ĐẠI HỌC QUỐC GIA HÀ NỘI

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Nguyễn Minh Tân**

# PHÂN TÍCH VÀ KHAI PHÁ DỮ LIỆU HỆ GEN LIÊN QUAN ĐẾN CÁC BỆNH DI TRUYỀN

Chuyên ngành: Hệ thống thông tin

Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ

**Hà Nội – 2023**

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: PGS.TS Nguyễn Hà Nam

TS. Trần Tiến Dũng

Phản biện:.....

Phản biện:.....

Phản biện:.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại .....

vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

# MỞ ĐẦU

## 1. Tính cấp thiết của nội dung nghiên cứu

Trong lĩnh vực sinh học phân tử việc dự đoán và xác định được chính xác các gen gây bệnh là rất quan trọng. Trước đây, việc xác định gen gây bệnh thường được thực hiện bằng những thực nghiệm sinh học. Phương pháp này được tiến hành với nhiều gen ứng viên trên vùng nhiễm sắc thể khả nghi, quá trình này gây tốn kém về mặt thời gian và chi phí. Để giải quyết vấn đề đó người ta tiến hành phân hạng các gen theo mức độ nhạy cảm để xác định gen gây bệnh. Sau khi phân hạng người ta sẽ xác định được một số lượng nhỏ các gen có thứ hạng cao để đưa vào thực nghiệm sinh học.

## 2. Mục tiêu nghiên cứu chính của luận án

Luận án tập trung nghiên cứu các kỹ thuật ứng dụng mạng phức hợp trong việc khai phá dữ liệu liên quan tới bệnh ung thư, để xác định các gen chỉ thị ung thư từ mạng sinh học phân tử bằng các kỹ thuật tính toán xác định các gen nhạy cảm, dễ bị đột biến trong lõi của mạng sinh học.

## 3. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu của luận án là phương pháp xác định gen chỉ thị gây bệnh ung thư, các kỹ thuật ứng dụng mạng phức hợp trong việc phát hiện các gen chỉ thị gây bệnh. Các thuật toán song song để nâng cao hiệu năng tính toán đối với những bộ dữ liệu lớn.

- Phạm vi áp dụng là một số bệnh ung thư được thể hiện trên 16 bộ dữ liệu được tải từ cơ sở dữ liệu KEGG và bộ dữ liệu khám nghiệm ung thư của bệnh viện K.

## 4. Phương pháp nghiên cứu

Luận án sử dụng các phương pháp tổng hợp lý thuyết, phân tích, đánh giá các nghiên cứu liên quan. Từ đó tiến hành mô hình hóa và triển khai thực nghiệm trên các bộ cơ sở dữ liệu. Luận án sử dụng các thuật toán xác định lõi của mạng R-core, K-core kết hợp kỹ thuật xếp hạng các nốt của mạng phức hợp là độ gần gũi thứ bậc (Hierarchical

Closeness) để xếp hạng gen bệnh từ đó xác định các gen nhậy cảm, dễ bị đột biến đó chính là các gen chỉ thị gây bệnh ung thư. Kết quả thực nghiệm được kiểm chứng, đánh giá, so sánh với kết quả của các nghiên cứu trước đó.

## 5. Đóng góp của luận án

Các đóng góp chính của luận án được thể hiện trong 3 công trình nghiên cứu. Trong đó có 01 công trình trên tạp chí Scopus, 01 công trình trên tạp chí SCIE, Q2.

**Đóng góp thứ nhất:** Mô hình hóa mạng sinh học bằng mạng phức hợp, khảo sát, đánh giá các thuật toán xếp hạng gen gây bệnh. Đề xuất phương pháp cải tiết thuật toán K-core theo hướng tiếp cận song song hóa để có thể xác định gen chỉ thị ung thư từ các mạng sinh học phân tử quy mô lớn. Thuật toán K-core xác định các lõi của mạng dựa vào mức độ kết nối của các nút mạng, thuật toán R-core xác định các lõi mạng dựa vào khả năng truy cập của các nút mạng liền kề. Với cải tiến này đã giúp cho việc xác định các gen chỉ thị ung thư trở lên chính xác hơn. Trên cơ sở đó, luận án đã xây dựng mô đun phần mềm C-Biomarker.net dùng để xác định gen chỉ thị ung thư từ mạng sinh học phân tử để tích hợp vào phần mềm Cytoscape. Kết quả của nghiên cứu được thể hiện trong công trình [CT3] và [CT5].

**Đóng góp thứ ba:** Nghiên cứu đề xuất một kỹ thuật mới là phân cụm mạng phức hợp trong khai phá bộ dữ liệu tầm soát ung thư, một loại bệnh di truyền bằng phương pháp mạng lưới. Áp dụng trên bộ dữ liệu khám nghiệm ung thư tại bệnh viện K giúp phát hiện ra các quy luật xã hội về ung thư, hỗ trợ trong công tác phòng và điều trị ung thư. Kết quả chính của nghiên cứu này được trình bày trong công trình [CT4].

## 6. Cấu trúc của luận án

Nội dung chính của luận án được cấu trúc gồm 3 chương.

- Chương 1. Tổng quan.
- Chương 2. Xác định gen chỉ thị ung thư bằng mạng phức hợp.

- Chương 3. Khai phá dữ liệu khám nghiệm ung thư bằng mạng phức hợp.

# CHƯƠNG 1. TỔNG QUAN

## 1.1. Giới thiệu

Xác định gen chỉ thị ung thư là quá trình phân tích và xác định các biến thể di truyền trong gen của một cá nhân để đánh giá nguy cơ ung thư và cung cấp thông tin hỗ trợ cho việc chẩn đoán, điều trị và quản lý bệnh ung thư.

## 1.2. Gen chỉ thị ung thư và vấn đề nghiên cứu phát hiện gen chỉ thị ung thư

### *1.2.1. Gen chỉ thị ung thư*

Gen là một đoạn xác định của phân tử axit nuclêic (ADN hoặc ARN) có chức năng di truyền nhất định. Gen có thể thu nạp các đột biến sinh học nằm trong trình tự của chúng, dẫn đến những biến thể. Những gene bị đột biến có thể gây ra bệnh.

### *1.2.2. Lợi ích của việc xác định gen chỉ thị ung thư*

Việc dự đoán và xác định chính xác các gen gây bệnh là rất quan trọng trong lĩnh vực y sinh và sinh học phân tử. Việc phân hạng gen giúp xác định gen gây bệnh sẽ rút ngắn thời gian và giảm chi phí rất nhiều.

## 1.3. Tổng quan về các kỹ thuật phát hiện gen chỉ thị ung thư

Hiện nay có nhiều nghiên cứu về việc xác định các gen gây bệnh, tuy nhiên có thể phân thành một số nhóm các phương pháp khác nhau để phân hạng gen gây bệnh. Các phương pháp đó bao gồm:

### *1.3.1. Phương pháp thống kê dựa trên độ tương tự*

Phương pháp thống kê dựa trên độ tương tự là việc xác định các biểu hiện gen khác biệt giữa các mẫu ung thư và mẫu bình thường.

Bằng cách so sánh mẫu ung thư và mẫu bình thường, các nghiên cứu đã tìm ra những gen có biểu hiện khác biệt đáng kể giữa hai loại mẫu này.

### *1.3.2. Phương pháp dựa trên kỹ thuật học máy*

Quá trình xác định gen chỉ thị ung thư thông qua học máy thường bắt đầu bằng việc thu thập dữ liệu gen từ bệnh nhân ung thư, bao gồm các mẫu tế bào hoặc mẫu máu. Sau khi dữ liệu gen được thu thập và tiền xử lý, các phương pháp học máy được áp dụng để phân tích và xác định gen chỉ thị ung thư. Các thuật toán học máy như học không giám sát, học có giám sát và học tăng cường được sử dụng để tạo ra mô hình dự đoán.

### *1.3.3. Phương pháp dựa trên mạng phức hợp*

Phương pháp dựa trên các mạng là sử dụng các mạng sinh học như mạng tương tác protein để phân tích và phân hạng các nút được sử dụng khá phổ biến và mang lại hiệu quả cao do cơ sở dữ liệu về sự tương tác protein ngày càng được bổ sung đầy đủ và tiến tới bao phủ được toàn bộ hệ gen. Phương pháp này được tiến hành dựa trên việc quan sát thấy rằng các gen liên quan đến cùng một bệnh hoặc những bệnh tương tự thường có xu hướng nằm gần nhau trong mạng tương tác protein (hay còn gọi là mô đun bệnh).

# CHƯƠNG 2.
# XÁC ĐỊNH GEN CHỈ THỊ UNG THƯ BẰNG MẠNG PHỨC HỢP

## 2.1. Giới thiệu về mạng phức hợp

Khoa học phức hợp là bộ môn khoa học nghiên cứu về các hệ thống phức hợp. Nói đơn giản, một hệ thống là phức hợp nếu nó chứa nhiều thành phần con tương tác với nhau và biểu hiện những tính chất,

những hành vi mà chúng ta không thể hiển nhiên suy ra từ sự tương tác giữa những thành phần cấu tạo nên nó.

**2.2. Các thành phần cơ bản trên mạng phức hợp**

Các thành cơ bản của mạng gồm: Nút mạng, liên kết, thông tin.

**2.3. Đặc trưng chung trên mạng phức hợp**

Có 3 đặc trưng sau; Đặc trưng không quy mô (Scale-free), đặc trưng thế giới nhỏ (Small-world), đặc trưng tập lõi.

**2.4. Một số tính chất cơ bản của mạng phức hợp**

***2.4.1. Kích thước mạng***

Kích thước của một mạng liên quan đến số lượng *V* nút, hoặc số lượng các cạnh E.

***2.4.2. Mật độ mạng***

Mật độ của mạng là một thuộc tính quan trọng ảnh hưởng đến các tính chất cấu trúc. Mật độ có thể được xác định trên đồ thị *G(V, E)* bằng công thức

$$\frac{|E|}{|V| \times (|V| - 1)} \tag{1.5}$$

***2.4.3. Trung bình bậc***

Bậc *k* của một nút là số cạnh được kết nối với nó, đến hoặc đi. Liên quan chặt chẽ đến mật độ của mạng là trung bình bậc

$$\langle k \rangle = \frac{2E}{V} \tag{1.6}$$

hoặc, trong trường hợp đồ thị có hướng,

$$\langle k \rangle = \frac{E}{V} \tag{1.7}$$

***2.4.4. Kết nối mạng***

Có 4 kiểu kết nối mạng: Mạng hoàn thiện (Complete Graph), thành phần "khổng lồ" (Giant Component), thành phần kết nối yếu, thành phần kết nối mạnh.

### *2.4.5. Độ bền vững của mạng*

Độ bền vững của mạng, là khả năng của mạng duy trì những chức năng khi đối mặt với những xáo trộn hoặc chịu tác động. Công thức tính:

$$\gamma(G) = \frac{1}{n|S|}\sum_{v\in V}\sum_{s\in S} I((s) = (s\vec{v})) \quad (1.8)$$

Trong đó: S là toàn bộ trạng thái mạng, I() là hàm chỉ thị và =1 nếu I true hoặc = 0 nếu I false.

### *2.4.6. Hệ số phân cụm mạng*

Hệ số phân cụm của nút thứ *i* được tính bằng:

$$k_i(k_i - 1) \quad (1.9)$$

Ở đây $k_i$ là số nút lân cận của nút thứ *i* và $e_i$ là số lượng kết nối giữa các nút lân cận này.

### *2.4.7. Tính mô đun*

Tính mô đun (Modularity) là mức độ mà các thành phần của hệ thống có thể được tách ra và kết hợp lại.

### *2.4.8. Cấu trúc lõi ngoại biên.*

Chúng bao gồm một lõi dính kết dày đặc và một vùng ngoại vi thưa thớt, lỏng lẻo.

## 2.5. Trung tâm mạng và phép đo

Các phép đo trung tâm gán mọi đỉnh một giá trị số thực, đỉnh $v_1$ được cho là trung tâm hơn hoặc quan trọng hơn đỉnh $v_2$ nếu $C(v_1) > C(v_2)$.

### *2.5.1. Trung tâm mạng*

Trung tâm mạng của một nút được định nghĩa là số lượng tương tác trực tiếp đến hoặc đi của nút đó với các nút khác trong mạng và được định nghĩa là:

$$C_{deg}(v) = |\{(v,w)|(v,w) \in E\}| + |\{(w,v)|(w,v) \in E\}| \quad \mathbf{(1.11)}$$

### *2.5.2. Trung tâm gần gũi*

Trung tâm gần gũi xác định mức độ gần gũi của một nút với các nút khác trong mạng bằng cách đo tổng khoảng cách ngắn nhất giữa nút đó và tất cả các nút khác trong mạng và được định nghĩa như sau:

$$C_{clo}(v) = \frac{1}{\sum_{w \in V\{v\}} d(v,w)} \quad \textbf{(1.12)}$$

Trong đó *d(v, w)* là khoảng cách của đường đi ngắn nhất từ nút *v* đến nút *w*.

### *2.5.3. Trung tâm giữa*

Phép đo trung tâm giữa giới thiệu khả năng đo lường của một đỉnh trong việc giám sát giao tiếp giữa các đỉnh khác.

$$C_{bet}(v) = \sum_{s,t \in V\setminus\{v\}, s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad \textbf{(1.14)}$$

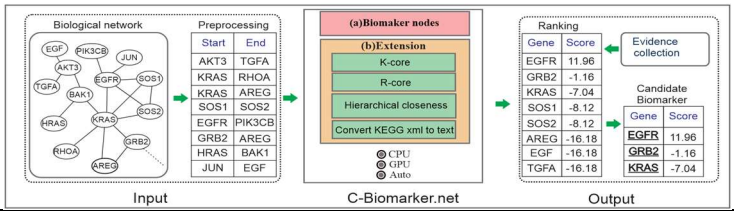
$\sigma_{st}$ biểu thị tổng số đường đi ngắn nhất giữa hai đỉnh *s* và *t* và $\sigma_{st}(v)$ biểu thị số đường đi ngắn nhất đi qua *v* mà từ *s* và *t*.

## 2.6. Mô hình mạng phức hợp của hệ gene

Dữ liệu sinh học phân tử có thể được mô hình hóa thành các mạng phức hợp sinh học phân tử của hệ gen như mạng tương tác protein (protein-protein interaction network), mạng điều hòa gen (gene regulatory network), mạng tín hiệu tế bào (cellular signaling network), mạng trao đổi chất (metabolic network).

## 2.6. Phương pháp đề xuất

### *2.6.1 Mô hình hệ thống*




Hình 0.1 Sơ đồ thiết kế phần mềm C-Biomarker.net

C-Biomarker.net được thiết kế để hoạt động với tất cả các loại mạng: có hướng, vô hướng và không đồng nhất. Chức năng chính của phần mềm là xác định các nút dấu ấn sinh học. Ngoài ra, các chức năng mở rộng có thể phát hiện lõi K/R của mạng, xếp hạng các nút theo HC để xác định các nút nhạy cảm với đột biến.

***2.6.2 Dữ liệu đầu vào***

Dữ liệu đầu vào là một tệp dạng txt gồm 3 cột: start, end và direction. Trong đó start là nút nguồn, end là nút đích và direction cho biết kiểu kết nối của cạnh, nếu direction bằng 0 là kết nối vô hướng, bằng 1 là kết nối có hướng.

***2.6.3. C-Biomarker.net***

C-Biomarker.net bao gồm các chế độ tính toán tuần tự trên CPU, song song trên CPU, hoặc song song trên GPU tùy theo quy mô mạng đầu vào. Nếu quy mô mạng đầu vào là nhỏ ta chọn chế độ tính toán tuần tự trên CPU, nếu quy mô mạng đầu vào là vừa ta chọn chế độ tính toán song song trên CPU còn quy mô mạng đầu vào là lớn ta sẽ chọn chế độ tính toán song song trên GPU.

**2.8. Xếp hạng gen chỉ thị ung thư**

***2.8.1. Thuật toán K-core***

Thuật toán phân rã K-core thường được sử dụng để xác định một tập con đặc biệt của mạng, trong đó k đại diện cho mức độ của lõi. *Phân tách K-lõi* dựa trên cấp độ nút thường được sử dụng để xác định các tập hợp con cụ thể của mạng, được gọi là $K$ - *lõi* ( $k \geq 1$), trong đó $k$ biểu thị cấp độ lõi. *Lõi k* của mạng $G$ bao gồm một tập hợp con các nút trong mạng $G$, có được bằng quy tắc cắt xén sau: cho một mạng, tất cả các nút có bậc $< k$ được loại bỏ, cùng với các tương tác của chúng, khỏi mạng. Quá trình loại bỏ này được lặp lại cho đến khi bậc của mọi nút trong mạng còn lại $\geq k$. Lõi $k$ biểu thị tập hợp các nút còn lại và do đó, $k_1$ lõi *là tập* con của $k_2$ *lõi* nếu $k_1 \geq k_2$.

### *2.8.2. Thuật toán R-core*

Thuật toán *R-core*, sử dụng quy tắc cắt tỉa tương tự như phương pháp phân tách *K-core* ngoại trừ R(v) được sử dụng thay vì số bậc của node. Nói cách khác, tất cả các node có R(v) < R và các tương tác của chúng sẽ bị loại bỏ ở mỗi bước cắt tỉa. Kết quả là, *R-core* phân tách mạng có hướng thành các mạng con được gọi là *R-core*, trong đó các node có R cao là các node có thể truy cập đến nhiều node khác trong mạng nhất. Theo định nghĩa phân tách, K và R core tương ứng đại diện cho số bậc và số lượng node truy cập được của node. Hơn nữa, *K-shell* (*R-shell*) là phần bù của *K-core* (*R-core*).

Thuật toán *R-core* được mô tả như sau:

**Algorithm** ParR-core algorithm

```
1   Procedure ParR-core(reachability)
2   i ← 0                          //Global variable
3   l ← 0; s ← 0; e ← 0   // Thread-local variables
4   Initialize a thread-local array buff of size n/n_t
5   while (i<n) do
6       for (v = 0 to n - 1) in parallel do
7           if(reachability[v] = l) then
8               buff[e] ← v; e ← e + 1
9       while (s < e) do   // process local buff
10          v ← buff[s]; s ← s + 1
11          for (u ∈Adj(v)) do     // Adj(v): nodes are the
    neighbours of v
12              if (reachability[u] > l) then
13                  a ← atomicSub(reachability[u],1)
14                  if (a = l ) then
15                      buff[e] ← u; e ← e + 1
16                  if (a < l) then
17                      atomicAdd(reachability[u], 1)
18      Wait for all threads done
19      atomicAdd(i, e)
```

### *2.8.3. Thuật toán HC*

HC của một nút v được xác định bằng cách kết hợp các phép đo khả năng truy cập và độ lân cận như sau:

$$C_{hc}(v) = N_R(v) + C(v)$$

Trong đó NR(v) ∈ [0,|V|-1] là khả năng tiếp cận của một nút v được xác định bởi NR(v)=|{w∈V | ∃ một đường đi từ v tới w}|; $C(v)$ là closeness centrality được chuẩn hóa về [0, 1].

Thuật toán xác định thứ bậc của một nút *v* là $N_R(v)$ và được tính như sau:

Thuật toán 2. 1. Thuật toán xác định trung tâm sự gần gũi thứ bậc

```
1   Procedure DFSN_R(v):
2   Set count to zero;
3   Label v as discovered;
4    for all vertex w in adjacent List(v) do
5   if vertex w is not labeled as discovered then
6      if size of adjacent List(w) > 0 then
7          count ← count +  DFSN_R(w);
8              end if
9                  count ← count + 1;
10   end if
11    end for
12        return count;
13  End
```

### *2.8.4. Kết quả đầu ra*

Kết quả đầu ra là một danh sách đỉnh gồm 3 cột: tên nút, R-/K-core và HC. Danh sách này cho biết thứ hạng của các nút theo K-core và theo HC. Từ danh sách này có thể chọn top 3 hoặc top 10 các nút

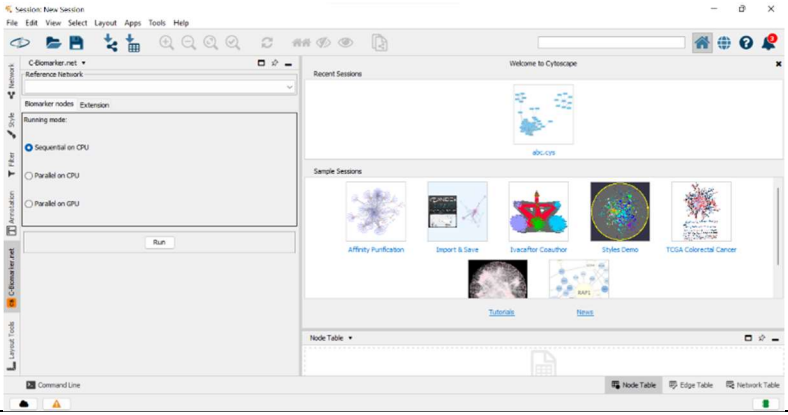
có điểm cao nhất theo core và HC. Đó là các ứng viên gen chỉ thị ung thư.

### 2.9. Thực nghiệm và kết quả

#### *2.9.1. Môi trường và cài đặt hệ thống*

C-Biomarker.net được tích hợp vào Cytoscape như một ứng dụng bổ trợ thông qua ngôn ngữ lập trình Java. Phần mềm này có thể hoạt động rất hiệu quả với các mạng quy mô lớn bằng chế độ chạy song song trên GPU.

Giao diện phần mềm được thể hiện ở Hình 2.2 dưới đây.




Hình 0.2 Giao diện chính của C-Biomarker.net

#### *2.9.2. Triển khai thực nghiệm*

Để tiến hành phân tích mạng, luận án đã sử dụng 16 mạng ung thư từ cơ sở dữ liệu KEGG (www.genome.jp/kegg)

#### *2.9.3. Kết quả thực nghiệm*

Kết quả kiểm tra 3 gen đứng đầu được xếp theo nhân R cao nhất, sau đó là xếp cao nhất theo giá trị HC đã cho ra danh sách các gen và những gen này hoàn toàn trùng khớp với các kết quả nghiên cứu trước đây.

### 2.10. Đánh giá hiệu suất của hệ thống

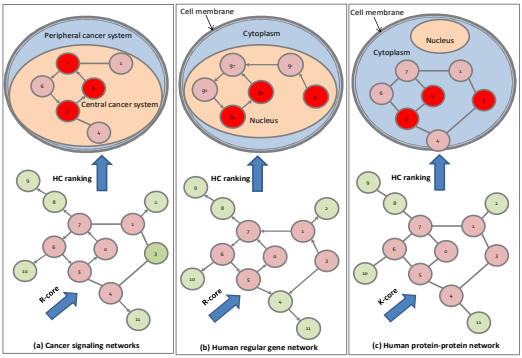
Kết quả cho thấy tính song song trên GPU cho kết quả khác biệt đáng kể trên mạng lớn (với quy mô trên 1.000 nút), còn đối với mạng nhỏ thì tuần tự sử dụng vẫn nên được sử dụng. Lý do là vì việc tải dữ liệu lên GPU sẽ lâu hơn thời gian tính toán, hoặc quá trình tạo luồng để tính toán song song trên CPU có lẽ mất nhiều thời gian hơn thời gian tính toán.

**Bảng 0.1 Đánh giá hiệu suất của phần mềm**

| Network | Số nút | Số cạnh | Running time (ms) | | | | |
|---|---|---|---|---|---|---|---|
| | | | Tuần tự (A) | Song song CPU (B) | Tăng tốc (A/B) | Song song GPU (C) | Tăng tốc (A/C) |
| Bladder cancer | 29 | 58 | 23 | 296 | 0.078 | 73 | 0.315 |
| Acute myeloid leukemia | 66 | 183 | 61 | 500 | 0.122 | 59 | 1.034 |
| Breast cancer | 144 | 773 | 63 | 152 | 0.414 | 118 | 0.534 |
| Human signaling network | 1.561 | 5.089 | 42.4 | 15.8 | 2.684 | 15.6 | 2.718 |
| Human PPI | 7.533 | 22.052 | 15.959.355 | 2.453.332 | 6.505 | 2.376.315 | 6.716 |

Đánh giá của phần mềm (Hình 2.5) trên các mạng phân tử sinh học khác nhau đã xác nhận rằng 3 nút xếp hạng HC cao nhất ở lõi trong

cùng (nút màu hồng) là gen đánh dấu sinh học (nút màu đỏ). (a) Các thí nghiệm trên mạng truyền tín hiệu ung thư đã xác nhận phần mềm có thể xác định chính xác các lõi mạng (được phát hiện bằng cách phân hủy lõi R) và các nút dấu ấn sinh học trên các mạng không đồng nhất. (b) Thử nghiệm trên mạng điều hòa gen của người khẳng định phần mềm có thể xác định chính xác các lõi mạng (được phát hiện bằng cách phân hủy lõi R) và các nút dấu ấn sinh học trên các mạng có hướng. (c) Thử nghiệm trên mạng tương tác protein ở người khẳng định phần mềm có thể xác định chính xác các nút dấu ấn sinh học trên các mạng vô hướng; tuy nhiên, các lõi mạng (được phát hiện bởi sự phân hủy lõi K) không trùng với lõi của tế bào con người như sự phân hủy lõi R.

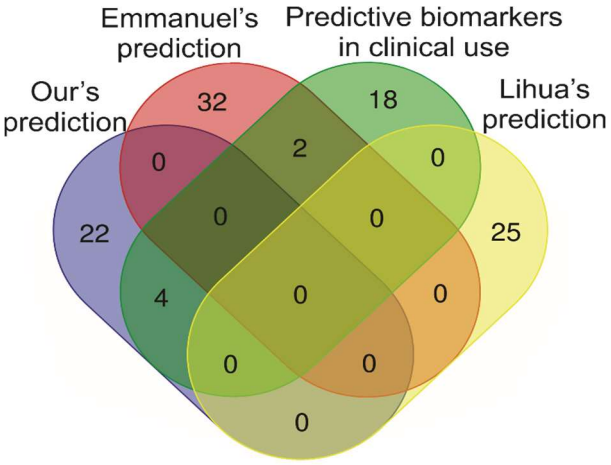



**Hình 0.3 Đánh giá phần mềm trên các mạng sinh học phân tử**

### 2.11. So sánh với các nghiên cứu khác

Luận án đã so sánh phương pháp của mình với ba phương pháp trước đó. Số biomarkergen tìm được của 4 phương pháp tham gia so sánh lần lượt là 34, 25, 24, 26, tỉ lệ xấp xỉ nhau (P-value> 0,05). Các phương pháp này là phương pháp tiếp cận dựa trên mạng để dự đoán các gen đánh dấu sinh học của các bệnh ung thư phổ biến. Để so sánh với các phương pháp trên, luận án đã sử dụng 3 dự đoán hàng đầu của luận án trong Bảng 2.2 bao gồm danh sách 26 yếu tố duy nhất cho các bệnh ung thư phổ biến. Biểu đồ Venn trong Hình 2.13 cho thấy dự

đoán của luận án và dấu ấn sinh học dự đoán trong sử dụng lâm sàng có số lượng yếu tố giao nhau lớn nhất, tức là 4 gen. Các gen giao nhau bao gồm KIT, KRAS, EGFR, MET.




**Hình 0.4 Biểu đồ so sánh các nghiên cứu**

## CHƯƠNG 3. KHAI PHÁ DỮ LIỆU KHÁM NGHIỆM UNG THƯ BẰNG MẠNG PHỨC HỢP

### 3.1. Đặt vấn đề

Phân cụm mạng là quá trình phân chia mạng thành các mô-đun mạng, mỗi mô-đun là tập hợp các đỉnh (bản ghi) kết nối chặt chẽ trong mỗi mô-đun và lỏng lẻo giữa các mô-đun. Thuật toán phân cụm mạng đã được áp dụng thành công trong một số bài toán. Ở đây luận án sử dụng thuật toán tối ưu hóa mô-đun mạng để phát hiện các mô-đun có ít tham số hơn so với phương pháp K-means truyền thống.

### 3.2. Các nghiên cứu liên quan

Các nghiên cứu gần đây đã chỉ ra rằng một số gen đánh dấu sinh học nhất định là nguyên nhân về nguồn gốc của nhiều bệnh ung thư với sự hiển diện của đột biến gen.

### 3.3. Phương pháp đề xuất

Trong phần này luận án giới thiệu một thuật toán phân cụm mạng nổi tiếng được sử dụng hiệu quả với thời gian thực hiện phân cụm ngắn đó là thuật toán tối ưu hoá phân cụm. Cho mạng G(V,E), giá trị Modularity Q được tính theo công thức sau:

$$Q = \frac{1}{2m}\sum_{vw}[A_{vw} - \frac{k_v k_w}{2m}]\frac{S_v S_w + 1}{2} \qquad (3.1)$$

Trong đó: n là tổng số nút của mạng, m là tổng số liên kết của mạng, kv và kw là bậc của nút v và w, Avw = 0 nếu không có cạnh giữa nút v và w, ngược lại Avw = 1 nếu có 1 cạnh nối giữa v và w, Sv và Sw chia mô hình thành 2 cụm: Sv = 1 nếu v thuộc cụm 1, Sv = -1 nếu v thuộc cụm 2. Trong trường hợp có nhiều hơn hai cụm, quá trình được lặp lại trên mỗi cụm. Việc tìm kiếm mô-đun mạng tối đa là một bài toán khó. Do đó, luận án đã nâng cấp thuật toán tối ưu hóa để tìm kiếm hàm mô-đun tối đa của các phân vùng mạng ngẫu nhiên

### 3.4. Dữ liệu khám sàng lọc ung thư

#### *3.4.1. Mô tả bộ dữ liệu*

Bộ dữ liệu được thu thập từ phần mềm quản lý hồ sơ các bệnh nhân, trải qua thăm khám và điều trị tại Bệnh viện ung bướu quốc gia Việt Nam (Bệnh viện K) từ 2/2009 - 6/2014. Sau khi tổng hợp dữ liệu, chúng ta có được một quan hệ ban đầu R1 với hơn 177.000 hồ sơ bệnh nhân (bản ghi) với 15 thuộc tính.

#### *3.4.2. Tiền xử lý dữ liệu*

Sau bước tiền xử lý ta thu được bộ dữ liệu *R* gồm 43.629 bản ghi được mô tả bằng 07 thuộc tính: *ID, Dòng họ, Tuổi, Giới tính, Địa chỉ, Kết luận, Topological (Top).*

### 3.5. Thực nghiệm

#### *3.5.1. Thực nghiệm 1: Xây dựng ma trận khoảng cách tỉnh thành*

Trước hết cần sắp xếp tên các địa danh theo thứ tự alphabet, và căn cứ vào ký tự đầu của địa danh để gán ID lần lượt từ *1..n, n* là số lượng

địa danh. Sau đó, dựa trên bản đồ địa lý, lập bảng danh sách liền kề các tỉnh với nhau theo quy luật các tỉnh liền kề với nhau được nối với nhau, và tỉnh này nối với tỉnh kia một lần duy nhất, không nối theo chiều ngược lại.

Bước tiếp theo, chúng tôi cải tiến thuật toán Breadth-First-Search (BFS) để duyệt một đồ thị gồm *n* đỉnh/tỉnh theo chiều rộng, từ đó tính toán và đưa ra giá trị khoảng cách giữa các cặp tỉnh. Một ma trận đối xứng $BFS_{nxn}$ được tạo ra, trong đó *BFS[i, j]* biểu thị khoảng cách giữa tỉnh "*i*" và tỉnh "*j*". Cụ thể, ma trận khoảng cách $BFS_{63x63}$ biểu diễn khoảng cách giữa các tỉnh/thành Việt Nam

### *3.5.2. Thực nghiệm 2: Mã hóa cặp thuộc tính bản ghi kiểu dữ liệu văn bản*

Trường *Kết luận* chứa các đoạn văn bản mô tả kết luận của bác sĩ về tình trạng bệnh của bệnh nhân. Chúng tôi sử dụng thuật toán *BoW(Bag-of-Words)* sau để số hóa một cặp xâu ký tự $s_1$, $s_2$ (kiểu Text như trường *Kết luận*) dựa trên độ tương tự ngữ nghĩa của chúng.

### *3.5.3. Thực nghiệm 3: Tính toán mức độ tương tự cặp bản ghi*

**Thuật toán 4.5 Tính toán mức độ tương tự của cặp bản ghi**

**Đầu vào:** Cho quan hệ $R(A_1, A_2, \ldots, A_n)$, trong đó dữ liệu của các trường $A_i$ ($i = 1..n$) được quy định thuộc một trong các kiểu dữ liệu sau: *Nhãn*, *Địa chỉ*, *Văn bản*, và *kiểu số*; cho hai bản ghi *a,bÍR*.

**Đầu ra:** Độ tương tự của một cặp bản ghi *a, b*

**Bước 1.** Mã hóa cặp bản ghi *a*, *b* về dạng vector *x*, *y* theo quy luật sau:

$$\vec{x} = [x_1, x_2, x_3, \ldots x_n]$$

$$\vec{y} = [y_1, y_2, y_3 \ldots y_n]$$

- Nếu *a. dongho°b.dongho* thì

$$\vec{x} = \begin{Bmatrix} 1,\ a.tuoi,\ a.giotinh,\ 1,\ a.XY.z, \\ BoW(a.ketluan, b.ketluan) \end{Bmatrix}$$

$$\vec{y} = \begin{Bmatrix} 1,\ b.tuoi,\ b.gioitinh,\ \text{BFS}[a.diachi, b.diachi], \\ b.\text{XY.z}, \text{BoW}(a.ketluan, b.ketluan) \end{Bmatrix}$$

- Ngược lại, Nếu *a.dongho≠b.dongho* thì (4.3)

$$\vec{x} = \begin{Bmatrix} 1,\ a.tuoi,\ a.giotinh,\ 1,\ a.XY.z, \\ BoW(a.ketluan, b.ketluan) \end{Bmatrix}$$

$$\vec{y} = \begin{Bmatrix} 0,\ b.tuoi,\ b.gioitinh,\ \text{BFS}[a.diachi, b.diachi], \\ b.\text{XY.z}, \text{BoW}(a.ketluan, b.ketluan) \end{Bmatrix}$$

**Bước 2.** *S*ử dụng phép đo Euclide để xác định khoảng cách giữa các cặp vector trong không gian *n* chiều, khoảng cách giữa hai điểm/đối tượng bất kỳ trong không gian tương ứng với độ dài của đoạn thẳng nối hai điểm đó và được xác định bằng công thức Euclide:

$$(x, y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + ... + (y_n - x_n)^2}$$

ở đây *n* là số chiều, $y_i$ và $x_i$ lần lượt là giá trị các thành phần của vector *y* và *x*. Nếu *d* lớn, khoảng cách giữa hai điểm càng xa nhau, ngược lại, hai điểm càng gần nếu *d* tiến về *0*, có nghĩa là hai bản ghi có nội dung giống nhau 100% nếu *d=0*.

**Bước 3.** Chuẩn hóa giá trị đo mức độ tương tự các cặp vector về khoảng [0; 1] theo công thức sau:

$$\alpha = \frac{1}{1+d}$$

ở đây, $\alpha$ tiến dần về 0 biểu thị mức độ tương tự giảm trong khi $\alpha$ tiến về gần 1 biểu thị mức độ tương tự tăng, $a$=1 có nghĩa là hai vector đồng dạng 100%.

Thuật toán 3. 1. Tính toán mức độ tương tự của cặp bản ghi

Thuật toán tính toán mức độ tương đồng giữa cặp bản ghi đề xuất, được sử dụng để xây dựng mạng từ một quan hệ.

### *3.5.4. Thực nghiệm 4: Xây dựng mạng từ các cặp bản ghi*

Để xây dựng mạng dữ liệu cho mục tiêu phân cụm, chúng tôi sử dụng giải thuật sau:

**Thuật toán 3.6. Thuật toán xây dựng mạng**

```
1   function [G(Start, End)] NetworkConstruction(R (A1, A2,..., An), α∈[0,1])
    //Input: + R (A1, A2,..., An) là một tập liên kết;
    // + α là ngưỡng được chọn cho mỗi cặp bản ghi theo mức độ tương đồng giữa chúng
    //Output: danh sách kề biểu diện mạng
2   G←new Relation(Start, End)
3   foreach a in R
4       foreach b in R
5           t ← Similarity(a,b)
6           if(t≥α)
7               G← G ∪ (a, b)
8       end for
9   end for
10  return G
11  End
```

Thuật toán 3. 2. Thuật toán xây dựng mạng

### 3.6. Kết quả và thảo luận

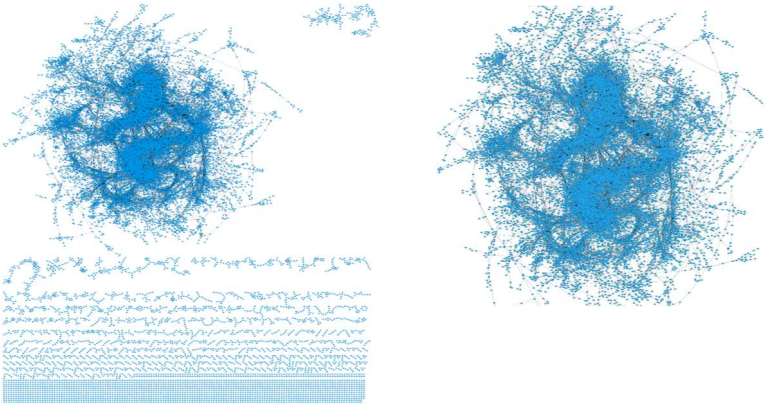
Áp dụng Thuật toán 3.6 trên toàn bộ dữ liệu quan hệ *R* gồm 43.629 bản ghi. Ngưỡng α được chọn để xác định các cặp có giá trị tương tự cao nhất hoặc gần giống nhau nhất. Trong nghiên cứu này, ngưỡng $\alpha$ được chọn để phân cụm theo tiêu chí đảm bảo đặc tính *Scale-free* của mạng và các cặp bản ghi trong mỗi cụm trùng nhau ở mức 03 thuộc tính, tương tương 50% số thuộc tính phân tích.

Bảng 4. 1. Kết quả thẩm tra giá trị tương đồng giữa các cặp bản ghi

| ID | Dòng họ | Tuổi | Giới tính | Địa chỉ | Vị trí | α |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| 607181 | mai | 41 | 1 | 42 | C50.9 | 1 |
| 607207 | mai | 41 | 1 | 42 | C50.9 | |
| 121054649 | nguyen | 39 | 1 | 24 | C53.9 | 0,5 |
| 121058256 | nguyen | 39 | 1 | 62 | C53.9 | |
| 679 | mai | 50 | 1 | 60 | C50.9 | 0,414 |
| 1186516 | nguyen | 50 | 1 | 60 | C50.9 | |
| 679 | mai | 50 | 1 | 60 | C50.9 | 0,333 |
| 656740 | hoang | 50 | 1 | 63 | C50.9 | |
| Trường hợp α = 0,333. Hai bản ghi trùng nhau ở mức dữ liệu 03 thộc tính | | | | | | |

Dựa trên việc khảo sát các giá trị $\alpha$ của các cặp bản ghi trong Bảng 4.2, ta chọn ngưỡng bằng *0,333*. Nói cách khác, chúng ta đã giảm số lượng các cặp có mức độ tương tự xuống [0,333; 1] và thu được một mạng *G với 23,308* nút *và 144,749* cạnh. Mô hình hóa bởi công cụ phân tích mạng Cytoscape, Hình 3.7A.



**A** *B*

Hình 3. 1. Mạng của tập dữ liệu và thành phần liên thông cực đại

Sau khi mạng được tạo, luận án sử dụng một thuật toán trong Cytoscape để tách ra thành phần mạng liên thông cực đại với 18,595 nút và 140,770 cạnh. Xem Hình 4.3B. Thành phần kết nối liên thông cực đại là cụm dữ liệu lớn nhất trong mạng và các đặc điểm của vùng mạng này thường được chọn để đại diện cho các thuộc tính của toàn mạng. Cuối cùng, thuật toán tối ưu hóa mô-đun đã được áp dụng để phát hiện 49 mô-đun (cụm) từ các thành phần mạng liên thông cực đại, trong đó các cụm chứa các đối tượng tương tự và mỗi quy tắc ung thư có thể được trích xuất từ mỗi cụm.

Tổng hợp thông tin từ các cụm ta thu được tri thức từ bộ dữ liệu như sau:

- Nữ có nguy cơ mắc ung thư cao hơn Nam. Một số loại ung thư phổ biến ở nữ như tuyến giáp, vú, cổ tử cung và buồng trứng như.

- Các dòng họ phổ biến hơn, đồng nghĩa có số ca mắc cao hơn. Tuy nhiên, tỷ lệ ung thư theo dòng họ không có sự khác biệt lớn giữ các dòng họ trên tổng số 100 trường hợp ở mỗi dòng học được chỉ định sinh thiết khi thăm khám. Kết quả cho thấy dòng họ Mai có tỷ lệ cao nhất (34,26%), trong khi tỷ lệ này của dòng họ Trương và họ Hà thấp nhất (29,88%).

- Tỷ lệ mắc bệnh ung thư diễn ra ở mọi lứa tuổi, phổ biến nhất là sau 44 tuổi.

- Tỉnh Điện Biên có tỷ lệ mắc bệnh ung thư cao hơn các tỉnh khác, khoảng 33,5–36%. Các tỉnh khác ở mức 31–33% trên tổng số 100 người được chỉ định sinh thiết khi thăm khám. Tỉnh có số ca mắc ung thư cao nhất là Hà Nội.

- Ung thư C37,9 (thận) có tỷ lệ cao nhất (41,57%). Tiếp theo là C48.0 (phúc mạc) (38,23%), và C17,9 (ruột, đại tràng, trực tràng) (37,73%). Thấp nhất là C54,9 (khối u ác tính cổ tử cung) (26,8%). Nhìn chung, các loại ung thư dao động trên 30% trên tổng số 100 người bệnh ở mỗi loại được chỉ định sinh thiết sau kết quả thăm khám ban đầu. Các bệnh ung thư phổ biến ở cả nam và nữ bao gồm ung thư

tuyến giáp, hạch bạch huyết, vòm họng, thực quản, dạ dày, phế quản và phổi. Phụ nữ có nguy có mắc cao hơn nam giới ở một số loại ung thư như tuyến giáp, hạch bạch huyết và vòm họng. Ngược lại, nam có tỷ lệ mắc cao hơn nữ ở một số bệnh ung thư gồm: thực quản, dạ dày, phế quản và phổi.

- Trong tổng số 190 loại ung thư được tìm thấy trong bộ dữ liệu, 16 loại ung thư phổ biến đã được phát hiện, chiếm 75% tổng số bệnh ung thư được ghi nhận, gồm: ung thư vú, tử cung, buồng trứng, gan, dạ dày, thực quản, ruột, thận, tuyến

## KẾT LUẬN

Luận án đã 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 phương xác định gen gây bệnh bằng phương pháp mạng lưới. Luận án đã tiến hành thực nghiệm trên các bộ dữ liệu để đánh giá hiệu quả.

Luận án đã thu được một số kết quả chính như sau:

- Đề xuất thuật toán song song R-core dựa trên việc cải tiến thuật toán K-core để xác định gen chỉ thị ung thư từ các mạng sinh học phân tử quy mô lớn.

- Xây dựng phần phần mềm xác định gen chỉ thị ung thư từ mạng sinh học phân tử để tích hợp vào phần mềm Cytoscape.

- Khai phá bộ dữ liệu xét nghiệm một bệnh liên quan tới di truyền là bệnh ung thư bằng phương pháp mạng lưới.

**Hướng nghiên cứu tiếp theo**

Trong tương lai có thể tiếp tục phát triển phần mềm C-Biomarker.net thêm tính năng visualize các biomarker gen tìm thấy được trên mạng lưới, đồng thời tiến hành phân tích core của các mạng sinh học có hướng. Ngoài ra có thể tiếp tục thử nghiệm trên mạng dữ liệu lớn như mạng tương tác protein và mạng dữ liệu cho các bệnh khác.

## DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ

[CT1]. Nguyễn Minh Tân, Trần Tiến Dũng (2020), "Ứng dụng mạng phức hợp trong khai phá dữ liệu tương tác người dùng", Hội nghị khoa học Quốc gia Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), trang 649-655.

[CT2]. Trần Tiến Dũng, Nguyễn Minh Tân (2022), "A network-based analysis of a workflow at Hanoi University of Industry", Hội nghị khoa học Quốc gia Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), trang 171-180.

[CT3]. Minh Tan Nguyen, Tien-Dzung Tran (2022), "Network approaches for identification of human genetic disease genes", Vietnam J. Sci. Technol., vol. 60, no. 4, pp. 700–712, Aug. 2022.

[CT4]. Minh Tan Nguyen, Duc Tinh Pham, Viet Ha Tran, and Tien-Dzung Tran (2022), "Identification of cancer rules in Vietnam by network modularity", Vietnam J. Sci. Technol., vol. 60, no. 6, pp. 1134–1148, Dec. 2022 (Scopus).

[CT5]. Tien-Dzung Tran, Minh Tan Nguyen (2023), Tien-Dzung Tran, Minh Tan Nguyen (2022), C-Biomarker.net: A Cytoscape app for identification of cancer biomarker genes from cores of large biomolecular networks, BioSystems, Volume 226, April 2023, 104887 (SCIE, Q2).

Danh mục gồm 05 công trình, trong đó 01 công trình thuộc Scopus, 01 công trình thuộc SCIE, Q2.