Tạp chí Tin học và Điều khiển học, T.30, S.2 (2014), 117–126

CÁC ĐỘ ĐO THÔNG TIN TƯƠNG HỖ ĐA BIẾN CÓ ĐIỀU KIỆN

NGUYỄN QUỲNH DIỆP[1], PHẠM THỌ HOÀN[1], HỒ TÚ BẢO[2]

[1] *Trường Đại học Sư phạm Hà Nội, 136 Xuân Thủy, Cầu Giấy, Hà Nội, Việt Nam*

[2] *Viện Khoa học và Công nghệ tiên tiến Nhật Bản, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

Tóm tắt. Thông tin tương hỗ (*Mutual Information*-MI) giữa hai biến đã được sử dụng để phát hiện mối quan hệ giữa hai biến; khi độ đo này lớn thì sự phụ thuộc giữa hai biến cũng lớn và ngược lại. Tuy nhiên, thông tin tương hỗ lại không cho ta biết mối quan hệ giữa các biến là trực tiếp hay gián tiếp. Để phát hiện quan hệ tương tác là trực tiếp hay gián tiếp, chúng ta có thể sử dụng thông tin tương hỗ có điều kiện đối với biến thứ ba (*Conditional Mutual Information*-CMI).

Trong các nghiên cứu trước đây, chúng tôi đã đề xuất các độ đo thông tin tương hỗ đa biến. Có rất nhiều độ đo thông tin tương hỗ khi số biến nhiều hơn hai, mỗi độ đo thể hiện một loại quan hệ có thể tồn tại giữa các biến. Tuy nhiên, cũng như thông tin tương hỗ của hai biến, các độ đo thông tin tương hỗ đa biến chỉ cho ta biết tồn tại hay không một mối quan hệ đa biến; nhưng không cho ta biết mối quan hệ đó là trực tiếp hay gián tiếp. Trong nghiên cứu này, chúng tôi đề xuất các độ đo thông tin tương hỗ đa biến có điều kiện và sử dụng chúng để phát hiện các mối quan hệ đa biến là trực tiếp hay gián tiếp thông qua biến điều kiện.

Từ khóa: Lý thuyết thông tin, entropy, thông tin tương hỗ, tái tạo mạng sinh học.

Abstract. Mutual information of two variables is a measure of relationship between two variables; the larger this measure, the stronger the dependence, and vice visa. However, mutual information does not indicate if the relationship between the variables is direct or indirect. To detect "direct mutual relations", we can use conditional mutual information.

In the previous studies, we have proposed the mutual information measures of multiple variables. There are many mutual information measures when the number of variables is greater than two. Each of them is sensitive to a kind of relationships that may exist among the multiple variables. However, as mutual information of two variables, the multivariate mutual information measures do not show if a multivariate relationship are direct or indirect. In this study, we propose the multivariate conditional mutual information measures and illustrate that they can detect indirect multivariate relationships through conditional variables.

Key words. Information theory, entropy, mutual information, biological network reconstruction.

1. GIỚI THIỆU

Thông tin tương hỗ giữa hai biến là một độ đo, đo mối quan hệ tương tác giữa hai biến [3]. Độ đo này đã được sử dụng để phát hiện các tương tác gien trong mạng điều hòa gien, tương tác protein trong mạng protein [1, 8]. Một số nghiên cứu sau đó đã chỉ ra rằng, thông tin tương hỗ giữa hai biến không thể phân biệt được các tương tác gián tiếp và tương tác trực tiếp [13, 14]. Trong nghiên cứu đó, tác giả đã đề xuất độ đo thông tin tương hỗ có điều kiện (CMI) của hai biến trên tập các biến còn lại để loại bỏ các tương tác gián tiếp giữa hai biến. Kết quả thực nghiệm cho thấy, tỷ lệ phát hiện đúng các tương tác gien tăng lên đáng kể nhờ loại bỏ được các tương tác gián tiếp trong mạng các gien.

Trong các nghiên cứu gần đây [10, 11], chúng tôi đã đề xuất mở rộng độ đo thông tin tương hỗ từ hai biến lên nhiều biến. Chúng tôi đã chỉ ra rằng, trong trường hợp hai biến, chỉ có duy nhất một loại quan hệ giữa chúng. Trong trường hợp ba biến trở lên, sẽ tồn tại nhiều loại quan hệ như quan hệ cặp đôi, quan hệ đồng thời giữa các biến và cả quan hệ bộ phận giữa chúng. Chúng tôi đã đề xuất các công thức khác nhau, mỗi công thức đặc trưng cho một loại quan hệ đa biến đó. Các độ đo thông tin tương hỗ đa biến đã được kiểm chứng về khả năng phát hiện tương tác tham gia bởi nhiều thành phần từ dữ liệu mô phỏng và dữ liệu thực.

Tuy nhiên, giống như thông tin tương hỗ của hai biến, các độ đo thông tin tương hỗ đa biến cũng không phân biệt được các tương tác trực tiếp với các tương tác gián tiếp. Việc phát hiện các tương tác đa biến gián tiếp sẽ giúp ta có một cái nhìn đầy đủ và chính xác hơn về mối quan hệ giữa các biến trong mạng sinh học. Trong nghiên cứu này, chúng tôi đề xuất mở rộng độ đo thông tin tương hỗ có điều kiện cho trường hợp đa biến và sử dụng chúng để xác nhận các quan hệ đa biến gián tiếp. Việc phát hiện các tương tác đa biến gián tiếp là tương đối phức tạp. Ý tưởng của chúng tôi là dùng các độ đo thông tin tương hỗ để phát hiện các tương tác gồm cả trực tiếp và gián tiếp. Sau đó, sử dụng thông tin tương hỗ đa biến có điều kiện để xác nhận hoặc loại bỏ các tương tác không phải là trực tiếp.

Nội dung tiếp theo của bài báo được trình bày theo thứ tự sau: phần 2 giới thiệu các kiến thức về thông tin tương hỗ và thông tin tương hỗ có điều kiện của hai biến trên biến thứ ba. Trong 2.3, chúng tôi đưa ra những đánh giá các độ đo này với phương pháp sử dụng hệ số tương quan và hệ số tương quan bộ phận trong việc phát hiện mối quan hệ giữa hai biến. Phần 3 giới thiệu các đề xuất mở rộng độ đo thông tin tương hỗ và thông tin tương hỗ có điều kiện trong trường hợp đa biến. Một số ví dụ được trình bày trong phần này nhằm kiểm chứng khả năng phát hiện các quan hệ đa biến là trực tiếp hay gián tiếp của thông tin tương hỗ có điều kiện. Cuối cùng là một ứng dụng các độ đo chúng tôi đề xuất trong việc phát hiện quan hệ gián tiếp trong mạng trao đổi chất ở người.

2. THÔNG TIN TƯƠNG HỖ CỦA HAI BIẾN, THÔNG TIN TƯƠNG HỖ CÓ ĐIỀU KIỆN CỦA HAI BIẾN

2.1. Thông tin tương hỗ của hai biến

Thông tin tương hỗ (MI) của hai biến ngẫu nhiên X và Y là độ đo trong Lý thuyết thông tin phản ánh quan hệ giữa chúng. Khi X và Y là các biến ngẫu nhiên rời rạc, MI được định

nghĩa như sau [2]:

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x).p(y)} \tag{2.1}$$

Trong đó, $p(x)$ và $p(y)$ lần lượt là hàm phân phối biên duyên của X và của Y; $p(x,y)$ là hàm phân phối xác suất đồng thời của hai biến X và Y.

Khi các biến X và Y là liên tục, phép tính tổng trong công thức trên được thay bởi phép tính tích phân trên miền giá trị của X và Y.

Chúng ta có thể biểu diễn thông tin tương hỗ qua entropy như sau:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.2}$$

Trong đó, $H(X)$, $H(Y)$ và $H(X,Y)$ lần lượt là entropy của biến X, biến Y và (X,Y).

Thông tin tương hỗ đã được xác nhận là độ đo hữu ích trong việc phát hiện sự tồn tại quan hệ giữa hai biến [8, 10, 11]. Tuy nhiên, độ đo này không thể phân biệt được đó là quan hệ trực tiếp giữa hai biến hay là quan hệ gián tiếp thông qua một hoặc nhiều biến trung gian. Ở đây, hai biến được gọi là có quan hệ trực tiếp nếu chúng cùng tham gia vào một sự kiện (phản ứng hoặc cơ chế điều hòa gien) và gọi là có quan hệ gián tiếp nếu chúng quan hệ với nhau thông qua một hoặc một dãy biến trung gian.

2.2. Thông tin tương hỗ có điều kiện của hai biến

Khi các biến X và Y không trực tiếp quan hệ với nhau nhưng có mối quan hệ gián tiếp thông qua biến thứ ba, MI sẽ phát hiện sự tồn tại quan hệ giữa X và Y. Nếu quan sát thêm được biến Z, ta có thể biết thêm thông tin về mối quan hệ này. Bằng cách lấy trung bình thông tin tương hỗ của hai biến X và Y trên biến Z, ta có thể biết được X và Y có quan hệ gián tiếp thông qua Z (ký hiệu, $X \leftrightarrow Z \leftrightarrow Y$) hay không. Độ đo trung bình thông tin tương hỗ của hai biến trên biến thứ ba được gọi là *thông tin tương hỗ có điều kiện* (CMI) và được định nghĩa như sau:

$$\begin{aligned} MI(X,Y \mid Z) &= \sum_{z\in Z} p(z) \sum_{x\in X}\sum_{y\in Y} p(x,y \mid z)\log\frac{p(x,y \mid z)}{p(x \mid z).p(y \mid z)} &(2.3)\\ &= \sum_{z\in Z} p(z) MI(X,Y \mid Z=z) &(2.4)\end{aligned}$$

Trong đó, $p(x \mid z)$ là hàm phân phối xác suất của biến X với điều kiện Z; $p(x,y \mid z)$ là hàm phân phối xác suất đồng thời của hai biến X và Y với điều kiện Z.

MI của hai biến có thể tăng lên hoặc giảm đi khi có sự xuất hiện của biến thứ ba. Trong khi $MI(X,Y \mid Z)$ đo mức độ tương hỗ trung bình giữa hai biến X và Y trên các giá trị của Z thì $MI(X,Y)$ đo mức độ tương hỗ trên không gian dữ liệu của hai biến X và Y. Có thể $MI(X,Y)$ lớn nhưng $MI(X,Y \mid Z)$ nhỏ vì khi quan sát trên hai biến X và Y ta chỉ nhìn được hình chiếu của dữ liệu trên không gian hai chiều X và Y. Nhưng khi quan sát cả ba biến X,Y,Z, mối quan hệ giữa X và Y có sự chi phối của Z, khi đó ta có thể biết được sự phụ thuộc gián tiếp $X \leftrightarrow Z \leftrightarrow Y$.

Ví dụ 2.1. Lặp 10 lần quá trình sinh ba biến rời rạc X,Y,Z theo xích Markov $X \to Z \to Y$. Cụ thể, trong ví dụ này, chúng tôi sử dụng Matlab để sinh ba biến theo qui tắc sau: $Z =$

$X + noise_1$ và $Y = Z + noise_2$. Kết quả tính toán các giá trị MI và CMI được trình bày trong Bảng 1.

***Bảng** 1.* Quan hệ gián tiếp $X \leftrightarrow Z \leftrightarrow Y$ với dữ liệu rời rạc

n	MI(X,Y)	MI(Y,Z)	MI(Z,X)	MI(Y,Z|X)	MI(Z,X|Y)	MI(X,Y|Z)
1	0.7179	0.8678	0.8022	0.1505	0.0850	0.0007
2	0.4602	0.6628	0.6636	0.2085	0.2093	0.0059
3	0.5031	0.6391	0.7599	0.1389	0.2598	0.0030
4	0.5626	0.7552	0.7132	0.1957	0.1536	0.0030
5	0.4400	0.7116	0.5954	0.2779	0.1618	0.0063
6	0.5395	0.6766	0.7622	0.1401	0.2257	0.0031
7	0.6343	0.6845	0.9256	0.0518	0.2929	0.0016
8	0.5460	0.7706	0.6640	0.2275	0.1209	0.0029
9	0.5695	0.7574	0.7261	0.1909	0.1596	0.0030
10	0.5811	0.7261	0.6758	0.1658	0.1154	0.0027

Quan sát các giá trị trong cột $MI(X, Y \mid Z)$ của Bảng 1, ta thấy chúng rất nhỏ so với các giá trị CMI khác. Hơn nữa, sự chênh lệch giữa $MI(X, Y)$ và $MI(X, Y \mid Z)$ rất lớn so với các cặp (Y, Z) và (Z, X). Trong trường hợp ba biến X, Y, Z là liên tục, ta cũng có kết quả tương tự như trường hợp rời rạc (xem Bảng 2).

***Bảng** 2.* Quan hệ gián tiếp $X \leftrightarrow Z \leftrightarrow Y$ với dữ liệu liên tục

n	MI(X,Y)	MI(Y,Z)	MI(Z,X)	MI(Y,Z|X)	MI(Z,X|Y)	MI(X,Y|Z)
1	1.1160	1.2275	1.6746	0.1168	0.5639	0.0053
2	1.0465	1.2705	1.5215	0.2240	0.4751	0.0000
3	1.1301	1.3662	1.6871	0.2395	0.5603	0.0033
4	1.0948	1.2949	1.4992	0.2034	0.4076	0.0032
5	1.1314	1.2995	1.6412	0.1693	0.5111	0.0013
6	1.0682	1.3292	1.6274	0.2687	0.5669	0.0077
7	0.9284	1.0772	1.6978	0.1531	0.7737	0.0043
8	1.1725	1.3710	1.6396	0.1993	0.4679	0.0008
9	1.1974	1.4622	1.7728	0.2740	0.5846	0.0092
10	1.0436	1.1358	1.6151	0.0998	0.5791	0.0075

2.3. Ứng dụng thông tin tương hỗ có điều kiện của hai biến để phát hiện tương tác gián tiếp

Trong một nghiên cứu trước đây [14], Zhang và cộng sự đã đề xuất thuật toán *Path Consistency* (PC) để phát hiện quan hệ giữa hai biến. Trong thuật toán đó, tác giả dùng MI của hai biến để phát hiện quan hệ giữa chúng, sau đó sử dụng CMI để loại bỏ quan hệ gián

tiếp. Tuy nhiên, thuật toán PC chỉ xét đến trường hợp $MI(X,Y)$ lớn hơn ngưỡng nào đó, còn trường hợp $MI(X,Y)$ nhỏ hơn ngưỡng thì tác giả coi như giữa X và Y không tồn tại quan hệ. Như vậy, phương pháp PC có thể đã bỏ sót các quan hệ gián tiếp $X \leftrightarrow Z \leftrightarrow Y$.

Ngoài việc sử dụng cặp MI và CMI để tìm các tương tác thực sự giữa các biến và loại bỏ các tương tác gián tiếp như trong bài báo trên, một ý tưởng tương tự đó là dùng hệ số tương quan và hệ số tương quan bộ phận để phát hiện tương tác giữa các biến [7]. Trong nghiên cứu này, tác giả đã sử dụng hệ số tương quan để phát hiện các tương tác giữa hai biến và dùng hệ số tương quan bộ phận (tức là hệ số tương quan giữa hai biến sau khi loại bỏ tương quan gián tiếp thông qua một biến thứ ba) để phát hiện tương quan giữa hai biến có là gián tiếp không. Nhược điểm của phương pháp này là hệ số tương quan và hệ số tương quan bộ phận chỉ có thể phát hiện được các kiểu quan hệ tuyến tính [4]. Chẳng hạn, khi hai biến có quan hệ phi tuyến, $y = x^2$, thì hệ số tương quan giữa chúng bằng 0. Như vậy, hệ số tương quan không thể phát hiện được sự phụ thuộc phi tuyến, trong khi thông tin tương hỗ lại làm được điều này.

3. THÔNG TIN TƯƠNG HỖ ĐA BIẾN, THÔNG TIN TƯƠNG HỖ ĐA BIẾN CÓ ĐIỀU KIỆN

3.1. Thông tin tương hỗ đa biến

Trong trường hợp đa biến, ngoài mối quan hệ tương tác giữa hai biến, còn có thêm mối quan hệ đồng thời giữa ba biến (gọi là quan hệ tổng hợp) và mối quan hệ giữa một biến với cặp hai biến còn lại (gọi là quan hệ bộ phận). Từ các phân tích đó, chúng tôi đã đề xuất công thức MI tổng quát trong trường hợp đa biến như sau [11]:

Định nghĩa 3.1. Thông tin tương hỗ của n biến $\{X_1, \ldots, X_n\}$ với phân hoạch $\{D_1, \ldots, D_k\}$ được định nghĩa:

$$MI_{\{D_1,\ldots,D_k\}}(X_1, \ldots, X_n) = H(D_1) + \ldots + H(D_k) - H(X_1, \ldots, X_n) \tag{3.5}$$

trong đó, $\{X_1, \ldots, X_n\} = D_1 \oplus \ldots \oplus D_k$.

Trong trường hợp ba biến, chúng ta có các độ đo thông tin tương hỗ như sau:

$$MI(X,Y,Z) = TC(X,Y,Z) = H(X) + H(Y) + H(Z) - H(X,Y,Z) \tag{3.6}$$

$$MI(X,[Y,Z]) = H(X) + H(Y,Z) - H(X,Y,Z) \tag{3.7}$$

$$MI(Y,[Z,X]) = H(Y) + H(Z,X) - H(X,Y,Z) \tag{3.8}$$

$$MI(Z,[X,Y]) = H(Z) + H(X,Y) - H(X,Y,Z) \tag{3.9}$$

Tuy nhiên, như phân tích trong phần 2.2, độ đo thông tin tương hỗ đa biến không thể cho ta biết mối quan hệ giữa các biến là quan hệ trực tiếp hay gián tiếp thông qua các biến trung gian. Vì vậy, trong phần tiếp theo, chúng tôi đề xuất các công thức mở rộng thông tin tương hỗ có điều kiện trong trường hợp đa biến.

3.2. Thông tin tương hỗ đa biến có điều kiện

Từ định nghĩa CMI trong trường hợp hai biến, chúng tôi đề xuất một mở rộng của độ đo CMI là *độ đo thông tin tương hỗ đa biến có điều kiện* như sau:

Định nghĩa 3.2. Thông tin tương hỗ có điều kiện của n biến $\{X_1, \ldots, X_n\}$ với phân hoạch $\{D_1, \ldots, D_k\}$ trên điều kiện C được định nghĩa:

$$\begin{aligned} MI_{\{D_1,\ldots,D_k\}}(X_1, \ldots, X_n \mid C) &= H(D_1 \mid C) + \ldots + H(D_k \mid C) - \\ &- H(X_1, \ldots, X_n \mid C) \end{aligned} \tag{3.10}$$

trong đó, $\{X_1, \ldots, X_n\} = D_1 \oplus \ldots \oplus D_k$.

Trong trường hợp ba biến X, Y, Z, ta có các phân hoạch sau:

a. $D_1 = \{X\}, D_2 = \{Y\}, D_3 = \{Z\}$

b. $D_1 = \{X\}, D_2 = \{Y, Z\}$

c. $D_1 = \{Y\}, D_2 = \{Z, X\}$

d. $D_1 = \{Z\}, D_2 = \{X, Y\}$

Do đó, tương ứng với 4 kiểu phân hoạch trên, theo công thức (3.10) ta có các độ đo thông tin tương hỗ có điều kiện của ba biến X, Y, Z trên biến thứ tư T như sau:

- *Thông tin tương hỗ tổng hợp của ba biến X, Y, Z trên điều kiện T*

$$MI(X, Y, Z \mid T) = H(X \mid T) + H(Y \mid T) + H(Z \mid T) - H(X, Y, Z \mid T) \tag{3.11}$$

- *Thông tin tương hỗ bộ phận giữa một biến với cặp hai biến trên điều kiện T*

$$MI(X, [Y, Z] \mid T) = H(X \mid T) + H(Y, Z \mid T) - H(X, Y, Z \mid T) \tag{3.12}$$

$$MI(Y, [Z, X] \mid T) = H(Y \mid T) + H(Z, X \mid T) - H(X, Y, Z \mid T) \tag{3.13}$$

$$MI(Z, [X, Y] \mid T) = H(Z \mid T) + H(X, Y \mid T) - H(X, Y, Z \mid T) \tag{3.14}$$

Giống như CMI của hai biến, các CMI đa biến cũng có khả năng phát hiện các tương tác đa biến gián tiếp.

Ví dụ 3.1. Trong ví dụ này, chúng tôi sử dụng Matlab lặp 10 lần quá trình sinh bốn biến X, Y, Z, T theo qui tắc sau: hai biến liên tục Y, Z độc lập; biến T phụ thuộc vào Y và Z, giả sử $T = Y + Z + noise_3$; biến X phụ thuộc vào biến T, giả sử $X = T + noise_4$. Sau khi tính toán các giá trị CMI trên tất cả các biến điều kiện, ta có kết quả được trình bày trong Bảng 3. Cột $MI(Y, Z)$ cho ta thấy hai biến Y và Z độc lập. Giá trị trong cột $MI(X, [Y, Z]|T)$ rất nhỏ so với các giá trị CMI trên các điều kiện biến X, Y, Z. Như phân tích trong Ví dụ 2.1, điều đó có nghĩa rằng, giữa X và (Y, Z) có mối quan hệ gián tiếp thông qua biến T (kiểu $X \leftrightarrow T \leftrightarrow [Y, Z]$).

Như vậy, dựa vào thông tin tương hỗ đa biến và thông tin tương hỗ đa biến có điều kiện, không những chúng ta có thể biết được mối quan hệ tồn tại giữa các biến mà còn có thể biết được mối quan hệ đó là quan hệ trực tiếp hay gián tiếp.

***Bảng** 3.* Quan hệ gián tiếp $X \leftrightarrow T \leftrightarrow [Y, Z]$ với dữ liệu liên tục

n	MI(Y,Z)	MI(T,[Y,Z])	MI(X,[Y,Z])	MI(T,[Y,Z]|X)	MI(X,[Y,Z]|T)
1	0.0019	2.6680	1.4355	1.2326	0.0001
2	0.0002	2.6469	1.5718	1.0890	0.0138
3	0.0020	2.6661	1.4642	1.2159	0.0140
4	0.0040	2.7685	1.5094	1.2775	0.0184
5	0.0012	2.6481	1.6885	0.9748	0.0152
6	0.0003	2.6646	1.3962	1.2699	0.0015
7	0.0020	2.7581	1.4426	1.3253	0.0097
8	0.0054	2.5400	1.2922	1.2746	0.0269
9	0.0051	2.6122	1.4902	1.1236	0.0017
10	0.0002	2.7149	1.6008	1.1272	0.0130

3.3. Ước lượng entropy, MI và CMI

Từ các công thức tính MI và CMI, ta thấy các độ đo này được định lượng dựa trên entropy, entropy được định lượng dựa trên hàm mật độ. Nếu dữ liệu là rời rạc, ta có thể dễ dàng ước lượng hàm mật độ dựa trên thống kê tần suất. Trong trường hợp dữ liệu liên tục, bài toán trở nên khó khăn hơn. Các phương pháp ước lượng được chia thành hai loại [6, 12]: ước lượng tham số (Bayesian, Maximum Likelihood, Edgeworth,...) và ước lượng phi tham số (Histogram, B-spline, Kernel density, k-nearest neighbours,...). Đối với các phương pháp tham số, tư tưởng chính của phương pháp này là giả định hàm mật độ thuộc một họ hàm nhất định với một tập các tham số kèm theo. Mục đích của phương pháp là tìm các giá trị thích hợp cho các tham số để phù hợp với dữ liệu đầu vào. Trong khi đó, phương pháp phi tham số lại không cần giả định hàm mật độ phải thuộc một họ hàm nhất định. Hiện nay, các phương pháp ước lượng phi tham số được sử dụng rộng rãi vì phương pháp này mang tính tự nhiên hơn. Thật khó để biết trước dữ liệu có phân bố thuộc dạng nào trong khi ta đang cần ước lượng phân bố đó.

Trong nghiên cứu này, chúng tôi sử dụng phương pháp ước lượng entropy, MI và CMI theo phân bố xác suất Gaussian được mô tả trong [1]:

$$P(X_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{(2\pi)^{n/2} \mid C \mid^{n/2}} exp\big(-\frac{1}{2}(X_j - X_i)^T C^{-1}(X_j - X_i)\big) \tag{3.15}$$

Trong đó, C là ma trận hiệp phương sai của biến X; N là số lượng mẫu; n là số lượng biến trong C.

Với ước lượng xác suất trong công thức (3.15), ta có entropy được biểu diễn như sau [14]:

$$H(X) = \log\big[(2\pi e)^{n/2} \mid C \mid^{1/2}\big] = \frac{1}{2}\log(2\pi e)^n \mid C \mid \tag{3.16}$$

Do đó, MI và CMI cũng được ước lượng như sau:

$$MI(X, Y) = \frac{1}{2}\log\frac{\mid C(X) \mid . \mid C(Y) \mid}{\mid C(X, Y) \mid} \tag{3.17}$$

$$MI(X,Y \mid Z) = \frac{1}{2}\log\frac{\mid C(X,Z)\mid .\mid C(Y,Z)\mid}{\mid C(Z)\mid .\mid C(X,Y,Z)\mid} \tag{3.18}$$

Tương tự, các công thức chúng tôi đề xuất (3.6)-(3.9) và (3.11)-(3.14) cũng được biểu diễn như sau:

$$MI(X,Y,Z) = \frac{1}{2}\log\frac{\mid C(X)\mid .\mid C(Y)\mid .\mid C(Z)\mid}{\mid C(X,Y,Z)\mid} \tag{3.19}$$

$$MI(X,[Y,Z]) = \frac{1}{2}\log\frac{\mid C(X)\mid .\mid C(Y,Z)\mid}{\mid C(X,[Y,Z])\mid} \tag{3.20}$$

$$MI(X,Y,Z \mid T) = \frac{1}{2}\log\frac{\mid C(X,T)\mid .\mid C(Y,T)\mid .\mid C(Z,T)\mid}{\mid C(T)\mid^2 .\mid C(X,Y,Z)\mid} \tag{3.21}$$

$$MI(X,[Y,Z] \mid T) = \frac{1}{2}\log\frac{\mid C(X,T)\mid .\mid C([Y,Z],T)\mid}{\mid C(T)\mid .\mid C(X,[Y,Z],T)\mid} \tag{3.22}$$

3.4. Ứng dụng thông tin tương hỗ đa biến có điều kiện để phát hiện tương tác gián tiếp

Trong phần này, chúng tôi sẽ trình bày một ứng dụng của các độ đo đề xuất trong việc phát hiện các quan hệ tương tác gián tiếp. Chúng tôi áp dụng phương pháp đề xuất lên dữ liệu trao đổi chất *in silico* về sự chuyển hóa trong tế bào hồng cầu RBC [5, 9]. Đây là dữ liệu biểu diễn dưới dạng một ma trận 1000×39 mô tả nồng độ của 39 chất chuyển hóa tại 1000 điểm thời gian. Dữ liệu này được tải về tại địa chỉ *http://menem.com/∼ilya/wiki/index.php/RBC_Metabolic_Network*. Mô hình RBC bao gồm 39 chất chuyển hóa và 44 phản ứng.

Chúng tôi đã lập trình trên Matlab, sử dụng các công thức ước lượng đã trình bày trong phần 3.3 để tính toán các giá trị MI. Sau khi đã phát hiện được mối quan hệ tương tác giữa các cặp bốn chất nhờ vào độ đo MI đa biến đã đề xuất, chúng tôi sử dụng độ đo CMI đa biến để phát hiện mối quan hệ tương tác gián tiếp giữa chúng.

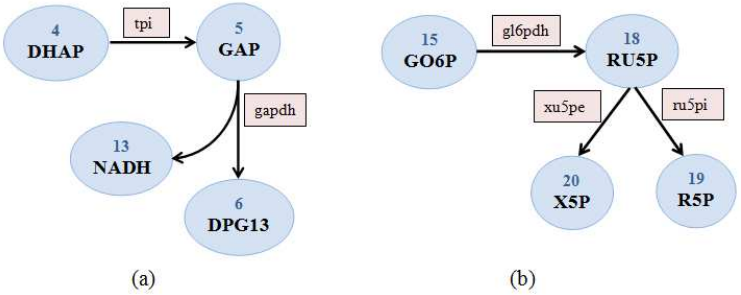
Chẳng hạn, với cặp bốn biến có giá trị thông tin tương hỗ lớn là (4,5,6,13) tương ứng với các chất $(DHAP, GAP, DPG13, NADH)$ liên quan đến hai phản ứng *tpi* và *gapdh*; (15,18,19,20) tương ứng với các chất $(GO6P, RU5P, R5P, X5P)$ liên quan đến phản ứng $gl6pdh, ru5pi$ và $xu5pe$, ta có các giá trị CMI được trình bày trong Bảng 4.

Bảng 4. Phát hiện các tương tác gián tiếp sử dụng CMI đề xuất

Cặp (4,5,6,13)	Cặp (15,18,19,20)
MI(4,[6,13]|5)=0.01	MI(15,[19,20]|18)=0.006
MI(6,[13,4]|5)=0.10	MI(19,[20,15]|18)=0.018
MI(13,[4,6]|5)=0.11	MI(20,[15,19]|18)=0.034
MI(4,[5,13]|6)=6.52	MI(18,[19,20]|15)=4.952
MI(4,[5,6]|13)=6.93	MI(18,[20,15]|19)=1.143

Quan sát các giá trị trong Bảng 4, ta có nhận xét: giá trị $MI(4,[6,13]|5)$ nhỏ nhất, điều đó có nghĩa là giữa $(DHAP, GAP, DPG13, NADH)$ có mối quan hệ tương tác gián tiếp kiểu

$DHAP \leftrightarrow GAP \leftrightarrow [DPG13, NADH]$. Tương tự, giữa $(GO6P, RU5P, R5P, X5P)$ có mối quan hệ gián tiếp $GO6P \leftrightarrow RU5P \leftrightarrow [R5P, X5P]$. Đối chiếu với mô hình RBC đã cho trong [5], ta thấy các quan hệ vừa tìm được hoàn toàn trùng khớp với các phản ứng được mô tả trong Hình 3.1(a) và Hình 3.1(b) của mô hình RBC.



Hình 3.1. Mô hình *RBC* tương ứng với: (a) phản ứng *tpi* và *gapdh*. (b) phản ứng *gl6pdh*, *ru5pi* và *xu5pe*. Trong đó, hình elip biểu diễn các chất, hình chữ nhật biểu diễn các phản ứng.

4. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất các độ đo thông tin tương hỗ đa biến có điều kiện. Bằng các ví dụ minh họa, chúng tôi đã chỉ ra rằng các độ đo thông tin tương hỗ đa biến có điều kiện là hữu ích trong việc xác định quan hệ đa biến gián tiếp thông qua biến trung gian. Khi số biến tăng lên, các loại quan hệ giữa các biến cũng đa dạng, việc xác định các loại quan hệ đa biến là gián tiếp thông qua các biến khác là vấn đề hết sức phức tạp.

Cũng trong nghiên cứu này, chúng tôi đã áp dụng độ đo thông tin tương hỗ đa biến có điều kiện trên dữ liệu trao đổi chất về sự chuyển hóa trong tế bào hồng cầu của người. Kết quả cho thấy, độ đo mà chúng tôi đề xuất có khả năng phát hiện chính xác các quan hệ tương tác gián tiếp mà các phương pháp trước đây có thể đã bỏ sót. Tuy nhiên, trong nghiên cứu này, chúng tôi mới chỉ kiểm chứng mối liên hệ giữa độ đo thông tin tương hỗ đa biến có điều kiện với tương tác gián tiếp chứ chưa chứng minh chặt chẽ về mặt mặt toán học là độ đo này có thể phát hiện tương tác gián tiếp. Đây là vấn đề chúng tôi sẽ còn tiếp tục nghiên cứu.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Quỹ Khoa học và Công nghệ Quốc gia (NAFOSTED), mã số đề tài 102.01-2001.05.

TÀI LIỆU THAM KHẢO

[1]. K. Basso, A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano, "Reverse engineering of regulatory networks in human B cells", *Nature Genetics*, 37 (2005), 382-390.

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

[3]. R. Fano, "A Statistical Theory of Communication", *MIT Press, Cambridge, Massachussets*, 1961.

[4]. F. He, R. Balling and A.P. Zeng, "Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present method and future perspectives", *Journal of Biotechnology*, 144 (2009), 190-203.

[5]. K.J. Kauffman, J.D. Pajerowski, N. Jamshidi, B.O. Palsson and J.S. Edwards, "Description and Analysis of Metabolic Connectivity and Dynamics in the Human Red Blood Cell", *Biophysical*, 83 (2002), 646-662.

[6]. Liam Paninski, "Estimation of Entropy and Mutual Information", *Neural Computation*, 15 (2003), 1191–1253.

[7]. P.M. Magwene and J. Kim, "Estimating genomic coexpression networks using first-order conditional independence", *Genome Biololy*, 5:R100 (2004), DOI:10.1186/gb-2004-5-12-r100.

[8]. P. Meyer, F. Lafitte and G. Bontempi, "A R/Bioconducto Package for Inferring Large Transcriptional Networks Using Mutual Information", *BMC Bioinformatics*, 9 (2008), DOI:10.1186/1471-2105-9-461.

[9]. 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 (2007), 102-115, DOI:10.1196/annals.1407.013.

[10]. Q.D. Nguyen, T.H. Pham, T.B. Ho, V.H. Nguyen and D.H. Tran, "Reconstruction of Triple-wise Relationships in Biological Networks from Profiling Data", *The* 9^{th} *International Conference on Computing and Information Technology-IC*^{2}*IT, Thailand, May.09-10, 2013*, 205-215, DOI:10.1007/978-3-642-37371-8_24.

[11]. T.H. Pham, T.B. Ho, Q.D. Nguyen, D.H. Tran and V.H. Nguyen, "Multivariate Mutual Information Measures for Discovering Biological Networks", *The* 9^{th} *IEEE - RIVF International Conference on Computing and Comunication Technologies Research, Ho Chi Minh city, Vietnam, Feb.27-Mar.01, 2012*, 103-108, DOI:10.1109/rivf.2012.6169834.

[12]. A.F. Villaverde, J. Ross and J.R. Banga, "Reverse Engineering Cellular Networks with Information Theoretic Methods", *Cells*, 2 (2013), 306-329, DOI:10.3390/cells2020306.

[13]. K. Wang, M. Saito, B. Bisikirska, M. Alvarez, W. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. Margolin, U. Klein, R. Favera and A. Califano, "Genome-wide identification of post-translational modulators of transcription factor activity in human B cells", *Nat. Biotechnol*, 27 (2009), 829-839.

[14]. X. Zhang, X. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.K. Hao, Z.P. Liu and L. Chan, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information", *Bioinformatics*, 28 (2012), 98-104.

Ngày nhận bàu 01 - 10 - 2013
Nhận lại sau sửa ngày 12 - 03 - 2014