Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00051

MỘT PHƯƠNG PHÁP CẢI TIẾN CHO BÀI TOÁN XÁC ĐỊNH CÁC GEN LIÊN QUAN ĐẾN BỆNH

Nguyễn Đại Phong[1], Đặng Vũ Tùng[2], Lê Đức Hậu[3], Từ Minh Phương[4]

[1] Viện Công nghệ thông tin và Truyền thông, Trường Đại học Bách Khoa Hà Nội
[2] Trung tâm Tin học, Học viện Thanh thiếu niên Việt Nam
[3] Trung tâm Tin học, Đại học Thủy Lợi
[4] Khoa Công nghệ thông tin, Học viện Công nghệ Bưu chính viễn thông

phongnd.hust@gmail.com, tung_dv@yahoo.com, hauldhut@gmail.com, phuongtm@ptit.edu.vn

TÓM TẮT *— Xác định gen gây bệnh thường bắt đầu việc phân hạng các gen ứng viên theo mức độ liên quan đến bệnh. Việc làm này nhằm mục đích thu hẹp tập gen liên quan đến bệnh cần xác định bởi các thực nghiệm y sinh chuyên sâu. Hiện nay, có rất nhiều phương pháp khác nhau được đề xuất để phân hạng các gen ứng viên dựa trên mối quan hệ giữa các protein trong mạng tương tác gen/protein. Trong đó, hầu hết các phương pháp đều dựa trên giả thiết "mô đun bệnh", tức là các gen liên quan đến cùng một bệnh có xu hướng nằm kề nhau trên các mạng tương tác. Các phương pháp này có xu hướng ưu tiên những gen ứng viên gần với gen gây bệnh đã biết trên mạng tương tác. Nhưng trong quá trình thực nghiệm, chúng tôi nhận thấy với nhiều bệnh, các gen liên quan đã biết cũng không hoàn toàn tạo thành một mô đun mà thậm chí chúng còn cách nhau rất xa trên mạng tương tác. Do đó, các phương pháp phân hạng hiện có không còn đạt hiệu quả cao. Để giải quyết vấn đề này, chúng tôi đề xuất một phương thức cải tiến mới nhằm mục đích ưu tiên cho các gen liên quan đến những bệnh có tính chất trên bằng cách tăng cường trọng số liên kết cho các gen ở xa các gen gây bệnh đã biết. Chúng tôi đã kiểm chứng hiệu quả phân hạng của phương pháp này với 148 bệnh trên mạng tương tác protein của người và so sánh hiệu quả phân hạng của phương pháp đề xuất với các phương pháp nổi trội hiện có như bước ngẫu nhiên có quay lại (RWR) và dựa trên xác suất liên kết (ERIN). Kết quả thực nghiệm đã chỉ ra rằng phương pháp của chúng tôi đạt độ chính xác là 95.3%, tốt hơn RWR (93.4%) và ERIN (89.8%). Thêm vào đó, sử dụng phương pháp của mình, chúng tôi đã xác định được một số gen mới liên quan đến bệnh ung thư tuyến tiền liệt.*

Từ khóa *— Disease genes prioritization, protein interaction network, random walk with restart algorithm, prostate cancer.*

I. GIỚI THIỆU

Xác định các gen mới có liên quan đến bệnh là một bài toán quan trọng trong nghiên cứu y sinh. Đây có thể coi là bước khởi đầu trong việc tìm ra phương pháp điều trị cho các bệnh phát sinh do yếu tố di truyền [1-3]. Trong giai đoạn trước đây, việc xác định gen gây bệnh được thực hiện chủ yếu bằng thực nghiệm sinh học để xác định các vùng nhiễm sắc thể khả nghi liên quan bệnh cần nghiên cứu [4, 5]. Tuy nhiên, những vùng nhiễm sắc thể này thường chứa hàng trăm gen ứng viên, trong khi chỉ có một số ít các gen thực sự liên quan đến bệnh [6]. Để xác định được chính xác các gen thực sự liên quan đến bệnh cần nghiên cứu, các nhà y sinh học phải tiến hành các thí nghiệm cho từng gen trong danh sách gen ứng viên thu được. Đây là công việc rất tốn kém về thời gian và kinh phí. Thách thức này hiện nay một phần đã được giải quyết bằng phương pháp phân hạng gen ứng viên liên quan đến bệnh trong Tin sinh học và nó đã trở thành trọng tâm trong lĩnh vực di truyền học.

Các phương pháp phân hạng gen gây bệnh dựa trên mạng thường căn cứ vào nguyên lý "mô đun bệnh" (nghĩa là, các gen/protein liên quan đến cùng một bệnh hoặc các bệnh tương tự nhau có xu hướng nằm kề nhau trong các mạng tương tác [5]) để tính toán độ tương tự tương giữa các gen ứng viên và các gen gây bệnh đã biết. Có rất nhiều phương pháp dựa trên mạng đã được đề xuất cho bài toán này như: dựa trên các láng giềng gần nhất, dựa trên các cụm trên mạng. Ngoài ra, các thuật toán phổ biến trong phân tích mạng xã hội và mạng Web dùng để đánh giá tầm quan trọng tương đối của nút như: HITS with priors, PageRank with priors, K-step Markov [7], RL_Rank [8] và ERIN [9] cũng đã được sử dụng cho bài toán phân hạng các gen ứng viên trên các mạng tương tác gen/protein. Trong số các phương pháp phân hạng gen dựa trên mạng, phương pháp sử dụng thuật toán bước ngẫu nhiên có quay lại RWR [10-12] được áp dụng phổ biến hơn các phương pháp khác vì thuật toán này xem xét toàn bộ các liên kết giữa các gen gây bệnh đã biết với các gen ứng viên trên mạng tương tác gen/protein, bao gồm cả các tương tác trực tiếp và gián tiếp. Không những đạt được hiệu quả cao trong bài toán phân hạng gen ứng viên liên quan đến bệnh, thuật toán này còn được sử dụng hiệu quả trong việc xác định các microRNA mới liên quan đến bệnh [13] cũng như các đích tác động mới của thuốc [14]. Tiếp nối thành công của thuật toán RWR cho bài toán phân hạng và tìm kiếm gen gây bệnh trên mạng tương tác gen/protein đồng nhất. Một phiên bản mới của thuật toán đã được đề xuất sử dụng trên mạng không đồng nhất kết hợp giữa mạng tương tác gen/protein và mạng kiểu hình bệnh [15] hoặc mạng tương tự bệnh [16] gọi là RWRH. Thuật toán này cho hiệu quả dự đoán tốt hơn RWR trên mạng protein đồng nhất.

Tuy nhiên, một thách thức cần được giải quyết là vấn đề nhiễu dữ liệu trong các mạng tương tác sinh học nói chung và sự tổng hợp chưa đầy đủ các liên kết giữa các gen trong mạng tương tác gen/protein dẫn đến mạng tương tác gen/protein hiện có chưa bao phủ hết toàn bộ các liên kết của hệ gen người. Cụ thể, khi làm thực nghiệm chúng tôi nhận thấy với nhiều bệnh, các gen liên quan đã biết cũng không hoàn toàn tạo thành một mô đun mà thậm chí chúng

còn cách nhau rất xa trên mạng tương tác. Vì vậy, các phương pháp phân hạng gen hiện có vẫn chưa đạt được hiệu quả cao. Để giải quyết vấn đề này, chúng tôi đề xuất một phương pháp kết hợp nhằm mục đích tìm kiếm các gen ứng viên gây bệnh có liên kết yếu hoặc ở xa những gen gây bệnh đã biết. Trong phương pháp này, chúng tôi tiến hành phân hạng tất cả các gen ứng viên bằng thuật toán dựa trên xác suất liên kết, sau đó trích chọn một tập các gen có độ liên quan cao nhất đối với các gen bệnh đã biết. Tập gen còn lại sẽ được tăng cường trọng số liên kết bằng phương pháp RWRH để xác định thêm các gen có khả năng liên quan đến bệnh đã biết. Kết quả thực nghiệm cho thấy phương pháp đề xuất tốt hơn đáng kể so với các phương pháp đã được sử dụng trong việc tìm kiếm gen ứng viên gây bệnh.

Các phần còn lại của bài báo được bố cục như sau: Phần 2 mô tả dữ liệu, các nghiên cứu liên quan và phương pháp đề xuất. Phần 3 trình bày các kết quả thực nghiệm. Cuối cùng là phần kết luận nêu các đóng góp chính của bài báo và đề xuất các hướng cải tiến mới.

II. DỮ LIỆU VÀ PHƯƠNG PHÁP

A. Dữ liệu

Để có thể thực nghiệm với các thuật toán phân hạng dựa trên mạng, chúng tôi cần một mạng tương tác gen/protein và các bệnh đã biết một số gen liên quan. Cụ thể, chúng tôi đã sử dụng mạng tương tác gen/protein từ [11, 17]. Đây là một mạng vô hướng, có trọng số (biểu thị độ tương tự về chức năng giữa các gen/protein) gồm 11.886 gen và 111.943 liên kết. Thêm vào đó, chúng tôi sử dụng các cơ sở dữ liệu về bệnh và các gen liên quan đã biết từ OMIM [18]. Kết quả thu được 622 bệnh với tổng số 3246 gen liên quan, trong đó 148 bệnh có từ 2 gen liên quan trở lên đã được phát hiện. Với mỗi bệnh, tập các gen đã biết được sử dụng như là tập gốc trong quá trình phân hạng bởi các thuật toán.

B. Các phương pháp phân hạng dựa trên đồ thị

Trong bài báo này, mạng tương tác gen/protein được biểu diễn bởi một đồ thị vô hướng, có trọng số $G = (V, E)$ trong đó, tập các nút V là các gen/protein và tập các cạnh E thể hiện liên kết tương tác giữa các gen/protein. Giả sử, cho trước S ($S \subseteq V$) là tập các gen bệnh đã biết (còn gọi là tập hạt giống hay tập nút gốc), tức là một số lượng nhỏ các gen đã được phát hiện có liên quan đến bệnh trong các nghiên cứu trước đó, C ($C \subseteq V$) là tập các gen ứng viên có liên kết với các nút trong S. Mục tiêu của bài toán phân hạng gen là tính toán điểm số cho các gen trong tập C theo độ liên quan với S. Các điểm số này sau đó được xếp hạng và căn cứ vào đó để đề xuất các gen gây bệnh mới.

1. Thuật toán dựa trên xác suất liên kết (ERIN)

Thuật toán dựa trên xác xuất liên kết [9] là một phương pháp mới trong phân tích mạng xã hội và đã được ứng dụng cho bài toán phân hạng gen gây bệnh [19] đạt kết quả khả quan. Thuật toán này xác định tất cả các đường đi không chu trình từ một nút (hoặc tập nút gốc) tới các nút còn lại trên đồ thị. Bắt đầu từ một nút gốc s chuyển tới các nút láng giềng bằng phương pháp tìm kiếm theo chiều sâu (DFS). Tại mỗi bước, nó tính tổng xác suất đường đi từ nút gốc tới nút được thăm hiện hành. Quá trình sẽ dừng khi tổng xác suất đường đi nhỏ hơn một ngưỡng giá trị δ cho trước. Điều này có nghĩa là đường đi tới các nút chưa thăm không còn quan trọng bởi vì xác xuất đường đi từ nút gốc tới các nút này là quá nhỏ.

Xác suất di chuyển từ nút v_i tới nút láng giềng v_j biểu thị độ liên quan của nút v_j với v_i và được xác định theo công thức:

$$P(v_i, v_j) = \begin{cases} (1-f)\frac{e(v_i,v_j)}{\sum_{v_k \in neighbor(v_i)} e(v_i,v_k)} & v_i \neq v_j \\ 1 & v_i = v_j \end{cases} \quad (1)$$

trong đó: $e(v_i,v_j)$, $e(v_i,v_k)$ là trọng số các cạnh tương ứng giữa nút v_i với các nút láng giềng v_j và v_k ; f là hệ số giảm trừ ($0 < f < 1$). Ở đây cần chú ý rằng việc lựa chọn giá trị hệ số f dựa trên 2 tiêu chí: 1) giá trị f phải bảo toàn được thuộc tính của phương pháp bước ngẫu nhiên; 2) cho phép xác suất hội tụ ở mức chấp nhận được. Về nguyên tắc, hệ số f càng nhỏ càng tốt nhưng khi đó thời gian tính toán sẽ tăng lên đáng kể.

Xác suất đường đi từ một nút khởi đầu s đến nút kết thúc t biểu hiện mức độ liên quan giữa s và t được xác định theo công thức:

$$PathProb(s,t) = P(s,v_1)(\prod_{i=1}^{m-1} P(v_i, v_{i+1}))P(v_m, t) \quad (2)$$

trong đó, $P(v_i,v_j)$ được định nghĩa ở công thức (1). Rõ ràng xác suất đường đi *PPath* là một giá trị thuộc khoảng [0, 1] do các thừa số trong (2) cũng thuộc khoảng [0, 1].

Nếu từ nút khởi đầu s tới nút kết thúc t có nhiều con đường thì độ liên quan của t đối với s được xác định là tổng tất cả các xác suất đường đi từ nút s đến nút t và được xác định theo công thức:

$$I(t|s) = \sum_p PathProb_p(s,t) \quad (3)$$

trong đó $\forall p \subseteq G$, có điểm bắt đầu là s và điểm kết thúc t, $PathProb_p(s, t) \geq \delta$. Như vậy, nếu nút s có nhiều đường đi tới t, điều này cho thấy rằng t có độ liên quan cao đối với s.

Đối với tập hợp các nút truy vấn *S*, thuật toán sẽ thực hiện cho từng nút trong tập hợp. Độ liên quan trung bình của nút *t* so với một tập các nút truy vấn *S* được tính theo công thức sau:

$$I(t|S) = \frac{1}{|S|}\sum_{s\in S} I(t|s) \quad (4)$$

Khi áp dụng thuật toán này cho bài toán phân hạng và tìm kiếm gen gây bệnh, giả sử *s* là một gen gây bệnh đã biết và *t* là một gen ứng viên trên đồ thị mạng tương tác protein. Nếu các xác suất đường đi của tất cả các con đường từ *s* tới *t* đều nhỏ hơn ngưỡng giá trị δ thì gen *t* hầu như không liên quan tới bệnh. Vì vậy, chúng ta chỉ xem xét các gen có ít nhất một đường đi tới *s* có xác suất đường đi lớn hơn ngưỡng δ cho trước. Đối với một tập gen bệnh đã biết *S*, độ liên quan trung bình của mỗi gen ứng viên đối với *S* được sử dụng để xếp hạng các gen ứng viên. Cuối cùng, các gen ứng viên có điểm xếp hạng cao nhất sẽ được lựa chọn. Kết quả của thuật toán này cho chúng ta *k* gen có độ liên quan cao nhất với các gen bệnh đã biết. Trong đó, *k* là một số rất nhỏ so với tổng số gen trong đồ thị mạng tương tác gen/protein.

2. Thuật toán bước ngẫu nhiên có quay lại (RWR)

Bước ngẫu nhiên có quay lui là một biến thể của thuật toán bước ngẫu nhiên trên đồ thị. Theo thuật toán này, một thực thể xuất phát từ một nút khởi đầu. Sau đó, nó 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. Tại thời điểm *t* 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 γ thuộc khoảng [0, 1]. Giả sử $G = (V, E)$ là một đồ thị vô hướng có trọng số, trong đó $V = (v_1, v_2, ...,v_n)$ là tập các nút và $E = ((v_i, v_j) \mid v_i, v_j \in V)$ là tập các cạnh. Gọi $S \subseteq V$ là tập các nút gốc (nút khởi đầu), *W* là ma trận kề của đồ thị *G*. Thuật toán bước ngẫu nhiên có quay lại được mô tả như sau:

$$p_{t+1} = (1-\gamma)W'p_t + \gamma p_0 \quad (5)$$

trong đó: p_{t+1} là vector xác suất của tập các nút $|V|$ tại thời điểm *t*, phần tử thứ *i* biểu diễn xác suất của thực thể tại nút $v_i \in V$; *W'* là ma trận chuẩn hóa từ ma trận kề *W*, trong đó W'_{ij} (kí hiệu các phần tử (i, j) trong *W'*) biểu diễn xác suất mà thực thể di chuyển từ v_i tới v_j nằm trong tập $V\backslash\{v_i\}$; p_0 là vector xác suất khởi đầu trong đó các phần tử có giá trị là 0 (nếu chúng không thuộc tập nút gốc) và $1/|S|$ (nếu chúng thuộc tập nút gốc).

Khi áp dụng RWR cho bài toán phân hạng gen ứng viên dựa trên mạng [10, 12], tập hợp các nút gốc *S* là các gen bệnh đã biết và các gen ứng viên là các gen còn lại trên mạng tương tác gen/protein. Ma trận chuẩn hóa *W'* được xác định như sau:

$$W'_{ij} = \frac{(W_G)_{ij}}{\sum_j (W_G)_{ij}} \quad (6)$$

trong đó W_G là ma trận kề của đồ thị mạng tương tác gen/protein.

Tất cả các gen ứng viên cuối cùng được phân hạng khi vector xác suất p_∞ đạt trạng thái ổn định sau một số bước lặp (tức là chênh lệch giữa p_{t+1} và p_t nhỏ hơn một giá trị tới hạn, ở đây chúng tôi chọn là 10^{-6}).

3. Thuật toán kết hợp xác suất liên kết và bước nhẫu nhiên có quay lại (CRWR)

Dựa trên ý tưởng của giải thuật RWRH cho mạng không đồng nhất đã được áp dụng để dự đoán các gen gây bệnh dựa trên mạng không đồng nhất kiểu gen - kiểu hình [15] và kiểu gen - bệnh tương tự [16]. Trong nghiên cứu này chúng tôi đề xuất một phương pháp phân hạng gen mới bằng cách kết hợp thuật toán dựa trên xác suất đường đi với RWRH gọi là CRWR (Complex Random Walk with Restart), nhằm mục đích tăng cường tầm quan trọng/độ liên quan cho các gen ở xa các gen bệnh đã biết trong cùng một mạng tương tác gen/protein.

Đầu tiên, từ mạng tương tác gen/protein đồng nhất, chúng tôi sử dụng thuật toán dựa trên xác suất đường đi để tách mạng tương tác gen/protein thành hai mạng con G_H và G_L. Ở đây, G_H chứa các gen có độ liên quan cao nhất đối với các gen bệnh đã biết, G_L chứa các gen còn lại trong mạng tương tác gen/protein.

Bước tiếp theo, chúng tối áp dụng thuật toán RWRH để tăng cường trọng số cho các mạng trên. Cụ thể là ma trận *W'* được xác định như sau:

$$W' = \begin{bmatrix} W'_H & W'_{HL} \\ W'_{LH} & W'_L \end{bmatrix} \quad (7)$$

trong đó: W'_H, W'_L, lần lượt là các ma trận chuẩn hóa của các mạng con G_H và G_L. W'_{HL}, W'_{LH} là các ma trận chuẩn hóa của các liên mạng con. Giả sử, λ là xác suất chuyển ngẫu nhiên của thực thể từ mạng con G_H sang mạng con G_L hoặc ngược lại. Khi đó, các ma trận được xác định như sau:

$$(W'_{HL})_{i,j} = \begin{cases} \frac{(\lambda W_{HL})_{ij}}{\sum_j (W_{HL})_{ij}} & \text{nếu } \sum_j (W_{HL})_{ij} \neq 0 \\ 0 & \text{ngược lại,} \end{cases} \quad (8)$$

$$(W'_{LH})_{i,j} = \begin{cases} \frac{(\lambda W_{HL})_{ji}}{\sum_j (W_{HL})_{ji}} & \text{nếu } \sum_j (W_{HL})_{ji} \neq 0 \\ 0 & \text{ngược lại,} \end{cases} \quad (9)$$

$$(W'_{H})_{i,j} = \begin{cases} \frac{(W_H)_{ij}}{\sum_j (W_H)_{ij}} & \text{nếu } \sum_j (W_{HL})_{ij} = 0 \\ \frac{(1-\lambda)(W_H)_{ij}}{\sum_j (W_H)_{ij}} & \text{ngược lại,} \end{cases} \quad (10)$$

$$(W'_{L})_{i,j} = \begin{cases} \frac{(W_L)_{ij}}{\sum_j (W_L)_{ij}} & \text{nếu } \sum_j (W_{HL})_{ij} = 0 \\ \frac{(1-\lambda)(W_L)_{ij}}{\sum_j (W_L)_{ij}} & \text{ngược lại,} \end{cases} \quad (11)$$

ở đây, W_H, W_L và W_{HL} là các ma trận kề của đồ thị mạng G_H, G_L và mạng hỗn hợp.

Vì thuật toán chỉ áp dụng cho một mạng gen/protein thuần nhất nên vector khởi đầu p_0 vẫn được xác định như thuật toán RWR theo công thức sau:

$$p_0 = \begin{cases} \frac{1}{|S|} & \text{nếu } v_i \in S \\ 0 & \text{ngược lại,} \end{cases} \quad (12)$$

C. Phương pháp đánh giá

Để đánh giá hiệu suất của phương pháp phân hạng, đối với mỗi bệnh chúng tôi sử dụng phương pháp kiểm tra chéo bỏ-ra-một (LOOCV: Leave-one-out cross validation). Theo đó, với mỗi lần lặp, một gen bệnh đã biết được lấy ra và coi như là một gen ứng viên bình thường, các gen còn lại được sử dụng như các gen gốc làm dữ liệu đầu vào cho thuật toán. Cụ thể như sau: với tập gen bệnh đã biết S và tập gen ứng viên C (là tất cả các gen còn lại trên mạng), một gen $s \in S$ được lấy ra và chúng tôi tiến hành phân hạng tập gen $C \cup \{s\}$ theo thuật toán đề xuất với tập $S \backslash \{s\}$ được sử dụng như tập các nút gốc. Quá trình này được lặp lại cho tất cả các gen bệnh đã biết. Sau đó chúng tôi thay đổi ngưỡng τ từ 1 cho đến $|C \cup \{s\}|$. Giá trị *sensitivity* và *1-specificity* được tính toán theo các công thức:

$$sensitivity = \frac{TP}{TP+FN} \quad (13)$$

$$1 - specificity = \frac{FP}{FP+TN} \quad (14)$$

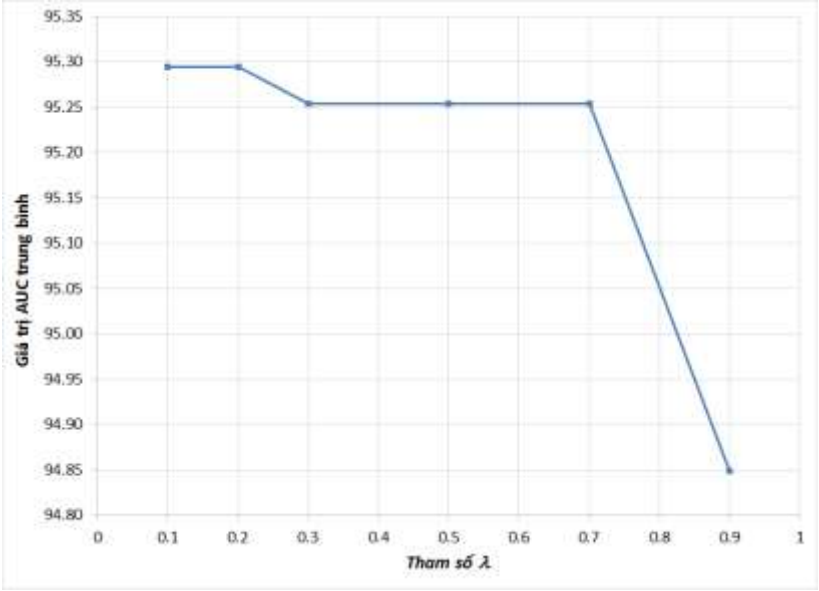
trong đó *TP* (true positive) là số trường hợp thử nghiệm mà thứ hạng của $s \leq \tau$, *FN* (false negative) là số trường hợp thử nghiệm mà thứ hạng của $s > \tau$, *FP* (false positive) là số trường hợp thử nghiệm mà thứ hạng của $c \leq \tau$ (với mỗi $c \in C$) và *TN* (true negative) là số trường hợp thử nghiệm mà thứ hạng của $c > \tau$ (với mỗi $c \in C$). Một cặp giá trị *sensitivity* và *1-specificity* tương ứng với một điểm trên đường cong ROC. Tiếp đó, hiệu suất của phương pháp phân hạng được xác định bằng cách tính toán giá trị AUC (Area Under ROC Curve) là phần diện tích dưới đường cong ROC.

III. THỰC NGHIỆM VÀ KẾT QUẢ

Trong phần này, chúng tôi đánh giá ảnh hưởng của các tham số tới sự ổn định cũng như hiệu quả của phương pháp đề xuất. Đồng thời lựa chọn bộ tham số tốt nhất để so sánh hiệu quả phân hạng với RWR và ERIN trên cùng một bộ dữ liệu theo giá trị AUC. Cuối cùng, chúng tôi ứng dụng phương pháp này để xác định các gen mới liên quan tới bệnh ung thư tuyến tiền liệt.

A. Ảnh hưởng của tham số

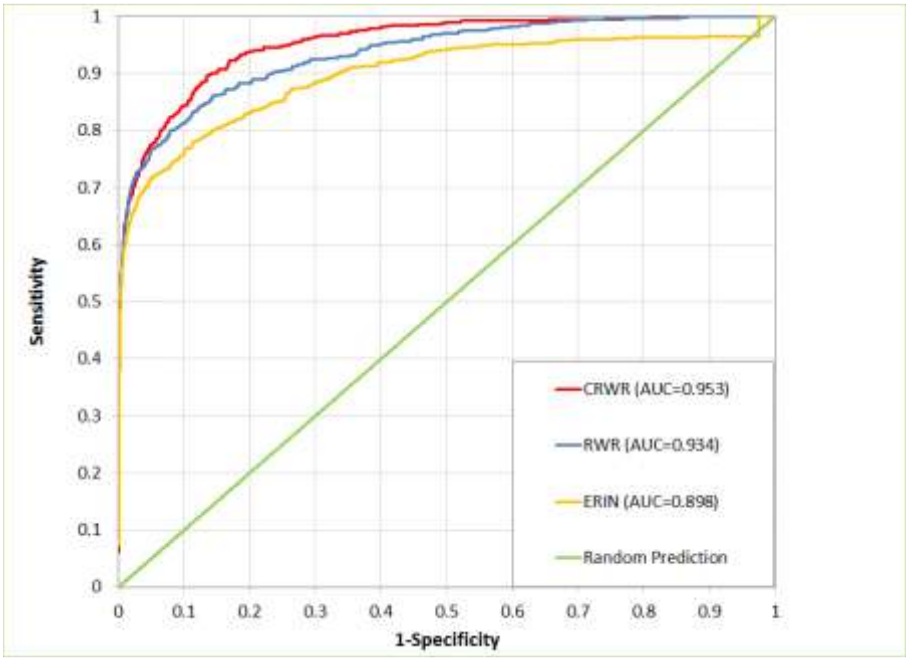
Thực nghiệm đầu tiên được chúng tôi tiến hành để xác định ảnh hưởng của tham số λ tới hiệu quả của phương pháp đề xuất. Chúng tôi lần lượt thay đổi giá trị tham số λ từ 0.1 đến 0.9 sau đó đó tính toán giá trị AUC cho từng trường hợp theo phương pháp đã đề xuất ở phần II.C. Kết quả thực nghiệm được mô tả trong Hình 1. Dựa trên kết quả thử nghiệm, chúng tôi nhận thấy khi giá trị tham số $\lambda \leq 0.2$, thuật toán đề xuất cho giá trị AUC cao nhất; trong trường hợp $0.2 < \lambda \leq 0.7$, kết quả phân hạng hầu như không thay đổi và $\lambda > 0.7$, kết quả phân hạng giảm một cách đáng kể. Điều này cho thấy rằng, với giá trị λ được lựa chọn phù hợp (cụ thể là $\lambda = [0.1, 0.2]$), thuật toán đề xuất sẽ ưu tiên cả những gen ứng viên nằm trong phân vùng cách xa các gen gây bệnh đã biết, dẫn đến đạt hiệu quả tốt hơn.



Hình 1. Biểu diễn các giá trị AUC trung bình trên 148 bệnh với tham số β tăng từ 0.1 đến 0.9

B. So sánh với RWR và ERIN

Để khẳng định hiệu quả phương pháp đề xuất, chúng tôi tiến hành thực nghiệm và so sánh kết quả phân hạng với các phương pháp RWR, ERIN trên cùng một bộ dữ liệu đã mô tả trong phần II.A và phương pháp đánh giá LOOCV. Dựa trên kết quả phân hạng gen gây bệnh của [10-12] và [9], chúng tôi thiết lập giá trị các tham số $\gamma = 0.7$ cho phương pháp RWR và $\delta = 10^{-6}$, $f = 0.1$ cho phương pháp ERIN. Đối với mỗi phương pháp, chúng tôi tiến hành vẽ đường cong ROC và tính giá trị AUC trung bình cho tất cả 148 bệnh bằng cách tính các giá trị *sensitivity* và *1-specificity* cho từng bệnh, sau đó tính giá trị *sensitivity* và *1-specificity* trung bình của 148 bệnh tại các ngưỡng τ. Hình 2 biểu diễn đường cong ROC và giá trị AUC trung bình của cả ba phương pháp được so sánh.



Hình 2. Đường cong ROC biểu diễn kết quả thực thi của các thuật toán

Với kết quả thực nghiệm này, chúng tôi nhận thấy rằng so với các phương pháp phân hạng dựa trên mạng tương tác protein khác như RWR, ERIN thì phương pháp đề xuất của chúng tôi đạt được hiệu suất cao hơn rõ rệt. Điều này cho thấy rằng độ chính xác của CRWR là tốt hơn bởi sự ưu tiên trọng số liên kết của các gen nằm ở phân vùng xa các gen gây bệnh đã biết.

C. Dự đoán các gen liên quan đến bệnh ung thư tuyến tiền liệt

Trong phần này, chúng tôi kiểm chứng khả năng xác định các gen mới liên quan đến bệnh của phương pháp đề xuất bằng cách áp dụng phương pháp này cho một bệnh cụ thể. Để thực hiện điều này, chúng tôi tiến hành xác định các

gen mới liên quan đến bệnh ung thư tuyến tiền liệt (prostate cancer) có mã MIM là 176807 và thu thập các bằng chứng y văn của các gen có thứ hạng cao trong kết quả phân hạng.

Ung thư tuyến tiền liệt xảy ra khi những tế bào bất thường phát triển trong tuyến tiền liệt. Những tế bào này có thể tiếp tục nhân lên một cách không kiểm soát và đôi khi lan ra ngoài tuyến tiền liệt sang những bộ phận kế cận hay xa hơn của cơ thể. Tra cứu trong cơ sở dữ liệu OMIM, chúng tôi thu thập được 22 gen đã được chứng minh là có liên quan tới bệnh. Trong đó có 7 gen không có liên kết trong mạng tương tác gen/protein chúng tôi sử dụng để làm thực nghiệm. Tập 15 gen còn lại được sử dụng như là tập gốc trong quá trình phân hạng, các gen còn lại trong mạng tương tác gen/protein được coi là các gen ứng viên và phân hạng theo phương pháp đã đề xuất… Thông tin về các gen liên quan tới bệnh được trình bày trong Bảng 1

Bảng 1. Các gen gây bệnh ung thư tuyến tiền liệt và số liên kết trong mạng gen/protein

TT	Ký hiệu của gen	Mã Entrez của gen	Số liên kết PPI
1	367	AR	108
2	675	BRCA2	42
3	3732	CD82	17
4	999	CDH1	64
5	11200	CHEK2	80
6	60528	ELAC2	3
7	3029	HAGH	25
8	6928	HNF1B	44
9	408259	HPC3	0
10	408260	HPC4	0
11	619402	HPC5	0
12	100188789	HPC6	0
13	347747	HPCQTL19	0
14	9566	HPCX	0
15	1316	KLF6	3
16	8379	MAD1L1	6
17	4481	MSR1	14
18	4601	MXI1	10
19	7834	PCAP	0
20	5728	PTEN	30
21	7991	TUSC3	8
22	463	ZFHX3	3

Sau khi phân hạng, chúng tôi lựa chọn 30 gen có thứ hạng cao nhất và tiến hành thu thập bằng chứng về mối quan hệ giữa các gen này với bệnh ung thư tuyến tiền liệt từ cơ sở dữ liệu PubMed [19]. Thông tin về các gen và mã các văn y chứng minh sự liên quan giữa các gen này với bệnh được trình bày trong Bảng 2. Các gen còn lại mặc dù chưa có bằng chứng trực tiếp liên quan đến bệnh cần nghiên cứu nhưng chúng cũng là nguyên nhân gây ra các bệnh ung thư khác như: ung thư tuyến giáp, tuyến mô, đại tràng,… Các gen này chúng tôi đề xuất với các nhà y sinh học nghiên cứu và tìm kiếm thêm các chứng cứ liên quan đến bệnh bằng các thí nghiệm y sinh chuyên sâu.

Bảng 2. Các gen liên quan tới bệnh ung thư tuyến tiền liệt trong số 30 gen có thứ hạng cao nhất

Xếp hạng	Ký hiệu của gen	Mã Entrez của gen	Mã y văn tham khảo trên PubMed
2	4602	MYB	26089205
3	10401	PIAS3	11071847
4	1487	CTBP1	23097625
7	1051	CEBPB	25772238
9	688	KLF5	24931571
15	6184	RPN1	19064571
21	6185	RPN2	17220478
24	7157	TP53	25827447
25	9611	NCOR1	23129261
26	4609	MYC	25973080
28	10608	MXD4	15862967

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một phương pháp kết hợp giữa thuật toán dựa trên xác suất đường đi và bước ngẫu nhiên có quay lại áp dụng cho bài toán phân hạng gen với mục đích tìm kiếm các gen ứng viên gây bệnh nằm xa các gen bệnh đã biết trên đồ thị mạng tương tác gen/protein. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt được hiệu quả tốt hơn so với các phương pháp được sử dụng một cách đơn lẻ trước đó. Chúng tôi cũng đã áp dụng phương pháp đề xuất để tìm kiếm các gen liên quan đến bệnh ung thư tuyến tiền liệt và thu được kết quả khả quan. Với kết quả này, có thể thấy rằng khả năng dự đoán gen bệnh mới dựa trên độ liên quan/ tầm quan trọng tương đối của chúng so với các gen bệnh đã biết là hoàn toàn khả thi. Các gen ứng viên thứ hạng cao có thể được đề xuất cho các nhà nghiên cứu y, sinh học kiểm tra bằng các thí nghiệm sinh học chuyên sâu. Phương pháp đề xuất trong bài báo có thể được phát triển thành một phần mềm ứng dụng, triển khai trong các cơ sở nghiên cứu y sinh học phục vụ công tác nghiên cứu và đào tạo. Đồng thời cũng có thể ứng dụng để phát hiện các gen liên quan đến những căn bệnh di truyền cụ thể. Đây cũng là bước tiền đề cho việc tìm ra các phương pháp điều trị thích hợp cho các bệnh liên quan đến gen. Trong các nghiên cứu tiếp theo, chúng tôi sẽ thực hiện thêm các kiểm nghiệm kết quả bằng cách tìm các bằng

chứng y văn về mối liên quan giữa các gen ứng viên có thứ hạng cao và bệnh đang được xem xét. Đồng thời, chúng tôi cũng sẽ thử nghiệm phương pháp đề xuất với các đồ thị mạng sinh học khác như: mạng trao đổi chất, mạng điều hòa gen, mạng tương tác di truyền… để khẳng định thêm về hiệu quả thuật toán.

TÀI LIỆU THAM KHẢO

[1] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine," Bioinformatics, vol. 27, pp. 1741-1748, 2011.

[2] K. Reynolds, "Achieving the Promise of Personalized Medicine," Clinical Pharmacology & Therapeutics, vol. 92, pp. 401-405, 2012.

[3] D. Jones, "Steps on the road to personalized medicine," Nature Reviews Drug Discovery, vol. 6, pp. 770-771, 2007.

[4] M. ML, M. JC, L. AC, A.-B. M, C. ME, and e. al, "Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35," American Journal of Human Genetics, vol. 75(2), pp. 161-173, 2004.

[5] S. R, U. I, and S. R, "Network-based prediction of protein function," Molecular Systems Biology, vol. 3(88), 2007.

[6] J. LB, "Linkage disequilibrium and the search for complex disease genes," Genome Research, vol. 10(10), pp. 1435-1444, 2000.

[7] C. J., A. B., and J. A., "Disease candidate gene identification and prioritization using protein interaction networks," BMC Bioinformatics, vol. 10, 2009.

[8] Đ. V. Tùng, D. A. Trà, L. Đ. Hậu, and T. M. Phương, "Phân hạng gen gây bệnh sử dụng học tăng cường kết hợp với xác suất tiền nghiệm," Tạp chí Công nghệ thông tin & Truyền thông, vol. 13(33), pp. 55-66, 2015.

[9] H. Wang, C. K. Chang, H.-I. Yang, and Y. Chen, "Estimating the Relative Importance of Nodes in Social Networks," Journal of Information Processing Society of Japan, vol. 21(3), pp. 414-422, 2013.

[10] D.-H. Le and Y.-K. Kwon, "GPEC: A Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection," Computational Biology and Chemistry, vol. 37, pp. 17-23, 2012.

[11] D.-H. Le and Y.-K. Kwon, "Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization," Computational Biology and Chemistry, vol. 44, pp. 1-8, 2013.

[12] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," The American Journal of Human Genetics, vol. 82, pp. 949-958, 2008.

[13] D.-H. Le, "Network-based ranking methods for prediction of novel disease associated microRNAs," Computational Biology and Chemistry, vol. 58, pp. 139-148, 2015.

[14] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," Molecular BioSystems, vol. 8, pp. 1970-1978, 2012.

[15] L. Y and P. JC, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," Bioinformatics, vol. 26, pp. 1219-1224, 2010.

[16] D.-H. Le and V.-T. Dang, "Ontology-based disease similarity network for disease gene prediction," Vietnam J Comput Sci, p. 9, 2016.

[17] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," Genome Biology, vol. 10, 2009.

[18] J. Amberger, C. A. Bocchini, A. F. Scott and A. Hamosh, "McKusick's Online Mendelian Inheritance in Man (OMIM®)", Nucleic Acids Research, 37 (2009), pp. D793-D796.

[19] J. D. Osborne, S. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila and R. L. Chisholm, "GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM", Oxford University Press (2006).

AN IMPROVED METHOD FOR DETERMINING DISEASE-RELATED GENES

Nguyen Dai Phong, Dang Vu Tung, Le Duc Hau, Tu Minh Phuong

ABSTRACT *— In computational biology, the identification of disease genes often begins with prioritizing candidate genes according to their relevance to a disease phenotype. This helps to narrow the set of disease-related genes which need to be identified by intensive biomedical experiments. Currently, many different methods have been proposed to prioritize candidate genes based on the relationships between proteins, which are encoded in gene/protein interaction networks. Most of these methods are based on the assumption of "module disease", i.e. genes relating to the same disease tend to be located next to each other on the interaction network. These methods prioritize candidate genes which are close to known disease genes on the interaction network. However, during the course of experiments, we found that for many diseases, the known genes do not completely form a module, but are located far from each other on the interaction network. In such cases, the existing methods for gene prioritization are no longer effective. In this paper, we propose an improved method to prioritize genes related to the abovementioned diseases by increasing the linking weights for genes which are located away from known disease genes. We experimentally evaluate the efficiency in prioritizing genes of this method on 148 diseases on human's interaction network and compare its performance with that of other significant methods, such as Random walk with restart (RWR) algorithm and method based on the probability of association (ERIN). The experiment results show that our proposed method achieves high performance of 95.3%, which is better than RWR (93.4%) and ERIN (89.8%). In addition, by using such method, we are able to identify a number of new genes which are related to prostate cancer.*