Một phương pháp phân hạng gen gây bệnh mới dựa trên tổng xác suất liên kết trong mạng tương tác protein

A Novel Candidate Disease Genes Prioritization Method based on the Total Probability Links Protein Interaction Network

Đặng Vũ Tùng, Nguyễn Đại Phong, Lê Đức Hậu, Từ Minh Phương

***Abstract**: Prioritizing candidate disease-related genes using computational methods and biological networks data is an important problem in bioinformatics. Random walk with restart (RWR) algorithm is widely used for this problem due to its relatively high accuracy. However, RWR is computationally expensive as it considers every node in a network. Here we propose to use a new method for prioritizing candidate genes, in which genes with low probability of association with disease genes are excluded from further consideration, thus reducing computational complexity. Experiments on real protein interaction networks show that the proposed method was computationally efficient, and more accurate than RWR, as measured by AUC scores. We applied the proposed method to prioritizing candidate genes for human diabetes type 2. The results were promising: among top 20 ranked genes, 11 are associated with diabetes, as reported in the biomedical literature.*

***Keywords**: Protein Interaction Network, Genes Prioritization, Type 2 Diabetes, RWR.*

I. MỞ ĐẦU

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 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. Các khó khăn này hiện nay đã được giải quyết một phần bằng phương pháp phân hạng gen ứng viên liên quan đến bệnh trong Tin sinh học.

Mục đích của việc phân hạng các gen ứng viên theo mức độ liên quan đến một căn bệnh là để xác định các gen mới có liên quan đến bệnh. Cho đến nay, đã có nhiều phương pháp tính toán được phát triển nhằm mục đích phân hạng các gen ứng viên liên quan đến các bệnh di truyền [7-11]. Trong giai đoạn đầu, các phương pháp tính toán chủ yếu dựa trên dữ liệu chú giải chức năng. Theo đó, mức độ liên quan của gen ứng viên và bệnh nghiên cứu căn cứ vào độ tương tự về hồ sơ chức năng được xây dựng từ các dữ liệu chú giải của gen ứng viên và các gen bệnh đã biết [7, 9, 10]. Tuy nhiên, hạn chế của các phương pháp này đó là các dữ liệu chú giải chức năng thường không đầy đủ cho tất cả các gen/protein. Điều này ảnh hưởng đến việc xây dựng các hồ sơ chức năng cho tất cả các gen.

Gần đây, các phương pháp tính toán được chuyển theo hướng dựa trên các mạng sinh học do dữ liệu về tương tác giữa các gen/protein ngày càng đầy đủ và có thể bao phủ toàn bộ hệ gen. Các phương pháp này 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 [4]) 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 [12], RL_Rank [13] và bước ngẫu nhiên có quay lui (RWR) [14] cũng đã được sử dụng cho bài toán phân hạng các gen ứng viên trên các mạng tương tác protein. Trong số đó, phương pháp RWR được đánh giá là phương pháp nổi trội nhất [15]. Phương pháp này khai thác cấu trúc tổng thể của mạng dựa vào hành vi của một chuyển động ngẫu nhiên trên một mạng hay đồ thị. 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. 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 lui (back-probability). Các nút trên đồ thị được thăm nhiều hơn sẽ được xem là có độ quan trọng lớn hơn, đại lượng này đánh giá tầm quan trọng tương đối/độ liên quan của các nút còn lại so với tập các nút gốc. Khi áp dụng thuật toán này cho bài toán phân hạng gen gây bệnh, các gen gây bệnh đã biết đóng vai trò như các nút khởi đầu, các gen còn lại trên mạng được xem là các ứng viên. Kohler và cộng sự [14] đã áp dụng thuật toán này trên các mạng tương tác protein để xác định các gen gây bệnh mới. Kết quả thử nghiệm trên một tập gồm 110 bệnh cho thấy phương pháp này đạt được hiệu năng dự đoán tốt và cao hơn so với các phương pháp dựa trên dữ liệu chú giải chức năng. Không những đạt được hiệu năng cao trong bài toán phân hạng gen ứng viên liên quan đến bệnh, thuật toán này còn được sử dụng hiệu quả trong việc các định các microRNA mới liên quan đến bệnh [16] cũng như các đích tác động mới của thuốc [17]. Tuy nhiên, RWR phải duyệt qua tất cả các nút trên đồ thị thông qua các phép nhân ma trận, do đó nó có độ phức tạp tính toán cao đối với các đồ thị lớn như các mạng sinh học.

Trong bài báo này, chúng tôi sử dụng một phương pháp phân tích mạng xã hội của HeyongWang và cộng sự [18] cho bài toán phân hạng gen gây bệnh. Phương pháp này thực hiện tính toán xác suất liên kết giữa các gen ứng viên và các gen gây bệnh đã biết. Đồng thời, thiết lập một ngưỡng ý nghĩa để xác định những liên kết quan trọng nhất. Do đó khi duyệt, chúng tôi có thể bỏ qua rất nhiều gen ứng viên không đạt độ liên quan cần thiết để xác định một cách hiệu quả các gen ứng viên có độ liên quan cao nhất đối với các gen gây bệnh đã biết. Thuật toán được cài đặt và thử nghiệm cho bài toán phân hạng và tìm kiếm gen gây bệnh dựa trên bộ dữ liệu mạng tương tác gen/protein. Kết quả thực nghiệm cho thấy độ chính xác và thời gian thực hiện của phương pháp sử dụng tốt hơn so với phương pháp RWR trên cùng bộ dữ liệu thử nghiệm. Chúng tôi cũng đã áp dụng phương pháp để dự đoán các gen bệnh mới liên quan đến bệnh tiểu đường tuýp 2 (Diabetes Type 2) và xác định được 11 gen trong số 20 gen có thứ hạng cao có bằng chứng về sự liên quan giữa chúng với bệnh này từ các tài liệu y văn đã công bố.

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 ứng dụng. 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

II.1. 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 [19, 20]. 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 [21]. Kết

quả thu được 3.284 bệnh, trong đó mỗi bệnh có từ 1 đến 31 gen liên quan 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.

II.2. Bài toán phân hạng gen dựa trên mạng

Tính toán tầm quan trọng/độ liên quan của các nút trên đồ thị mạng là vấn đề đã được nghiên cứu trong một thời gian dài, đặc biệt là các mạng xã hội, mạng phân tích liên kết và mạng sinh học. Hầu hết các nghiên cứu tập trung vào việc đánh giá độ liên quan của các nút với một nút (hoặc một số nút) trung tâm còn gọi là các nút gốc dựa vào liên kết giữa các nút. Giả sử $G = (V, E)$ là một đồ thị vô hướng, có trọng số với V là tập các nút, E là tập các cạnh. Cho S ($S \subseteq V$) là tập các nút gốc và C ($C \subseteq V$) là tập các nút có liên kết với S. Yêu cầu của bài toán đặt ra là tính toán độ liên quan của các nút trong C đối với S.

Khi áp dụng mô hình này cho bài toán phân hạng gen, mạng tương tác gen/protein sẽ được biểu diễn bởi đồ thị G, 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; S là tập các gen bệnh đã biết, C là tập các gen ứng viên có liên kết với các gen trong S. Sơ đồ tổng quan của bài toán phân hạng gen dựa trên mạng được biểu diễn như Hình 1. Sau khi tính toán điểm số và xếp hạng cho các gen ứng viên, các gen có thứ hạng cao sẽ là các gen có khả năng liên quan tới bệnh.

II.3. Các phương pháp phân hạng liên quan

II.3.1. Thuật toán dựa trên xác suất đường đi

Trong phần này, chúng tôi trình bày thuật toán dựa trên xác suất đường đi. Thuật toán này được chúng tôi ứng dụng cho bài toán phân hạng gen gây bệnh đã giới thiệu. Do các mạng sinh học trên thực tế có các đặc tính cấu trúc tương đồng với các mạng xã hội và mạng web như "kích thước tự do" (scale-free) và "thế giới nhỏ" (small-world) [22], nhiều nghiên cứu đã áp dụng các thuật toán được sử dụng để phân hạng các nút trong mạng xã hội và mạng Web cho bài toán phân hạng các gen/protein trong các mạng sinh học [23]. Theo [18], một số khái niệm được định nghĩa như sau:

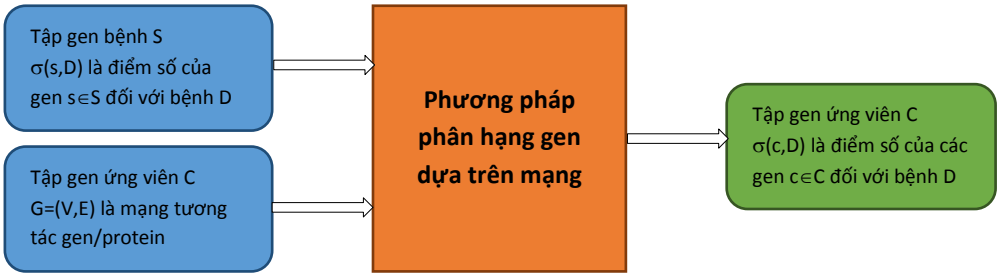
Đường đi không chu trình là đường đi không có bất kỳ nút nào được lặp lại. Giả sử p là một đường đi không chu trình trên đồ thị $G = (V, E)$, nó được mô tả như sau:

$$p = ((v_1, v_2, \dots, v_m) \mid \forall i, j: 1 \le i, j \le m, v_i \in V \text{ và } v_i \ne v_j \text{ nếu } i \ne j) \quad (1)$$

Cần lưu ý rằng trong trường hợp một nút truy vấn s chỉ có một nút láng giềng, mô hình bước ngẫu nhiên trên đồ thị sẽ coi s và láng giềng của nó có độ liên quan đến s như nhau. Để chắc chắn rằng nút s quan trọng hơn nút láng giềng, một hệ số giảm trừ f được sử dụng trong mô hình bước ngẫu nhiên và nó có thể được hiểu như là sự mất mát thông tin trong quá trình lan truyền. Như vậy, độ liên quan của một nút láng giềng u đối với nút s được định nghĩa là xác suất từ s chuyển ngẫu nhiên tới u với hệ số giảm trừ f ($0 < f < 1$) và xác định như sau:

$$P(s,u) = \begin{cases} (1-f)\dfrac{e(s,u)}{\sum_{v \in neighbor(s)} e(s,v)} & s \ne u \\ 1 & s = u \end{cases} \quad (2)$$

trong đó: $e(s,u)$, $e(s,v)$ là trọng số các cạnh tương ứng giữa nút s với các nút láng giềng u và v. 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.



Hình 1. Sơ đồ phân hạng gen dựa trên mạng

Khi quyết định đường đi *p* là quan trọng, một khái niệm mới gọi là xác suất đường đi của *p* được xác định để tính toán độ liên quan của nút kết thúc *t* đối với nút truy vấn *s* thông qua con đường xác suất *p* này. Xác suất đường đi $PPath_p(s, t)$ của đường đi từ *s* đến *t* trong đồ thị theo một đường đi không chu trình $p = (s, v_1, v_2, ..., t)$, được xác định như sau:

$$PPath_p(s,t) = P(s,v_1)\left(\prod_{i=1}^{m-1} P(v_i, v_{i+1})\right)P(v_m,t) \quad (3)$$

trong đó, $P(v_i,v_{i+1})$ được định nghĩa ở công thức (2). Rõ ràng xác suất đường đi *PPath* là một giá trị thuộc khoảng [0, 1] do các thừa số trong (3) cũng thuộc khoảng [0, 1]. Mỗi đường đi bao gồm một số liên kết giữa nút bắt đầu *s* và nút kết thúc *t*, nếu xác xuất đường đi là lớn, đường đi này sẽ biểu hiện độ liên quan cao giữa nút bắt đầu và nút kết thúc.

Ngưỡng ý nghĩa δ là một giá trị nằm trong khoảng [0, 1]. Đường đi có ý nghĩa từ nút *s* đến nút *t* là đường đi có xác suất lớn hơn hoặc bằng ngưỡng giá trị δ:

$$PPath_p(s,t) \geq \delta \quad (4)$$

Khi đó, độ liên quan của một nút *t* đối với một nút *s* được xác định là tổng tất cả các xác xuất đường đi có ý nghĩa từ nút *s* đến nút *t*:

$$I(t|s) = \sum_p PPath_p(s,t) \quad (5)$$

trong đó $\forall p \subseteq G$, có điểm bắt đầu là *s* và điểm kết thúc *t*, $PPath_p(s, t) \geq \delta$.

Độ liên quan trung bình của nút *t* so với một tập các nút truy vấn *S* được tính theo công thức sau:

$$I(t|S) = \frac{1}{|S|}\sum_{s\in S} I(t|s) \quad (6)$$

nếu nút *s* có nhiều đường đi có ý nghĩa tới *t*, điều này cho thấy rằng *t* có độ liên quan cao đối với *s*.

Mục tiêu của bài toán phân hạng gen là tìm kiếm và phân hạng tất cả các gen có liên quan tới gen bệnh đã biết, sau đó trích chọn *k* gen có độ liên quan cao nhất để các nhà y sinh học làm các thực nghiệm y sinh để khẳng định thêm khả năng liên quan đến bệnh. Trong đó, *k* là một số rất nhỏ so với tổng số gen trong đồ thị mạng tương tác gen/protein. Nếu xác suất đường đi của các con đường từ gen bệnh *s* tới gen *t* đều nhỏ hơn ngưỡng δ thì gen *t* hầu như không liên quan với gen bệnh *s*. Thuật toán tính tổng xác suất đường đi của mỗi gen ứng viên tới gen bệnh *s* đã biết được mô tả như sau:

```
void SingPathSum(node, δ, PPath[])
{
    Initialize: danh_dau_da_duoc_duyet(node);
    PPath[node] = 1;
    EnStack(node);
    While (Stack != rong)
    {
        s = DeStack();
        For (u là lang gieng cua s)
        {
            If (u chua duoc duyet)
            {
                PPath = PPath[s] * (1-
f)*e(s,u)/tong_trong_so_lang_gieng_cua_s;
                If (PPath < δ)
                    {
                    continue;
                    }
                EnStack(u);
                danh_dau_da_duoc_duyet(u)
                PPath[u] = PPath;
            }
        }
    }
}

For (u' la lang gieng của s)
{
    SingPathSum(u', δ, PPath[]);
}
```

Hình 2. Thuật toán SigPathSum tính tổng xác suất đường đi của mỗi gen ứng viên tới gen bệnh s

Đối với tập hợp các gen bệnh *S* đã biết của một bệnh, thuật toán sẽ thực hiện cho từng gen trong tập hợp. Độ liên quan trung bình của mỗi gen ứng viên đối với tập hợp các gen bệnh *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.

II.3.2. Thuật toán bước ngẫu nhiên có quay lui (RWR)

Trong bài báo này, chúng tôi so sánh hiệu năng phân hạng của phương pháp được đề xuất với phương pháp dựa trên thuật toán được đánh giá là tốt nhất hiện nay cho bài toán phân hạng gen ứng viên liên quan đến bệnh, đó là RWR. 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:

$$p_{t+1} = (1 - \gamma)Wp_t + \gamma p_0 \quad (7)$$

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 $v_i \in V$, 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, tập hợp các nút gốc S là các gen bệnh đã biết và các gen ứng viên là các gen còn lại trên mạng. Tất cả các gen trong mạng 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}).

II.4. 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 $S \setminus \{s\}$ 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:

$$sensitivity = \frac{TP}{TP+FN} \quad (8)$$

$$1 - specificity = \frac{FP}{FP+TN} \quad (9)$$

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 tính ổn định của thuật toán đồng thời lựa chọn bộ tham số tối ưu nhất cả về độ chính xác và thời gian thực hiện. Sau đó, chúng tôi so sánh hiệu suất của phương pháp đề xuất với phương pháp RWR trên cùng bộ dữ liệu theo giá trị AUC và thời gian thực hiện. Do sử dụng phương pháp LOOCV để đánh giá hiệu năng của các phương pháp phân hạng, nên mỗi bệnh phải có ít nhất hai gen liên quan và các gen này đều phải nằm trên mạng tương tác đã thu thập. Từ cơ sở dữ liệu, chúng tôi lọc ra được 398 bệnh phù hợp cùng với các gen liên quan của chúng dùng cho các thực nghiệm đánh giá. Hiệu năng của mỗi phương pháp là giá trị AUC trung bình trên 398 bệnh này. Cuối cùng, chúng tôi áp dụng phương pháp được đề xuất để tìm kiếm các gen gây bệnh mới liên quan đến bệnh tiểu đường tuýp 2.

III.1. Ảnh hưởng của các tham số

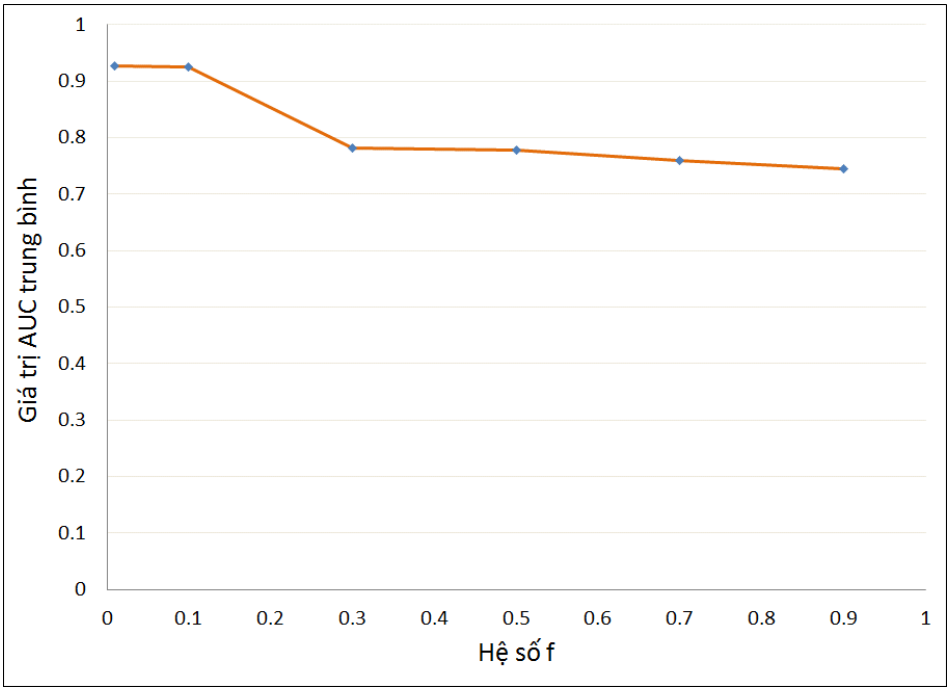
Thực nghiệm đầu tiên được chúng tôi tiến hành để xác định ảnh hưởng của tham số f tới hiệu quả của phương pháp đề xuất ứng dụng. Với một ngưỡng δ cố định ($\delta = 10^{-6}$), chúng tôi nhận thấy khi $f \leq 0.1$, giá trị AUC không thay đổi nhiều nhưng khi $f > 0.1$, số lượng gen bị loại bỏ nhiều, dẫn đến giá trị này giảm rất nhanh. Mặt khác, thời gian thực hiện trung bình khi $f = 0.1$ gần gấp đôi thời gian thực hiện trung bình khi $f = 0.3$. Kết quả thực nghiệm được ghi nhận trong Bảng 1 và Hình 3. Để thỏa mãn cả 2 tiêu chí về thời gian thực hiện nhanh và hiệu năng theo AUC cao, trong các thực nghiệm còn lại của bài báo này, chúng tôi lựa chọn $f = 0.1$.

Bảng 1. Kết quả thực hiện thuật toán với giá trị f thay đổi, tính trung bình trên 398 bệnh

f	*Thời gian thực hiện*	*Số gen được duyệt*	*Giá trị AUC*
0.01	6690.82s	3889	0.927
0.1	6013.35s	3614	0.925
0.3	3090.19s	1933	0.781
0.5	2860.59s	1256	0.778
0.7	1416.40s	1191	0.759
0.9	1238.25s	544	0.745

Bảng 2. Kết quả thực hiện thuật toán với giá trị δ thay đổi, tính trung bình trên 398 bệnh

δ	*Thời gian thực hiện*	*Số gen được duyệt*	*Giá trị AUC*
10^{-6}	6013.35s	3614	0.925
10^{-5}	3239.93s	1866	0.882
10^{-4}	1410.71s	643	0.827
10^{-3}	614.27s	132	0.740



Hình 3. Đường biểu diễn các giá trị AUC trung bình khi thay đổi giá trị f

Tiếp theo, chúng tôi thiết lập các giá trị ngưỡng δ khác nhau. Đối với mỗi ngưỡng, chúng tôi tiến hành phân hạng các gen ứng viên và tính giá trị AUC trung bình trên 398 bệnh, đồng thời tính số lượng các gen ứng viên được duyệt và thời gian thực hiện cho từng trường hợp, kết quả thực nghiệm được cho trong Bảng 2. Chúng tôi nhận thấy rằng, khi giá trị ngưỡng δ giảm, số lượng các gen được duyệt tăng dẫn đến kết quả phân hạng cũng tăng. Tuy nhiên, thời gian thực hiện thuật toán cũng tăng một cách đáng kể (từ 614.27s với $\delta = 10^{-3}$ đến 6013.35s với $\delta = 10^{-6}$). Việc chọn giá trị ngưỡng δ đóng một vai trò rất quan trọng trong phương pháp tiếp cận này. Với giá trị ngưỡng δ được lựa chọn phù hợp, thuật toán có thể đạt được sự tối ưu cả về độ chính xác và thời gian thực hiện.

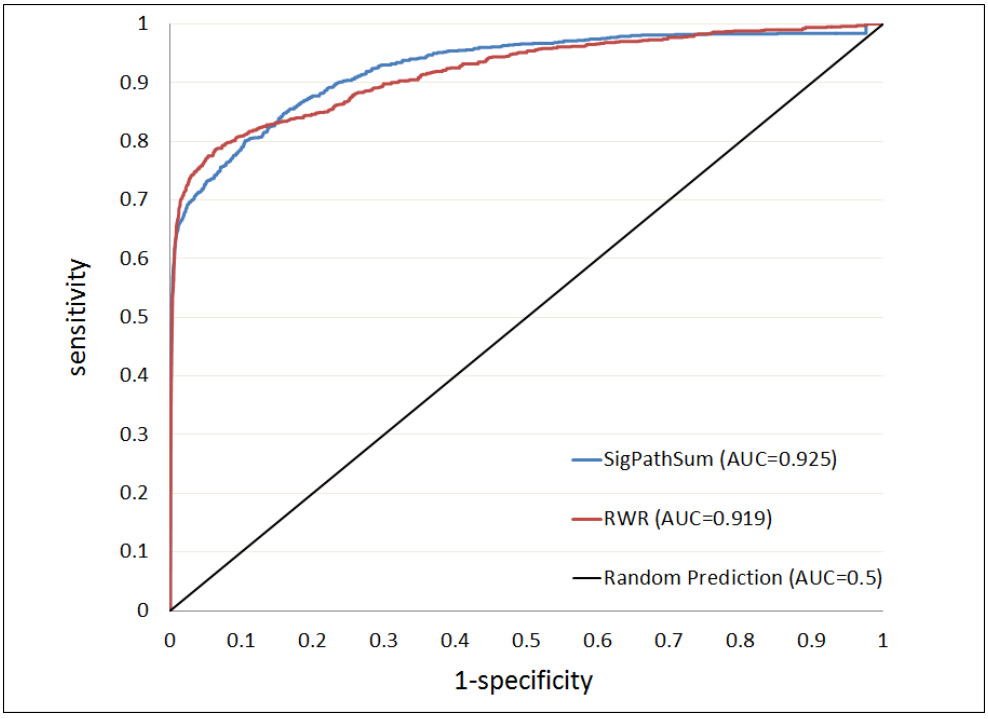
III.2. So sánh với RWR

Để khẳng định hiệu quả của phương pháp đề xuất, chúng tôi thiết lập giá trị các tham số $f = 0.1$, $\delta = 10^{-6}$ và so sánh kết quả phân hạng với phương pháp RWR. Theo [19] thì RWR đạt được hiệu quả lớn nhất với xác suất quay lui $\gamma = 0.7$.

Kết quả thực nghiệm trong Bảng 3 và Hình 4 cho thấy với $\delta = 10^{-6}$, giá trị AUC đạt được lớn hơn một chút so với phương pháp RWR nhưng thời gian thực hiện chỉ bằng 1/6 thời gian thực hiện RWR.

Bảng 3. Kết quả thực hiện SigPathSum với $f = 0.1$, $\delta = 10^{-6}$ và RWR với $\gamma = 0.7$, tính trung bình trên 398 bệnh

Thuật toán	*Thời gian thực hiện*	*Số gen được duyệt*	*Giá trị AUC*
SigPathSum	6013.35s	3614	0.925
RWR	37133.98s	11592	0.919



Hình 4. Biểu diễn đường cong ROC của SigPathSum và RWR

Bảng 4. Danh sách các gen gây bệnh tiểu đường tuýp 2 và số liên kết trong mạng PPI

TT	Ký hiệu của gen	Mã Entrez của gen	Số liên kết PPI
1	6833	ABCC8	6
2	208	AKT2	92
3	54901	CDKAL1	23
4	5167	ENPP1	7
5	2642	GCGR	2
6	2645	GCK	13
7	2820	GPD2	21
8	3159	HMGA1	8
9	6927	HNF1A	16
10	6928	HNF1B	44
11	3172	HNF4A	69
12	10644	IGF2BP2	2
13	3569	IL6	242
14	3667	IRS1	99
15	8660	IRS2	45
16	3767	KCNJ11	8
17	3990	LIPC	19
18	9479	MAPK8IP1	26
19	4760	NEUROD1	18
20	50982	NIDDM3	0
21	100188782	NIDDM4	0
22	5078	PAX4	13
23	3651	PDX1	0
24	5468	PPARG	27
25	5770	PTPN1	24
26	56729	RETN	2
27	6517	SLC2A4	27
28	169026	SLC30A8	5
29	6934	TCF7L2	13
30	7422	VEGFA	249
31	7466	WFS1	0
Tổng số liên kết trong mạng PPI			**1120**

Từ kết quả thực nghiệm thu được, chúng tôi nhận thấy: với các đồ thị có kích thước lớn như mạng tương tác protein của người, phương pháp RWR có chi phí tính toán cao cả về thời gian và không gian lưu trữ cần thiết. Khi đó, phương pháp đề xuất ứng dụng là một lựa chọn tối ưu hơn so với RWR.

III.3. Dự đoán các gen bệnh mới liên quan đến bệnh tiểu đường tuýp 2

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 tiểu đường tuýp 2 (Diabetes type 2) có mã OMIM 125853. Tiểu đường tuýp 2 là một nhóm bệnh rối loạn chuyển hóa cacbohydrat khi hoóc môn insulin của tụy bị thiếu hay giảm tác động trong cơ thể, biểu hiện bằng mức đường trong máu luôn cao. Đây là một trong những nguyên nhân chính của nhiều căn bệnh hiểm nghèo khác, điển hình là bệnh tim mạch vành, tai biến mạch máu não, mù mắt, suy thận, hoại thư, v.v..

Theo OMIM, có 31 gen đã được xác định là liên quan đến bệnh tiểu đường tuýp 2, trong đó có 27 gen nằm trên mạng tương tác gen/protein đã thu thập được sử dụng như các nút gốc. Danh sách các gen này được liệt kê trong Bảng 4. Chúng tôi coi các gen còn lại trên mạng đều là các gen ứng viên và tiến hành phân hạng dựa vào thuật toán đã đề xuất.

Sau khi tất cả các gen ứng viên đều được phân hạng, chúng tôi chọn ra 20 gen có thứ hạng cao nhất và thu thập các bằng chứng y văn được công bố trong cơ sở dữ liệu PubMed [24] về sự liên quan của các gen này với bệnh tiểu đường tuýp 2. Từ kết quả tra cứu thu thập được, chúng tôi thấy rằng có 11 gen đã được báo cáo có liên quan trực tiếp đến bệnh tiểu đường tuýp 2 (các gen đánh dấu * trong Bảng 5). Ví dụ gen INSR, mã hóa các thụ thể insulin, là một gen ứng viên cho bệnh tiểu đường type 2 [25]. Hơn nữa, khi phân tích DNA trong tế bào máu của 128 bệnh nhân tiểu đường tuýp 2 người Iran, Bahram Kazemi và cộng sự [26] cho thấy kết quả có 26% bệnh nhân bị đột biến gen INSR. Deniz Rende và cộng sự [27] thông qua phân tích cấu trúc mô đun bệnh đã chứng minh rằng gen CREBBP có liên quan mật thiết tới bệnh tiểu đường tuýp 2. Trong nghiên cứu của mình, Stephen A. Myers và cộng sự [28] đã nêu bật vai trò của hệ thống vận chuyển kẽm (các gen SLC30Ax) và vai trò sinh học của chúng trong quá trình phát sinh bệnh tiểu đường tuýp 2 v.v..

Bảng 5. Danh sách các gen có thứ hạng cao và các y văn liên quan

TT	Ký hiệu của gen	Mã Entrez của gen	Điểm phân hạng	Mô tả	Tài liệu y văn tham khảo trên PubMed
1	3764	KCNJ8	0.474112	Gen này liên quan đến đột biến bệnh tiểu đường ở trẻ sơ sinh	[32]
2	3759	KCNJ2	0.351654	Đột biến gen này có liên quan với hội chứng Andersen, được đặc trưng bởi tình trạng tê liệt tuần hoàn, rối loạn nhịp tim.	[33]
3	3175	ONECUT1	0.319095	Đột biến gen này gây ra bệnh ung thư tuyến tụy.	[34]
4	3643	INSR*	0.309135	Gen này mã hóa các thụ thể insulin, gây kháng insulin, là một ứng viên cho bệnh tiểu đường tuýp 2	[25], [26]
5	3670	ISL1	0.306483	Mã hóa gen này liên quan tới bệnh tiểu đường tuýp 1	[35]
6	1387	CREBBP*	0.285394	Gen này cho thấy mối liên quan giữa bệnh tiểu đường tuýp 2 và bệnh thần kinh cơ	[27]
7	7779	SLC30A1*	0.263269	Hệ thống vận chuyển kẽm đóng vai trò quan trọng trong việc tổng hợp, bài tiết và hoạt động của insulin.	[28]
8	2033	EP300*	0.237035	Gen này tương tác với phloridzin là tác nhân gây ra bệnh tiểu đường tuýp 2	[36]
9	6514	SLC2A2*	0.236819	Các đa hình trong gen này liên quan tới việc dung nạp gluco và điều tiết insulin. Biến thể di truyền có nguy cơ gây mắc bệnh tim mạch	[37], [38]
10	6667	SP1*	0.230503	Gen này được chứng minh có liên quan đến bệnh béo phì và tiểu đường tuýp 2	[39]
11	148867	SLC30A7*	0.224545	Hệ thống vận chuyển kẽm đóng vai trò quan trọng trong việc tổng hợp, bài tiết và hoạt động của insulin.	[28]
12	5451	POU2F1*	0.222273	Gen này nằm trên vùng nhiễm sắc thể 1q24 có mối liên kết với bệnh tiểu đường tuýp 2	[40]
13	59084	ENPP5	0.216666	Gen này mã hóa một màng glycoprotein type-I, đóng vai trò truyền thông của các tế bào thần kinh.	[41]
14	3110	MNX1	0.209235	Là nguyên nhân gây ra bệnh tiểu đường ở trẻ sơ sinh	[42]
15	1080	CFTR	0.197098	Đột biến gen này có liên quan với bệnh xơ nang và viêm tụy	[43]
16	207	AKT1	0.19335	Các rối loạn của gen này dẫn đến các bệnh như ung thư, tiểu đường, tim mạch và các bệnh về thần kinh.	[44]
17	1906	EDN1	0.192944	Liên quan tới bệnh lý võng mạc của bệnh nhân tiểu đường type 2	[45]
18	55532	SLC30A10*	0.192821	Hệ thống vận chuyển kẽm đóng vai trò quan trọng trong việc tổng hợp, bài tiết và hoạt động của insulin.	[28]
19	3766	KCNJ10*	0.188767	Gen này được xác định có liên quan đến bệnh tiểu đường tuýp 2 ở người da đỏ Pima và sáu nhóm người khác	[46]
20	8091	HMGA2*	0.187472	Các nucleotide polymorphisms (SNPs) trong loci HMGA2 liên quan đến bệnh nhân tiểu đường tuýp 2 ở người Nhật Bản	[47]

Các gen còn lại mặc dù không có bằng chứng trực tiếp liên quan đến bệnh nhưng chúng là nguyên nhân gây ra các bệnh tiểu đường tuýp 1, viêm tụy, ung thư tuyến tụy, rối loạn sản sinh insulin, kháng insulin và bệnh xơ nang. Các bệnh này cũng đã được chứng minh là có liên quan tới bệnh tiểu đường tuýp 2 [29-31]. Đối với các gen này, chúng tôi xem là những đề xuất cho các nhà y sinh học nghiên cứu và tìm kiếm các bằng chứng liên quan đến bệnh trong các phòng thí nghiệm.

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất ứng dụng một thuật toán mới trong phân tích mạng xã hội, mạng web để phân hạng và tìm kiếm các gen ứng viên có độ liên quan cao nhất đối với các gen bệnh đã biết dựa trên tổng xác suất đường đi giữa hai gen/protein trong mạng. Thực nghiệm cho thấy khi sử dụng một giá trị ngưỡng nhất định ($\delta = 10^{-6}$) kết quả phân hạng đạt được tốt hơn so với phương pháp dựa trên thuật toán RWR nhưng với thời gian thực hiện ít hơn. Chú ý rằng, mạng tương tác gen/protein có thể được hình thành bởi các tương tác vật lý giữa chúng hoặc có thể được xây dựng dựa trên độ tương tự về chức năng giữa các gen/protein trên mạng. Dẫn đến, các mạng gen/protein có thể có kích thước rất lớn để phản ánh đầy đủ mối quan hệ chức năng phức tạp giữa các thành phần trong tế bào. Phương pháp này được đề xuất để áp dụng trên các mạng tương tác gen/protein có kích thước lớn trong khi vẫn đảm bảo hiệu năng dự đoán cao. Kết quả thực nghiệm cũng cho thấy ngoài đạt được hiệu năng tổng thể cao, phương pháp này còn có thể sử dụng để xác định các gen mới liên quan đến một bệnh cụ thể. Các gen có thứ hạng cao nhưng chưa có bằng chứng y sinh trực tiếp về mối liên quan giữa chúng với bệnh xem xét có thể được đề xuất để các nhà nghiên cứu y sinh học tiếp tục nghiên cứu thực nghiệm.

Với các kết quả nghiên cứu và thực nghiệm đã thu được, chúng tôi hy vọng có thể phát triển phương pháp đề xuất ứng dụng thành công cụ tìm kiếm gen gây bệnh trong tương lai như [48]. Thêm vào đó, với sự gia tăng không ngừng của các dữ liệu sinh học, nhiều mạng sinh học cũng đươc cấu thành dựa trên các dữ liệu này. Việc tích hợp nhiều loại dữ liệu liên quan đến bệnh sẽ cải thiện hiệu năng của các thuật toán dựa trên mạng, cũng như tạo động lực để đề xuất các thuật toán mới hiệu quả hơn [19]. Thật vậy, bằng việc tích hợp thêm dữ liệu về độ tương tự giữa các kiểu hình bệnh, Li và cộng sự [49] đã sử dụng thuật toán bước ngẫu nhiên có khởi động lại cho mạng không đồng nhất bằng cách kết hợp mạng gen và mạng kiểu hình.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Quỹ phát triển khoa học và công nghệ quốc gia (NAFOSTED) trong đề tài mã số 102.01-2014.21.

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, 27 (2011), pp. 1741-1748.

[2] D. JONES, "*Steps on the road to personalized medicine"*, Nature Reviews Drug Discovery, 6 (2007), pp. 770-771.

[3] K. REYNOLDS, "*Achieving the Promise of Personalized Medicine"*, Clinical Pharmacology & Therapeutics, 92 (2012), pp. 401-405.

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

[5] 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, 75(2) (2004), pp. 161-173.

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

[7] A. EA, A. RR, E. KL, P. DJ and P. BS, "*Suspects: enabling fast and effective prioritization of positional candidates"*, Bioinformatics, 22 (2006), pp. 773-774.

[8] H. JE, K. AT, M. HL and P. MA, "*Candid: a flexible method for prioritizing candidate genes for complex human traits"*, Genetic Epidemiology, 32 (2008), pp. 779-790.

[9] A. S, L. D, M. S, V. L. P, C. B and E. AL, "*Gene prioritization through genomic data fusion"*, Nature Biotechnology, 24 (2006), pp. 537-544.

[10] C. J, X. H, A. BJ and J. AG, "*Improved human disease candidate gene prioritization using mouse phenotype"*, BMC Bioinformatics, 8 (2007).

[11] S. D, S. JM and S. M, "*Genedistiller - distilling candidate genes from linkage intervals"*, PLoS ONE,, 3 (2008).

[12] C. J., A. B. and J. A., "*Disease candidate gene identification and prioritization using protein interaction networks"*, BMC Bioinformatics, 10 (2009).

[13] 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, 13(33) (2015), pp. 55-66.

[14] 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, 82 (2008), pp. 949-958.

[15] S. NAVLAKHA and C. KINGSFORD, "*The power of protein interaction networks for associating genes with diseases."*, Bioinformatics 26 (2010), pp. 1057-1063.

[16] D.-H. LE, "*Network-based ranking methods for prediction of novel disease associated microRNAs"*, Computational Biology and Chemistry, 58 (2015), pp. 139-148.

[17] X. CHEN, M.-X. LIU and G.-Y. YAN, "*Drug–target interaction prediction by random walk on the heterogeneous network"*, Molecular BioSystems, 8 (2012), pp. 1970-1978.

[18] 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, 21(3) (2013), pp. 414-422.

[19] D.-H. LE and Y.-K. KWON, "*Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization"*, Computational Biology and Chemistry, 44 (2013), pp. 1-8.

[20] 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, 10 (2009).

[21] 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.

[22] D. J. WATTS and S. H. STROGATZ, "*Collective dynamics of small-world networks"*, Nature 393(1) (1998), pp. 440-442.

[23] B. H. JUNKER, D. KOSCHÜTZKI and F. SCHREIBER, "*Exploration of biological network centralities with CentiBiN"*, BMC Bioinformatics, 7:219 (2006).

[24] 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).

[25] B. D, S. M, G. S, M. PP, R. MR, M. V and R. V, "*Association of His1085His INSR gene polymorphism with type 2 diabetes in South Indians"*, Diabetes Technol Ther, 14 (2012), pp. 696-700.

[26] B. KAZEMI, N. SEYED, E. MOSLEMI, M. BANDEHPOUR, M. B. TORBATI, N. SAADAT, A. EIDI, E. GHAYOOR and F. AZIZI, "*Insulin Receptor Gene Mutations in Iranian Patients with Type II Diabetes Mellitus"*, Iranian Biomedical Journal, 13 (2009), pp. 161-168.

[27] D. RENDE, N. BAYSAL and B. KIRDAR, "*Complex Disease Interventions from a Network Model for Type 2 Diabetes"*, PLoS One, 8 (2013).

[28] S. A. MYERS, A. NIELD and M. MYERS, "*Zinc Transporters, Mechanisms of Action and Therapeutic Utility: Implications for Type 2 Diabetes Mellitus"*, Journal of Nutrition and Metabolism, 2012 (2012), pp. 13.

[29] C. S. C. RICHARD I. G. HOLT, ALLAN FLYVBJERG, BARRY J. GOLDSTEIN, *Textbook of Diabetes*, Wiley-Blackwell, 2010.

[30] L. PORETSKY, *Principles of Diabetes Mellitus*, Springer New York Dordrecht Heidelberg London, 2010.

[31] R. TAYLOR, "*Insulin Resistance and Type 2 Diabetes"*, Diabetes, 61 (2012), pp. 778-779.

[32] M. WINKLER, R. LUTZ, U. RUSS, U. QUAST and J. BRYAN, "*Analysis of two KCNJ11 neonatal diabetes mutations, V59G and V59A, and the analogous KCNJ8 I60G substitution: differences between the channel subtypes formed with SUR1."*, J Biol Chem, 284 (2009), pp. 6752-6762.

[33] K.-P. A, P.-C. A, P. P, B. K, M.-K. M, B. P, S. K, L. HY, Q. E, P. R, K. A and P. LJ, "*Andersen-Tawil syndrome: report of 3 novel mutations and high risk of symptomatic cardiac involvement"*, Muscle Nerve, 51 (2015), pp. 192-196.

[34] X. JIANG, W. ZHANG, H. KAYED, P. ZHENG, N. A. GIESE, H. FRIESS and J. KLEEFF, "*Loss of ONECUT1 expression in human pancreatic cancer cells"*, Oncol Rep, 19 (2008), pp. 157-163.

[35] P. HOLM, B. RYDLANDER, H. LUTHMAN and I. KOCKUM, "*Interaction and Association Analysis of a Type 1 Diabetes Susceptibility Locus on Chromosome 5q11-q13 and the 7q32 Chromosomal Region in Scandinavian Families"*, Diabetes, 53 (2004), pp. 1584-1591.

[36] V. RANDHAWA, P. SHARMA, S. BHUSHAN and G. BAGLER, "*Identification of Key Nodes of Type 2 Diabetes Mellitus Protein Interactome and Study of their Interactions with Phloridzin"*, OMICS: A Journal of Integrative Biology, 17 (2013), pp. 302-317.

[37] A. BORGLYKKE, N. GRARUP, T. SPARSØ, A. LINNEBERG, M. FENGER, J. JEPPESEN, T. HANSEN, O. PEDERSEN and T. JØRGENSEN, "*Genetic Variant SCL2A2 Is Associated with Risk of Cardiovascular Disease – Assessing the Individual and Cumulative Effect of 46 Type 2 Diabetes Related Genetic Variants"*, PLoS One, 7 (2012).

[38] O. LAUKKANEN, J. LINDSTRÖM, J. ERIKSSON, T. T. VALLE, H. HÄMÄLÄINEN, P. ILANNE-PARIKKA, S. KEINÄNEN-KIUKAANNIEMI, J. TUOMILEHTO, M. UUSITUPA and M. LAAKSO, "*Polymorphisms in the SLC2A2 (GLUT2) Gene Are Associated With the Conversion From Impaired Glucose Tolerance to Type 2 Diabetes: The Finnish Diabetes Prevention Study"*, Diabetes, 54 (2005), pp. 2256-2260.

[39] J. CHEN, Y. MENG, J. ZHOU, M. ZHUO, F. LING, Y. ZHANG, H. DU and X. WANG, "*Identifying Candidate Genes for Type 2 Diabetes Mellitus and Obesity through Gene Expression Profiling in Multiple Tissues or Cells"*, J Diabetes Res, 2013 (2013).

[40] N. MC, L. VK, T. CH, C. AW, S. WY, M. RC, Z. BC, W. MM, M. WW, H. C, W. CR, T. PC, J. WP and C. JC, "*Association of the POU class 2 homeobox 1 gene (POU2F1) with susceptibility to Type 2 diabetes in Chinese populations"*, Diabetic Medicine, 27 (2010), pp. 1443-1449.

[41] REFSEQ, *ENPP5 ectonucleotide pyrophosphatase/phosphodiesterase 5*, 2014.

[42] B. A, V. E, P. J, S. B, L. S, Y. L, H. M, C. H, B. K, S. R, P. M, A.-R. M, F. P and V. M, "*Transcription factor gene MNX1 is a novel cause of permanent neonatal diabetes in a consanguineous family"*, Diabetes Metab, 39 (2013), pp. 276-280.

[43] S. KONDO, K. FUJIKI, S. B. H. KO, A. YAMAMOTO, M. NAKAKUKI, Y. ITO, N. SHCHEYNIKOV, M. KITAGAWA, S. NARUSE and H. ISHIGURO, "*Functional characteristics of L1156F-CFTR associated with alcoholic chronic pancreatitis in Japanese"*, American Journal of Physiology - Gastrointestinal and Liver Physiology, 309 (2015), pp. 260-269.

[44] I. HERSA, E. E. VINCENT and J. M. TAVARÉ, "*Akt signalling in health and disease"*, Cellular Signalling, 23 (2011), pp. 1515-1527.

[45] H. LI, J. W. C. LOUEY, K. W. CHOY, D. T. L. LIU, W. M. CHAN, Y. M. CHAN, N. S. K. FUNG, B. J. FAN, L. BAUM, J. C. N. CHAN, D. S. C. LAM and C. P. PANG, "*EDN1 Lys198Asn is associated with diabetic retinopathy in type 2 diabetes"*, Molecular Vision, 2008 (2008), pp. 1698-1704.

[46] V. S. FAROOK, R. L. HANSON, J. K. WOLFORD, C. BOGARDUS and M. PROCHAZKA, "*Molecular Analysis of KCNJ10 on 1q as a Candidate Gene for Type 2 Diabetes in Pima Indians"*, Diabetes, 51 (2002), pp. 3342-3346.

[47] T. OHSHIGE, M. IWATA, S. OMORI, Y. TANAKA, H. HIROSE, K. KAKU, H. MAEGAWA, H. WATADA, A. KASHIWAGI, R. KAWAMORI, K. TOBE, T. KADOWAKI, Y. NAKAMURA and S. MAEDA, "*Association of New Loci Identified in European Genome-Wide Association Studies with Susceptibility to Type 2 Diabetes in the Japanese"*, PLoS One, 6 (2011).

[48] 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, 37 (2012), pp. 17-23.

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

Nhận bài ngày: 13/03/2016

SƠ LƯỢC VỀ TÁC GIẢ

ĐẶNG VŨ TÙNG

Sinh năm 1972.

Tốt nghiệp ĐH Tổng hợp Hà Nội năm 1995; Thạc sỹ chuyên ngành Hệ thống thông tin năm 2011; NCS khóa 2013 tại Học viện Công nghệ bưu chính viễn thông.

Hiện công tác tại bộ môn Tin học, Học viện Thanh thiếu niên Việt Nam.

Lĩnh vực nghiên cứu: hệ thống thông tin, tin sinh học.

Điện thoại: 0913542479
Email: tung_dv@yahoo.com

NGUYỄN ĐẠI PHONG

Sinh năm 1993.
Sinh viên trường ĐH Bách Khoa Hà Nội.
Lĩnh vực nghiên cứu: lập trình Matlab, tin sinh học.
Điện thoại: 0973794518
Email: phongnd.hust@gmail.com

LÊ ĐỨC HẬU

Sinh năm 1979.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 2002; Thạc sỹ khoa học ĐH Bách Khoa Hà nội năm 2008; Bảo vệ Tiến sĩ năm 2012 tại ĐH Ulsan, Hàn Quốc.

Hiện công tác tại Trung tâm Tin học, ĐH Thủy Lợi.

Lĩnh vực nghiên cứu: học máy và khai phá dữ liệu, tin sinh học và ứng dụng, phân tích và khai phá mạng xã hội, lập trình song song trên GPU với CUDA và OpenCL.

Điện thoại: 0912324564
Email: hauldhut@gmail.com

TỪ MINH PHƯƠNG

Sinh năm 1971.

Tốt nghiệp ĐH Bách khoa Taskent, Uzobekistan; Bảo vệ tiến sĩ năm 1995 tại Viện Điều khiển học thuộc Viện hàn lâm khoa học Uzobekistan.

Hiện công tác tại Khoa CNTT, Học viện Công nghệ Bưu chính viễn thông.

Lĩnh vực nghiên cứu: ứng dụng của học máy, tin sinh học.

Điện thoại: 0913507508
Email: phuongtm@ptit.edu.vn