FSM: A Fast Similarity Measurement for Gene Regulatory Networks via Genes' Influence Power

Zhongzhou Liu[1] and **Wenbin Hu**[1,2*]

[1]School of Computer Science, Wuhan University, China
[2]Shenzhen Research Institute, Wuhan University, China
{qed, hwb}@whu.edu.cn

Abstract

The problem of graph similarity measurement is fundamental in both complex networks and bioinformatics researches. Gene regulatory networks (GRNs) describe the interactions between the molecules in organisms, and are widely studied in the fields of medical AI. By measuring the similarity between GRNs, significant information can be obtained to assist the applications like gene functions prediction, drug development and medical diagnosis. Most of the existing similarity measurements have been focusing on the graph isomorphisms and are usually NP-hard problems. Thus, they are not suitable for applications in biology and clinical research due to the complexity and large-scale features of real-world GRNs. In this paper, a fast similarity measurement method called FSM for GRNs is proposed. Unlike the conventional measurements, it pays more attention to the differences between those influential genes. For the convenience and reliability, a new index defined as influence power is adopted to describe the influential genes which have greater position in a GRN. FSM was applied in nine datasets of various scales and is compared with state-of-art methods. The results demonstrated that it ran significantly faster than other methods without sacrificing measurement performance.

1 Introduction

Gene regulatory network (GRN) analysis has attracted increasing attention in bioinformatics and data mining, as some complex diseases like cancer or diabetes are usually caused by dysfunction of relevant networks or network communities rather than mutations of individual molecules [de Souza Jacomini *et al.*, 2016]. GRNs illustrate the interactions between molecular regulators in organisms, such as DNA, RNA, proteins, and other chemicals. Researchers often use data mining techniques on GRNs to uncover the internal relationships as well as essential pathways of diseases, particularly cancer[Ruan *et al.*, 2015; Denitto *et al.*, 2015].

*Corresponding Author

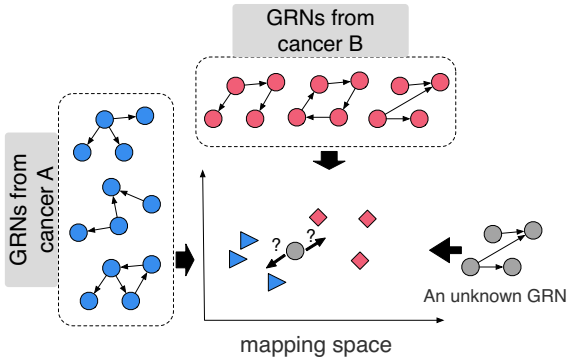

Figure 1: An example of the application for GRN similarity measurement. The class of unknown GRN is determined by its distance to other known GRNs.

A central problem in GRNs is how to measure the similarities between two networks, each of which is under a specific condition. For example, given several GRN obtained from clinical data, some of which are labeled as specific cancers, can we have a "ruler" to measure the distance between any two of them and determine if the rest of the samples are also suffered from the same cancer or not (such as the example illustrated in Figure 1)? Or given a sequence of GRNs obtained from patients to be diagnosed, can we quickly scan them, capture the similarities and differences between them, and detect the anomalies? Moreover, by searching and measuring such similarities, the function of unknown molecules or biological pathways associated with specific diseases can be determined. Similarity problem in GRNs is more complex than it is in other domains such as social networks. In social networks, we focus more on the structural similarities and consider nearly no additional assumptions or properties on nodes and edges. However, when we consider similarity measurement problems in GRNs, we actually consider the functional similarity of two networks. Properties of nodes and edges are highly respected, including the properties which are hand-designed or obtained from biological experiments.

Various computational methods have been proposed to compare graphs by learning a distance function [Liu *et al.*, 2019]. One widely used method is graph edit distance (GED). GED calculates the minimal cost of transforming a graph into another. Generally, optimization of the cost function is NP-

hard [Zeng *et al.*, 2009]. another state-of-art method is graph kernel. Graph kernels compare the substructure of two networks by the following steps: first it divides networks into sets of characteristic substructure patterns, and the cut-set of the pattern sets is determined [Horváth *et al.*, 2004]. Then, the similarity of two networks is calculated by kernel function $k : \varphi(X) \times \varphi(X') \to \mathbb{R}$. φ is a decomposition algorithm for graph $X \to \mathbb{H}$ into a Hilbert space, such that $k(X, X') = I(\varphi(X), \varphi(X'))$, where I is a isomorphism algorithm for substructures. Graph kernels usually have good performance in some application of real-world networks, because they consider local or global properties of networks. However, finding such a similarity using graph kernels is usually a NP-complete problem [Gärtner *et al.*, 2003].

Accordingly, computational complexity is a major challenge in graph similarity measurement. Existing methods are not usually suitable for practical applications in biology and clinical research because real-world GRNs are highly complex systems with thousands of nodes. To measure the similarity between GRNs faster and credibly, a method called fast similarity measurement (FSM) is proposed. This method is inspired by the concept of generalized hamming distance (GHD) which is capable of identifying subtle variations in the topology of paired graphs [Mall *et al.*, 2017]. Besides, the concept of influence power (IP) is introduced to identify influential nodes in a GRN. FSM can fast measure the similarity of paired GRNs by detecting the differences in nodes' IP between two GRNs. The contributions of this study are as follows:

- FSM captures the subtle variations between influential nodes and provide a more credible GRN similarity measurement.
- The concept of nodes' influence power (IP) is introduced. It is inspired by real-world phenomena and can effectively recognize influential nodes among the GRN.

The rest of this paper is organized as follows. Section 2 introduces some preliminary concepts and describes IP and GHD. In Section 3, IP and the FSM method are described in detail. In Section 4, the method is evaluated on a large number of datasets of varying size and sources. Finally, Section 5 concludes the paper.

2 Preliminaries

2.1 Definitions and Problem Statement

Let $G = (V, E)$ be a directed unweighted graph, where V is the node set and E is the edge set. In a GRN, each gene is represented as a node in V. And a directed edge between genes, e.g., $(u, v) \in E$ implies that gene u regulates gene v in the regulatory network. For the node $v \in V$, $inN^k(v)$ is defined as the k-th order incoming neighborhood set of v ($k = 1, 2, \dots$). When $k = 1$, for each node in the set there exists a directed edge pointing at v, i.e., $inN^1(v) = \{\forall u | (u, v) \in E\}$. Furthermore, $inN^2(v) = \{\forall u | \exists w, (u, w) \in E, (w, v) \in E\}$. Similarly, the k-th order outgoing neighborhood set of v is denoted by $outN^k(v)$. That is, for all $u \in outN^k(v)$, the distance from v to u is k in the directed graph. Finally, $|\cdot|$ denotes the number of elements in a set. For example, the number of nodes in a GRN is $|V|$, and the in-degree and the out-degree of node v are $\left|inN^1(v)\right|$ and $\left|outN^1(v)\right|$, respectively.

Based on the definitions above, we formalize the problem as follows.

Problem 1. supposing $G_A = (V, E_A)$ and $G_B = (V, E_B)$ are two different GRNs sharing the same node set V, and the permutation of nodes keeps unchanged. The distance between the two networks is $D(G_A, G_B)$, where D is a distance metric. Let $p = P(y = 1|D(G_A, G_B))$ be the conditional probability that G_A and G_B belong to the same class given the distance $D(G_A, G_B)$, and $(1 - p) = P(y = 0|D(G_A, G_B))$ be the probability of the opposite situation. We aim to find a proper distance metric D, satisfying:

$$\min_D H(p) = -p \log p - (1 - p) \log(1 - p)$$

In our experiment, the dataset contains networks labelled as different classes (e.g. class1, class 2, ...). First, We randomly chose two networks with same scale from two different classes respectively, denoted as G_A and G_B. Then, a series of permutated networks (namely $G_{A1}, G_{A2}, \cdots$ and $G_{B1}, G_{B2}, \dots$)are generated by randomly swap some links in G_A and G_B. $D(G_{Ai}, G_A)$ and $D(G_{Bi}, G_A)$ $(i = 1, 2, \dots)$ are calculated. For the former case, we label $y = 1$, and for the latter case, we label $y = 0$. Then, a 10-fold cross-validation SVM was performed to check whether D can correctly measures the similarity between the networks and satisfy the minimum entropy. The above process is repeated several times to ensure that as many samples as possible in the dataset are covered.

2.2 Influence Power and Generalized Hamming Distance

Since many real-world GRNs are thought to be scale-free [Caldarelli *et al.*, 2002], a large proportion of genes are regulated by only a few hub genes. If each gene is thought of as a person in a community, and the regulatory relationship between genes as exchange of economic goods, the differences in income and expenses will result in social class inequality. A gene's "social class" is associated with the level at which it regulates other genes and the level at which it is regulated by other genes. Here, a metric called *influence power* (IP) is proposed to quantify the degree of regulation based on the topological structure of GRNs. IP measures a node's influence in terms of its capacity to influence its outgoing neighbors (IP_O), and the total influence by its incoming neighbors (IP_I). The significance of IP stems from the fact that in the evolution of GRNs, particularly in cancer, the mutation of gene functions is usually caused by the localized rewiring of influential genes [Ruan *et al.*, 2015]. To compare two GRNs, the differences in their functions are compared, and a larger functional gap implies larger distance between them. Therefore, paying more attention to high-IP genes will lead to a more effective and credible result. The detailed description of IP is discussed in Section 3.

GHD is a metric that measures the distance of two networks proposed by [Ruan *et al.*, 2015]. Given the adjacency

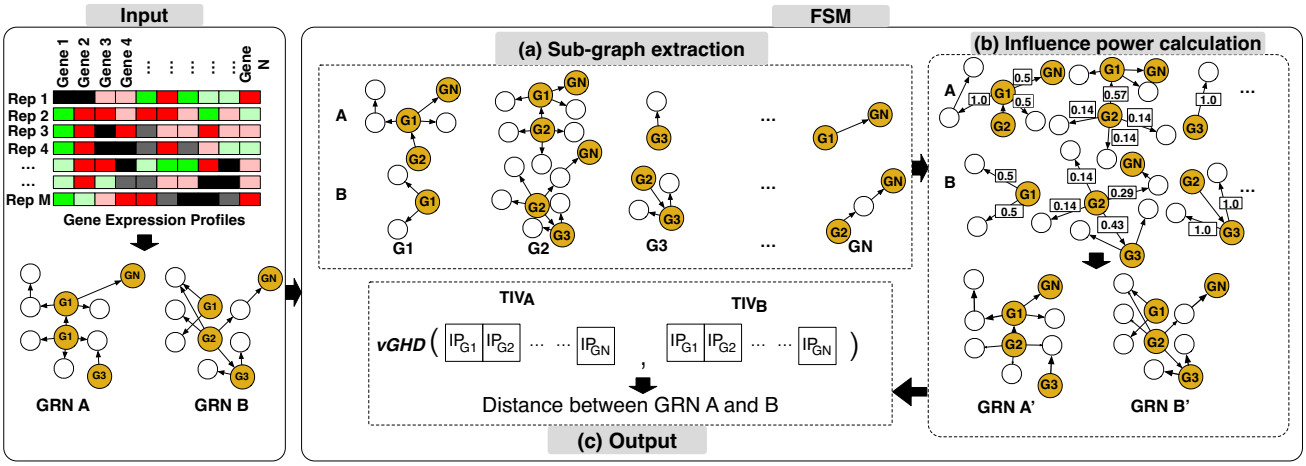

Figure 2: FSM framework. For a pair of GRNs inferred from gene expression profiles, (a) sub-graphs are extracted for each gene; (b) the IP of a gene on its neighbors in every sub-graph is calculated, and then these sub-graphs are integrated into the original GRN (the width of the edge indicates the strength of IP); (c) by analyzing the changes in IP, the similarity between two GRNs is calculated.

matrices of two networks A and B, the GHD of them is defined as:

$$GHD(A,B) = \frac{1}{N(N-1)} \sum_{i,j,i\neq j} \left(a_{ij}^{'} - b_{ij}^{'}\right)^2 \qquad (1)$$

where $N = |V|$ is the scale of A and B, $a_{ij}^{'}$ and $b_{ij}^{'}$ are mean-centered edge weights which are based on the original edge weights a_{ij} and b_{ij}.

3 Methodologies

In this Section, the details of FSM method is described. The framework of FSM is shown in Figure 2. It can be divided into two parts. First, The influence power from one gene to another is calculated according to their local topological structure. In the second part, each node's IP_I and IP_O are derived according to a specific rule and assembled into a vector which is termed as total influence power vector (TIV). Finally, the similarity of the GRNs is outputted. Herein, the calculation of IP and the previously mentioned similarity measurement will be detailed.

3.1 Influence Power

Even though the original GHD can measure the difference between two GRNs, it has certain weaknesses: (1) It is difficult to apply to unweighted GRNs, and thus weights should be manually assigned to each edge. (2) For large-scale GRNs, calculations involving large adjacency matrices should be performed, which requires considerable memory space and CPU resources. Accordingly, IP is proposed to determine efficiently influential genes with the largest contribution to the differences between GRNs. As stated in Section 2.2, the basic principle of IP is to rank genes in a GRN based on their influence during the regulation process. Hence, the most influential genes should be determined. The value of IP of u on v is obtained in terms of the available influence power from u and the influence component from u to v, which are denoted by *available power* ($AP(u)$) and *influence component* ($IC(u \rightarrow v)$), respectively. As different neighbors of u have different local topological structure, the intensity of influence is also different. The former measures the "power" of gene u, whereas the latter measures the probability of gene u influencing v.

If u has a large number of incoming neighbors, then u is regulated by several different sources. These incoming neighbors are regarded as the power providers. An influential gene can exert more "power" on other genes. Hereby, the numbers of genes in node sets $inN^1(u)$ and $inN^2(u)$ are used to present the influence of the provider.

Definition 1. Given a directed GRN $G = (V, E)$, the *available power* of a gene u is denoted by $AP(u)$ and is defined as follows:

$$AP(u) = 1 + \left|inN^1(u)\right| + \left|inN^2(u)\right| \qquad (2)$$

In Equation 2, $\left|inN^1(u)\right|$ represents the width of pre-nodes of u, while $\left|inN^2(u)\right|$ represents the depth of pre-nodes of u. When considering the available power of a specific gene, its pre-nodes' depth property and width property are equally important.

Considering the gene $G2$ of GRN A in Figure 2, the volume of "powers" it distributed to its neighbors must not be the same due to the different topological structures of these neighbors. To quantify the intensity with which a gene uses its available power to influence its neighbors, we make an assumption that a gene tends to indirectly influence more other genes in a network. This assumption is very intuitive if we imagine the influence power as a flow in network and flow seeks its own level. For two genes u and v, given two gene sets $\phi_1 = outN^1(u) + outN^2(u)$ and $\phi_2 = v + outN^1(v)$, ϕ_1 and ϕ_2 would together determines the chance that gene v would be regulated by u.

Definition 2. Given a directed GRN $G = (V, E)$, the *influence component* (IC) between two genes u and v is defined

as the ratio of their degree distributions, that is,

$$IC(u \to v) = \begin{cases} 1 & \left|outN^1(u)\right| = 1 \\ \frac{1+\left|outN^1(v)\right|}{\left|outN^1(u)\right|+\left|outN^2(u)\right|} & else \end{cases} \quad (3)$$

Finally, by considering these two factors, the IP of gene u on gene v is modeled as follows.

Definition 3. Given a directed GRN $G = (V, E)$, for two genes with regulation relationship $(u, v) \in E$, the $influce\ power$ (IP) from u to v is defined by

$$IP(u \to v) = AP(u) \times IC(u \to v) \quad (4)$$

As the differences in GRN functions, the IPs of the same gene across different GRNs are also different. Hence, from the differences of the IPs, the distance between two network can be estimated. Then the FSM method based on the enhanced GHD is introduced to measure this type of distance.

3.2 Implementations of FSM based on Influence Power

To overcome the second weakness described in Section 3.1, the size of input data should be reduced. A desirable GRN similarity measurement metric should satisfy two requirements: First, it should handle real-world GRNs with acceptable time overhead and spatial cost. Secondly, the distance between two different GRNs should be normally distributed and be able to capture the subtle variations. For the first requirement, a total IP vector (TIV) that represents IPs of all nodes is constructed to replace the original GRN adjacency matrix for dimension reduction. For the second requirement, an enhanced GHD for vectors, called *vGHD*, is proposed to represent as the similarity.

Definition 4. Given each edge's IP, IP_O and IP_I are defined as follows:

$$\begin{aligned} IP_O(i) &= \sum_{p_1}\sum_{q_1} IP(p_1 \to q_1) + IP(i \to p_1), \\ IP_I(i) &= \sum_{q_2}\sum_{p_2} IP(p_2 \to q_2) + IP(q_2 \to i) \end{aligned} \quad (5)$$

where $p_1 \in outN^1(i)$ and $q1 \in outN^1(p_1)$, approximately, $q_2 \in inN^1(i)$, $p_2 \in inN^1(q_2)$.

After we obtain the two vectors, IP_O and IP_I. TIV is defined as follows:

$$TIV = S(IP_O) + S(IP_I) \quad (6)$$

where $S(\cdot)$ is the softmax function for normalization which smooth the scale gap between IP_O and IP_I.

Given a GRN, the steps of computing TIV are given in Algorithm 1. TIV is a vector storing each gene's IP. Compared to the raw adjacency matrix, it has smaller size and its elements are the weights of different genes.

Definition 5. Given two GRNs G_A, G_B and their scale N, their similarity, denoted by *vGHD*, is defined as follows:

$$vGHD(G_A, G_B) = \frac{1}{N}\sum_{i=1}^{N}\left(TIV'_{A_i} - TIV'_{B_i}\right)^2, \quad (7)$$

Algorithm 1 TIV Calculation

Input: A GRN $G = (V, E)$
Parameter: $N = |V|$
Output: all genes' IPs in G

```
Initialize TIV to be a zero vector with 1 × N elements.
Initialize a temp matrix Λ with N × N elements.
for gene u in V do
    Let O* be the set of outgoing neighbors of gene u.
    for gene v in O* do
        Calculate IP(u → v) by Equations (2)—(4)
        Λ(u, v) = IP(u → v)
    end for
end for
Calculating IP_O based on Λ following Equation (5).
Calculating IP_I based on Λ following Equation (5).
TIV = S(IP_O) + S(IP_I)
return TIV
```

where

$$\begin{aligned} TIV'_{A_i} &= TIV_{A_i} - \frac{1}{N}\sum_i TIV_{A_i} \\ TIV'_{B_i} &= TIV_{B_i} - \frac{1}{N}\sum_i TIV_{B_i} \end{aligned} \quad (8)$$

TIV_{A_i} is the i-th component of TIV derived from A. TIV_{B_i} is analogously defined.

3.3 Distribution Analysis for *vGHD*

For measurement purposes, the similarity results should be credible and interpretable. Assuming that there is a typical GRN A from brain cancer, and a new incoming GRN B from an unkonwn person. To determine whether B is also brain cancer (as shown in Figure 1), the distribution of $vGHD(A, B)$ should be analyzed.

First, the null hypothesis is set as follows:

$$H_0 :\ networks\ G_A\ and\ G_B\ are\ independent.$$

Under the null hypothesis, the definition of t*vGHD* in Equation (9) can be rewritten as follows:

$$vGHD(A, B) = c - \frac{2}{N}\sum_{i=1}^{N}\left(TIV'_{A_i} \times TIV'_{B_i}\right) \quad (9)$$

where c is a constant. By the transformation of Equation (7), well-known sufficient conditions for asymptotic normality can be used, which can also be easily verified in practice. With the sufficient conditions in [Friedman *et al.*, 1983; Pham *et al.*, 1989], it can be proved that the distribution of *vGHD* in Equation (9) is approximately normal, that is,

$$\frac{vGHD(A, B) - \mu}{\sigma} \sim N(0, 1),$$

where μ and σ are the mean and standard deviation, respectively. The simulation results shown in Figure 3 demonstrate that the distribution of *vGHD* is nearly normal distribution.

As *vGHD(A,B)* obeys follows a normal distribution, given a threshold α, if the p-values exceed α (e.g. 0.05), then H_0

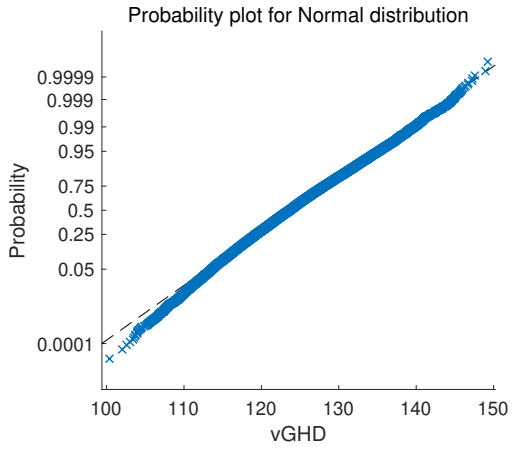

Figure 3: Normal probability plot that compares the distribution of *vGHD* to the normal distribution. The *vGHDs* were calculated from 1000 permutated networks which were generated based on a real GRN from OC. The data points of *vGHDs* appear along the reference line indicates that the *vGHDs* are very close to normal distribution.

cannot be rejected, and thus A and B are independent. Higher p-values indicate a higher probability that B is not brain cancer. Hence, the distribution of *vGHD* provides strong evidence of FSM's interpretability and credibility.

3.4 Computational Complexity

The proposed method considers the local topology structure of the genes when calculating IP. Its computational complexity depends on the density of the input graph. Let k be the number of average neighbors of a gene, and k^2 be the number of second-order neighbors. According to Equations (2)—(5), the computational complexity for calculating each node's IP is $O(k^2n)$. As k is constant and significantly less than n, it is in fact nearly $O(n)$. Furthermore, according to Section 2.2, the computational complexity of the original GHD algorithm is $O(n^2)$. However, after the improvement in the form of Equation 7, it reduces to $O(n)$. The total computational complexity of the proposed method is $O(k^2n + n) \approx O(n)$, $k \ll n$.

4 Experiment

4.1 Datasets

Several public datasets of different scales and sources were used; they are all related to gene regulation: AIDS [Riesen and Bunke, 2008], BZR [Sutherland *et al.*, 2003], COX2 [Sutherland *et al.*, 2003], NCI1 [Wale *et al.*, 2008], ENZYMES [Borgwardt *et al.*, 2005], and PROTEINS [Borgwardt *et al.*, 2005]. Furthermore, there are three real-world GRNs derived from gene expression profiles were used, namely, IDH, OC [Zhang *et al.*, 2016], and RAT [Stevenson *et al.*, 2007]. IDH contains 12717 gene regulation relationships from brain tumor data collected from TCGA panglioma samples, and is divided into two subtypes: IDH-wildtype and IDH-mutation. OC comprises GRNs of patients with ovarian cancer and consists of 11750 genes. There are two classes in this dataset according to the treatment response (platinum-sensitive and platinum-resistant). RAT comprises GRNs with 8799 genes of two groups of mice: one is the experimental group that was exposed to smoke for about 200 days, and the other is the control group that lived in natural environment.

4.2 Performance Evaluation

In the experiment, the proposed method was compared with these baselines and the following state-of-art methods under the experiment setting described in Section 2.1:

- Random walk kernel (RW) [Vishwanathan *et al.*, 2010], it counts common walks in two graphs.
- Shortest path kernel (SP) [Borgwardt and Kriegel, 2005], it counts shortest paths of equal length between pairs of nodes.
- PM [Nikolentzos *et al.*, 2017], it finds an approximate correspondence between the sets of vectors of the two graphs.
- Propagation kernel (PK) [Neumann *et al.*, 2016], a general graph-kernel framework for efficiently measuring the similarity of graphs. It is based on monitoring how information spreads through a set of given graphs.
- Hash graph kernel (HG) [Morris *et al.*, 2016], where continuous attributes are continuously turned into discrete labels using randomized hash functions.

Table 1 shows the average prediction accuracy and standard deviation. Table 2 shows the time required for calculating the distance between two networks.

The experiments were conducted on GRNs and biological networks that, to some extent, have relationships with the process of gene regulatory. The results shown in Table 1 demonstrate two points: (1) In large networks with higher average degree, such as ENZYMES, PROTEINS, and IDH, the proposed method performs the best. The obtained results conforms to the intuition that influential genes should be weighted more in network comparison. Furthermore, the genes' contributions to the mutation of the entire network are different, i.e., influential genes (usually hub genes) contribute more than other marginal genes. (2) In some small networks with lower average degree, although the proposed method is not the best, it still performs well. This is because in those networks, the aggregation effect is not pronounced. Figure 4 shows the comparison of degree distribution between OC, RAT, COX2, and BZR. It can be seen that in OC and RAT, there are only a few genes with large degree distributions that control other nodes. By contrast, in COX2 and BZR, the degree distribution is even across all nodes. The essence of FSM is to detect the differences of those few genes that control the remaining genes. These are assigned higher weights. When the connectivity of higher-position genes changes, a higher distance score is added to them. Hence, the relative accuracy of FSM varies across different datasets.

In terms of runtime, FSM is significantly faster than the other methods. Compared to the RW, FSM is nearly 300 times as fast. This enables FSM to handle large-scale network similarity measurement problems, particularly real-

Method \ Dataset	AIDS	BZR	COX2	NCI1	ENZYMES	PROTEINS	IDH	OC	RAT
RW	98.50 ±0.29	69.76 ±2.11	73.05 ±2.74	56.89 ±0.24	31.60 ±1.30	72.61 ±0.53	-	-	-
SP	98.07 ±0.34	72.83 ±1.87	73.97 ±2.33	62.02 ±0.17	40.75 ±0.81	75.50 ±0.80	-	-	-
PM	**98.89** ±0.22	69.75 ±0.31	75.16 ±1.37	**69.73** ±0.41	42.17 ±2.02	69.81 ±1.23	65.02 ±6.91	65.42 ±5.81	27.50 ±9.79
PK	96.50 ±0.74	**82.06** ±0.13	**77.71** ±0.78	52.09 ±2.18	46.00 ±0.30	68.37 ±1.18	55.08 ±7.85	69.70 ±2.45	59.02 ±2.87
HG	97.12 ±0.36	69.81 ±0.33	74.92 ±0.77	57.95 ±1.46	**66.73** ±0.91	75.14 ±0.47	66.71 ±1.13	66.44 ±0.98	58.25 ±2.06
FSM	98.41 ±0.52	71.96 ±1.35	75.15 ±1.23	64.21 ±0.98	63.70 ±1.03	**77.16** ±0.93	**69.00** ±2.43	**72.06** ±3.02	**61.70** ±0.82

Table 1: Classification accuracy (± standard deviation) in 10-fold validation of the proposed FSM and several baselines and state-of-art methods on various datasets. The symbol "-" indicates that the computation did not finish in 24 h. The best accuracy for each dataset is reported in bold.

Method \ Dataset	AIDS	BZR	COX2	NCI1	ENZYMES	PROTEINS	IDH	OC	RAT
RW	1h27'45"	4'15"	6'43"	2h47'55"	4'57"	4h19'11"	-	-	-
SP	4'03"	1'17"	1'55"	8'13"	1'27"	8'57"	-	-	-
PM	1'49"	8.21"	13.37"	3'21"	8.23"	32.19"	5'33"	5'29"	4'24"
PK	**6.09"**	2.47"	3.13"	2'39"	2.57"	**8.29"**	3'02"	3'03"	2'15"
HG	1'06"	13.02"	15.03"	47.50"	1'11"	1'16"	13'50"	14'04"	11'02"
FSM	10.21"	**1.20"**	**≤1"**	**24.30"**	**1.08"**	11.27"	**28.24"**	**29.56"**	**22.31"**

Table 2: CPU runtime for computation in s ($x^{''}$), min ($x^{'}$), or h (xh). The symbol "-" indicates that the computation did not finish in 24 h.

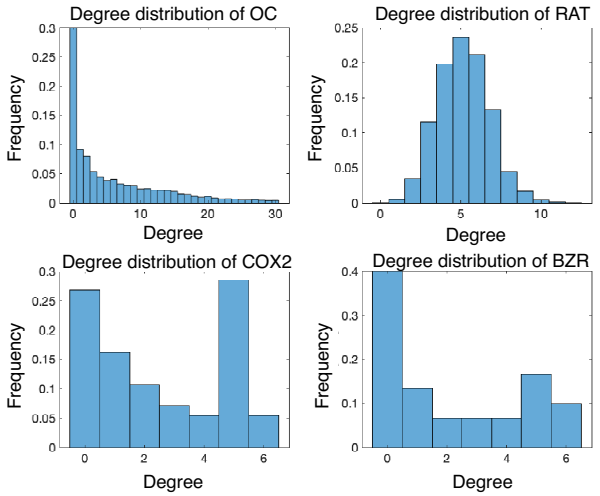

Figure 4: Degree distribution of four datasets. It is clear that the aggregation effect of the top two datasets is more typical and remarkable than that of the bottom two. Hence, FSM performed better in OC and RAT as well.

world GRNs such as OC, RAT, and IDH, which usually contain thousands of nodes. In smaller-size networks such as BZR and COX2, FSM also performs well in terms of computation time. However, in AIDS and PROTEINS, PK is a little faster than FSM. It seems that the speeds of PK and FSM are tied in small datasets, but FSM is more suitable in handling larger data just like IDH, OC and RAT.

5 Conclusion

The FSM method was proposed for fast similarity measurement between GRNs via *vGHD* and influence power. A new index (IP) was defined to describe influential genes in the network. Inspired by common phenomena, the basic principle of FSM is that more influential genes contribute more to the differences of the GRNs. The proposed method exhibits good performance in GRNs and other related bio-networks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61572369), in part by the Key Projects of Guangdong Natural Science Foundation (No. 2018B030311003)

References

[Borgwardt and Kriegel, 2005] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[Borgwardt *et al.*, 2005] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47–i56, 2005.

[Caldarelli *et al.*, 2002] Guido Caldarelli, Andrea Capocci, Paolo De Los Rios, and Miguel A Munoz. Scale-free networks from varying vertex intrinsic fitness. *Physical review letters*, 89(25):258702, 2002.

[de Souza Jacomini *et al.*, 2016] Ricardo de Souza Jacomini, David Correa Martins Jr, Felipe Leno da Silva, and Anna Helena Reali Costa. A framework for scalable inference of temporal gene regulatory networks based on clustering and multivariate analysis. In *BAI@ IJCAI*, pages 7–13, 2016.

[Denitto *et al.*, 2015] Matteo Denitto, Alessandro Farinelli, and Manuele Bicego. Biclustering gene expressions using factor graphs and the max-sum algorithm. In *IJCAI*, pages 925–931, 2015.

[Friedman *et al.*, 1983] Jerome H Friedman, Lawrence C Rafsky, et al. Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 11(2):377–391, 1983.

[Gärtner *et al.*, 2003] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines*, pages 129–143. Springer, 2003.

[Horváth *et al.*, 2004] Tamás Horváth, Thomas Gärtner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 158–167. ACM, 2004.

[Liu *et al.*, 2019] W. Liu, D. Xu, I. W. Tsang, and W. Zhang. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):408–422, Feb 2019.

[Mall *et al.*, 2017] Raghvendra Mall, Luigi Cerulo, Halima Bensmail, Antonio Iavarone, and Michele Ceccarelli. Detection of statistically significant network changes in complex biological networks. *BMC systems biology*, 11(1):32, 2017.

[Morris *et al.*, 2016] Christopher Morris, Nils M Kriege, Kristian Kersting, and Petra Mutzel. Faster kernels for graphs with continuous attributes via hashing. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1095–1100. IEEE, 2016.

[Neumann *et al.*, 2016] Marion Neumann, Roman Garnett, Christian Bauckhage, and Kristian Kersting. Propagation kernels: efficient graph kernels from propagated information. *Machine Learning*, 102(2):209–245, 2016.

[Nikolentzos *et al.*, 2017] Giannis Nikolentzos, Polykarpos Meladianos, and Michalis Vazirgiannis. Matching Node Embeddings for Graph Similarity. In *AAAI*, pages 2429–2435, 2017.

[Pham *et al.*, 1989] Dinh Tuan Pham, Joachim Möcks, and Lothar Sroka. Asymptotic normality of double-indexed linear permutation statistics. *Annals of the Institute of Statistical Mathematics*, 41(3):415–427, 1989.

[Riesen and Bunke, 2008] Kaspar Riesen and Horst Bunke. Iam graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 287–297. Springer, 2008.

[Ruan *et al.*, 2015] Da Ruan, Alastair Young, and Giovanni Montana. Differential analysis of biological networks. *BMC bioinformatics*, 16(1):327, 2015.

[Stevenson *et al.*, 2007] Christopher S Stevenson, Cerys Docx, Ruth Webster, Cliff Battram, Debra Hynx, June Giddings, Philip R Cooper, Probir Chakravarty, Irfan Rahman, John A Marwick, et al. Comprehensive gene expression profiling of rat lung reveals distinct acute and chronic responses to cigarette smoke inhalation. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 293(5):L1183–L1193, 2007.

[Sutherland *et al.*, 2003] Jeffrey J Sutherland, Lee A O'brien, and Donald F Weaver. Spline-fitting with a genetic algorithm: A method for developing classification structure- activity relationships. *Journal of chemical information and computer sciences*, 43(6):1906–1915, 2003.

[Vishwanathan *et al.*, 2010] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

[Wale *et al.*, 2008] Nikil Wale, Ian A Watson, and George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, 14(3):347–375, 2008.

[Zeng *et al.*, 2009] Zhiping Zeng, Anthony KH Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. Comparing stars: On approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1):25–36, 2009.

[Zhang *et al.*, 2016] Xiao-Fei Zhang, Le Ou-Yang, Xing-Ming Zhao, and Hong Yan. Differential network analysis from cross-platform gene expression data. *Scientific reports*, 6:34112, 2016.