BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Đức Tĩnh

NGHIÊN CỨU PHÁT TRIỂN MÔ HÌNH ĐỘNG LỰC CẠNH TRANH TRONG MẠNG THÔNG TIN PHỨC HỢP VÀ ỨNG DỤNG DỰ ĐOÁN GEN ĐIỀU TRỊ UNG THƯ

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Mã số: 9480104

Hà Nội – 2024

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Người hướng dẫn khoa học:
1. Người hướng dẫn 1: TS. Trần Tiến Dũng, Trường Đại học Công nghiệp Hà Nội
2. Người hướng dẫn 2: TS. Hoàng Đỗ Thanh Tùng, Viện Công nghệ Thông tin, Học viện Khoa học và Công nghệ.

Phản biện 1: …

Phản biện 2: …

Phản biện 3: ….

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện, họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi ……. giờ ……, ngày …….. tháng …….. năm ……..

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Hiện nay, việc xác định các gen đột biến gây ra bệnh hay còn được gọi là gen bệnh được thực hiện chủ yếu bằng các thực nghiệm xét nghiệm sinh học lâm sàng trên các mẫu bệnh phẩm [3]. Công việc này thường được thực hiện thủ công trong phòng thí nghiệm cho hàng nghìn gen ứng viên nằm trên một vùng nhiễm sắc thể khả nghi và cho độ chính xác cao nhưng đòi hỏi nhiều thời gian và chi phí [4]. Để giảm khối lượng mẫu cho thực nghiệm lâm sàng, các hướng tiếp cận công nghệ đã được giới thiệu như thống kê và học máy, bao gồm cả học sâu. Tuy có những đóng góp quan trọng nhưng hai hướng này gặp hạn chế là không hiểu được tương tác gen và cần tập mẫu lớn, trong khi việc xác định tập mẫu vẫn còn là thách thức.

Nhìn từ góc độ đồ thị mạng lưới, dữ liệu sinh học có thể được mô hình hóa thành các mạng phức hợp, mà ở đó các đỉnh được hiểu là các gen hoặc sản phẩm của gen, liên kết thể hiện sự tương tác giữa chúng [11]. Vì vậy, việc khai phá dữ liệu sinh học có thể được quy về bài toán khai phá dữ liệu trên mạng phức hợp. Cách tiếp cận này thường dẫn đến việc đề xuất các mô hình tính toán trên mạng [13], từ đó đưa ra bảng xếp hạng các đỉnh (gen) theo thuộc tính nào đó, các đỉnh có thứ hạng cao được cho là quan trọng và có thể liên quan đến mục tiêu dự đoán [13]. Sau khi xếp hạng, một số lượng nhỏ các đỉnh (gen/protein) có thứ hạng cao được đưa vào thực nghiệm lâm sàng, tìm kiếm minh chứng, để khẳng định chức năng của gen có liên quan đến bệnh hay không.

2. Mục tiêu nghiên cứu của luận án

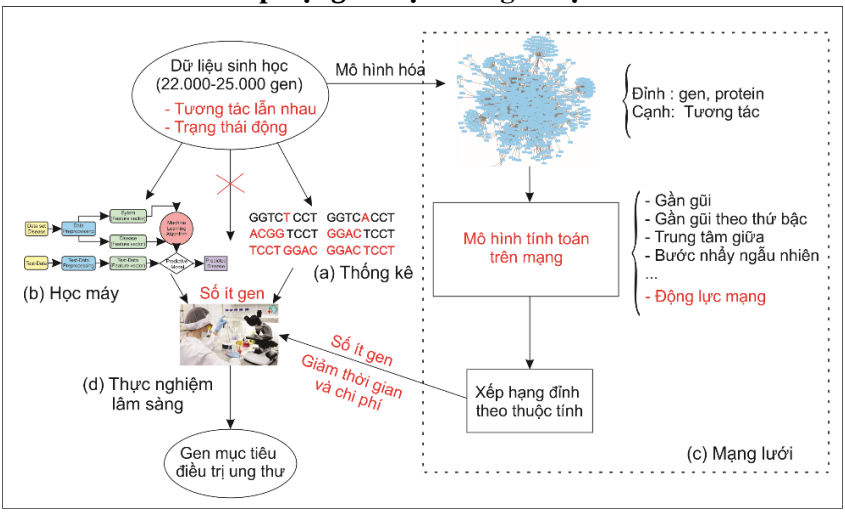
Mục tiêu nghiên cứu và phát triển các mô hình động lực cạnh tranh trong các mạng thông tin phức hợp, xác định thành phần điều khiển mạng, ứng dụng trên các mạng sinh học dự đoán gen mục tiêu điều trị ung thư.

3. Nội dung nghiên cứu

Hệ thống hoá các kiến thức cơ bản về lý thuyết đồ thị, lý thuyết mạng phức hợp, dữ liệu và mô hình hóa dữ liệu mạng sinh học, mô hình động lực cạnh tranh mạng, các mô hình và thuật toán phân hạng dự đoán chức năng của các đỉnh trên mạng phức hợp.

Chương 1. TỔNG QUAN VỀ PHÂN HẠNG ĐỂ DỰ ĐOÁN GEN MỤC TIÊU ĐIỀU TRỊ UNG THƯ

1.1. Bài toán xếp hạng để dự đoán gen bệnh



Hình 1.1. Bức tranh tổng quan dự đoán gen mục tiêu điều trị ung thư trên các mạng sinh học.

(a) hướng tiếp cận thống kê, (b) hướng tiếp cận học máy, (c) hướng tiếp cận dựa trên mạng, (d) thực nghiệm lâm sàng.

Sau đây, luận án phát biểu bài toán xếp hạng để dự đoán gen mục tiêu điều trị ung thư:

- Phát biểu bài toán: Cho một mạng sinh học, dự đoán các gen mục tiêu điều trị ung thư bởi thuốc.

- Đầu vào: Cho mạng sinh học $G = (V, E)$, với V là tập đỉnh (gen/protein) ($V = \{v_1, v_2, \dots, v_n\}$), E là tập cạnh (tương tác các gen) ($E = \{(v_i, v_j) | v_i, v_j \in V, i, j = 1, .., n\}$).

- Đầu ra: Một mối quan hệ $R^*(V,F)$, trong đó V là tập đỉnh; $F \in R^*$ cho biết khả năng đột biến của v gây ra ung thư và là mục tiêu điều trị.

1.2. Cơ sở lý thuyết

1.2.1. Lý thuyết đồ thị

1.2.2. Biểu diễn đồ thị trên máy tính

1.2.2.1. Ma trận kề

1.2.2.2. Ma trận trọng số

1.2.2.3. Danh sách cạnh

1.2.3. Mạng phức hợp

1.2.3.1. Các thành phần cơ bản của mạng phức hợp

1.2.3.2. Đặc trưng trên mạng phức hợp

1.2.3.3. Tính chất cơ bản của mạng phức hợp

1.2.3.4. Trung tâm mạng

1.2.3.5. Phân cụm mạng

1.2.4. Dữ liệu và mô hình hoá dữ liệu mạng sinh học

1.3. Các phương pháp và nghiên cứu liên quan dự đoán gen điều trị bệnh dựa trên mạng phức hợp

1.3.1. Thuộc tính gần gũi của một đỉnh

1.3.2. Thuộc tính gần gũi theo thứ bậc của đỉnh

1.3.3. Thuộc tính trung tâm giữa của một đỉnh

1.3.4. Thuật toán bước nhảy ngẫu nhiên có quay lại

1.3.5. Thuật toán ORIENT

1.3.6. Thuật toán sử dụng xác xuất tiền nhiệm PRINCE

1.4. Tổng quan về mạng quy mô lớn

1.4.1. Khái niệm mạng quy mô lớn

1.4.2. Một số hướng nghiên cứu trên mạng quy mô lớn

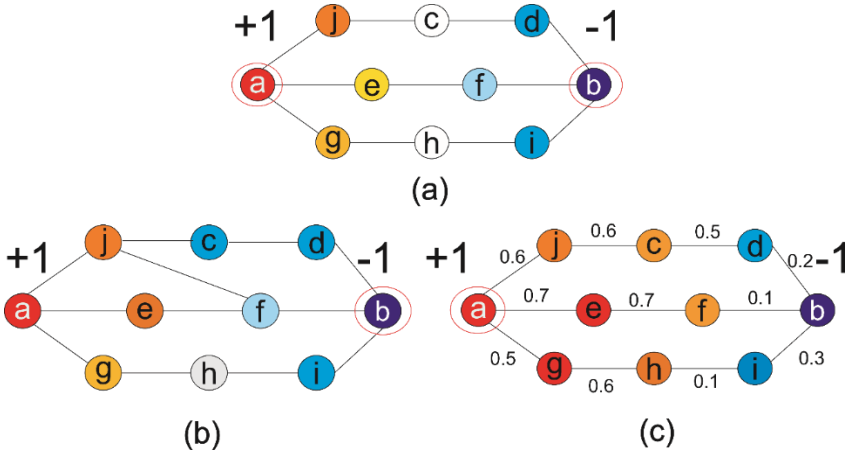
1.5. Mô hình động lực mạng

Chương 2. MÔ HÌNH ĐỘNG LỰC CẠNH TRANH TRÊN MẠNG PHỨC HỢP ỨNG DỤNG TRONG DỰ ĐOÁN GEN ĐIỀU TRỊ UNG THƯ

2.1. Mô hình động lực cạnh tranh trên mạng phức hợp

Zhao và cộng sự [104] đã giới thiệu một mô hình động lực học cạnh tranh trên các mạng phức hợp. Mô hình thể hiện sự cạnh tranh giữa hai tác nhân (đỉnh) bên trong mạng về khả năng kiểm soát hay ảnh hưởng của các tác nhân còn lại trong mạng với tác nhân đó.

Cho mạng trọng số *G(V,E)*, *n* đỉnh và *m* liên kết. Tập đỉnh $V = \{1,2,...,n\}$ và kiến trúc mạng được mô tả bởi ma trận kề $A = (a_{kl})_{nxn}$; nếu *k* tương tác trực tiếp với *l*, thì có một liên kết nối từ *k* đến *l* và $a_{kl} > 0$; ngược



Hình 2.1: Một ví dụ về mô hình động lực cạnh tranh trong trên mạng phức hợp [82].

(a) mạng vô hướng gồm 10 đỉnh với trọng số các cạnh bằng nhau, cuộc cạnh tranh giữa đỉnh a và đỉnh b kết thúc hòa. (b) mạng có nguồn gốc từ mạng (a) và được thêm một cạnh giữa đỉnh j và đỉnh f, kết quả cạnh tranh đỉnh b chiến thắng. (c) mạng có cấu trúc giống như mạng (a) nhưng có trọng số cạnh khác nhau, dẫn đến đỉnh a chiến thắng.

lại $a_{kl} = 0$. Giả sử có một cuộc cạnh tranh trong mạng giữa đỉnh *i* và đỉnh *j* mà có trạng thái cố định và khác nhau được biểu diễn bởi công thức (2.1).

$$x_i(t) = +1, x_j(t) = -1, \forall t \geq 0; i, j \in V \quad (2.1)$$

Khi đó, mỗi tác nhân bình thường còn lại trong mạng điều chỉnh trạng thái của mình theo một giao thức đồng thuận phân tán, thể hiện sự ảnh hưởng của từng tác nhân bình thường đến mỗi tác nhân cạnh tranh và dự đoán tác nhân cạnh tranh nào sẽ giành chiến thắng. Trạng thái các tác nhân bình thường được biểu diễn bởi công thức (2.2).

$$x_k(t+1) = x_k(t) + \varepsilon \sum_{\substack{l=1 \\ l \neq k \\ l \in V\{k\}}}^{n} a_{kl}\left(x_l(t) - x_k(t)\right) \quad (2.2)$$

Khi đó trạng thái của mỗi tác nhân bình thường cuối cùng sẽ đạt đến trạng thái ổn định, tức là $t \to \infty$ và được tính bởi công thức (2.3):

$$X_{norm}(t) \to \overline{X} = (\overline{D} - \overline{A})^{-1}\left[c_i c_j\right]\begin{bmatrix} +1 \\ -1 \end{bmatrix} \quad (2.3)$$

$X_{norm} \in R^{n-2}$ đại diện vector trạng thái hội tụ của các tác nhân bình thường.

Dấu của trạng thái ổn định thể hiện sự "thiên vị" của tác nhân đó. $\overline{x}_k > 0$ $(\overline{x}_k < 0)$ ngụ ý rằng tác nhân *k* cuối cùng sẽ hỗ trợ đối thủ cạnh tranh *i* (*j*), và $|\overline{x}_k|$ tương ứng với mức độ hỗ trợ hay ảnh hưởng. $\overline{x}_k = 0$ ngụ ý tác nhân *k* là tác nhân trung lập. Ta có công thức (2.4).

$$\Phi_{ij} = \sum_{\substack{k=1 \\ k \in V \setminus \{i,j\}}}^{n} sign(\overline{x}_k) \quad (2.4)$$

trong biểu thức trên, sign() là dấu của hàm. Nếu $\Phi_{ij} > 0$, thì tác nhân cạnh tranh *i* sẽ chiến thắng, nếu $\Phi_{ij} < 0$ tác nhân đối lập *j* sẽ dành chiến thắng, nếu $\Phi_{ij} = 0$ cạnh tranh kết thúc với tỷ số hoà.

Nghiên cứu không xem xét trường hợp một đối thủ cạnh tranh ở bên trong mạng, trong khi đối thủ còn lại ở bên ngoài mạng. Ngoài ra, việc chỉ

xem xét các tương tác trực tiếp từ các đỉnh đến mỗi đỉnh mạng có thể chưa hiệu quả với mạng lớn.

2.2. Đề xuất mô hình động lực cạnh tranh ngoài trên mạng phức hợp

Cho trước một mạng phức hợp *G(V, E)*, với *n* là số tác nhân (đỉnh) và *m* là số liên kết giữa chúng. Tập các tác nhân được mô tả là $V = \{1,2,\ldots,n\}$, và kiến trúc của mạng được mô tả bởi một ma trận kề trọng số $W=w(u,v)_{nxn}$; nếu tác nhân *u* liên kết trực tiếp với tác nhân *v* thì $w_{uv} > 0$, ngược lại $w_{uv} = 0$. Giả sử trạng thái ban đầu của các đỉnh trong mạng $x_u(t_0)=0$, $u \in V$. Chúng tôi giả định rằng đỉnh $\alpha \in V$ là một tác nhân điều khiển (ví dụ như gen đích của thuốc) và đỉnh $\beta \notin V$ là đối thủ cạnh tranh bên ngoài (tác nhân môi trường, thuốc), trong đó trạng thái của các đỉnh điều khiển và tác nhân đối thủ có các trạng thái cố định và khác nhau:

$$x_\alpha(t) = +1, x_\beta(t) = -1, x_u(t_0) = 0, \forall t \geq 0, \alpha, u \in V, \beta \notin V \qquad (2.5)$$

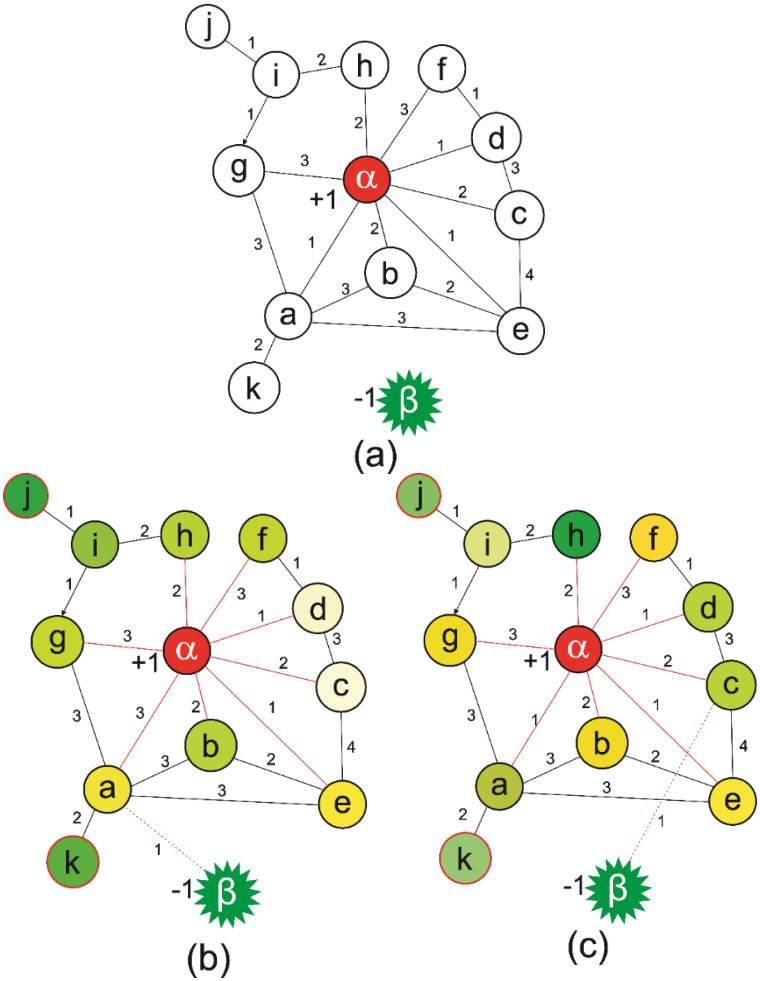
Bất cứ khi nào có một liên kết tạm thời có thể kết nối từ β đến bất kỳ đỉnh γ nào trong mạng để gây nhiễu loạn đối với α, bất cứ khi nào γ điều chỉnh trạng thái của nó. Mọi tác nhân còn lại được gọi là đỉnh bình thường và ký hiệu là $u \in V/\{\alpha, \beta\}$ có trạng thái ở thời điểm *t* là $x_u(t)$ và cập nhật trạng thái của nó ở thời điểm *t+1* là $x_u(t+1)$ theo công thức (2.6):

$$x_u(t+1) = x_u(t) + \varepsilon \sum_{\substack{v=1 \\ v \in V\{u\}}}^{n} w(u,v) * \left(x_v(t) - x_u(t)\right) \qquad (2.6)$$

tham số $0 < \varepsilon < Deg_{\max}^{-1}$ nắm giữ mức độ tương tác của các đỉnh gần kề, cùng với Deg_{max} là bậc ngoài lớn nhất của các đỉnh trong mạng; và $V\{u\}=\{v \in V | w(u, v) > 0\}$ là tập hợp các đỉnh lân cận của đỉnh *u* có thể ảnh hưởng trực tiếp đến đỉnh *u*. Với $t \to \infty$, trạng thái của mỗi đỉnh thông thường *u* hội tụ đến một giá trị ổn định $\bar{x}_u$, là tổ hợp lồi của các trạng thái đối thủ trong cuộc cạnh tranh. Dấu (âm/dương) của trạng thái ổn định của mỗi đỉnh bình thường: $\bar{x}_u > 0$ ($\bar{x}_u < 0$) ngụ ý rằng đỉnh *u* cuối cùng sẽ tác động hoặc ảnh

hưởng bởi đỉnh điều khiển α *hoặc* β, và $|\bar{x}_u|$ tương ứng với mức độ tác động/ảnh hưởng, $\bar{x}_u = 0$ nếu đỉnh u là trung lập. Xem hình 2.2.

Biểu thức tính tổng trạng thái tác động/ảnh hưởng của các tác nhân bình thường dành cho mỗi tác nhân α chống lại sự nhiễu loạn từ β được đề xuất bởi công thức (2.8).



Hình 2.2. Một ví dụ về mô hình động lực cạnh tranh ngoài.

$$ToS(\alpha) = \sum_{\substack{u=1 \\ u \in V \setminus \{\alpha,\beta\}}}^{n} sign(\bar{x}_u) \quad (2.8)$$

Đỉnh điều khiển của mạng được xác định bởi $C = \max_{\alpha \in V} ToS(\alpha)$.

Mạng có 12 đỉnh (gen/protein) và 19 tương tác, giả sử đỉnh α (đỏ) là đỉnh điều khiển có trạng thái được cố định bằng $+1$, β (xanh) là một tác nhân môi trường thiết lập trạng thái đối lập và cố định bằng -1. Ở thời điểm t một tương tác vô hướng được thêm tạm thời giữa tác nhân môi trường (thuốc) tới một đỉnh nào đó (đỉnh bình thường trong mạng), khi đó trạng thái của các đỉnh bình thường trong mạng sẽ thay đổi và hội tụ về một giá trị ổn định theo một giao thức đồng thuận phân tán là sự kết hợp lồi của trạng thái của các đối thủ cạnh tranh. Dải màu thể hiện mức độ ảnh hưởng của chúng tới đỉnh điều khiển bên trong mạng hay tác nhân bên ngoài. (a) trạng thái mạng ở thời điểm t_0, $x_u(t_0)=0$, $u \in V/(\alpha\beta)$. (b) trạng thái mạng ở thời điểm t. (c) trạng thái của mạng ở thời điểm $t+1$.

2.3. Xây dựng thuật toán của mô hình động lực học cạnh tranh ngoài

2.3.1. Ý tưởng của thuật toán

2.3.2. Chức năng, đầu vào, đầu ra của thuật toán

2.3.3. Sơ đồ luồng và mã giả của thuật toán

Mã giả của thuật toán

Thuật toán 2.1. Thuật toán của mô hình động lực cạnh tranh ngoài.

1	**function** OutsideCompetition(Graph *G(V,E)*, Node $\alpha \in V$) // $W=w(u,v)_{nxn}$ = *{start, end, direction, weight}*.
2	**begin**
3	*Epsilon = 2 * 1e-7f;*
4	**for** each Node in *V* **do**
5	**begin**

```
6    X0[Node] ← 0; //trạng thái ban đầu
7    end for
8    Xt[α] ← 1;    // Trạng thái của đỉnh điều khiển
9    Xt+1[α] ← 1;
10   Support ← new Dictionary<node, state>;
11   β ← new Node; //khởi tạo tác nhân ngoài
12   Xt[β] ← −1;   // trạng thái của tác nhân ngoài
13   Xt+1[β] ← −1;
14   NormalAgents ∈ V\{α, β};
15   for each γ in NormalAgents do
16   begin
17     e ← new Edge(β, γ);    //Tạo kết nối γ với β
18    E= E ∪{e};     // Bổ sung tập cạnh E
19    maxIterations ← n x m;  // n là số đỉnh và m số cạnh của G.
20    ε ← 1/Max(Deg(v), ∀v∈V);
21    t ← 0;
22    do
23     Converging ← 0;
24     for each u in V do
25      begin
26       if (u == α or u == β)
27       continue;
28        s ← 0;
29        for each v in Neighbors of u do
30         begin
31           s ← s + weight(u, v)*(Xt[v]- Xt[u]);
32         end for
33       Xt+1[u] ← Xt[u]+ ε * s;  // theo công thức 2.5
```

34	*Converging* ← *Converging* + Abs(*X_{t+1}[u]- X_t[u]);*
35	**end for**
36	*Temp* ← *X_t;*
37	*X_t* ← *X_{t+1};*
38	*X_{t+1}* ← *Temp;*
39	*t* ← *t +1;*
40	**while** *(Converging > Epsilon* and *t < maxIterations)*
41	*Support[γ]* ← *$\bar{X}$[γ];*
42	*E= E \ {e};* // Hủy kết nối γ đến β
43	**end for**
44	**return** *Support;* //trạng thái mạng ở thời điểm có kết nối với *β*
45	**end function.**
46	**function** ToS(Graph *G(V,E)*, Node *α∈V*)
47	**begin**
48	*Support* ← **new** *Dictionary<node,state>;*
49	*Support* ← *OutsideCompetition(G(V,E),* α*);*
50	*TotalSupport* ← *0;*
51	**for** each γ in *V - {α}* **do**
52	**begin**
53	*TotalSupport* ← *TotalSupport + Support[γ];*
54	**end for**
55	**return** *TotalSupport;* //tổng ảnh hưởng của các đỉnh đến α.
56	**end function**

Thuật toán gồm 2 hàm OutsideCompetition và hàm TOS. (a) hàm *OutsideCompetition (G(V,E), α ∈V)* tính toán sự ảnh hưởng của mỗi đỉnh đến đỉnh *α*, ở thời điểm mạng có kết nối với tác nhân ngoài *β* trong mô hình

động lực cạnh tranh ngoài. (b) hàm *ToS(G(V, E), α ∈ V)* tính tổng trạng thái ảnh hưởng của các đỉnh trong mạng đến đỉnh α.

2.4. Đánh giá độ phức tạp của thuật toán

Tổng hợp: Độ phức tạp tính toán của thuật toán động lực cạnh tranh ngoài là $O(n^3 \times m^2)$.

2.5. Xây dựng hệ thống dự đoán gen điều trị ung thư sử dụng mô hình động lực học cạnh tranh ngoài

2.5.1. Bài toán dự đoán gen mục tiêu điều trị ung thư

Đầu vào: Cho trước một mạng sinh học *G(V,E)*, với *V* là tập gen/protein (đỉnh) ($V = \{v_1, v_2, \dots, v_n\}$), *E* là tập tương tác các gen (cạnh) ($E = \{(v_i, v_j) | v_i, v_j \in V, i, j = 1, .., n\}$).

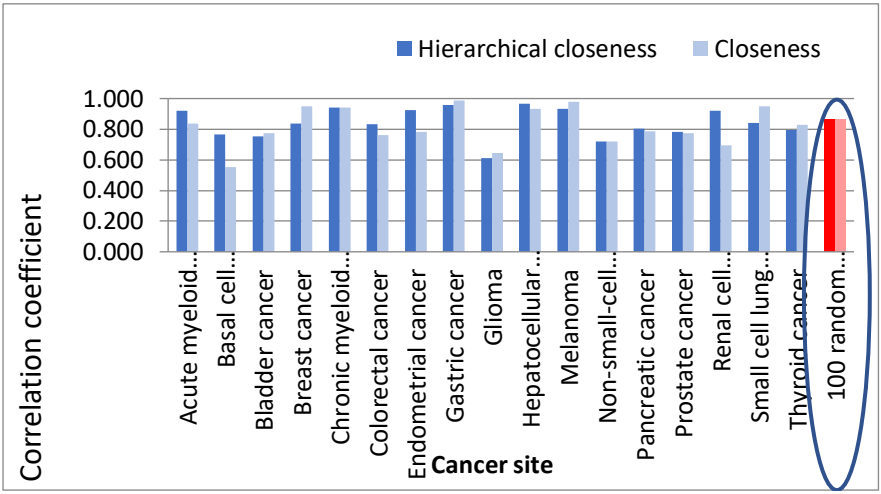
Đầu ra: Bảng xếp hạng các gen theo tổng trạng thái ảnh hưởng của các gen đến mỗi gen trong mạng. Các gen có thứ hạng cao được tìm kiếm minh chứng sinh học là các gen mục tiêu điều trị ung thư.

2.5.2. Dữ liệu thực nghiệm

Luận án sử dụng dữ liệu 17 mạng tín hiệu ung thư từ cơ sở dữ liệu KEGG (www.genome.jp/kegg) để tiến hành phân tích. Dữ liệu sau tiền xử lý có thể tải về tại đường dẫn sau https://github.com/tinhpd/NetCMD.git

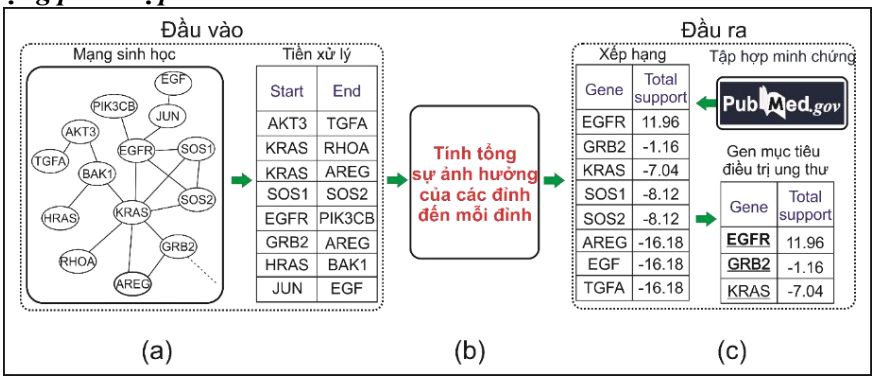
2.5.3. Sự tương quan giữa các phép đo

Thử nghiệm trên 17 mạng truyền tín hiệu ung thư và 100 mạng có hướng ngẫu nhiên được tạo ra bởi mô hình phát triển mạng Barabasi với | V | = 50 và 49 ≤ | E | ≤ 100, cho thấy rằng tổng sự ảnh hưởng của mỗi đỉnh tương quan với mức độ gần gũi và mức độ gần gũi theo thứ bậc của đỉnh, mà thứ hạng cao của hai phép đo này đã được sử dụng để dự đoán gen bệnh và gen chỉ báo ung thư, cũng thường là gen mục tiêu điều trị ung thư (Hình 2.5).



Hình 2.5. Sự tương quan phép đo gần gũi với tổng sự hỗ trợ

2.5.4. Mô hình tổng thể hệ thống chẩn đoán gen ung thư dựa trên mạng phức hợp



Hình 2.3: Mô hình tổng thể xác định gen mục tiêu điều trị ung thư.

Mô hình được thiết kế theo hướng tiếp cận mạng phức hợp. (a) Tiền xử lý dữ liệu đầu vào, (b) mô hình tính toán và thuật toán, tính toán thuộc tính đỉnh của mạng, (c) tổ chức dữ liệu đầu ra và đối sánh cơ sở dữ liệu để dự đoán gen tiềm năng cho các thực nghiệm tiếp theo.

2.5.5. Kết quả dự đoán gen mục tiêu điều trị ung thư

Thực nghiệm trên 17 mạng tín hiệu ung thư, kết quả 42/51 tương đương 82.36% các gen thuộc top 3 có tổng sự ảnh hưởng cao nhất là các gen mục tiêu điều trị ung thư. Các gen đậm gạch chân là đã được phê duyệt sản xuất thuốc, gen gạch chân đang thực nghiệm lâm sàng, các gen còn lại được coi là gen tiềm năng.

Bảng 2.1. Hiệu suất xác định gen mục tiêu điều trị ung thư bởi mô hình động lực cạnh tranh ngoài.

Mạng tín hiệu ung thư	**Các gen Top 3**		
	Cl	***C2***	***C3***
Bệnh bạch cầu dòng tủy cấp	GRB2	**FLT3**	**PML**
Ung thư biểu mô tế bào đáy	SUFU	**SMO**	GLI3
Ung thư bàng quang	RASSF1	**FGFR3**	HRAS
Ung thư vú	**LRP6**	LRP5	**WNT1**
Bệnh bạch cầu dòng tủy mãn	CRK	CRKL	GAB2
Ung thư đại trực tràng	**EGFR**	**GRB2**	KRAS
Ung thư nội mạc tử cung	EGF	EGFR	AXIN1
Ung thư dạ dày	LRP6	LRP5	WNT7A
U thần kinh	CALM1	CALML5	CALM2
Ung thư biểu mô tế bào gan	LRP6	WNT3A	WNT7A
Khối u ác tính	FGF2	FGF1	HGF
Ung thư phổi không phải tế bào nhỏ	**ALK**	EML4	KRAS
Ung thư tuyến tụy	**KRAS**	AKT2	AKT1
Ung thư tuyến tiền liệt	**IGF-1**	INS	PDGFB
Ung thư biểu mô tế bào thận	HGF	**MET**	EGLN2
Ung thư phổi tế bào nhỏ	ITGB1	COL4A1	LAMB3
Ung thư tuyến giáp	NTRK1	TPR	TPM3

Bảng 2.1 gồm các gen mục tiêu điều trị ung thư được xác định theo xếp hạng tổng trạng thái tác động. Trong bảng, C1, C2 và C3 biểu thị các ký hiệu gen NCBI của ba gen hàng đầu có tổng trạng thái tác động cao nhất. Các gen được gạch chân (42 trong số 51) trước đây đã được báo cáo là gen đích của thuốc chống ung thư. Trong số đó, có 12 gen gạch chân và bôi đậm là các gen đã được chấp nhận để sản xuất thuốc và 30 gen gạch chân không bôi đậm là các gen đang trong các giai đoạn thử nghiệm lâm sàng. Các gen còn lại không được gạch chân gồm 09 gen vẫn chưa được nghiên cứu đầy đủ có thể là gen đích của thuốc chống ung thư tiềm năng và có ý nghĩa tham khảo.

2.5.6. So sánh kết quả dự đoán

Cả hai nghiên cứu đều được tiến hành trên cùng bộ dữ liệu là 17 mạng tín hiệu ung thư KEGG. Kết quả thể hiện ở bảng 2.3.

Bảng 2.2: Kết quả so sánh giữa hai mô hình khác nhau trên cùng bộ dữ liệu.

3	**Số mạng dự đoán được**	**Tỷ lệ dự đoán trên top 3**	**Tổng thời gian thực thi (phút)**
Mô hình tính toán thuộc tính gần gũi theo thứ bậc [13, 99].	16/17	37/48 gen, tương đương 70,59%,	124
Mô hình động lực cạnh tranh ngoài	17/17	42/51 gen, tương đương 82,36%	126

Hệ thống thử nghiệm: ASUS X510U, Intel i5-8250U CPU, xung nhịp 1.6GHz (8CPUs), bộ nhớ DDRAM 8GB DDR IV, Rander NVIDIA GeFore 940MX 2GB, SSD M2 120GB Intel.

Chương 3. TƯƠNG TÁC GIÁN TIẾP TRONG MÔ HÌNH ĐỘNG LỰC CẠNH TRANH NGOÀI VÀ ỨNG DỤNG DỰ ĐOÁN GEN ĐIỀU TRỊ UNG THƯ

3.1. Đề xuất mô hình động lực cạnh tranh ngoài cải tiến

Luận án gọi F là ma trận ảnh hưởng (tác động/ tương tác giữa các phần tử trong mạng), trong đó mỗi phần tử của ma trận F mô tả sự ảnh hưởng của tác nhân (đỉnh) này lên tác nhân khác. Lưu ý rằng nếu có một liên kết trực tiếp từ tác nhân u đến tác nhân v, khi đó ta hiểu tác nhân v tương tác/ảnh hưởng trực tiếp đến tác nhân u. Trong trường hợp khác, không có liên kết trực tiếp từ u đến v, nghĩa là có một tương tác từ tác nhân u đến tác nhân γ và một tương tác từ tác nhân γ đến v, khi đó tác nhân v tác động gián tiếp đến tác nhân u thông qua tác nhân γ, Tác động gián tiếp như vậy thường sẽ yếu hơn tác động trực tiếp.

Giả sử, gọi $D=(d_{uv})_{NxN}$ là ma trận khoảng cách biểu diễn mạng.

Luận án định nghĩa ma trận $F=(f_{uv})_{NxN}$ là ma trận ảnh hưởng của mạng, thể hiện sự ảnh hưởng của tác nhân v đến tác nhân u, với $\forall u,v \in V$, và được tính bằng công thức (3.5).

$$f(u,v) = \frac{x(v)}{\left(d(u,v)\right)^2} \tag{3.5}$$

Trong đó, x_v là trạng thái của đỉnh v ở thời điểm t, $t \to \infty$; d_{uv} là khoảng cách đường đi ngắn nhất từ $u \to v$.

Gọi $f_{\alpha v}$ là phần tử của ma trận ảnh hưởng F trên hàng thứ α và cột thứ v. Khi đó v sẽ tác động / ảnh hưởng đến α một đại lượng nào đó, và biểu

thức tính tổng sự ảnh hưởng của các tác nhân *v* đến mỗi tác nhân điều khiển α được tính bởi công thức (3.6).

$$ToSF(\alpha) = \sum_{\substack{v=1 \\ v \in V \setminus \{\alpha,\beta\}}}^{n} sign(f(\alpha,v) - f(\beta,v)) \qquad (3.6)$$

Trong đó, *sign()* là dấu (+) hay (–) thể hiện sự ảnh hưởng/tác động đến đỉnh điều khiển α hay tác nhân bên ngoài cạnh tranh β. Nếu $f(\alpha,v)>f(\beta,v)$thì đỉnh *v* sẽ ảnh hưởng nhiều đến đỉnh điều khiển α hơn, ngược lại $f(\alpha,v) <f(\beta,v)$ có nghĩa là đỉnh *v* sẽ ảnh hưởng/tác động nhiều về tác nhân bên ngoài β hơn, nếu $f(\alpha,v) = f(\beta,v)$ thì đỉnh *v* là trung lập; $ToSF(\alpha)$ trả về mức độ ảnh hưởng/tác động của các đỉnh bình thường *v* trong mạng đến đỉnh điều khiển α trong mô hình động lực cạnh tranh ngoài cải tiến.

3.2. Xây dựng thuật toán tính toán tương tác gián tiếp động lực cạnh tranh ngoài

3.2.1. Thuật toán tính ma trận khoảng cách

Trong nội dung nghiên cứu này, luận án sử dụng thuật toán Floyd-Warshll [100] để tính ma trận khoảng cách giữa các đỉnh trên đồ thị mạng có trọng số. Thuật toán có 3 vòng lặp với *n* lần, vì vậy độ phức tạp của thuật toán là $O(n^3)$.

3.2.2. Thuật toán tính toán ma trận ảnh hưởng

```
1  function Matrix F[,] InfluenceMatrix(Graph G(V,E), Node α∈V)
   //đầu vào: ma trận kề có trọng số W=w(u,v)nxn; α
   //đầu ra: ma trận ảnh hưởng F=f(u,v)nxn
2    D ← DistanceMatrix(G(V,E)) // Tính ma trận khoảng cách
3    X ← OutsideCompetition(G(V,E),α) // Tính trạng thái các đỉnh
4    for each vertex u in V do // duyệt hàng
5        for each vertex v in V do // duyệt cột
6            if D[u,v]==0 then // Giá trị trên đường chéo
7                F[u,v] ← NA // Gán giá trị không xác định
```

```
8          else
9            F[u,v] ← X(v) / (D[u,v])^2 // Tính ảnh hưởng
10         end if
11       end for
12     end for
13   return F //ma trận ảnh hưởng F
14 end function
```

Độ phức tạp thời gian tính toán của hàm InfluenceMatrix là $O(n^3+m^2)$.

3.2.3. Thuật toán tính tổng sự ảnh hưởng trên mỗi đỉnh mạng

```
function ToSF(Graph G(V,E), Node α, out result)
1  // đầu vào: ma trận kề trọng số W, α.
   // đầu ra: tổng sự ảnh hưởng của các đỉnh đến đỉnh α
2  F ← InfluenceMatrix(G(V,E),α) // Tính ma trận ảnh hưởng
3  TotalSupportF ← 0 // Khởi tạo tổng sự ảnh hưởng
4  for each v in V - {α, β} do
5    TotalSupportF ← TotalSupportF + (F[α, v] – F[β, v])
     // theo công thức 3.6
6  end for
7    result ← TotalSupportF  // Tổng sự ảnh hưởng của các đỉnh
   đến đỉnh α.
8  end procedure
```

Độ phức tạp thời gian của hàm ToSF là $O(n^3+m^2)$.,

3.3. Tính toán hiệu năng cao cho mô hình động lực cạnh tranh ngoài

3.3.1. Xây dựng thuật toán tính toán hiệu năng cao cho mô hình

```
function Matrix DnF[,] ParFindDriverNode(Graph G(V,E))
1  //đầu vào:ma trận kề trọng số W=(w_uv)_nxn, {start, end, direction,
   weight};
```

	// đầu ra: tổng sự ảnh hưởng của các đỉnh trong mạng đến mỗi đỉnh mạng
2	*DnF = new Matrix[n, n]* // Khởi tạo biến kết quả
3	// Thực hiện tính toán song song cho mỗi đỉnh α trong V
4	**parallel for** each α in V **do**
5	*result* $\leftarrow 0$ // Biến cục bộ để lưu kết quả cho mỗi α
6	*ToSF(G(V,E),α, result)* // Gọi hàm ToSF để tính tổng sự ảnh hưởng
7	Wait for all works done // Chờ cho tất cả các công việc hoàn thành (synchronize)
8	*DnF[α,]* $\leftarrow$ *result* // Lưu kết quả vào ma trận DnF
9	**end parallel**
10	**return** *DnF* //Ma trận tổng sự ảnh hưởng của các đỉnh trong mạng đến mỗi đỉnh mạng
11	**end function**

Độ phức tạp thời gian phụ thuộc vào độ phức tạp thời gian của hàm ToSF, bao gồm tính toán ma trận ảnh hưởng với độ phức tạp là $O(n^3+m^2)$.

3.3.2 Thiết kế công cụ phần mềm tính toán hiệu năng cao

Phần mềm *Drivergen.net* được phát triển dựa trên mô hình động lực cạnh tranh ngoài với khả năng tính toán hiệu năng cao trên CPUs đa lõi. Nó được thiết kế hoạt động như một trình cắm Cytoscape, với giao diện đồ họa (GUI). Chi tiết về phần mềm cùng dữ liệu thực nghiệm có thể được tải xuống từ https://github.com/tinhpd/Drivergen.git

3.3.3. Đánh giá hiệu suất và tốc độ tính toán của thuật toán

Bảng 3.3 cho thấy kết quả thử nghiệm của phần mềm Drivergen.net với các chế độ tính toán khác nhau trên 04 mạng sinh học. Kết quả cho thấy tốc độ tăng tốc cải thiện đặt từ 51 – 145 lần tuỳ thuộc vào loại mạng cụ thể.

Bảng 3.3. Năng lực tính toán trên mạng quy mô lớn

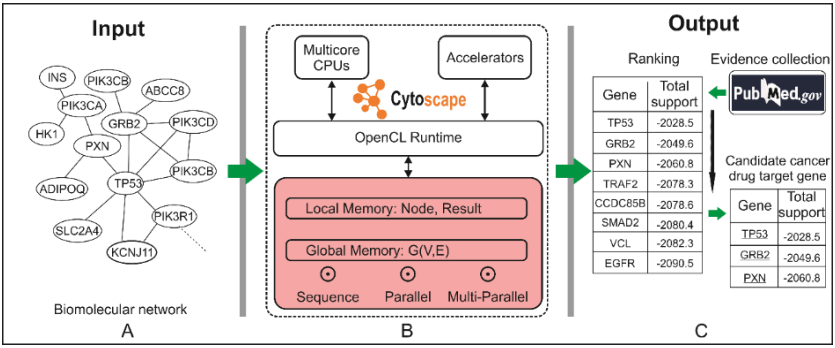
Mạng	Thuộc tính			Thời gian (phút)		Tăng tốc (lần)
	Loại mạng	***Số đỉnh***	***Số cạnh***	***Tuần tự***	***Song song***	
Mạng virus cytomegalo	Vô hướng	213	1214	5,7	0,11	51,8
Mạng tương tác protein vi khuẩn E. coli		850	1193	341	5	68,2
Mạng lưới điều hòa gen	Mạng có hướng	943	3917	207	7	29,5
Mạng tín hiệu tế bào	Mạng hỗn hợp	1549	5074	5092	35	145,5
Hệ thống thử nghiệm: Dell OptiPlex 5050, CPU Intel i7-7700 tám nhân với xung nhịp 3,6GHz, bộ nhớ DDRAM 32GB DDR IV.						

3.4. Thực nghiệm

3.4.1. Dữ liệu thực nghiệm

Luận án thực nghiệm trên 03 loại mạng sinh học quy mô lớn, được tải về từ những công bố uy tín trước đây. Dữ liệu sau tiền xử lý của 03 mạng được lưu trữ và có thể tải về tại đường dẫn https://github.com/tinhpd/Drivergen.git

3.4.2. Kiến trúc của mô hình dự đoán



Hình 3.1. Mô hình dự đoán gen mục tiêu điều trị ung thư trên mạng lớn

(a) dữ liệu mạng sinh học đầu vào, (b) kiến trúc thiết kế cho mô hình tính toán, (c) tổ chức dữ liệu đầu ra và tìm kiếm minh chứng đối sách. Dữ liệu thực nghiệm, phần mềm và hướng dẫn sử dụng của nghiên cứu này được lưu trữ và có thể tải về tại đường dẫn https://github.com/tinhpd/Drivergene

3.4.3. Kết quả dự đoán gen mục tiêu điều trị ung thư

Kết quả dự đoán trên 03 mạng sinh học quy mô lớn cho thấy 86,67 %, tức là 26 trong tổng số 30 gen top 10 có có tổng trạng thái ảnh hưởng cao nhất là các gen mục tiêu của thuốc trong liệu pháp điều trị ung thư.

Bảng 3.4. Xác định gen mục tiêu thuốc ung thư trên 3 mạng lớn.

Mạng sinh học	**Thuộc tính**			**Tên gen**	**Minh chứng từ cơ sở dữ liệu PubMed.gov**
	Loại mạng	***Số đỉnh***	***Số cạnh***		
Mạng lưới điều hòa gen	Mạng có hướng	943	3917	*NFKB1*	30205516
				RELA	Cho nghiên cứu tiếp theo
				JUN	32917236
				FOS	34610301
				MYC	22464321
				STAT1	33608980
				CCND1	29969496
				CREB1	30127997

				STAT3	24743777
				HIF1A	28358664
Mạng tín hiệu tế bào	Mạng hỗn hợp	1549	5074	*SRC*	11114744
				AR	24425228
				AKT	27232857
				SHC	Cho nghiên cứu tiếp theo
				SMAD3	20010874
				RAC1	32460002
				GAB2	22858987
				PI3K	30782187
				PKA	24212646
				SMAD4	29602802
Mạng tương tác protein	Mạng vô hướng	7279	21911	*TP53*	23115424
				GRB2	29550383
				PXN	34135128
				TRAF2	30294322
				DIPA	Cho nghiên cứu tiếp theo
				SMAD2	20010874
				VCL	Cho nghiên cứu tiếp theo
				EGFR	28368335
				SRC	11114744
				SMAD3	20010874

Ngoài ra, các gen top 10 trong bảng 3.4 được tìm thấy thuộc lõi K-core và R-core [49] trong cùng của mạng.

Bảng 3.5. Xác định lõi K-core và R-core

Loại mạng	**Loại lõi**	
	K-core	***R-core***
Mạng tín hiệu tế bào	80%	

Mạng điều hòa gen		70%
Mạng tương tác protein	60%	

Kết quả này phù hợp với kết quả của các nghiên cứu trước đây rằng các gen dấu ấn sinh học ung thư quan trọng thường nằm ở lõi trong cùng của mạng lưới sinh học [168-170].

3.4.4. So sánh kết quả dự đoán với các nghiên cứu khác

- So sánh được thực hiện giữa hai mô hình đề xuất ở chương 2 và chương 3 của luận án trên cùng một bộ dữ liệu.

Bảng 3.6. Kết quả dự đoán trên 2 mô hình với cái tiến gián tiếp

Mô hình động lực cạnh tranh ngoài	**Dữ liệu**	**Tỷ lệ dự đoán trên top 10**
Chỉ xét các tương tác trực tiếp (Chương 2)	- 01 mạng tín hệu tế bào - 01 mạng tương tác protein - 01 mạng điều hoà gen	82.36 %
Xét bổ sung tương tác gián tiếp (Chương 3)		86,67 %

- So sánh được tiến hành giữa các nghiên cứu độc lập với kết quả nghiên cứu của luận án. Luận án sử dụng kết quả dự đoán gồm danh sách các gen có minh ở Bảng 2.1 và Bảng 3.4. Kết quả cho thấy số gen dự đoán của luận án là lớn nhất, với 55 gen, và đồng thuận với 3/4 phương pháp, cùng với số gen giao thoa lớn nhất là 5 gen. Trong khi các phương pháp còn lại có số lượng gen dự đoán lớn nhất là 30 gen và số giao thoa lớn nhất là 4. Điều này ngụ ý rằng kết quả dự đoán của luận án tốt hơn các phương pháp tham gia so sánh.

Bảng 3.7. So sánh kết quả dự đoán với các nghiên cứu trước đó

Tác giả đại diện của nghiên cứu	**Số phương pháp đồng thuận**	**Số lượng gen không trùng lặp dự đoán được**	**Số gen giao thoa**
Luận án	3/4	55	5
Emig [126]	2/4	17	4

Wang [125]	1/4	25	1
Li [127]	2/4	16	2
Peng [128]	2/4	30	2

KẾT LUẬN VÀ KIẾN NGHỊ

Chẩn đoán và điều trị ung thư đã và đang đối mặt với nhiều thử thách khó khăn và thực tế vẫn chưa đạt được nhiều thành công trong thực tiễn. Một trong những cách tiếp cận trong điều trị bệnh ung thư là dự đoán được gen có khả năng đột biến gây ra bệnh, nhằm hướng tới phát triển phương thuốc điều trị hiệu quả. Nghiên cứu hướng tới đề xuất các mô hình tính toán động lực học cạnh tranh mới trên mạng phức hợp mà ứng dụng trên các mạng sinh học có thể giúp chẩn đoán được gen gây bệnh một cách chính xác. Đây là nghiên cứu có ý nghĩa thời sự, khoa học và thực tiễn.

Luận án đã trình bày những kiến thức cơ bản về mạng phức hợp, tiến hành khảo sát các phương pháp xác định gen gây bệnh, đánh giá hiệu quả của các phương pháp từ đó đề xuất phương pháp xác định gen gây bệnh bằng kỹ thuật mạng phức hợp. Luận án đã tiến hành thực nghiệm trên các bộ dữ liệu để đánh giá hiệu quả.

Hai kết quả chính đạt được của luận án gồm:

- Luận án đề xuất một mô hình động lực học cạnh tranh mới trên các mạng phức hợp, gọi là mô hình động lực học cạnh tranh ngoài. Mô hình mô tả sự cạnh tranh giữa các đỉnh (tác nhân) bên trong mạng (tác nhân điều khiển) với tác nhân môi trường bên ngoài mạng (thuốc). Mô hình có thể xác định được các đỉnh điều khiển, nổi trội trong một mạng phức hợp bất kỳ. Ứng dụng mô hình đề xuất trên các mạng sinh học có khả năng dự đoán gen điều trị ung thư;

- Luận án đề xuất một mô hình động lực cạnh tranh ngoài cải tiến với khả năng xử lý tương tác gián tiếp giữa các đỉnh mô hình mạng phức

hợp, giúp nâng cao khả năng dự đoán gen mục tiêu điều trị ung thư, đặc biệt trên các mạng sinh học có kích thước lớn.

Ngoài ra, mạng phức hợp là một lĩnh vực nghiên cứu đa ngành và có sự hợp lưu giữa các loại mạng khác nhau, như mạng xã hội và mạng sinh học. Vì vậy, kết quả nghiên cứu của luận án có thể được áp dụng cho nhiều loại mạng khác nhau với bài toán cụ thể.

- Hướng nghiên cứu tiếp theo

Mô hình động lực học cạnh tranh ngoài và cạnh tranh ngoài cải tiến mà luận án đề xuất đã cho kết quả thực nghiệm khả quan trong việc dự đoán gen mục tiêu điều trị ung thư trên các mạng sinh học. Tuy nhiên, mô hình đề xuất mới xét trường hợp tại một thời điểm t hay $t+1$ chỉ có một liên kết (tương tác) từ tác nhân bên ngoài tới hệ thống. Trong tương lai, có thể tiếp tục nghiên cứu phát triển mô hình động lực cạnh tranh ngoài với trường hợp tại cùng một thời điểm có nhiều hơn một tương tác đến hệ thống (tác nhân có nhiều tương tác cùng lúc hay có nhiều tác nhân ngoài cùng tương tác đến hệ thống). Đây là trường hợp không hiếm gặp trong các bài toán thực tế, ví dụ như trong điều trị bệnh, các pháp đồ có thể được sử dụng kết hợp tại cùng một thời điểm (hoá trị, thuốc đích), hay một loại thuốc đích có thể có nhiều hoạt chất được tổng hợp hoặc dùng nhiều hơn một loại thuốc cùng lúc trong điều trị bệnh.

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN LIÊN QUAN ĐẾN LUẬN ÁN

1. Tien-Dzung Tran, **Duc-Tinh Pham**, 2021, Identification of anticancer drug target genes using an outside competitive dynamics model on cancer signaling networks, *Scientific Reports*, vol. 11, no. (1), p. 14095. (SCI Q1).

2. **Duc-Tinh Pham**, Tien-Dzung Tran, 2024, Drivergene.net: A Cytoscape app for the identification of driver nodes of large-scale complex networks and case studies in discovery of drug target genes, *Computers in Biology and Medicine*, ISSN: 1879-0534. Revised(SCIE Q1).

3. Nguyen, Trong-The, Thi-Kien Dao, **Duc-Tinh Pham**, and Thi-Hoan Duong. 2024. "Exploring the Molecular Terrain: A Survey of Analytical Methods for Biological Network Analysis" *Symmetry* 16, no. 4: 462, ISSN 2073-8994. (SCIE Q2)

4. **Duc-Tinh Pham**, Do-Thanh-Tung Hoang, Trong-The Nguyen, Thi-Kien Dao, Thi-Xuan-Huong Nguyen, 2024, A Hybridized Network Analysis and Community Detection for Unraveling Disease Spreading Covid-19 Pandemic Mechanisms, *Journal of Network Intelligence*, 2024 ISSN 2414-8105. (Scopus Q3).

5. **Duc-Tinh Pham**, Hoang Do Thanh Tung, Tien-Dzung Tran, 2021, Xác định gen mục tiêu thuốc ung thư bằng một mô hình động lực cạnh tranh mạng, *Kỷ yếu Hội thảo Quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông (@) – Thái Nguyên*, ISBN: 978-604-67-1744-7, trang 622-628.

6. **Duc-Tinh Pham**, Tien-Dzung Tran, 2020, Phân tích hệ gen virus nCoV bằng khoa học mạng lưới, *Kỷ yếu Hội thảo Quốc gia lần thứ XXIII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông (@) – Quảng Ninh*, ISBN: 978-604-67-1744-7, trang 382-387.