Vietnam J. Agri. Sci. 2017, Vol. 15, No. 1: 73-84 Tạp chí Khoa học Nông nghiệp Việt Nam 2017, tập 15, số 1: 73-84
www.vnua.edu.vn

NGHIÊN CỨU PHƯƠNG PHÁP DỰA TRÊN MẠNG SINH HỌC ĐỂ DỰ ĐOÁN CÁC GENE GÂY BỆNH

Vũ Thị Lưu[1*], Trần Thị Thu Huyền[1], Nguyễn Văn Hoàng[1], Nguyễn Thị Huyền[1], Lê Đức Hậu[2]

[1]Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam
[2]Khoa Công nghệ thông tin, Đại học Thủy lợi

Email[*]: luuvt207@gmail.com

Ngày gửi bài: 21.12.2016 Ngày chấp nhận: 23.02.2017

TÓM TẮT

Dự đoán gene gây bệnh là một trong những mục tiêu quan trọng trong nghiên cứu y sinh. Hiện nay có khá nhiều phương pháp được xây dựng để dự đoán các gene liên quan đến một số bệnh cụ thể. Tuy nhiên, do mối quan hệ phức tạp giữa các gene và bệnh nên rất nhiều các gene là nguyên nhân gây ra một số bệnh di truyền hiện vẫn chưa được phát hiện ra. Bài toán phân hạng gene để tìm ra các gene bệnh là một trong bài toán được nhiều nhà khoa học quan tâm nghiên cứu. Để tìm ra một phương pháp tốt với mục tiêu là dự đoán được các gene gây bệnh với hiệu suất cao, chúng tôi đã tiến hành khảo sát một số phương pháp phân hạng gene đã có dựa trên mạng sinh học, sau đó đề xuất một phương pháp dự đoán sử dụng mô hình mạng Boolean Network. Trong mạng sinh học các khuyết tật do đột biến về gene/protein có thể gây ra một bệnh nào đó ở người. Cũng chính vì thế, những đột biến của gene/protein này có thể ảnh hưởng đến gene/protein khác thông qua cấu trúc của các mạng sinh học. Phương pháp mới sử dụng mô hình Boolean này đánh giá sự phù hợp của những gene ứng viên đối với một bệnh nào đó quan tâm bằng cách đo mức độ ảnh hưởng đột biến từ gene gây bệnh đã biết tới các gene ứng viên. Dựa trên giá trị này để sắp xếp các gene ứng viên sao cho các gene có khả năng liên quan tới bệnh được nhận thứ hạng cao hơn. Sau khi phân hạng, một nhóm nhỏ các gene với thứ hạng cao sẽ được lựa chọn để kiểm nghiệm bằng thực nghiệm để xác định có liên quan đến bệnh. Các kết quả giả lập trên một bộ dữ lieu gene - bệnh đã cho thấy rằng phương pháp đề xuất của chúng tôi tốt hơn phương pháp dựa trên giải thuật ngẫu nhiên - Random Walk Restart. Sử dụng phương pháp đề xuất, kết quả thử nghiệm đã xác định được 27 gene có liên quan đến bệnh ung thư vú.

Từ khóa: Gene bệnh, mạng tín hiệu sinh học, Boolean động, phương pháp dựa trên mạng, thuật toán ngẫu nhiên.

Study Method Base on Biological Networks for Disease Candidate Gene Prediction

ABSTRACT

Predicting genes which may associate with disease is one of the important goals of biomedical research. There have been many computational methods developed to rank genes involved in a particular disease. However, due to the complex relationship between genes and the diseases, many genes that cause genetic diseases have not yet been discovered. The problem of ranking genes to identify the disease-associated gene has drawn attention of many researchers. To find a good method to predict target genes that cause diseases with high performance, we have conducted a survey of prediction methods based on biological network. We then proposed a new method using a Boolean network model. In biological network, defects by mutations on genes/proteins may cause a disease to occurin a person. Also, these mutations may affect other genes/proteins through structures of the biological networks. In this study, we proposed to use Boolean network model to assess the relevance of candidate genes to a disease of interest by measuring the degree of mutational effect from known disease-associated genes to candidate genes. Particularly, we mutated known disease-associated genes and measured the effect of this mutation on candidate genes based on Boolean dynamics of biological networks. Based on this measured value, candidate genes can be prioritized and finally top-ranked candidate genes can be selected as novel promising disease genes. Simulation results on a set of diseases showed that the proposed method is superior to a state-of-the-art one, which is based on a random walk with a restart algorithm. Using the proposed method, we have identified 27 genes associated with breast cancer with evidences from literature.

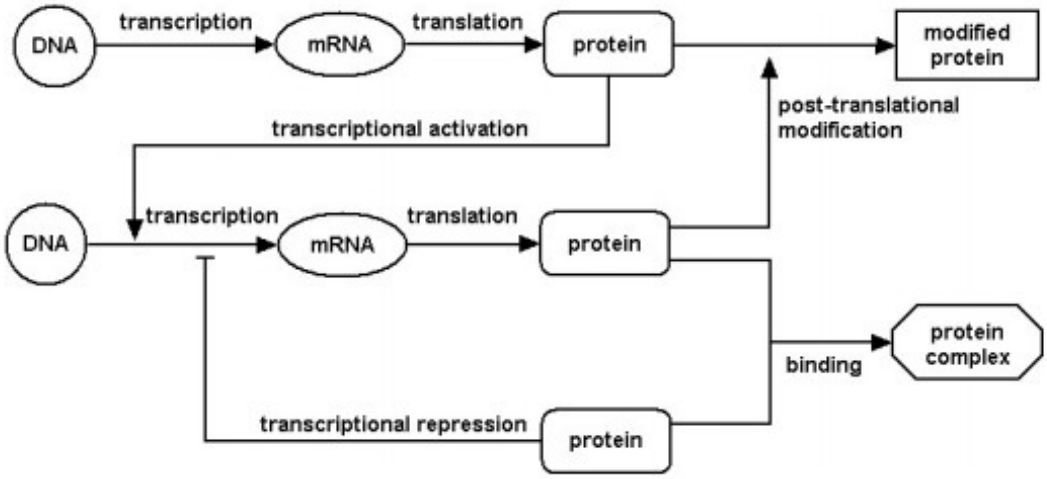
Keywords: Disease candidate gene prioritization, human signaling network, Boolean dynamics, network-based method, random walk with restart algorithm.

1. ĐẶT VẤN ĐỀ

Xác định gene gây bệnh là bài toán quan trọng trong y sinh học và sinh học phân tử. Để dự đoán gene bệnh đã có một số phương pháp được đề xuất (Kann, 2010). Trước đây, việc xác định gene gây bệnh được thực hiện chủ yếu bằng các thực nghiệm sinh học. Phương pháp này được thực hiện cho hàng trăm gene ứng viên nằm trên một vùng nhiễm sắc thể khả nghi nên đòi hỏi nhiều thời gian và chi phí rất cao. Phân hạng gene là sử dụng các phương pháp tính toán để sắp xếp các gene ứng viên sao cho các gene có khả năng liên quan tới bệnh được nhận thứ hạng cao hơn. Sau khi phân hạng, một nhóm nhỏ các gene với thứ hạng cao sau đó sẽ được lựa chọn để kiểm nghiệm bằng thực nghiệm.

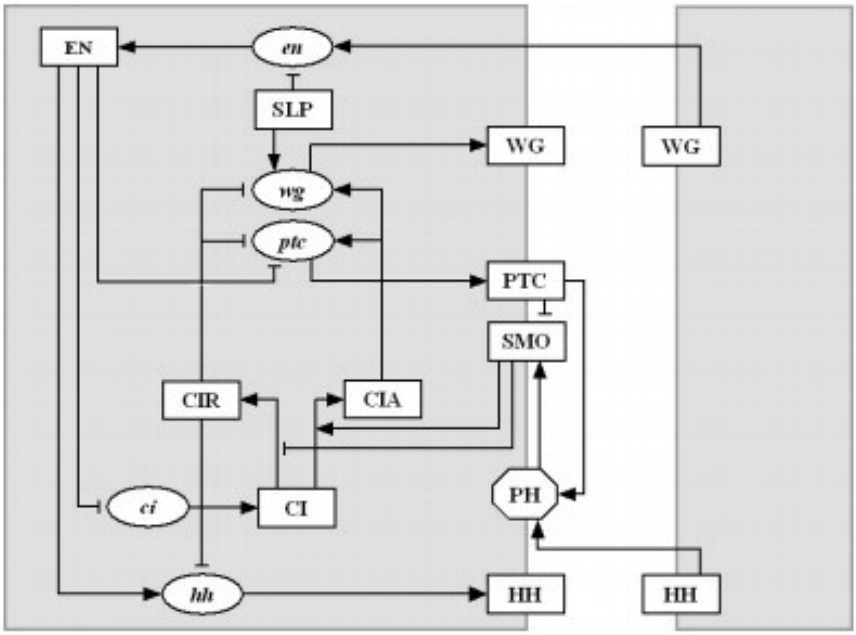
Các phương pháp phân hạng gene ứng viên đã được đề xuất có thể chia làm 3 hướng chính: i) Dựa trên đánh dấu nhãn chức năng; ii) dựa trên mạng và iii) dựa trên học máy. Trong đó, các phương pháp dựa trên đánh dấu nhãn chức năng phân hạng các gene ứng viên bằng cách đo mức độ tương tự của mỗi gene ứng viên tới một tập hợp các gene gây bệnh đã biết dựa trên các hồ sơ được xây dựng từ nhiều nguồn dữ liệu (Aerts, 2006). Do đó, những phương pháp này chủ yếu tập trung vào việc tích hợp nhiều bộ dữ liệu sinh học khác nhau để có được sự tương tự chính xác hơn để bao phủ toàn bộ hệ gene người. Bên cạnh các phương pháp dựa trên đánh dấu nhãn chức năng, các phương pháp dựa trên học máy với phân lớp nhị phân để xác định các gene bệnh tương ứng cũng đã được nghiên cứu. Ở thời kỳ đầu, những nghiên cứu dựa trên học máy thường tiếp cận dự đoán gene bệnh như bài toán phân lớp nhị phân. Một số kỹ thuật phân lớp nhị phân đã được đề xuất cho vấn đề này như: cây quyết định (Adie, 2005), k-láng giềng gần nhất (Li, 2006), phân loại Naïve Baysian (Calvo, 2006), mạng nơron nhân tạo (Sun, 2009) và máy vector hỗ trợ (Keerthikumar, 2009). Trong những nghiên cứu này, các mẫu học bao gồm mẫu huấn luyện tích cực và tiêu cực. Trong đó, mẫu huấn luyện tích cực được xây dựng từ các gene gây bệnh đã biết, mẫu huấn luyện tiêu cực là gene chưa xác định là liên quan với bệnh. Đây là hạn chế của các giải pháp phân lớp nhị phân cho bài toán dự đoán gene bệnh vì tập huấn luyện tiêu cực không thực sự là những gene không liên quan tới bệnh. Tuy nhiên, việc xây dựng tập dữ liệu này là gần như không thể trong các nghiên cứu y sinh bởi vì trong y sinh những trường hợp không quan sát được liên kết không có nghĩa là liên kết đó không tồn tại. Do vậy, để giảm sự không chắc chắn này của các phương pháp trước đó, phương pháp bán giám sát đã được đề xuất cho vấn đề, trong đó bộ phân loại được học từ cả hai: dữ liệu có nhãn (ví dụ, gene bệnh đã biết) và không có nhãn (ví dụ, các gene chưa biết). Tuy nhiên, các mẫu tiêu cực vẫn phải được xác định trong các nghiên cứu này. Để khắc phục những hạn chế của cả hai phương pháp nêu trên, các phương pháp dựa trên mạng để xác định gene bệnh đã được đề xuất (Wang, 2011). Những phương pháp này chủ yếu dựa trên các mạng sinh học như mạng tương tác protein được sử dụng khá phổ biến do dữ liệu tương tác protein/gene ngày càng đầy đủ và đa dạng. Ngoài ra, phương pháp này vượt trội so với hai tiếp cận trước là bởi vì nó dựa trên nguyên lý "module bệnh" (chẳng hạn các gene/protein gắn kết với cùng một bệnh hoặc các bệnh tương tự thường có xu hướng nằm gần nhau trong mạng tương tác gene/protein). Hơn nữa, phương pháp dựa trên mạng này nhắm tới bản chất của vấn đề dự đoán gene bệnh, phân hạng thay vì phân loại các gene ứng viên (ví dụ, gán nhãn cho một gene ứng viên là gene bệnh hoặc không) như các phương pháp dựa trên học máy thực hiện.

Theo lý thuyết, các khiếm khuyết trong đột biến trên gene/protein có thể là nguyên nhân gây ra một bệnh ở một người. Những đột biến này có thể ảnh hưởng đến gene/protein khác thông qua các cấu trúc của các mạng sinh học. Các đột biến trên gene/protein ảnh hưởng đến mạng sinh học được điều khiển bởi thuộc tính cấu trúc của mạng sinh học. Dưới đây là hình ảnh quá trình điều tiết gene trong mạng điều hòa.



Hình 1. Mô hình mạng điều hòa điều tiết hoạt động gene

Ghi chú: Các thành phần trong mạng có thể ảnh hưởng (tác động tích cực hoặc tiêu cực) tới nhau trong quá trình phiên mã và phiên dịch thành các chất tương ứng.



Hình 2: Một ví dụ về mạng lưới các tương tác giữa các gene phân đoạn của ruồi giấm

Nguồn: Albert, Boolean Modeling of Genetic Regulatory Networks, 2004.

Trên hình 2 thể hiện hình ảnh của các nút với các chất tương ứng: mRNAs (elip), proteins (chữ nhật), protein complexes (bát giác). Đường tín hiệu của mạng thể hiện phản ứng sinh hóa (phiên dịch) hoặc tương tác điều tiết (phiên mã). Đường mũi tên và đường có nét ngang đầu dòng đại diện tương tác tích cực và tiêu cực tương ứng

Dựa vào sự ảnh hưởng của các quá trình chuyển hóa trong mạng, đột biến trên một nút có thể làm ảnh hưởng đến các giá trị của các nút khác. Chính vì thế trong nghiên cứu này nhóm nghiên cứu đi khảo sát một số phương pháp dự trên mạng sinh học để phân hạng và tìm ra gene bệnh. Sau đó chúng tôi đi nghiên cứu và sử

dụng mô hình Boolean Network trong mạng điều hòa gene để đo những tác động từ gene bệnh đã biết (được gọi là gene đích) tới gene ứng viên là gene có nguy cơ liên quan đến bệnh. Dựa trên giá trị này, các gene ứng viên có thể được sắp xếp theo thứ tự ưu tiên và gene được xếp hạng đầu tiên có thể được lựa chọn như gene bệnh mới để thử nghiệm. Kết quả so sánh hiệu suất của phương pháp này với thuật toán RWR (Random Walk with Restart) trên một tập gồm 25 bệnh cho thấy phương pháp có sử dụng mô hình mạng Boolean Network thực hiện tốt hơn phương pháp RWR. Áp dụng phương pháp đề xuất để xác định gene liên quan đến bệnh ung thư và kết quả lựa chọn 27 trong số 50 gene ứng viên được xếp hạng cao và được chứng minh là liên quan đến bệnh ung thư vú. Kết quả nghiên cứu trên đạt được dựa trên các bộ dữ liệu về mạng sinh học và các liên kết gene bệnh đã biết được thu thập và nghiên cứu theo phương pháp được trình bày phần tiếp theo của bài báo.

2. CÁC NGHIÊN CỨU LIÊN QUAN VÀ GIẢI PHÁP PHÂN HẠNG GENE ĐỀ XUẤT

Trong phần này chúng tôi khảo sát một số phương pháp dựa trên mạng đã được các nhà nghiên cứu công bố trong và ngoài nước. Sau đó chúng tôi đi nghiên cứu cụ thể về mô hình Boolean network thuộc mạng điều hòa gene được đề xuất vào năm 1969 bởi Kauffmann và thử nghiệm trên bộ dữ liệu thu thập được. Sau đó so sánh hiệu suất với mô hình RWR đã được các nhóm nghiên cứu kiểm nghiệm (Le, 2012).

2.1. Các nghiên cứu liên quan đến bài toán dự đoán gene bệnh dựa trên phương pháp mạng sinh học

a. Phương pháp tính toán dựa vào tích hợp mạng tương tác protein, mạng tương tác microRNA-gene và các gene gây bệnh ung thư đã biết của tác giả Trần Thị Bích Phương và cộng sự đã nghiên cứu năm 2013 (Trần Thị Bích Phương, 2013)

Phương pháp này nhằm tích hợp thông tin từ nhiều nguồn dữ liệu khác nhau thành một mạng tổng thể, sau đó đưa ra một cách phân tích mạng tương tác này để biểu diễn các gene thông qua các gene khác có tương tác trực tiếp và gián tiếp với nó. Ý tưởng chính là tìm cách biểu diễn được sự tương tác của một gene với các gene khác để từ đó đoán nhận khả năng gây bệnh của nó. Điểm mới trong phương pháp này là tích hợp thông tin mạng tương tác microRNA-gene vào mạng tương tác protein và đưa ra một cách phân tích mạng tương tác dựa trên thuật toán tìm kiếm theo chiều rộng để biểu diễn các nút trên mạng. Phương pháp được thử nghiệm trên dữ liệu thực tế được download từ các trung tâm dữ liệu sinh học trên thế giới và sử dụng các phương pháp phân lớp phổ biến (SVM, C4.5, K-NN) để đánh giá hiệu quả. Kết quả kiểm nghiệm trên dữ liệu thực cho thấy độ chính xác của các phương pháp dự đoán được nâng lên. Điều này chứng tỏ thông tin về microRNA là hữu ích trong việc tiên lượng các gene gây bệnh. Cụ thể của kết quả được công bố trên bài báo trên tạp chí khoa học của trường Đại học Sư phạm Hà Nội (Trần Thị Bích Phương, 2013).

b. Cách tiếp cận khác sử dụng xác suất tiền nghiệm là PRINCE

PRINCE (PRIoritizatioN and Complex Elucidation) được phát triển bởi Vanunu *et al.* (2010). PRINCE sử dụng thuật toán lan truyền để dự đoán gene bệnh dựa vào thông tin tích hợp giữa kiểu hình bệnh và mạng tương tác protein. Phương pháp này tính toán mối liên quan giữa một bệnh và gene bệnh đã biết với một bệnh khác sử dụng hàm logistic dựa trên sự tương tự kiểu hình giữa hai bệnh. Gene liên quan tới bệnh sau đó được sử dụng như xác suất tiền nghiệm để xây dựng chức năng phân hạng gene.

c. Phương pháp phân hạng RL_Rank và RL_Rank with priors

Phương pháp phân hạng mới RL_Rank được đề xuất bởi Derhami *et al.* (2013) dựa trên sự liên kết của các nút trong đồ thị và khái niệm về học tăng cường để phân hạng các trang Web. Xuất phát từ sự thành công của các thuật toán trên trong việc sử dụng "thứ hạng đầu" hay xác suất

tiền nghiệm, để biến độ quan trọng tuyệt đối của các nút trong mạng thành độ quan trọng tương đối/độ tương tự của các nút đối với một tập các nút gốc. Sau đó Đặng Vũ Tùng và cs. (2015)sự đã cải tiến thuật toán RL_Rank thành thuật toán RL_Rank with priors bằng cách bổ sung thêm các xác suất tiền nghiệm nhằm mục đích nâng cao hiệu quảcủa thuật toán. Thuật toán này được cài đặt và thử nghiệm cho bài toán phân hạng và tìm kiếm gene gây bệnh dựa trên bộ dữ liệu mạng tương tác protein. Kết quả thực nghiệm cho thấy độ chính xác của phương pháp đề xuất tốt hơn so với phương pháp PageRank with priors trên cùng bộ dữ liệu thử nghiệm.

d. Thuật toán RWR (Random Walk Restart)

Thuật toán RWR khai thác cấu trúc tổng thể của mạng dựa trên hành vi của một chuyển động ngẫu nhiên trên một mạng hay đồ thị (Lovasz, 1996). Theo hành vi này, một thực thể xuất phát từ một nút khởi đầu sau đó di chuyển trên đồ thị bằng cách chuyển đến các nút lân cận một cách ngẫu nhiên với xác suất tỷ lệ với trọng số của các cạnh kết nối. Tập hợp các nút trong quá trình di chuyển là một chuỗi Markov và được gọi là một bước ngẫu nhiên trên đồ thị (random walk on graph) (Duc-Hau Le, 2012). Tại thời điểm bất kỳ trong quá trình di chuyển, thực thể cũng có thể quay lại nút khởi đầu với một xác suất nhất định được gọi là xác suất quay lại (back-probability) $\gamma \in (0, 1)$. Các nút được thăm nhiều hơn được coi là có độ quan trọng lớn hơn. Đại lượng này đánh giá tầm quan trọng tương đối/độ tương tự của các nút còn lại so với tập các nút gốc. Ưu điểm chính của phương pháp bước ngẫu nhiên là tốc độ thực hiện nhanh do đó có thể áp dụng cho các mạng có kích thước lớn. Khi áp dụng thuật toán này cho bài toán phân hạng gene gây bệnh, các gene gây bệnh đã biết đóng vai trò như các nút khởi đầu, các gene còn lại trên mạng được xem là các ứng viên. Kết quả thử nghiệm trên một tập gồm 25 kiểu hình bệnh cho thấy phương pháp này đạt được hiệu năng dự đoán tốt. RWR có thể được mô tả bằng công thức như sau:

$$P^{t+1} = (1 - \gamma)\, W'P^t + \gamma P^0$$

Trong đó, P^t là một vector xác suất N×1 của |V| nút tại bước thời gian *t*, trong đó, phần tử thứ *i* đại diện cho xác suất người đi đang ở nút $v_i \in V$ và P^0 là vector xác suất khởi tạo $N \times 1$ mà giá trị của một phần tử tương ứng là 0 hoặc $1/|S|$ tương ứng với nút đó không là nút nguồn hoặc là nút nguồn. S là tập nút nguồn. Ma trận *W'* là ma trận xác suất chuyển vị, vì thế $(W')_{ij}$ biểu thị xác suất người đi di chuyển từ v_i tới v_j trong $V \setminus \{v_i\}$ nút có thể. Về hình thức, đối với một mạng không trọng số, nó là được xác định như sau:

$$(W')_{ij} = \frac{1}{|(V_{out})_i|}$$

với (V_{out}) tập nút đi ra v_i.

2.2. Phân hạng gene ứng viên dựa trên mô hình mạng boolean động

2.2.1. Các bộ dữ liệu về mạng sinh học và các liên kết gene bệnh đã biết

Để đánh giá tác động đột biến trên các gene liên quan tới bệnh đã biết tới các gene ứng viên, bài báo sử dụng mạng tín hiệu của người với quy mô lớn thu được từ một nghiên cứu đã được

Cột 1 (Mã gene)	Cột 2 (Tương tác)	Cột 3 (Mã gene)	Mô tả tương tác
10971	-1	572	inhibition
10971	1	572	activation
10971	0	572	neutral
10971	-1	581	inhibition
…	…	…	…
7472	1	23509	activation
8882	1	29843	activation
8409	1	11143	activation

công bố (Cui, 2009). Mạng tín hiệu này gồm 1.539 nút và 4.754 tương tác. Ngoài ra, một tập bệnh và các gene liên quan đến bệnh đã biết được thu thập từ nghiên cứu về OMIM (Amberger, 2009). Do sử dụng phương pháp kiểm tra chéo bỏ ra một (LOOCV: leave-one-out cross validation) để đánh giá hiệu năng của các phương pháp phân hạng, nên mỗi bệnh phải có ít nhất hai gene liên quan. Do đó từ cơ cở dữ liệu mạng tương tác protein và quan hệ bệnh-gene, chúng tôi lọc ra được 25 bệnh gây ra bởi từ hai gene trở lên và các gene này có trong mạng tương tác protein. Dữ liệu về mạng tín hiệu được thể hiện dưới *bảng 1* như sau:

Chúng tôi sử dụng phần mềm Cytoscape phiên bản 2.7 để mô phỏng mạng tín hiệu trên.

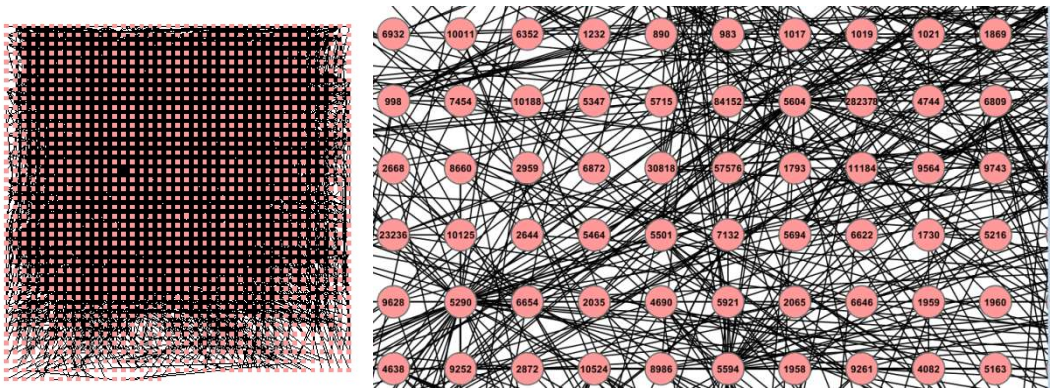
2.2.2. Phân hạng gene ứng viên dựa trên mô hình mạng boolean động

Để định nghĩa độ đo đánh giá những tác động đột biến trên các gene liên quan đến bệnh đã biết tới các gene ứng viên, trong nghiên cứu này chúng tôi sử dụng mô hình mạng Boolean Network, đây là mô hình đã được sử dụng rộng rãi để biểu diễn các mạng sinh học và đã diễn đạt thành công một vài thuộc tính sinh học (Kauffman, 2003; Kwon, 2011). Đặc biệt, nó cũng được sử dụng thường xuyên trong giả lập hành vi động của nhiều mạng tín hiệu khác nhau (Kwone *et al.*, 2011; Trinh, 2014). Dưới đây là mô hình mạng điều hòa gene thể hiện quá trình chuyển hóa và các tác động giữa các thành phần trong mạng.

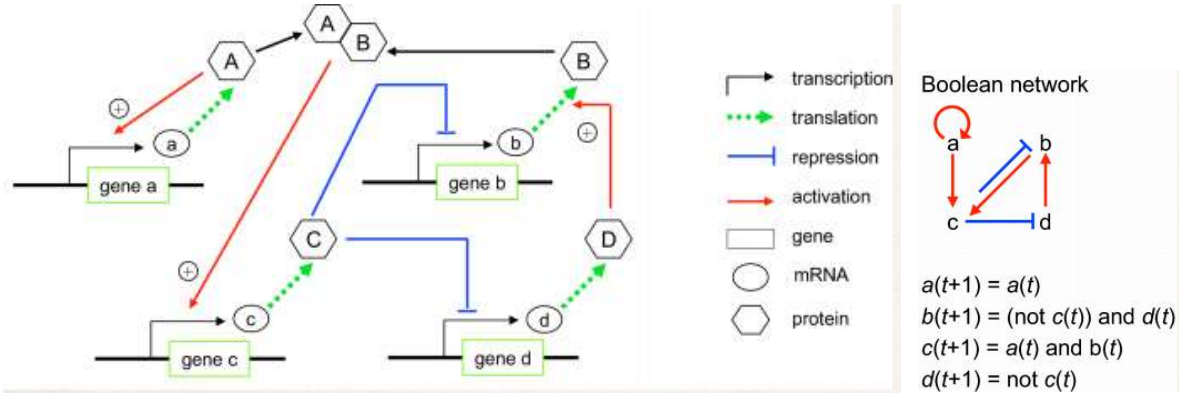
Nhìn vào sơ đồ mạng điều hòa gene trên ta có thể xây dựng được một đồ thị có hướng với các đường tín hiệu như ở hình 4. Trong đó protein A tác động tích cực đến quá trình gene a phiên mã tạo thành mRNA a. Protein B tác động tích cực đến quá trình gene c phiên mã tạo thành mRNA c. Protein C tác động tiêu cực (ức chế) quá trình gene b phiên mã tạo thành mRNA b và gene d phiên mã tạo thành mRNA d. Protein D tác động tiêu cực (ức chế) đến quá trình mRNA b tạo thành protein B.

Bảng 1. Dữ liệu mạng tín hiệu gồm 1539 nút và 4754 tương tác

Cột 1 (Mã gien)	Cột 2 (Tương tác)	Cột 3 (Mã gien)	Mô tả tương tác
10971	-1	572	inhibition
10971	1	572	activation
10971	0	572	neutral
10971	-1	581	inhibition
…	…	…	…
7472	1	23509	activation
8882	1	29843	activation
8409	1	11143	activation



Hình 3. Mô phỏng liên kết trong mạng tín hiệu gồm 1.539 nút và 4.754 tương tác

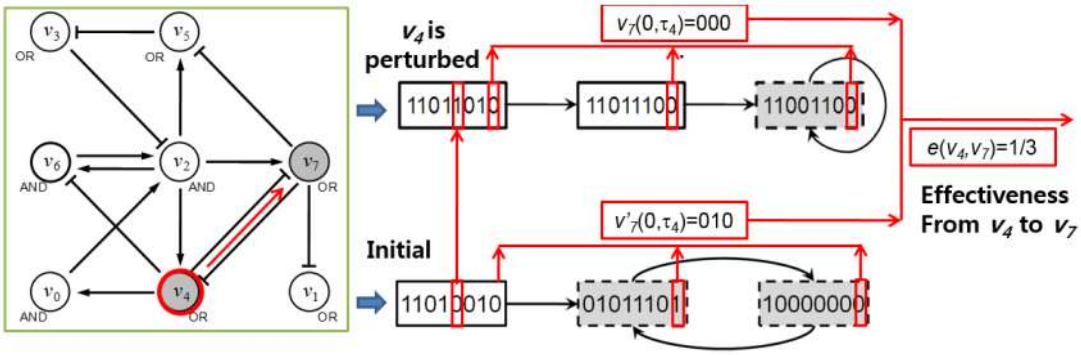


Hình 4. Một ví dụ về mô hình mạng điều hòa gene biểu diễn bằng mạng Boolean network với các luật cập nhật tương ứng

Mạng Boolean được biểu diễn bởi một đồ thị có hướng *G(V, E)*, mỗi $v_i \in V$ có giá trị là 1 ("bật") hoặc 0 ("tắt"), biểu diễn cho các trạng thái có thể của các phần tử tương ứng. Giá trị của mỗi biến v_i tại thời điểm *t+1* được xác định bởi giá trị của k_i các biến khác $v_{i_1}, v_{i_2}, \ldots, v_{i_{k_i}}$ liên kết đến v_i tại thời gian t bằng hàm Boolean f_i: $\{0,1\}^{k_i} \rightarrow \{0,1\}$. Do đó, có thể viết lại: $v_i(t+1) = f_i(v_{i_1}(t), v_{i_2}(t), \ldots, v_{i_{k_i}}(t))$, ở đây chúng tôi chọn ngẫu nhiên hoặc logic hội hoặc logic tuyển cho tất cả các liên hệ tín hiệu trong f_i với phân bố đều. Ví dụ, nếu biến Boolean *v* có quan hệ tích cực từ v_1, có quan hệ tiêu cực từ v_2, và có quan hệ tích cực với v_3 thì luật cập nhật theo logic hội và logic tuyển tương ứng là $v(t+1) = v_1(t) \cap \overline{v_2}(t) \cap v_3(t)$ và $v(t+1) = v_1(t) \cup \overline{v_2}(t) \cup v_3(t)$. Trong trường hợp logic hội, giá trị của v tại thời điểm *(t+1)* là 1 chỉ khi các giá trị của v_1, v_2, v_3 tại thời điểm *t* tương ứng là 1, 0, 1, trong khi trường hợp logic tuyển, giá trị của v tại thời điểm *t+1* là 1 nếu có ít nhất một trong các trạng thái của mệnh đề $v_1(t)$, $\overline{v_2}(t)$, và $v_3(t)$ là 1. Mặc dù có nhiều hàm logic khác có thể sử dụng ngoài logic hội và logic tuyển, trong nhiều nghiên cứu trước, các mạng sinh học đã được miêu tả thành công bằng mô hình Boolean sử dụng chỉ hai hàm logic hội và logic tuyển (Albert, 2004; Faure, 2006). Ngoài ra, tín hiệu của mỗi liên kết được xác định là tích cực hay tiêu cực một cách hoàn toàn ngẫu nhiên. Tuy nhiên trong mạng tín hiệu thu thập để thử nghiệm cho mô hình này thì kiểu tương tác là thực giữa các gene (activation, inhibition và neutral).

Cho một mạng Boolean với N biến Boolean, $v_1, v_2, \ldots, v_n$, chúng ta định nghĩa trạng thái mạng là một vector gồm giá trị của các biến Boolean: tất cả có 2^N trạng thái có thể. Trạng thái này chuyển sang trạng thái khác thông qua một tập *N* hàm Boolean cập nhật: $f_1, f_2, \ldots, f_n$. Chúng ta có thể xây dựng sơ đồ chuyển trạng thái biểu diễn sự dịch chuyển của mỗi trạng thái. Một lộ trình trạng thái bắt đầu từ trạng thái khởi tạo và cuối cùng hội tụ đến một điểm cố định hoặc một attractor vòng hữu hạn. Ngoài ra, chúng ta xác định một chuỗi chuyển giá trị của node v như sau: Khi mạng Boolean *G(V, E)* khởi tạo với $v_1(0), v_2(0), \ldots$, và $v_n(0)$ tại thời điểm khởi đầu *0*, $v_i(t_0, t_1)$ biểu diễn cho chuỗi các giá trị tạm thời của nút v_i trong khoảng thời gian từ t_0 đến t_1.

Trong mô hình trên hình 5 mô phỏng mạng Boolean với 8 nút và 14 liên kết, đường mũi tên và đường có nét ngang đầu dòng đại diện tương tác tích cực và tiêu cực tương ứng. 'AND' và 'OR' biểu thị các hàm cập nhật hội và tuyển tương ứng. Quỹ đạo bắt đầu từ trạng thái ban đầu (11010010) tới trạng thái khác (11011010) khi v_4 chịu sự thay đổi của trạng thái ban đầu. Trạng thái của mạng (chuỗi tám-bit trong hình chữ nhật) đại diện cho các giá trị của v_0 qua v_7 trong chuỗi và hình chữ nhật màu xám với các đường nét đứt. Các trạng thái và quỹ đạo được tính



Hình 5. Ví dụ minh họa về tính độ đo ảnh hưởng các nút trong mạng Boolean

toán từ mạng. Sau đó, tính toán độ đo ảnh hưởng từ v_4 để v_7. Luật cập nhật và các hàm chuyển trạng thái tương ứng các trường hợp khởi tạo ngẫu nhiên: *Ảnh hưởng từ nút tới nút trong mạng boolean ngẫu nhiên*

Trong các mạng Boolean, chúng tôi đề xuất một độ đo mới, gọi là effectiveness, để định lượng sự ảnh hưởng từ một node tới một node khác trong cư xử động của mạng. Để định nghĩa độ đo này, trước tiên chúng ta định nghĩa hai loại dao động, dao động trạng thái khởi tạo và dao động theo luật cập nhật. Cho một mạng Boolean khởi tạo với $v_0(0)$, $v_1(0)$,..., và $v_n(0)$, dao động khởi tạo tại nút $v_i \in V$ là việc chuyển từ $v_i(0)$ sang $\bar{v}_i(0)$. Mặt khác, dao động theo luật cập nhật tại $v_i \in V$ là việc thay đổi luật cập nhật ở v_i từ logic hội sang logic tuyển hoặc ngược lại, tùy thuộc vào kiểu hàm đang được sử dụng. Giả định dao động ở v_i, chúng ta định nghĩa độ đo effectiveness từ v_i tới v_j là $e(v_i, v_j)$, như sau:

1. Gọi $\hat{o}_i$ là thời gian hội tụ hợp lệ của v_i, $\tau_i = max\ \{T_i, T'_i\}$ trong đó T_i hoặc T'_i là các bước thời gian để mạng hội tụ tới một attractor khi v_i bị dao động và không dao động tương ứng.

2. Chúng ta nhận được hai chuỗi dịch chuyển khác nhau của v_j, $v_j(0, \hat{o}i)$ và $v'_j\ (0, \tau_i)$ tương ứng với khi v_j chịu dao động và không.

3. Sau đó, chúng ta tính $e(v_i, v_j) = d(v_j(0, \tau_i), v'_j(0, \tau_i))/\hat{o}_i$ trong đó $d(\cdot)$ là khoảng cách Hamming (tức là số bit có giá trị khác nhau) giữa hai chuỗi. Do đó, $e(v_i, v_j)$ thể hiện độ lớn của hành trình đối với v_j khi bị ảnh hưởng bởi dao động ở v_i. Điều này cũng đo lường ảnh hưởng đột biến của v_i tới v_j.

Vì vậy, effectiveness là độ đo ảnh hưởng ở mỗi nút khi bị ảnh hưởng bởi dao động ở nút khác trong khía cạnh động. Trong mạng Boolean, một nút được gọi là một nút chức năng quan trọng nếu một dao động tại nút đó hoặc không có dao động tại nút đó đều làm mạng hội tụ tới một attractor khác với attractor ban đầu. Trong lĩnh vực này, *các gene bệnh có thể được coi là nút quan trọng trong mạng tín hiệu* và effectivness trong các mạng Boolean có thể được sử dụng để biểu đạt sự ảnh hưởng trên các gene ứng viên khi các gene gắn kết với bệnh đã biết bị đột biết trong các mạng hiệu. Hình 1 cho thấy ví dụ về việc tính toán sự ảnh hưởng của v_4 tới v_7. Để tính $e(v_4, v_7)$, chúng ta nhận được hai chuỗi dịch chuyển của v_7 là $v_7\ (0, \hat{o}_4)$ và $v_7'(0, \hat{o}_4)$ tương ứng với khi v_4 bị dao động và không bị dạo động.

Ảnh hưởng từ một tập nút tới một nút trong mạng boolean ngẫu nhiên

Theo cách tương tự, cho một trạng thái khởi tạo, ảnh hưởng từ tập nút S tới một nút có thể được tính toán bằng cách áp dụng các dao động lên tất cả các nút trong tập một cách đồng thời, sau đó tính toán ảnh hưởng từ các nút đến một nút v_j là $e(S, v_j)$. Đối với tập các trạng thái ban đầu I_s, ảnh hưởng từ tập các nút đến một nút được xác định như sau:

$$\bar{e}(S, v_i) = \frac{\sum_{I_s} e(S, v_i)}{|I_s|}$$

Đối với việc xác định các gene liên quan bệnh, S là một tập gene đã biết của một bệnh nào đó, v_i là 1 gene trong tập các gene ứng viên, $\bar{e}(S, v_i)$ đo ảnh hưởng từ tập các gene bệnh đã

biết tới gene ứng viên. Do đó, các gene ứng viên có thể được xếp hạng dựa trên mức độ liên quan tới bệnh thông qua độ đo này.

2.2.3. Phương pháp đánh giá hiệu suất

Hiệu suất xếp hạng được đánh giá thông qua quá trình kiểm tra chéo bỏ ra một gọi tắt là LOOCV (leave-one-out cross-validation). Đối với mỗi kiểu hình bệnh (*d*), mỗi vòng LOOCV lấy ra một gene s liên quan tới bệnh d đã biết. Phần còn lại của tập gene liên quan tới bệnh d được đặc tả như tập các nút nguồn (tức là, *S\{s}).* Các gene ứng viên trong mạng tín hiệu ký hiệu W. Các gene bị lấy ra và các gene còn lại trong mạng tín hiệu vốn không biết là có liên quan với bệnh *d* hay không thì được xếp hạng theo hai phương pháp nêu trên. Sau đó, chúng tôi vẽ đồ thị đường cong *ROC* và tính toán diện tích dưới đường cong *AUC* (Hanley, 1982) để so sánh hiệu suất của hai phương pháp này. Đường cong này đại diện cho các mối quan hệ giữa độ nhạy (sensitivity) và độ đặc hiệu (specificity). Mỗi cặp giá trị sensitivity và (1-specificity) tương ứng với một điểm trên đường cong ROC, trong đó độ nhạy là tỷ lệ phần trăm của các gene liên quan đến bệnh đã biết được xếp hạng trên một ngưỡng cụ thể và độ đặc hiệu là tỷ lệ phần trăm của các gene vốn chưa biết là có liên với bệnh hay không được xếp hạng ở dưới ngưỡng này. Giá trị sensitivity và 1-specificity được tính toán theo các công thức:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

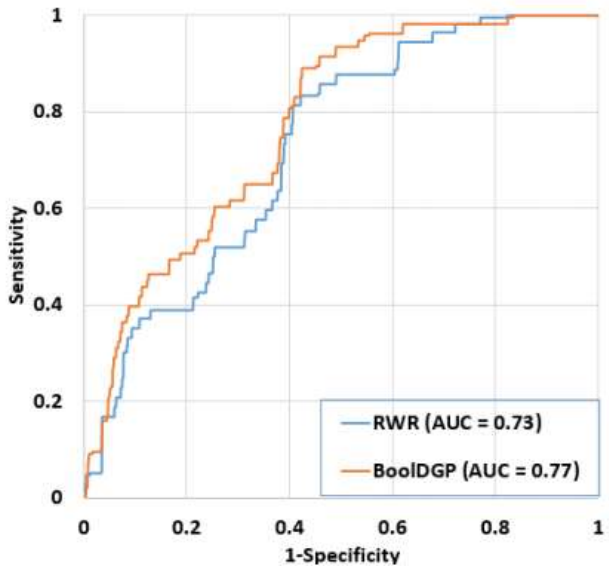
$$1 - \text{specificity} = \frac{FP}{FP + TN}$$

Trong đó: TP (true positive) là số trường hợp thử nghiệm mà thứ hạng của s ≤ ô, FN (false negative) là số trường hợp thử nghiệm mà thứ hạng của s > ô, FP (false positive) là sốtrường hợp thử nghiệm mà thứ hạng của w ≤ ô (với mỗi w ∈ W) và TN (true negative) là số trường hợp thử nghiệm mà thứ hạng của w > ô (với mỗi w∈ W). Một cặp giá trị sensitivity và 1-specificity tương ứng với một điểm trên đường cong ROC.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Đánh giá hiệu suất

Để đánh giá hiệu suất của phương pháp sử dụng mô hình Boolean Network này đối với bài toán xếp hạng các gene ứng viên liên quan tới bệnh, trong nghiên cứu thử nghiệm ta chọn ngẫu



Hình 6. So sánh hiệu suất giữa Boolean Network và RWR

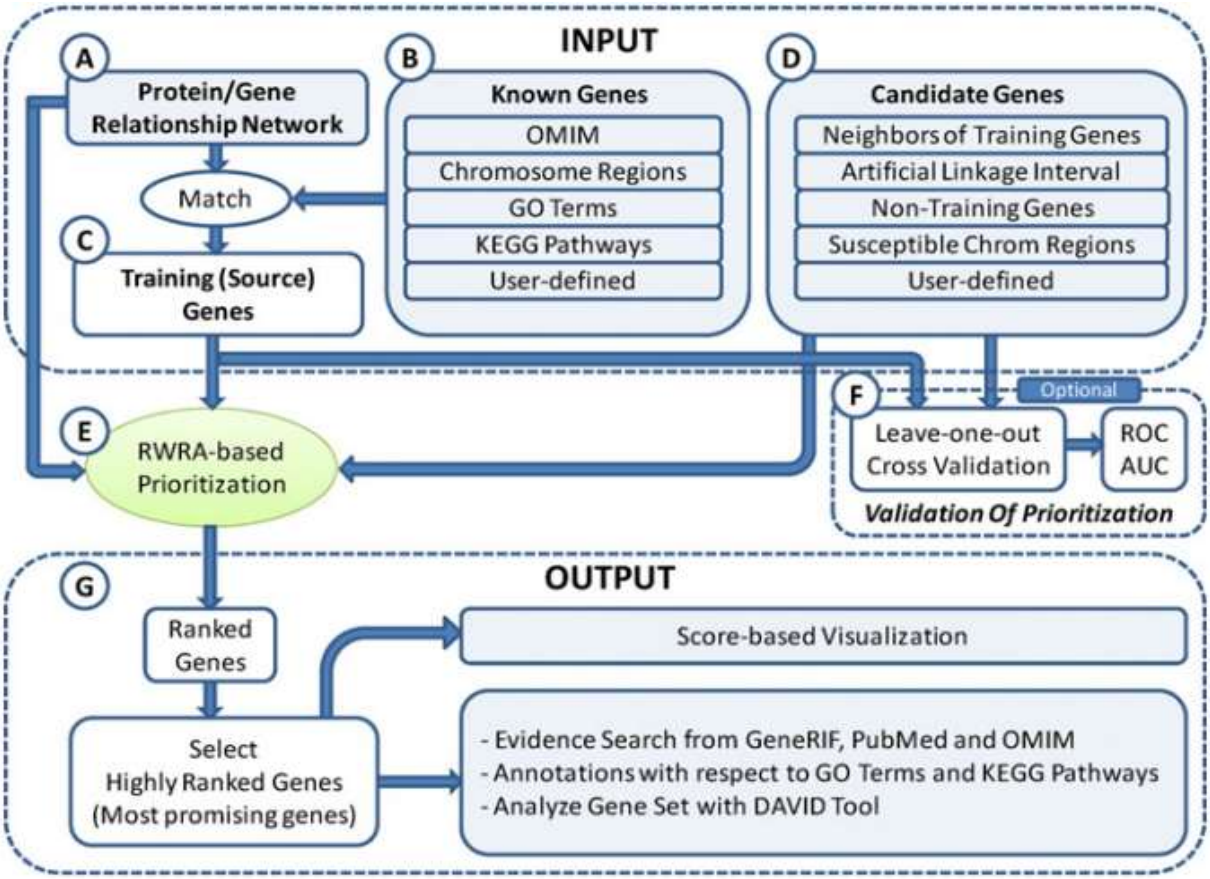
Ghi chú: Hiệu suất của hai phương pháp được đánh giá bằng phương pháp LOOCV trên tập 25 loại bệnh từ OMIM. Đối với Boolean Network, phương pháp dao động trạng thái ban đầu và tập 100 trạng thái ban đầu được sử dụng. Đối với RWR, xác suất đã được thiết lập 0,5.

nhiên I_s = 100 trạng thái khởi tạo và sử dụng LOOCV để vẽ đường cong ROC và tính toán giá trị AUC cho tập của 25 kiểu hình bệnh thu thập từ OMIM. Để so sánh hiệu năng của phương pháp đề xuất với phương pháp dựa trên RWR, nhóm nghiên cứu tiến hành một thủ tục tương tự như đã làm cho Boolean Network và thiết lập xác suất quay lui là 0,5 vì hiệu năng của phương pháp dựa trên RWR ổn định với sự thay đổi của tham số xác suất quay lui. Hình 6 cho thấy hiệu năng của hai phương pháp, rõ ràng rằng Boolean Network (AUC = 0,77) vượt trội so với dựa trên phương pháp RWR (AUC = 0,73). Kết quả này cho thấy vấn đề phân hạng gene bệnh có thể được tiếp cận hiệu quả với giả thuyết rằng đột biến trên các gene gây bệnh đã biết của một bệnh ảnh hưởng đến các gene khác và sự ảnh hưởng này có thể được sử dụng như mức độ gắn kết giữa các gene ứng viên và bệnh.

Trong nghiên cứu này, tất cả các nút trong mạng cuối cùng được xếp hạng dựa theo vector xác suất ở trạng thái ổn định P^{∞} thu được bằng cách lặp các vòng lặp cho đến khi $||P^{t+1}-P^{t}|| < 10^{-6}$.

Đối với việc xác định các gene liên quan đến bệnh, *S* là một tập các gene liên quan tới một loại bệnh cụ thể đã biết, một phần tử của P^{∞} đo lường mức độ liên quan của *S* tới một gene trong mạng. Vì vậy, nó có thể gọi đó chính là mức độ liên quan giữa một gene ứng viên và bệnh quan tâm.

Phương pháp RWR được tiến hành trên gói Plugin GPEC.jar chạy trên phân mềm Cytoscape 2.7 được nhóm tác giả Lê Đức Hậu và cộng sự nghiên cứu và phát triển vào năm 2012 (Le, 2012). Các bước tiến hành phương pháp này được thực hiện theo sơ đồ sau:



Hình 7. Sơ đồ thực hiện phương pháp RWR trên gói plugin GPEC.jar

3.2. Nghiên cứu thử nghiệm trên bệnh ung thư vú

Kết quả bài báo được thử nghiệm trên phương pháp đã đề xuất với bệnh ung thư vú (OMIM ID: 114.480), đây là một bệnh phức tạp với 22 gene có liên quan đến bệnh đã biết. Tuy vậy, chỉ có 10 gene trong số đó là có sẵn trên mạng tín hiệu. Sử dụng những gene này như là các nút nguồn và xem xét các gene khác trong mạng như các gene ứng viên, chúng tôi tính độ ảnh hưởng từ những gene đã biết liên quan đến bệnh này tới tất cả các gene ứng viên, các gene ứng viên sau đó được xếp hạng dựa trên độ đo ảnh hưởng đã được tính toán. Kết quả lựa chọn được 50 gene ứng viên được xếp hạng cao để tìm bằng chứng về sự liên quan với ung thư vú theo lý thuyết y sinh học. Sau khi tìm kiếm bằng chứng thì kết quả thu được là 27 gene trong số các gene đó có ít nhất một bằng chứng có liên quan tới bệnh ung thư vú (Bảng 2). Các gene ứng viên khác trong top 50 có thể sử dụng nhiều cho các nghiên cứu trong tương lai.

Bảng 2. Tổng 27 gene trong số 50 gene được xếp thứ hạng cao nhất có bằng chứng y văn liên quan đến bệnh ung thư vú ở người

Entrez Gene ID	Gene Symbol	PubMed ID
1111	CHEK1	20567916, 21401699, 21752283, 23844225
1647	GADD45A	15735726, 18350249, 19728081, 23158659, 23706118
1649	DDIT3	21741997, 23065795, 24625971
1839	HBEGF	17962208, 24013225
1843	DUSP1	15448190, 15590693, 19724859, 25377473
2150	F2RL1	16650817, 16925462, 19074826, 19543320, 19795460, 24177339, 24568471, 12813467, 15580312, 17631500, 17911267, 18612547, 18652687, 18787170, 18831768, 20030877, 20459645, 20479250,
2475	MTOR	21046231, 21963359, 22349822, 23991038, 24323026, 24630930, 24637915, 25659153
2908	NR3C1	15590693, 17512111, 18668364, 19875955, 21868756
3339	HSPG2	23436656
375	ARF1	18990689, 21478909, 24407288 15492785, 16807677, 16821086, 17259657, 17262178, 17592771, 17726138, 19671875, 20204503, 20428939, 20720556, 21409393, 21671140, 21872972, 24265520
4846	NOS3	21409393, 21671140, 21872972, 24265520
51085	MLXIPL	19252981
5111	PCNA	12088102, 22238610, 22622474, 23542172
51341	ZBTB7A	20394500, 21392388
5294	PIK3CG	17515959, 18625725, 18652687, 18725974, 19269083, 19471547, 20030877, 20226014, 20458733, 23500535
5313	PKLR	19655166
5333	PLCD1	11960991, 20657189
5524	PPP2R4	19890961, 24958351
5581	PRKCE	18317451, 20198332, 23562764, 24825907
6097	RORC	22404826, 24911119
6647	SOD1	16423367
7015	TERT	19501078, 19596972, 19787269, 20056641, 20225759, 21411498, 21526393, 21627565, 21911295, 21949822, 22037553, 22134622, 23065203, 23158658, 23629941, 23677713, 23741361, 24216762
7161	TP73	15450420, 15849742, 16814250, 17446929, 21127199, 21933556, 22535334, 23443851
79444	BIRC7	16026775, 17035597, 23524337
8915	BCL10	16280327
8976	WASL	17985201, 20880986, 22559840
9181	ARHGEF2	22002306

4. KẾT LUẬN

Trong bài viết này, chúng tôi đã trình bày các nghiên cứu về các phương pháp phân hạng gene ứng viên có khả năng liên quan đến bệnh, sau đó đề xuất một phương pháp dựa trên mạng để dự đoán các gene liên quan đến bệnh. Nghiên cứu này được dựa trên mạng sinh học boolean động dưới giả thuyết rằng sự đột biến về gene gây bệnh đã biết ảnh hưởng đến gene khác thông qua mạng và mức độ ảnh hưởng này có thể được sử dụng như là mức độ liên kết giữa các gene ứng viên với bệnh quan tâm. Các kết quả giả lập cho thấy rằng phương pháp của chúng tôi vượt trội hơn so với phương pháp dựa trên dựa trên mạng hiện tại. Sử dụng phương pháp đề xuất, chúng tôi cũng dự đoán 27 gene mới liên quan đến bênh ung thư vú. Trong nghiên cứu tương lai, bên cạnh Boolean động, chúng tôi sẽ tích hợp tính chất cấu trúc của mạng sinh học để đo liên kết giữa các gene ứng viên và bệnh vì một vài nghiên cứu đã cho thấy những tính chất cấu trúc đặc biệt của các gene bệnh trong các mạng sinh học.

TÀI LIỆU THAM KHẢO

Adie E., R. A. (2005). Speeding disease gene discovery by sequence based candidate. BMC Bioinformatics, 6: 55.

Aerts S., D. L. (2006). Gene prioritization through genomic data fusion. Nature Biotechnology, 24: 537-544.

Albert, R. (2004). Boolean Modeling of Genetic Regulatory Networks. Lecture Notes in Physics., 650: 459-481.

Albert, R. (2004). Boolean Modeling of Genetic Regulatory Networks. Lect. Notes Phys., 650: 459-481.

Amberger J., C. A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Research, 37: D793-D796.

Calvo S., M. J. (2006). Systematic identification of human. Nat Genet, 38: 576-582.

Cui Q., E. P. (2009). Protein evolution on a human. BMC Systems Biology, 3: 21.

Đặng Vũ Tùng, D. A. (2015). Phân hạng gen gây bệnh sử dụng học tăng kết hợp với xác suất tiền nghiệm. Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Tập V-1, Số13 (33).

Duc-Hau Le, Y.-K. K. (2012). A Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. Computational Biology and Chemistry, pp. 17-23.

Faure A., A. N. (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics, 22: e124-131.

Hanley JA, M. B. (1982). The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve. Radiology, 143: 29-36.

Kann, M. G. (2010). Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in Bioinformatics, 11: 96-110.

Kauffman S., C. P. (2003). Random Boolean network models and the yeast transcriptional. Proceedings of the National Academy of Sciences, 100: 14796-14799.

Keerthikumar S., S. B. (2009). Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. DNA Research, 16: 345-351.

Kwon, D.-H. L.-K. (2011). The effects of feedback loops on disease comorbidity in human signaling networks. Bioinformatics, 27: 1113-1120.

Kwon, D.-H. L.-K. (2013). A coherent feedforward loop design principle to sustain robustness of biological networks. Bioinformatics, 29: 630-637.

Kwon, D.-H. Le and Y.-K. (2011). NetDS: a Cytoscape plugin to analyze the robustness of dynamics and feedforward/feedback loop structures of biological networks. Bioinformatics, 27: 2767-2768.

Li, J. X. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 22: 2800-2805.

LOVASZ, L. (1996). Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty, 2: 353-398.

Sun J., J. C. (2009). Functional link artificial neural. In Neural Networks.

Trần Thị Bích Phương, N. V. (2013). Một phương pháp phân tích mạng tương tác protein để dự đoán gen gây bệnh ung thư. journal of science of hnue, 58: 38-46.

Trinh H.-C., D.-H. L.-K. (2014). PANET: A GPU-Based Tool for Fast Parallel Analysis of Robustness Dynamics and Feed-Forward/Feedback Loop Structures in Large-Scale Biologica lNetworks. PLoS ONE, 9: e103010.

Vali derhami, E. K. (2013). Applying reinforcement learning for web pages ranking algorithms. Applied Soft Computing, 13: 1686-1692.

Vanunu O, M. O. (2010). Associating genes and protein complexes with disease via network propagation. PLoSComput Biol, 6: e1000641.

Wang X., N. G. (2011). Network-based methods for human disease gene prediction. Briefings in Functional, 10: 280-293.