MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

Pham Duc Tinh

RESEARCH ON DEVELOPING A DYNAMIC COMPETITIVE MODEL IN COMPLEX INFORMATION NETWORKS AND APPLICATION IN PREDICTING CANCER TREATMENT GENES

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM

Code: 9 48 01 04

Hanoi - 2024

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy of Science and Technology.

Supervisors:

1. Supervisor 1: Dr. Tran Tien Dung, Hanoi University of Industry.
2. Supervisor 2: Dr. Hoang Do Thanh Tung, Institute of Information Technology, Graduate University of Science and Technology Library.

Referee 1: …

Referee 2: …

Referee 3: ….

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at…….. (time, date, year…)

This dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library of Vietnam

INTRODUCTION

1. The urgency of the thesis

Currently, identifying disease-causing or so-called pathogenic genes is primarily carried out through clinical biological testing experiments on disease samples [3]. This task is often manually performed in the laboratory for thousands of candidate genes located in a suspicious chromosomal region, ensuring high accuracy but requiring significant time and cost [4]. To reduce the sample volume for clinical experiments, technological approaches have been introduced such as statistics and machine learning, including deep learning. Although these approaches have made significant contributions, they face limitations such as not fully understanding gene interactions and requiring large sample sizes, while sample selection remains a challenge.

From the perspective of network graph theory, biological data can be modeled as complex networks, where vertices are understood as genes or gene products, and edges represent interactions between them [11]. Therefore, exploring biological data can be reduced to the problem of mining data on complex networks. This approach often leads to the proposal of computational models on networks [13], thereby ranking vertices (genes) based on certain attributes, with high-ranking vertices considered important and potentially related to the prediction target [13]. After ranking, a small number of high-ranking vertices (genes/proteins) are included in clinical experiments to search for evidence, affirming the function of genes related to the disease.

2. Research Objectives of the Thesis

The research objectives are to develop dynamic competitive models in complex information networks, identify network control components, and apply them to biological networks to predict target cancer treatment genes.

3. Research Content

Systematizing fundamental knowledge of graph theory, complex network theory, data and modeling of biological network data, dynamic competitive network models, models and algorithms for ranking the functional prediction of vertices on complex networks.

Chapter 1. AN OVERVIEW OF RANKING FOR PREDICTING TARGET CANCER TREATMENT GENES

1.1. The Ranking Problem for Predicting Disease Genes

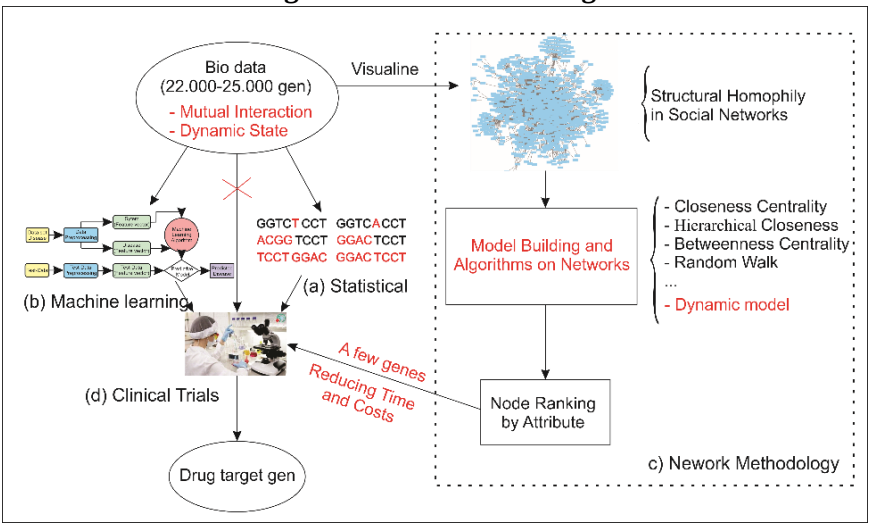

Figure 1.1. Overview of Predicting Target Cancer Treatment Genes on Biological Networks.

(a) Statistical approach, (b) Machine learning approach, (c) Network-based approach, (d) Clinical experiments.

The thesis states the ranking problem for predicting target cancer treatment genes as follows:

- Problem Statement: Given a biological network, predict the target cancer treatment genes for drugs.

- Input: Given a biological network $G=(V,E)$, where V is the set of vertices (genes/proteins) $V=\{v_1,v_2,...,v_n\}$, E is the set of edges (interactions between genes) $E=\{(v_i,v_j)/v_i,v_j \in V,\ i,j=1,...,n\}$.

- Output: A relationship $R^*(V,F)$, where V is the set of vertices, and $F \in R^*$ indicates the likelihood of mutation of v causing cancer and being a treatment target.

1.2. Theoretical Foundations

1.2.1. Graph Theory

1.2.2. Graph Representation on Computers

1.2.2.1. Adjacency Matrix

1.2.2.2. Weighted Matrix

1.2.2.3. Edge List

1.2.3. Complex Networks

1.2.3.1. Basic Components of Complex Networks

1.2.3.2. Characteristics of Complex Networks

1.2.3.3. Fundamental Properties of Complex Networks

1.2.3.4. Network Centers

1.2.3.5. Network Clustering

1.2.4. Data and Modeling of Biological Network Data

1.3. Methods and Related Research in Predicting Disease Treatment Genes Based on Complex Networks

1.3.1. Proximity Property of a Vertex

1.3.2. Degree Proximity Property of a Vertex

1.3.3. Betweenness Centrality Property of a Vertex

1.3.4. Random Walk Algorithm with Restart

1.3.5. ORIENT Algorithm

1.3.6. PRINCE Algorithm using Prior Probability

1.4. Overview of Large-Scale Networks

1.4.1. Concept of Large-Scale Networks

1.4.2. Some Research Directions on Large-Scale Networks

1.5. Dynamic Network Models

Chapter 2. DYNAMIC COMPETITIVE MODELS ON COMPLEX NETWORKS APPLIED IN PREDICTING CANCER TREATMENT GENES

2.1. Dynamic Competitive Models on Complex Networks

Zhao and colleagues [104] introduced a dynamic competitive model on complex networks. The model depicts the competition between two agents (vertices) within the network regarding their ability to control or influence other agents in the network with respect to that agent.

For a weighted network *G(V,E)* with *n* vertices and *m* links, where the vertex set *V={1,2,...,n}* and the network structure is described by an adjacency matrix $A = (a_{kl})_{nxn}$; if vertex *k* interacts directly with *l*, then there

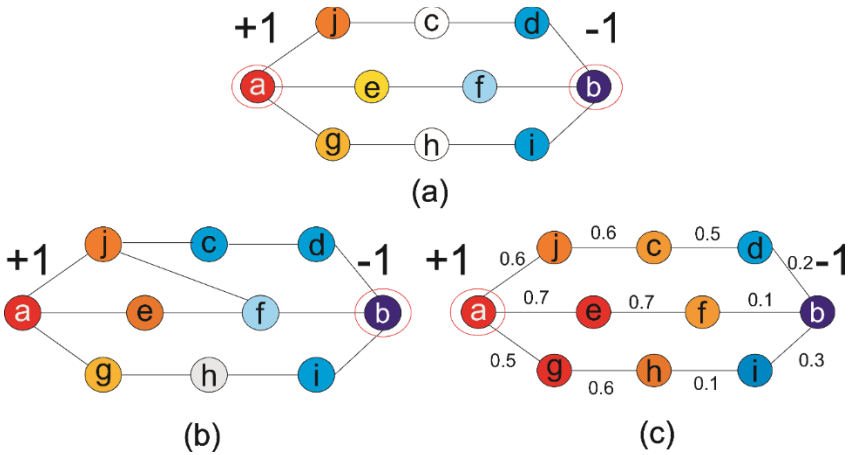

Figure 2.1: An example of dynamic competitive model on complex networks [82].

(a) An undirected network consisting of 10 vertices with equal edge weights, the competition between vertex a and vertex b ends in a tie. (b) A network derived from network (a) with an additional edge between vertex j and vertex f, resulting in vertex b winning the competition. (c) A network with a structure similar to network (a) but with different edge weights, leading to vertex a winning.

is a link from k to l and $a_{kl}>0$; otherwise $a_{kl}=0$. Suppose there is a competition in the network between vertex i and vertex j with fixed and different states, represented by formula (2.1).

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

In that case, each remaining normal agent in the network adjusts its state according to a distributed consensus protocol, reflecting the influence of each normal agent on each competing agent and predicting which competing agent will win. The states of the normal agents are represented by formula (2.2).

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

In that case, the state of each remaining normal agent will eventually reach a stable state, i.e., $t \to \infty$, and is calculated by formula (2.3).

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

$X_{norm} \in R^{n-2}$ represents the converged state vector of the normal agents.

The sign of the stable state indicates the "bias" of that agent. $\bar{x}_k > 0$ $(\bar{x}_k < 0)$ it implies that the last agent k will support the competing opponent i (or j), and $|\bar{x}_k|$ corresponds to the degree of support or influence. $\bar{x}_k = 0$ It implies that agent k is a neutral agent. We have formula (2.4).

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

In the expression above, sign() denotes the sign function. If $\Phi_{ij} > 0$ the competing agent i will win; if $\Phi_{ij} < 0$ the opposing agent j will win; if $\Phi_{ij} = 0$ the competition ends in a tie.

The study does not consider the case where one competing opponent is inside the network while the other is outside the network. Additionally, only considering direct interactions from each vertex to every other vertex in the network may not be efficient for large networks.

2.2. Proposed Model of External Competitive Dynamics on Complex Networks

Given a complex network $G(V,E)$ with n agents (vertices) and m links between them. The set of agents is described as $V=\{1,2,...,n\}$, and the network structure is described by a weighted adjacency matrix $W=w(u,v)_{n\times n}$; if agent u is directly linked to agent v, then $w_{uv}>0$, otherwise $w_{uv}=0$. Suppose the initial state of the vertices in the network is $x_u(t_0)=0,\ u\in V$. We assume that vertex $\alpha\in V$ is a control agent (such as a drug target gene), and vertex $\beta\in V$ is an external competing opponent (environmental agent, drug), where the states of control vertices and competing agents have fixed and different states:

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

Whenever there is a temporary link that can connect from β to any vertex γ in the network to disrupt α, whenever γ adjusts its state. All remaining agents are called normal vertices and denoted as $u\in V/\{\alpha,\beta\}$ with a state at time t as $x_u(t)$ and update its state at time $t+1$ as $x_u(t+1)$ according to formula (2.6):

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

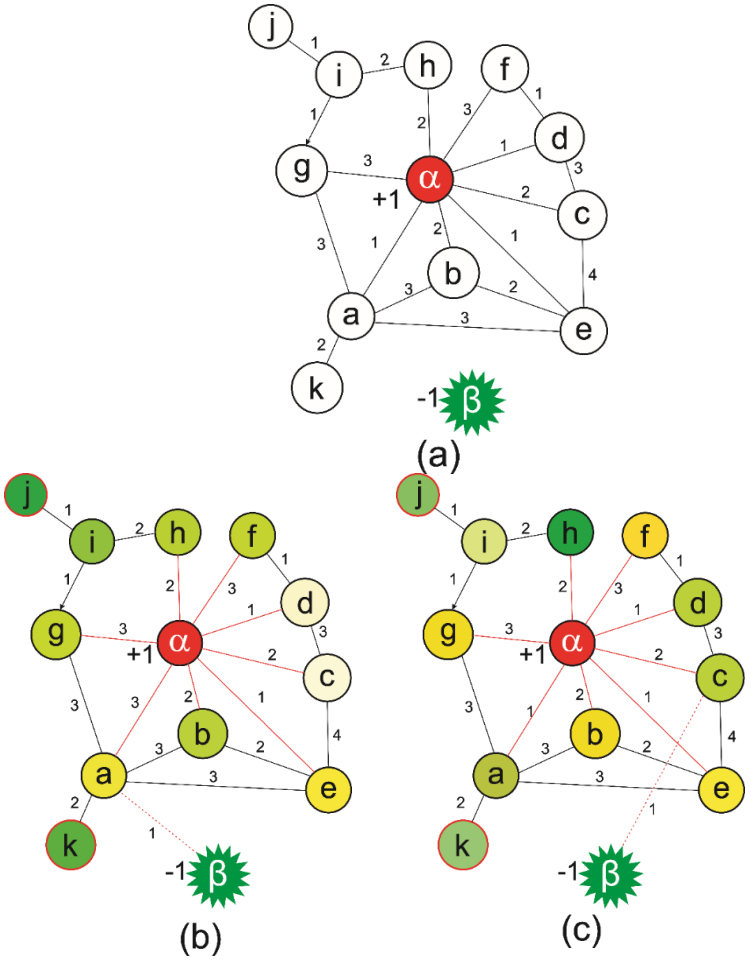

Figure 2.2. An example of external competitive dynamics model.

The network has 12 vertices (genes/proteins) and 19 interactions. Let's assume vertex α (red) is the control vertex with a fixed state of +1, and β (blue) is an environmental agent with an opposing and fixed state of -1. At time t, a temporary undirected interaction is added between the environmental agent (drug) and a vertex (normal vertex in the network), causing the states of normal vertices in the network to change and converge

to a stable value according to a distributed consensus protocol, which is a convex combination of opponent states. The color spectrum indicates their influence on the control vertex inside the network or external agents. (a) Network state at time t_0, $x_u(t_0)=0$, $u \in V/(\alpha\beta)$. (b) Network state at time t. (c) Network state at time $t+1$.

The parameter $0 < \varepsilon < Deg_{\max}^{-1}$ captures the degree of interaction of neighboring vertices, along with Deg_{max} being the maximum out-degree of vertices in the network; and $V(u)=\{v \in V / w(u,v)>0\}$ is the set of neighboring vertices of vertex u that can directly influence vertex u. As $t \to \infty$, the state of each normal vertex u converges to a stable value $\bar{x}_u$ which is a convex combination of opponent states in the competition. The sign (positive/negative) of the stable state of each normal vertex $\bar{x}_u>0$ ($\bar{x}_u<0$) implies that vertex u will eventually be influenced by the control vertex α or β, and $|\bar{x}_u|$ corresponds to the degree of influence $\bar{x}_u$ if vertex u is neutral. See Figure 2.2.

The expression to calculate the total influence state of normal agents for each control agent α against disruption from β is proposed by formula (2.8).

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

The control vertex of the network is determined by $C = \max_{\alpha \in V} ToS(\alpha)$.

2.3. Building the Algorithm of the External Competitive Dynamics Model

2.3.1. Algorithm Idea

2.3.2. Function, Input, Output of the Algorithm

2.3.3. Flowchart and Pseudocode of the Algorithm Pseudocode of the Algorithm Algorithm

Algorithm Algorithm

Algorithm 2.1 of the External Competitive Dynamics Model.

```
1   function OutsideCompetition(Graph G(V,E), Node α∈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;
7      end for
8      Xt[α] ← 1;
9      Xt+1[α] ← 1;
10    Support ← new Dictionary<node, state>;
11    β ← new Node;
12    Xt[β] ← −1;
13    Xt+1[β] ← −1;
14    NormalAgents ∈ V\{α, β};
15   for each γ in NormalAgents do
16    begin
17      e ← new Edge(β,
18     E= E ∪{e};
19     maxIterations ← n x m;
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)*(X_t[v]- X_t[u]);
32        end for
33       X_{t+1}[u] ← X_t[u]+ ε * s;
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] ← X̄[γ];
42    E= E \ {e};
43   end for
44    return Support; // The network state at the time connected to 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; // The total influence of all vertices on α
```

56	**end function**

The algorithm consists of two functions: OutsideCompetition and TOS. (a) The OutsideCompetition function *(G(V,E), $\alpha \in V$)* calculates the influence of each vertex on vertex α at the time when the network is connected to an external agent β in the external competitive dynamics model. (b) The TOS function *(G(V, E), $\alpha \in V$)* calculates the total influence state of vertices in the network on vertex α.

2.4. Assessing the Complexity of the Algorithm Summary:

The computational complexity of the external competitive dynamics algorithm is $O(n^3 * m^2)$.

2.5. Building a Prediction System for Cancer Treatment Genes Using the External Competitive Dynamics Model

2.5.1. Problem of Predicting Cancer Treatment Genes

Input: Given a biological network *G(V,E)*, where *V* is the set of genes/proteins (vertices) $V= \{v_1,v_2,...,v_n\}$, *E* is the set of gene interactions (edges) $E=\{(v_i,v_j) \mid v_i,v_j \in V,i,j=1,..,n\}$.

Output: Ranking table of genes based on the total influence state of genes on each gene in the network. Genes with high rankings are searched for biological evidence as cancer treatment target genes.

2.5.2. Experimental Data

The thesis utilizes data from 17 cancer signaling networks from the KEGG database (www.genome.jp/kegg) for analysis. The preprocessed data can be downloaded from the following link: https://github.com/tinhpd/NetCMD.git

2.5.3. Correlation between Measurements

Experiments on 17 cancer signaling networks and 100 randomly generated directed networks created by the Barabasi network development model with $|V|=50$ and $49 \leq |E| \leq 100$ show that the total influence of each vertex correlates with closeness centrality and degree centrality, where high

rankings of these two measurements have been used to predict disease genes and cancer biomarker genes, often also cancer treatment target genes (Figure 2.5).

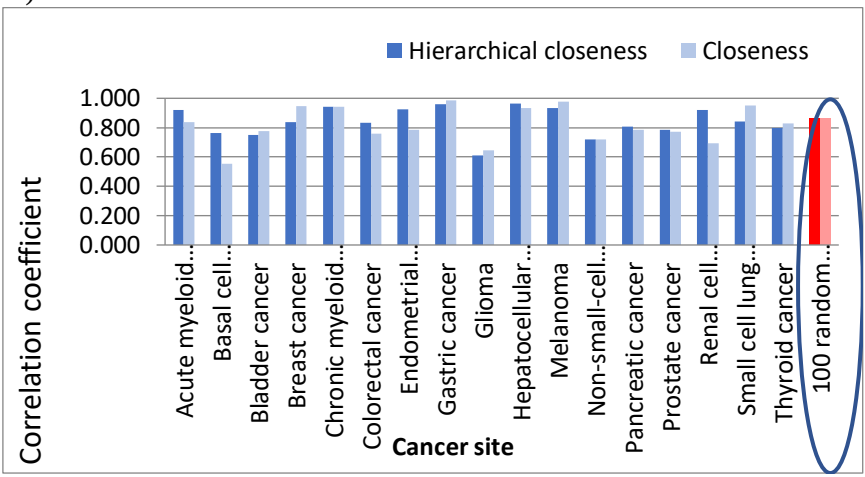

Figure 2.5. Correlation between Closeness Centrality and Total Support

2.5.4. Comprehensive Model of the Cancer Gene Diagnostic System Based on Complex Networks

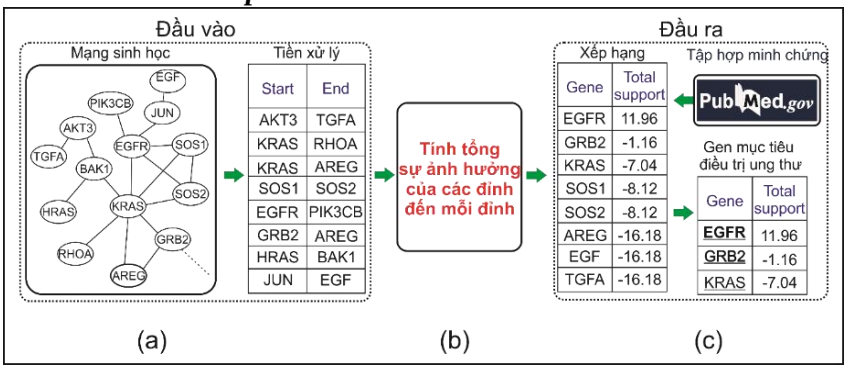

Figure 2.3: Overall model for identifying cancer treatment target genes. The model is designed following a complex network approach.

(a) preprocessing of input data, (b) computational model and algorithm, calculating vertex attributes of the network, (c) organization of

output data and database matching to predict potential genes for further experiments.

2.5.5. Prediction Results of Cancer Treatment Target Genes

In experiments conducted on 17 cancer signaling networks, the results showed that 42 out of 51, equivalent to 82.36%, of the top 3 genes with the highest total influence were cancer treatment target genes. The genes marked in bold have been approved for drug production, while those marked with underline are undergoing clinical trials. The remaining genes are considered potential target genes.

Table 2.1. Performance of target gene identification for cancer treatment by dynamic competition outside mode

Cancer signaling network	**Top 3 gens**		
	Cl	***C2***	***C3***
Acute myeloid leukemia	GRB2	**FLT3**	**PML**
Basal cell carcinoma	SUFU	**SMO**	GLI3
Bladder cancer	RASSF1	**FGFR3**	HRAS
Breast cancer	**LRP6**	LRP5	**WNT1**
Chronic myeloid leukemia	CRK	CRKL	GAB2
Colorectal cancer	**EGFR**	**GRB2**	KRAS
Endometrial cancer	EGF	EGFR	AXIN1
Gastric cancer	LRP6	LRP5	WNT7A
Glioma	CALM1	CALML5	CALM2
Hepatocellular carcinoma	LRP6	WNT3A	WNT7A
Melanoma	FGF2	FGF1	HGF
Nonsmall cell lung cancer	**ALK**	EML4	KRAS
Pancreatic cancer	**KRAS**	AKT2	AKT1
Prostate cancer	**IGF-1**	INS	PDGFB
Renal cell carcinoma	HGF	**MET**	EGLN2
Small cell lung cancer	ITGB1	COL4A1	LAMB3

Thyroid cancer	NTRK1	TPR	TPM3

Table 2.1 consists of target cancer treatment genes identified by ranking the overall impact status. In the table, C1, C2, and C3 represent the NCBI gene symbols of the top three genes with the highest overall impact status. The underlined genes (42 out of 51) were previously reported as drug-target genes for cancer. Among them, 12 underlined genes in bold have been accepted for drug production, and 30 underlined genes without bold are genes in clinical trial stages. The remaining non-underlined genes include 09 genes that are still under insufficient research but may serve as potential drug-target genes for cancer and are provided for reference purposes.

2.5.6. Comparison of Prediction Results

Both studies were conducted on the same dataset consisting of 17 cancer signaling networks from KEGG. The results are presented in Table 2.3.

Table 2.2: Comparison of results between two different models on the same dataset.

3	**The number of network predicted**	**The prediction accurary in the top 3 ranks**	**Execution time toal (minutes)**
Hierarchical Closeness Model [13, 99].	16/17	37/48 gen, equivalent 70,59%,	124
Outside competitive dynamic model network	17/17	42/51 gen, equivalent 82,36%	126

Test system: ASUS X510U, Intel i5-8250U CPU, clock speed 1.6GHz (8CPUs), 8GB DDR IV DDRAM memory, NVIDIA GeForce 940MX 2GB graphics card, Intel M2 120GB SSD.

Chapter 3. INDIRECT INTERACTIONS IN THE IMPROVED OUTSIDE COMPETITION DYNAMIC MODEL AND ITS APPLICATION IN PREDICTING CANCER TREATMENT TARGET GENES

3.1. Proposed Improved Outside Competition Dynamic Model

In the thesis, F is referred to as the influence matrix (interaction/impact between elements in the network), where each element of matrix F describes the influence of one agent (vertex) on another. It should be noted that if there is a direct link from agent u to agent v, then it is understood that agent v directly interacts/influences agent u. In other cases, if there is no direct link from u to v, it means there is an interaction from agent u to agent γ and an interaction from agent γ to v. In this case, agent v indirectly affects agent u through agent γ. Such indirect effects are usually weaker than direct effects.

Let's denote $D = (d_{uv})nxn$ as the distance matrix representing the network.

The thesis defines the matrix $F = (f_{uv})_{nxn}$ as the influence matrix of the network, representing the influence of agent v on agent u, for all $u, v \in V$, and it is calculated by the formula (3.5).

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

Where x_v is the state of vertex v at time t, as $t \to \infty$; d_{uv} is the shortest path distance from u to v.

Let $f(\alpha,v)$ denote the element of the influence matrix F in the α-th row and v-th column. Then, v will exert an influence on α by a certain

amount, and the expression for calculating the total influence of agents v on each control agent α is given by formula (3.6).

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

Where sign() denotes the sign (+ or –) indicating the influence/impact on the control vertex α or the external competing agent β. If $f(\alpha,v) > f(\beta,v)$, then vertex v will exert more influence on the control vertex α; conversely, if $f(\alpha,v) < f(\beta,v)$, it means vertex v will have a greater impact on the external agent β. If $f(\alpha,v) = f(\beta,v)$, then vertex v is neutral. ToSF(α) returns the degree of influence/impact of the normal vertices v in the network on the control vertex α in the improved outside competition dynamic model.

3.2. Developing an algorithm for computing indirect dynamic competitive interaction

3.2.1. Algorithm for calculating distance matrix

In this research, the thesis utilizes the Floyd-Warshall algorithm [100] to compute the distance matrix between vertices in a weighted graph network. The algorithm consists of three nested loops executed n times, resulting in a time complexity of $O(n^3)$.

3.2.2. Algorithm for computing the influence matrix

```
1  function Matrix F[,] InfluenceMatrix(Graph G(V,E), Node α∈V)
   //input: Adjacency weight matrix W=w(u,v)nxn; α
2    D ← DistanceMatrix(G(V,E))
3    X ← OutsideCompetition(G(V,E),α)
4    for each vertex u in V do
5        for each vertex v in V do
6            if D[u,v]==0 then
7                F[u,v] ← NA
8            else
```

```
9           F[u,v] ← X(v) / (D[u,v])^2
10        end if
11      end for
12    end for
13   return F // Influence matrix F
14 end function
```

The time complexity of the InfluenceMatrix function is $O(n^3 + m^2)$, where *n* is the number of vertices and *m* is the number of edges in the graph.

3.2.3. Algorithm for computing the total influence on each network vertex

```
1  function ToSF(Graph G(V,E), Node α, out result)
   // input: Adjacency weight matrix W, α.
2    F ← InfluenceMatrix(G(V,E),α)
3    TotalSupportF ← 0
4    for each v in V - {α, β} do
5      TotalSupportF ← TotalSupportF + (F[α, v] – F[β, v])
6    end for
7    result ← TotalSupportF  // The total influence of vertices on vertex α
8  end procedure
```

The time complexity of the ToSF function is $O(n^3 + m^2)$.

3.3. High-performance computation for the dynamic external competitive model

3.3.1. Developing an algorithm for high-performance computation for the model

```
1  function Matrix DnF[,] ParFindDriverNode(Graph G(V,E))
   // input: weight matrix W=(w_uv)_nxn, {start, end, direction, weight};
2  DnF = new Matrix[n, n]
4    parallel for each α in V do
```

5	*result* ← *0*
6	*ToSF(G(V,E),α, result)*
7	Wait for all works done
8	*DnF[α,]* ← *result*
9	**end parallel**
10	**return** *DnF* // "The matrix of total influence of each vertex in the network on every other vertex in the network."
11	**end function**

The time complexity depends on the time complexity of the ToSF function, which includes computing the influence matrix with a complexity of $O(n^3 + m^2)$.

3.3.2. Designing a high-performance computing software tool

The software Drivergen.net is developed based on the dynamic external competitive model with the capability of high-performance computing on multi-core CPUs. It is designed to function as a Cytoscape plugin, featuring a graphical user interface (GUI). Details about the software along with experimental data can be downloaded from https://github.com/tinhpd/Drivergen.git

3.3.3. Performance evaluation and computation speed of the algorithm

Table 3.3 presents the test results of the Drivergen.net software with different computation modes on 04 biological networks. The results indicate a speedup improvement ranging from 51 to 145 times depending on the specific network type.

Table 3.3. Computational Capability on Large-Scale Networks

Name Network	**Attribute**			**Time (minutes)**		**Speedup**
	Type	***Node***	***Edge***	***Sequentially***	***Parallel***	

virus cytomegalo network	Indirect network	213	1214	5,7	0,1 1	51,8
E. coli protein interaction network		850	1193	341	5	68,2
Gen Gene gegulatory network	Direct network	943	3917	207	7	29,5
Cell signaling network	Mix network	1549	5074	5092	35	145,5
Test system: Dell OptiPlex 5050, Intel Core i7-7700 octa-core CPU with a clock speed of 3.6GHz, 32GB DDR4 RAM						

3.4. Experiment

3.4.1. Experimental Data

The thesis conducts experiments on three types of large-scale biological networks, downloaded from reputable publications. The preprocessed data of these three networks are stored and can be downloaded from the following link: https://github.com/tinhpd/Drivergen.git

3.4.2. Architecture of the Prediction Model

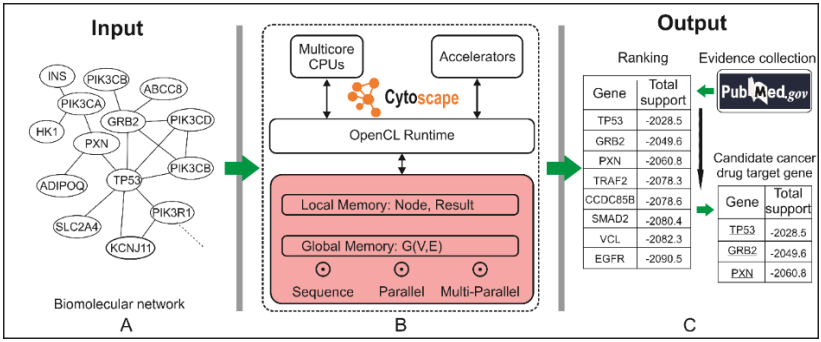

Figure 3.1. Prediction Model for Target Cancer Therapy Genes on Large-Scale Network

(a) Input Biological Network Data, (b) Design Architecture for Computational Model, (c) Output Data Organization and Policy Evidence Search. The experimental data, software, and usage instructions for this study are stored and can be downloaded at https://github.com/tinhpd/Drivergene.

3.4.3. Prediction Results of Cancer Therapy Target Genes

The prediction results on three large-scale biological networks show 86.67%, i.e., 26 out of the top 30 genes with the highest total influence states are target genes of drugs in cancer therapy.

Table 3.4. Identification of Cancer Therapy Target Genes on 3 Large Networks.

Biology network	**Attribute**			**Gen name**	**Evidence from the PubMed.gov database**
	Network type	***Number node***	***Number edge***		
Gene regulatory network	Direct network	943	3917	*NFKB1*	30205516
				RELA	
				JUN	32917236
				FOS	34610301
				MYC	22464321
				STAT1	33608980
				CCND1	29969496
				CREB1	30127997
				STAT3	24743777
				HIF1A	28358664
Cell signaling network	Mix network	1549	5074	*SRC*	11114744
				AR	24425228
				AKT	27232857
				SHC	

				SMAD3	20010874
				RAC1	32460002
				GAB2	22858987
				PI3K	30782187
				PKA	24212646
				SMAD4	29602802
Protein interation network	Indirect netwok	7279	21911	*TP53*	23115424
				GRB2	29550383
				PXN	34135128
				TRAF2	30294322
				DIPA	
				SMAD2	20010874
				VCL	
				EGFR	28368335
				SRC	11114744
				SMAD3	20010874

Additionally, the top 10 genes in Table 3.4 are found to belong to the K-core and R-core [49] cores of the network.

Table 3.5. Identification of K-core and R-core Cores

Network type	**Core type**	
	K-core	***R-core***
Cell signaling network	80%	
Gene regulatory network		70%
Protein interation network	60%	

This result is consistent with previous studies' findings that important cancer hallmark genes tend to reside in the innermost core of the biological network [168-170].

3.4.4. Comparison of Prediction Results with Other Studies

- A comparison is made between the two proposed models in Chapter 2 and Chapter 3 of the thesis using the same dataset.

Table 3.6. Prediction Results on 2 Models with Incremental Indirect Interaction

Dynamic model of external competition	**Data**	**Prediction accuracy in the top 10 ranks**
Only considering direct interactions (Chapter 2)	01 cell signaling network 01 protein interaction network 01 gene regulatory network	82.36 %
Including additional indirect interactions (Chapter 3)		86,67 %

- A comparison is conducted between independent studies and the research results of the thesis. The thesis uses the prediction results, including the list of genes supported in Table 2.1 and Table 3.4. The results show that the number of predicted genes in the thesis is the largest, with 55 genes, and is consistent with 3 out of 4 methods, along with the largest intersection of 5 genes. Meanwhile, the other methods have the largest number of predicted genes at 30 genes and the largest intersection of 4 genes. This implies that the prediction results of the thesis outperform the methods involved in the comparison.

Table 3.7. Comparison of Prediction Results with Previous Studies

Representative author of the research	**The number of consensus methods**	**The number of non-redundant predicted genes**	**The number of overlapping genes**
Thesis	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

CONCLUSION AND FUTURE DEVELOPMENT

Diagnosis and treatment of cancer have been facing numerous challenges, and in reality, have not achieved much success in practice. One approach in cancer treatment is to predict the mutation-prone genes causing the disease, aiming towards developing effective therapeutic drugs. Research focuses on proposing novel competitive dynamic modeling approaches on complex networks that can aid in accurately diagnosing disease-causing genes. This research is of current, scientific, and practical significance.

The thesis presented fundamental knowledge about complex networks, surveyed methods for identifying disease-causing genes, evaluated the effectiveness of these methods, and proposed a method for identifying disease-causing genes using complex network techniques. The thesis conducted experiments on datasets to evaluate effectiveness.

The two main achievements of the thesis are:

Proposing a novel competitive dynamic modeling approach on complex networks, termed as the outside competitive dynamic model. The model describes the competition among vertices (agents) within the network (controller agents) with the environmental agents outside the network (drugs). The model can identify prominent controller vertices in any complex network. Applying the proposed model on biological networks can predict cancer treatment genes.

Proposing an improved outside competitive dynamic model capable of handling indirect interactions among vertices in complex network models, enhancing the ability to predict target cancer treatment genes, especially in large-scale biological networks.

Furthermore, complex networks are a multidisciplinary research field that converges various types of networks, such as social networks and biological networks. Hence, the research results of the thesis can be applied to various types of networks with specific problems.

Future research directions: The outside competitive dynamic model and its improved version proposed in the thesis yielded promising experimental results in predicting target cancer treatment genes on biological networks. However, the proposed models currently consider the case where at time t or t+1, there is only one link (interaction) from outside agents to the system. In the future, further research may continue to develop the outside competitive dynamic model with the case where at the same time there are more than one interactions to the system (multiple agents with simultaneous interactions or multiple external agents with interactions to the system). This is a common scenario in real-world problems, for example, in disease treatment, combination therapies may be used simultaneously (chemotherapy, targeted therapy), or a targeted drug may have multiple active ingredients synthesized or used simultaneously in disease treatment.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

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.