can infer quartic or higher interactions using the Apriori property (see *M*ethods). As can be seen in Table 5, the quartic interactions (or higher ones) can, however, only partially match the original reactions. There are two reasons of why it is very hard to detect those complete interactions. Firstly, the number of k-subsets from n metabolites is C_n^k and it increases exponentially when k increase. For example, the number of triplets in RBC with 39 metabolites is $C_{39}^3 = 9139$, and the number of quartets is $C_{39}^4 = 82251$. Secondly, metabolic networks are often complex, i.e. a reaction may consists of some metabolites, and a metabolite is usually controlled by some reactions. Moreover, reactions are often active at the same time. However, $MI^{(3)}$ can detect groups of metabolites that are highly cohesive, i.e. they often concern with a reaction or two adjacent reactions or a set of closed reactions.

4. CONCLUSIONS

The computational reconstruction of pairwise interactions in networks from high-throughput profiling data is one of difficult problems in systems biology. Nevertheless, the multivariate

interaction reconstruction is more difficult. We proposed a novel interpretation of ternary from the view of interactions of variables and then illustrated that it is a very good measure to capture the multivariate interactions that involve the same mechanism.

We developed a method based on (secondary and ternary) mutual information to capture reactions involving two or more than two substrates/products such as pairwise interactions, triple interactions, quartic interactions, those often involve with the same reaction or close adjacent reactions. When applying the proposed method to *i*n silico metabolome data, the reconstruction accuracy is high. We can conclude that secondary and ternary mutual information are an interesting measurement relevant for detecting multivariate interactions.

Acknowledgements

This work was supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED Project No. 102.03.21.09).

REFERENCES

[1] R. Agrawal, T. Imielinski, and A.N. Swami. Mining Association Rules between Sets of Items in Large Databases. *SIGMOD*, 22(5):207–216, 2003.

[2] T. Aho, O.P. Smolander, J. Niemi, and O. Yli-Harja. RMBNToolbox: Random Models for Biochemical Networks. *BMC Systems Biology*, 1:22, doi:10.1186/1752-0509-1-22, 2007.

[3] D. Anastassiou. Computational Analysis of the Synergy among Multiple Interacting Genes. *Molecular Systems Biology*, pages 3:83, doi:10.1038/msb4100124, 2007.

[4] M. Bansal, G.D. Gatta, and D. Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

[5] A. Butte and I. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[6] T. Cakir, M. Hendriks, J. Westerhuis, and A.K. Smilde. Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, doi: 10.1007/s11306-009-0156-4, 2009.

[7] D. Camach, P.V. Licona, P. Mendes, and R. Laubenbacher. Comparison of Reverse-Engineering Methods Using an in Silico Network. *Ann. N.Y. Acad. Sci.*, doi: 10.1196/annals.1407.006, 2007.

[8] T.M. Cover and J.A. Thomas. Elements of Information Theory (Second edition). *Molecular Systems Biology*, Wiley-Interscience, A John wiley & Sons, Inc., Publication, 2006.

[9] R.M. Fano. Transmission of Information. *New York NY, USA: MIT press*, 1961.

[10] N. Friedman, M. Linial, I. Nachman, and D. Peer. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.*, 7:601–620, 2000.

[11] T.S. Gardner, D¿ Bernardo, D. Lorenz, and J.J. Collins. Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling . *Science*, 301:102–105, 2003.

[12] Michael Heckera, Sandro Lambecka, Susanne Toepferb, Eugene Somerenc, and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic models - A review. *Biosystems*, 96(1):86–103, 2009.

[13] A. Jakulin and I. Bratko. Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy. *CoRR*, cs.AI/0308002, http://arxiv.org/abs/cs.AI/0308002, 2004.

[14] N. Jamshidi, J.S. Edwards, T. Fahland, G.M. Church, and B.O. Palsson. Dynamic Simulation of the Human Red Blood Cell Metabolic Network. *Bioinformatics*, 17(3):286–287, 2001.

[15] S.M Keating, B.J. Bornstein, A. Finney, and M. Hucka M. Sbmltoolbox: an sbml toolbox for matlab users. *Bioinformatics*, 22(10):1275?277, 2006.

[16] Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev.*, 10.1103/PhysRevE.69.066138, 2004.

[17] L. Leydesdorff. Interaction Information: Linear and Nonlinear Interpretations. *Int. J. General Systems*, 2010.

[18] W. Lu, E. Kimball, and J.D. Rabinowitz. A High-performance Liquid Chromatography-tandem Mass Spectrometry Method for Quantitation of Nitrogen-containing Intracellular Metabolites. *J. Am. Soc. Mass Spectrom.*, 17:37–50, 2006.

[19] A.A. Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustag, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. *Nature Protocols*, 1:662 – 671, 2006.

[20] I. Nemenman, G.S. Escola, W.S. Hlavacek, P.J. Unkefer, C.J. Unkefer, and M.E. Wall. Reconstruction of Metabolic Networks from High-throughput Metabolite Profiling Data: in silico Analysis of Red Blood Cell Metabolism. *Ann N. Y. Acad Sci.*, 1115:102–115, doi: 10.1196/annals.1407.013, 2007.

[21] B.G. Olivier and J.L. Snoep. Web-based kinetic modelling using JWS Online. *Bioinformatics*, 20:2143–2144, 2004.

[22] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470, 1995.

[23] N. Soranzo, G. Bianconi, and C. Altafini. Comparing Association Network Algorithms for Reverse Engineering of Large-scale Gene Regulatory Networks: Synthetic versus Real Data. *Bioinformatics*, 23(13):1640 – 1647, 2007.

[24] R. Steuer, J. Kurths, C.O. Daub, J. Weise, and J. Selbig. The Mutual Information: Detecting and Evaluating Dependencies between Variables. *Bioinformatics*, 2002(18, Suppl 2):S231–S240, 2002.

[25] A.V. Werhli and D. Husmeier. Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, Vol. 6 : Iss. 1, Article 15, 2007.

Received on December 10 - 2010
Revised on March 22 - 2011