TINTIN: Exploiting Target Features for Signaling Network Similarity Computation and Ranking

Huey Eng Chua, Sourav S. Bhowmick
Complexity Institute, Nanyang Technological University, Singapore
hechua|assourav@ntu.edu.sg

Lisa Tucker-Kellogg
Duke-NUS Medical School, National University of Singapore, Singapore
lisa.tucker-kellogg@duke-nus.edu.sg

ABSTRACT

Network similarity ranking attempts to rank a given set of networks based on its "similarity" to a reference network. State-of-the-art approaches tend to be general in the sense that they can be applied to networks in a variety of domains. Consequently, they are not designed to exploit domain-specific knowledge to find similar networks although such knowledge may yield interesting insights that are unique to specific problems, paving the way to solutions that are more effective. We propose TINTIN which uses a novel *target feature-based* network similarity distance for ranking similar signaling networks. In contrast to state-of-the-art network similarity techniques, TINTIN considers both topological *and* dynamic features in order to compute network similarity. Our empirical study on signaling networks from *BioModels* with real-world curated outcomes reveals that TINTIN ranking is different from state-of-the-art approaches.

1 INTRODUCTION

The *network similarity problem* scores the resemblance between a pair of related networks. We can broadly classify network similarity approaches for biological networks into two classes, namely *node mapping-based* and *network feature-based*. The former is based on graph isomorphism [15, 19] as node mappings are performed using different measures and the extent of the network similarity is dependent on the mappings. In contrast, the latter class of approaches do not assume the existence of such node mappings. These approaches [4, 17] generally employ network structure-based *local* or *global* measures to determine similarity between networks. Specifically, *global measures*, such as those based on network features, are derived from the entire biological network and tend to be biased as biological networks are inherently noisy and incomplete [17]. In contrast, *local measures*, which are typically derived from regions of networks that are well-studied, are deemed as more appropriate [17]. For instance, graphlet degree distribution (GDD) [17] and NETSIMILE [4] both use local measures to determine network similarity. A common theme that runs through these network feature-based approaches is their generality and applicability to other types of networks such as social networks. Consequently, they are not designed to exploit domain-specific knowledge to find similar networks, although such knowledge may yield interesting insights that are unique to specific problems. *In this paper, we present a novel network similarity technique called TINTIN, which is designed for signaling networks and leverages such domain-specific knowledge to identify networks that are similar in terms of target features.*

Signaling networks model biological systems as networks of interacting molecules. When biological processes are dysregulated due to diseases, the activities of downstream molecule(s) (referred to as *disease node(s)* in the paper) are affected and typically manifest themselves as phenotypic changes related to the disease. For instance, in the MAPK-PI3K network, the hyperactivity of activated ERK, a downstream molecule, is often linked to cell proliferation [21], a cancer hallmark. This has led to increasing popularity of targeted therapeutic strategies to tackle such diseases where drugs are designed to hit molecules in a signaling network crucial for tumor growth and progression [22]. These molecules (referred to as *targets*) modulate the disease node(s) directly (*e.g.*, MEK [21]) or indirectly (*e.g.*, Raf [21]).

In this paper, we exploit the network features associated with targets in a signaling network to compute similarity between signaling networks. Intuitively, in *target feature-based network similarity* problem we deem a pair of signaling networks as *similar* if their targets have *similar* characteristics. Hence, our network similarity technique called TINTIN (**T**arget-based S**I**gnaling **N**e**T**work S**I**milarity Computatio**N**) is driven by similarity of features of targets that modulate specific disease-related nodes (disease nodes) associated with a pair of signaling networks. It takes as input (a) a *reference network* (e.g., Ras activation), its disease node (e.g., Ras) and an associated set of known targets; and (b) a set of *candidate networks* along with their disease nodes, known targets and features. It ranks these candidate networks based on their *degree of similarity* to the reference network w.r.t. the features of the known targets. TINTIN is potentially useful for several real-world applications such as network-based target prioritization [8][1] and clustering signaling networks based on the feature similarity of targets.

In summary, this paper makes the following key contributions: (a) We introduce the novel problem of *target feature-based network similarity* for signaling networks (Section 4). (b) We present TINTIN, a target-driven approach that is, to the best of our knowledge, the world's first algorithm to address this problem (Section 5). (c) We conduct an empirical study on real signaling networks and drug target data to analyze the similarities and differences of TINTIN's rankings compared to that of the state-of-the-art generic network similarity techniques (Section 6).

ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.
 ISBN 978-1-4503-4722-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3107411.3107470

[1] In contrast to this work, the study in [8] does not focus on details of the network similarity problem. Instead, it *assumes* the existence of such a network similarity technique, which is utilized for target prioritization.

2 RELATED WORK

Compared to generic *target-unaware* network similarity techniques such as GDD [17] and NETSIMILE [4], TINTIN differs in the following ways. First, the network similarity problems are defined differently in existing work. In GDD [17], two networks are deemed similar when they share similar *graphlet degree distribution* that is measured using the *GDD agreement* (topology-based feature); in NETSIMILE, network similarity is measured using a feature vector consisting of seven topological features. In contrast, we define similarity as the likelihood of two network sharing targets with similar topological and dynamic characteristics. Second, GDD was designed for undirected networks (*e.g.*, protein-protein interaction) and is not immediately applicable to directed networks such as cell signaling networks. Third, we consider a wider variety of network features inclusive of both *topological* and *dynamic* features. Both GDD and NETSIMILE use only topological features. Traditionally, the topological representation of signaling networks is considered static [12] as they capture specific observation under particular condition (*e.g.* equilibrium condition). Hence, topological characteristics (*e.g.*, betweenness centrality) that are derived from the network topology are static in nature as well. However, biological systems change with time and time-dependent response of molecular species are typically captured as dynamic time series data (*e.g.*, concentration-time profiles). Lastly, GDD and NETSIMILE are generic techniques. That is, they are not designed to exploit domain-specific knowledge (e.g., knowledge of disease nodes in a signaling network) to find similar networks. In contrast, our similarity measure is "disease node-aware" and is designed specifically for disease-related signaling networks.

3 BACKGROUND

In this section, we introduce key concepts necessary to understand this paper. We first describe the graphical representation of signaling networks and ordinary differential equation (ODE) models. Next, we introduce a recently proposed dynamic feature called *profile shape similarity distance* [5]. Lastly, we briefly describe a *target characterization* framework called TENET [6].

3.1 Graph Model of Signaling Networks

A biological signaling network describes biochemical reactions (with reactants and products) that affect the concentrations of molecular species in the network. Graphically, this reaction is typically represented as a directed hyperedge connecting one set of nodes to another set [14, 20]. Hence, a signaling network is naturally represented as a directed hypergraph $G = (V, E)$. Analysis of directed hypergraphs is generally more complex than graphs and many graph algorithms cannot be used directly on hypergraphs. Hence, hypergraphs are often transformed into graphs containing simple edges for analysis using techniques such as bipartite and substrate graph representation [14]. In this paper, we convert signaling network hypergraphs into bipartite graphs by adopting the method in [10]. We chose the bipartite graph representation as it retains the original structural information of the hypergraphs [14]. Note that the transformed bipartite graph is used to compute the topological features.

We refer to a node in a signaling network as a *candidate target* if when perturbed, it modulates the activity of a specific node (referred to as *disease node*). A *disease node* is a protein that is either involved in some biological processes which may be deregulated, resulting in manifestation of a disease, or be of interest due to its potential role in the disease. Given a signaling network $G = (V, E)$ and a disease node $x \in V$, let the set of nodes having a path leading to x be denoted as $V_x \subseteq V$. Then, the set of *candidate target* nodes in G relevant to x is denoted as $T_x \subseteq V_x$.

Each reaction (edge) in a signaling network is further annotated with dynamic information associated with the biochemical process. In signaling networks, numerous ordinary differential equations (ODEs) containing various reaction kinetics and initial concentrations for every species are used to model the production and consumption rates of different molecular species [2]. The determination of these reaction kinetics can be technically challenging. Hence, a large proportion of these kinetics are usually estimated using parameter estimation techniques [18]. Despite this uncertainty, these under-determined ODE systems can still model real, observable biological behaviour, providing valuable means for quantitative study. Note that ODE-based models of signaling networks are expected to grow further in the future and become an important and accepted way of representing biological knowledge [2]. In this paper, we use hypergraphs containing ODEs for simulation to obtain *concentration-time series profiles* (i.e., plots of concentration against time) of nodes.

3.2 Profile Shape Similarity Distance (PSSD)

In signaling networks, signal responses to perturbation are typically measured in terms of phosphoprotein concentrations dynamics represented as concentration-time profiles. There are certain considerations in comparing these phosphoprotein concentration-time profiles. In signaling networks, reactions occur at different and non-uniform rates [1] resulting in profiles with variable time delays. Hence, a distance measure based on one-to-one alignment on a time axis (Euclidean) is ineffective at detecting similarity in these profiles. A non-linear measure, such as *dynamic time warping* (DTW) distance, allows a more intuitive alignment between profiles [13] and is more suitable for biological time series data [1].

*Definition 3.1. Given two discrete time series φ_u and φ_v, the **dynamic time warping distance** between them is defined recursively as:*

$$DTW(\varphi_u, \varphi_v) = \xi(F(\varphi_u), F(\varphi_v)) + Min \begin{cases} DTW(\varphi_u, Rest(\varphi_v)) \\ DTW(Rest(\varphi_u), \varphi_v) \\ DTW(Rest(\varphi_u), Rest(\varphi_v)) \end{cases}$$

where $F(\varphi_u) = \{\varphi_{u[1]}\}$, $Rest(\varphi_u) = \{\varphi_{u[2]}, \varphi_{u[3]}, \cdots, \varphi_{u[n]}\}$, $\xi(\varphi_{u[i]}, \varphi_{v[j]}) = (\varphi_{u[i]} - \varphi_{v[j]})^2$ and $\varphi_{u[i]}$ is the value of φ_u at time point i [13].

*Definition 3.2. Given a concentration-time profile ζ_u having n time points, denoted as $\varphi_u = \{\varphi_{u[0]}, \cdots, \varphi_{u[n]}\}$, let m be the median value of φ_u. The corresponding **inverted profile** is denoted as $\varphi'_u = \{\varphi'_{u[0]}, \cdots, \varphi'_{u[n]}\}$ where $\varphi'_{u[i]} = 2 \times m - \varphi_{u[i]}$.*

*Definition 3.3. Given a signaling network $H = (V_H, E_H)$, let φ_u, φ_v be the Z-normalized concentration-time profiles of $u, v \in V_H$. The **profile shape similarity distance** [5] of u with respect to v is defined as:*

$$\Phi_{(u,v)} = Min(\text{DTW}(\varphi_u, \varphi_v), \text{DTW}(\varphi'_u, \varphi_v))$$

In summary, PSSD identifies the most relevant upstream regulators by assessing the similarity of the concentration-time series profiles of a target and its upstream regulators.

3.3 Target Characterization using TENET

In this paper, we leverage on a target characterization approach (TENET [6]) to generated the ground truth for assessing the performance of TINTIN. TENET is a network-centric, *in silico* target characterization system, which uses signaling networks having known targets from publicly-available signaling network repositories (e.g., *BioModels*) to learn for each network, a set of topological features that are predictive of targets and a *characterization model* that can be used to generate topological feature-based (TFB) rankings of targets. The characterization model specifies which topological features are important for discriminating the targets in a signaling network and how these features should be combined to quantify the likelihood of a node being a target. It generates different characterization models for different networks as it is unlikely for one characterization model to generalize the characteristics of known targets in all networks due to the complexity and diversity of signaling networks.

4 TARGET FEATURE-BASED NETWORK SIMILARITY PROBLEM

Network similarity measures the *similarity* between a pair of networks. In the literature, there are different strategies to measure such similarity for both directed (e.g., signaling networks) and undirected (e.g., PPI networks) networks. Hence, a set of networks can be *ranked* with respect to a *reference* network based on their similarity degrees to the latter. Formally, the *similarity-based network ranking problem* can be defined as follows.

*Definition 4.1. Given a reference network G and a set of candidate networks $\mathcal{L} = \{L_1, \cdots, L_N\}$, the **similarity-based network ranking problem** computes **network similarity distance** $D(G, L_i)$ between G and each $L_i \in \mathcal{L}$ and ranks them in ascending (or descending) order of $D(.)$. Given the networks L_i and L_j where $L_i, L_j \in \mathcal{L}$, L_i is more similar to G if $D(G, L_i) < D(G, L_j)$. The **best matched network** $L_k \in \mathcal{L}$ of G is the network with the smallest network similarity distance. That is, $\forall i\ D(G, L_k) < D(G, L_i)$.*

Recall that the network similarity distance in TINTIN is measured using target features. Hence, we now formally introduce it. The network similarity distance $D(G, L)$ between a pair of signaling networks, G and L, is based on similarity of the targets in G and L with respect to their network features. Table 1 lists the set of topological and dynamic features that we consider for computing $D(G, L)$ (referred to as *target features*). That is, given a reference network G and two candidate networks L_i and L_j, G is more *similar* to L_i if the feature distribution of the targets in G is more *similar* to that of the targets in L_i across all network features being considered.

*Definition 4.2. Given a reference network G, its disease node x_G and an associated set of known targets T_G; a set of candidate networks $\mathcal{L} = \{L_1, \cdots, L_N\}$, their disease nodes $\bigcup_{i=1}^{|\mathcal{L}|} x_{\mathcal{L}[i]}$ and associated sets of known targets $\bigcup_{i=1}^{|\mathcal{L}|} T_{\mathcal{L}[i]}$; and a set of target features $\mathcal{X}$, the **distribution similarity** of G and $\mathcal{L}_i$ for a target feature $\mathcal{X}_j$ is defined as*

$$p_{G,\mathcal{L}_i,X_j} = \mathcal{F}(\mathcal{G}(G, x_G, T_G, \mathcal{X}_j), \mathcal{G}(\mathcal{L}_i, x_{\mathcal{L}_i}, T_{\mathcal{L}_i}, \mathcal{X}_j)) \quad (1)$$

*where $\mathcal{G}(G, x_G, T_G, \mathcal{X}_j)$ is a function that retrieves target feature $\mathcal{X}_j$ for a set of nodes T_G in a given network G with disease node x_G, and $\mathcal{F}(A, B)$ is a statistical function that computes the similarity of two distributions A and B. Then, the **target feature-based similarity distance** between G and $\mathcal{L}_i$ is defined as*

$$D(G, \mathcal{L}_i) = C(\{p_{G,\mathcal{L}_i,X_j} \mid 1 \leq j \leq |X|\}) \quad (2)$$

where $C(Z)$ is a function that aggregates the set of items Z.

Symbol	Description	Type
θ_u	Degree centrality of node u. The in, out and total degree centralities are denoted as $\theta_{in(u)}$, $\theta_{out(u)}$ and $\theta_{total(u)}$, respectively.	T
α_u	Eigenvector centrality of node u.	T
β_u	Closeness centrality of node u.	T
γ_u	Eccentricity centrality of node u.	T
δ_u	Betweenness centrality of node u.	T
π_u	Bridging coefficient of node u.	T
ζ_u	Bridging centrality of node u.	T
κ_u	Clustering coefficient of node u. The undirected, in, out, cycle and middleman clustering coefficients are denoted as $\kappa_{undir(u)}$, $\kappa_{in(u)}$, $\kappa_{out(u)}$, $\kappa_{cyc(u)}$ and $\kappa_{mid(u)}$, respectively.	T
μ_u	Proximity prestige of node u.	T
ω_u	Target downstream effect of node u [5].	T
$\Phi_{(u,v)}$	Profile shape similarity distance (PSSD) between nodes u and v [5].	D

Table 1: Target features. T: Topological, D: Dynamic

In this work, we use the target features specified in Table 1. Hence, the functions ($\mathcal{G}(G, x_G, T_G, \mathcal{X}_j)$) used for retrieving the target features correspond directly to the formula for computing these features. For example, to compute PSSD, we use the formula given in Definition 3.3. The formula for the remaining target features can be found in [6]. In particular, we use two statistical functions ($\mathcal{F}(A, B)$), namely, Wilcoxon Rank-Sum (Wilcoxon) and Kolmogorov-Smirnov (KS) statistical measures[2] to assess distribution similarity. Hence, distribution similarities are obtained as p-values. We use *Stouffer's method* (Stouffer) as the aggregation function ($C(Z)$) to aggregate the p-values in Equation 2. Note that Stouffer's method can be applied to combine dependent p-values by introducing some degree of dependence (correlation) between pairs by following the approach in [11]. A larger aggregated p-value implies the null hypotheses (a closer target feature-based similarity) are true for every test.

5 THE TINTIN ALGORITHM

Algorithm 1 outlines the TINTIN algorithm. Given a reference signaling network G, its disease node x_G and known targets T_G, a set of candidate networks $\mathcal{L}$, their disease nodes $x = \bigcup_{i=1}^{|\mathcal{L}|} x_{\mathcal{L}[i]}$ and known targets $T = \bigcup_{i=1}^{|\mathcal{L}|} T_{\mathcal{L}[i]}$, it identifies the best matched network $G_{best} \in \mathcal{L}$ of G and a ranked list of $\mathcal{L}$ in two phases. TINTIN provides an optional relaxation parameter p_r to configure the criteria for filtering out dissimilar networks.

[2]The Wilcoxon and KS tests are nonparametric and are suitable for features with distribution that are unknown a priori.

Algorithm 1 Algorithm TINTIN

Require: Reference network G and its disease node x_G and known targets T_G; set of candidate networks $\mathcal{L}$, their disease nodes x and known targets T; relaxation parameter p_r (optional).
Ensure: Best matched network G_{best} and network ranked list r.
1: $\mathcal{X}_{best}$, P_{best}, H, $Truth$ ←LEARNBESTVARIANT($\mathcal{L}$,x,T)
2: p_t ←LEARNPTHRESHOLD($Truth$, P_{best}, p_r)
3: G_{best}, r ←GETBESTNETWORK(G, x_G, T_G, $\mathcal{L}$, $\mathcal{X}_{best}$, p_t, H)

Algorithm 2 Procedure LEARNBESTVARIANT

Require: Set of training networks $\mathcal{L}$, their disease nodes x and known targets T.
Ensure: Best feature type $\mathcal{X}_{best}$; matrix of p-values for training networks using the best feature type P_{best}; matrix of all features values in all training networks H; and the ground truth $Truth$.
1: $\mathcal{A}$ ←INITIALIZE($\{\mathcal{X}_{all}$,$\mathcal{X}_T$,$\mathcal{X}_D\}$)
2: H ←EXTRACTFEATURES($\mathcal{L}$,$\mathcal{X}_{all}$)
3: $Truth$ ←GETGROUNDTRUTH($\mathcal{L}$,H)
4: **for** iteration i=1 to $|\mathcal{L}|$ **do**
5: **for** iteration j=i+1 to $|\mathcal{L}|$ **do**
6: **for** iteration k=1 to $|\mathcal{A}|$ **do**
7: M ←GETRELEVANTFEATURES(H,$\mathcal{A}_k$)
8: P_W ←WILCOX($\mathcal{L}_i$, $\mathcal{L}_j$, M)
9: P_{KS} ←KS($\mathcal{L}_i$, $\mathcal{L}_j$, M)
10: $PVal_i(k, \mathcal{L}_j)$ ←COMBINEP(P_W,P_{KS})
11: **end for**
12: **end for**
13: **end for**
14: **for** iteration i=1 to $|\mathcal{A}|$ **do**
15: **for** iteration j=1 to $|\mathcal{L}|$ **do**
16: $Rank_{i,j}$ ←RANK($PVal$, i, j)
17: **end for**
18: $Distance_i$ ←SUM(SPEARMAN($Rank$, i),KENDALL($Rank$, i),$Truth_i$)
19: **end for**
20: $\mathcal{X}_{best}$, P_{best}, H, $Truth$ ←GETBESTVARIANT($Distance$,$PVal$,$\mathcal{A}$)

The Learning Phase. In this phase, TINTIN learns the best variant for finding G_{best} (Lines 1-2 in Algorithm 1). In particular, three variants (Table 2) of TINTIN utilizing different feature sets (only topological features, only dynamic feature and combination of topological and dynamic features) are considered. We begin by identifying the *ground truth* (true order of the ranking) of the candidate networks. The ground truth reflects the *actual* similarity of the target characteristics of the networks based on prior knowledge or empirical results. Hence, it can either be provided by experts familiar with the candidate networks and its associated targets or generated automatically by using models that characterize the known targets in these networks. We adopt the latter strategy by utilizing TENET[6]. Specifically, in order to generate the ground truth, we exploit TENET in the following way. Given a reference network G and two candidate networks L_1 and L_2, L_1 is more similar to G if the characterization model of L_1 produces a better characterization of known targets in G compared to L_2. That is, the characterization model of L_1 achieves a larger AUROC (area under ROC curve) for known targets in G compared to that of L_2. Note that AUROC is typically used to assess classifier performance as the metric is robust for imbalanced datasets [9]. The ground truth can be interpreted as the ordering of the candidate networks based on decreasing AUROC.

Variant	All Features	T. Features	D. Features	Stouffer
TINTIN$_A$	√			√
TINTIN$_S$		√		√
TINTIN$_D$			√	√

√ marks inclusion in the variant. T.=Topological, D.=Dynamic

Table 2: Variants of our network similarity strategy.

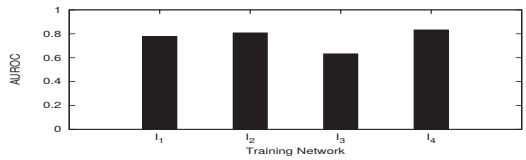

Figure 1: Ground truth found using AUROC.

Next, all the features for targets in all candidate networks are extracted and the combined p-value for the Wilcoxon and the KS tests for each variant is computed. Then, for each candidate network $\mathcal{L}_i$, three ranked lists (denoted as r), each corresponding to one variant (see Table 2), are obtained by ordering the remaining candidate networks in decreasing order of the combined p-value. Next, the disagreement between r and the ground truth is measured using the *Spearman footrule distance* and *Kendall-Tau distance*, denoted as $\Upsilon(.)$ and $\Lambda(.)$, respectively. Given two complete rankings r_1 and r_2 of a set of N individuals, let $r_1(i)$ be the rank of $i \in N$ in the ranked list r_1. Then, the distances are calculated as:

$$\Upsilon(r_1, r_2) = \sum_{i \in N} |r_1(i) - r_2(i)|$$

$$\Lambda(r_1, r_2) = |\{i, j\} : r_1(i) < r_1(j) \text{ and } r_2(i) > r_2(j)|$$

For each variant, the Spearman footrule and Kendall-Tau distances for all the candidate networks are aggregated to obtain the overall distance. The best variant is the one yielding the least overall distance. Furthermore, this phase determines a p-value threshold (denoted as p_t) that shall be used to filter off dissimilar network (Algorithm 1, Line 2). First, it obtains a mapping between the combined p-values and "poor"[3] AUROC. Then, these combined p-values are averaged to obtain p_m. The value of p_t is set as the minimum of p_m and p_r since combined p-values less than p_t are removed.

The Ranking Phase. In the next phase, the TINTIN algorithm identifies the best matched network G_{best} and rank of the candidate networks with combined p-values greater than p_t. First, values of predictive topological features is extracted for G. Then, for each pair of (G,$\mathcal{L}_i$), the Wilcoxon and KS tests are performed for each of these features and the p-values obtained are combined. Finally, the candidate networks with combined p-values greater than or equal to p_t are ranked in order of decreasing combined p-values to obtain r. The top-ranked network is G_{best}.

Observe that it is possible for our ranking strategy to return no best matched network if all the combined p-values are less than p_t. This indicates that none of the given candidate networks are similar. In this case, we can either explore additional candidate networks to identify G_{best} or relax p_t using p_r to obtain a suboptimal G_{best}.

THEOREM 5.1. *The worst-case time complexity of Algorithm 1 is* $O(|\mathcal{L}|(|\mathcal{L}|-1)(\mathcal{G}(\mathcal{X}_{all}) + |\mathcal{A}|(|V_{\mathcal{L}[i]}||V_{\mathcal{L}[j]}|)^2))$ *time in the worst*

[3] We deem AUROC<0.5 as "poor" since it indicates performance worse than random prioritization.

case, where $\mathcal{G}(X_{all})$ is the worst time complexity for extracting all features, $|V_{\mathcal{L}[i]}|$ is the number of nodes of the i^{th} network in $\mathcal{L}$ and $|\mathcal{A}|$ is the number of TINTIN variants.

The proof of the above theorem is given in [7].

6 EXPERIMENTS

TINTIN is implemented using Java. In this section, we investigate its performance. All experiments are performed on a computer system using a 64-bit operating system with 8GB RAM and a dual core processor running at 3.60GHz.

Datasets. We use a reference network (I_0) and four candidate networks (I_1 to I_4) for our experiments as shown in Table 3. Note that although there are more than 600 curated signaling networks in *Biomodels*, we restrict our study to only five signaling networks. This is because we need to identify known targets of signaling networks for validating our experimental results. Unfortunately, to the best of our knowledge, there is no publicly available technique that can automatically identify known targets from signaling networks by analyzing biomedical literature. Hence, we are confined to manual target curation from a large volume of biomedical literature, a time-intensive process. Also, although larger signaling networks are desirable, to the best of our knowledge, no publicly-available large signaling networks (e.g., human cancer signaling network) contain dynamic information of *all* edges (ODES), preventing us to exploit dynamic features such as PSSD. The targets of I_0 are curated from [16] (Ca^{2+}, EGF:EGFR, EGFR, activated EGFR and Ras) and from [3] (dimerized EGFR). The curated targets of the candidate networks are given in [7].

Best TINTIN variant. First, we identify the best variant of TINTIN. Specifically, we examine how various variants (Table 2) perform on the given set of candidate networks (I_1 to I_4). Then, we examine if the best performing variant for the candidate networks is also effective in identifying the best matched network for the reference network. Finally, we analyze their runtime performance.

Figure 2 reports the results. Interestingly, we observe that the variant using only dynamic feature performs better (smaller distance between ground truth and TINTIN ranking) than variants that either use only topological feature set or a combination of topological and dynamic feature sets. *This underscores the importance of considering dynamic feature. The best performing variant obtained from the learning phase (learnBestVariant) is TINTIN_D.*

Next, we examine the effect of applying different variants on the reference network to validate if the learning phase yields the desired best performing variant. Indeed, TINTIN_D is the top performing variant for the reference network (Figure 2, top, reference network). In fact, majority of the TINTIN variants identify I_4 as the best matched network (Figure 2, middle).

In terms of the runtime performance (Figure 2, bottom), the learning process (Step 2) consumes the bulk of its execution time. Hence, learning can be performed offline to improve the runtime performance. The improvement is up to 2 orders of magnitude for certain variants (*e.g.*, TINTIN_S for comparison of networks with fewer than 100 nodes). In addition, we observe that the runtime performance is dependent on the size of the network, the types and number of features used by the variant. In subsequent experiments, we shall use TINTIN_D.

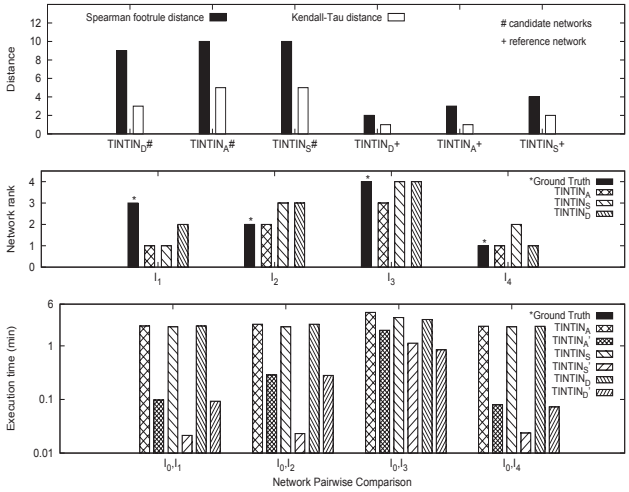

Figure 2: Performance of TINTIN variants. Top: distance between ground truth and rankings from variants in candidate networks and reference network, middle: individual network ranking, bottom: runtime performance.

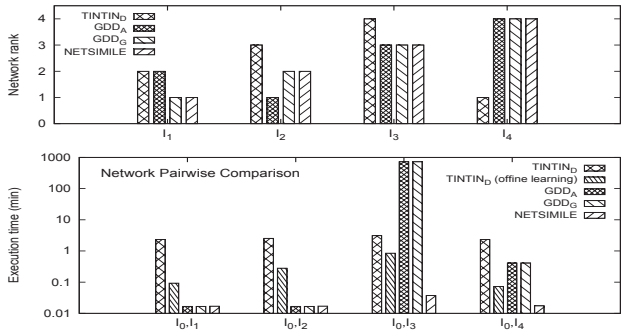

Figure 3: Performance of different network similarity-based ranking approaches.

Performance of the LEARNPTHRESHOLD procedure. In this set of experiments, we identify the threshold p-value p_t learnt from the candidate networks. Table 4 shows the combined p-values of each pair of candidate networks using TINTIN_D and the corresponding AUROC when applying the characterization model of one candidate network to another. p_t is the average combined p-value of the cells marked with ♯. That is, $p_t = \frac{2.2x10^{-16}+2.2x10^{-16}+0.045}{3} = 0.01$. When we apply $p_t = 0.01$ to the ranked list of I_1 to I_4 obtained using TINTIN_D, we note that I_3 (AUROC=0.632, combined p-value=$2.2x10^{-16}$) is considered as a dissimilar network and filtered off. Note that p_t affects the number of networks being ranked, but not the actual rank of the network. Hence, it is possible to have no best matched network if the given set of candidate networks is considered dissimilar to the unseen network.

Comparison with state-of-the-art. Lastly, we compare TINTIN_D against GDD [17] and NETSIMILE [4] in terms of network ranking and runtime performance. We consider both the arithmetic and geometric versions of GDD which are denoted as GDD_A and GDD_D, respectively. Since these target-unaware network similarity approaches define similarity differently from TINTIN (Section 2), it is

Network notation	I_0	I_1	I_2	I_3	I_4
Data set (BioModel ID)	Ras activation (0000000161)	MAPK-PI3K (0000000146)	glucose-stimulated insulin secretion (0000000239)	endomesoderm gene regulatory (0000000235)	glucose metabolism (0000000244)
Disease node	RasGTP_{PM}	ERKPP	$\text{ATP}_{mitochondrial}$	Protein_E_Endo16	acetate
No. of nodes	46	36	59	622	47
No. of hyperedges	43	34	45	778	109
No. (%) of targets	5 (10.9%)	9 (25%)	6 (10.2%)	206 (33.1%)	16 (34%)

Table 3: Dataset.

	I_1 Model	I_2 Model	I_3 Model	I_4 Model
I_1	-	$6.52\text{x}10^{-4}$ [0.55]	$2.2\text{x}10^{-16}$ [$0.47^{\sharp}$]	0.128 [0.64]
I_2	$6.52\text{x}10^{-4}$ [0.62]	-	$2.2\text{x}10^{-16}$ [0.54]	0.045 [$0.43^{\sharp}$]
I_3	$2.2\text{x}10^{-16}$ [0.65]	$2.2\text{x}10^{-16}$ [0.60]	-	$2.2\text{x}10^{-16}$ [$0.48^{\sharp}$]
I_4	0.128 [0.61]	0.045 [0.51]	$2.2\text{x}10^{-16}$ [0.55]	-

Table 4: Summary of result for LEARNPTHRESHOLD procedure (Algorithm 1). The $(i,j)^{th}$ cell entry is of the form $x[y]$ where x is the combined p-value for the (I_i,I_j) pair and y is the AUROC when characterization model of I_j is applied to I_i. $^{\sharp}$ refer to cases with low AUROC (i.e., AUROC<0.5).

Rank comparison	S	K
TINTIN,GDD_A	6	4
TINTIN,GDD_G	6	3
TINTIN,NETSIMILE	6	3

Rank comparison	S	K
GDD_A,GDD_G	2	1
GDD_A,NETSIMILE	2	1
GDD_G,NETSIMILE	0	0

Table 5: Summary of comparison of ranks obtained using different approach. S and K indicate Spearman footrule and Kendall-Tau distances, respectively.

not possible to have a good network ranking benchmark that can be used as ground truth for comparison. Instead, we compare the similarities and differences in the rankings derived from TINTIN as compared to that from other approaches. As observed in Figure 3 (top), our approach differs in ranking of networks I_1 to I_4 when compared to traditional target-unaware network similarity approaches. The differences in ranking is more significant when we compare TINTIN against target-unaware approaches (Table 5, left) versus a comparison among traditional target-unaware approaches only (Table 5, right). In particular, I_4 was ranked best by TINTIN and worst by the target-unaware approaches (Figure 3, top).

The runtime performance is affected by the network size (Figure 3, bottom), of which the most significant[4] impact is experienced by the GDD-based approaches. In particular, NETSIMILE performs the best and scales well even for larger networks. The order of the approaches in terms of runtime performance is NETSIMILE $\prec$ TINTIN_D (offline learning) $\prec$ $\text{TINTIN}_D \prec$ GDD. Hence, *our approach has moderate runtime performance and the ranks produced are markedly different from state-of-the-art approaches.*

7 CONCLUSIONS & FUTURE WORK

In this paper, we present TINTIN, a target feature-based signaling network similarity computation and ranking technique by exploiting the topological and dynamic characteristics of the targets. It is interesting to note that the empirical study highlights a single dynamic feature (PSSD) as being more important than a set of topological features in identifying the best matched network. This signals the importance of considering dynamic features in measuring network similarity. However, to the best of our knowledge, majority of the work on network feature focus on topological features instead. Hence, as part of future work, we intend to explore novel dynamic network features that could be used to study signaling networks. In addition, our empirical results demonstrate the differences in terms of network ranking of TINTIN compared to state-of-the-art target-unaware network similarity techniques.

Acknowledgments. Huey Eng Chua and Sourav S Bhowmick were partially supported by MOE AcRF Tier-1 Grant RGC 1/13.

[4]The runtime performance of GDD degrades by about 3 orders of magnitude when applied to I_3 (622 nodes) as compared to I_4 (47 nodes).

REFERENCES

[1] J. Aach et al. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495-508, 2001.
[2] B. Aldridge et al. Physicochemical modelling of cell signalling pathways. *Nat. Cell Biol.*, 8(11):1195-1203, 2006.
[3] J. Bangham. Therapeutics: Cetuximab constricts conformational contortionist. *Nat. Rev. Cancer*, 5:421, 2005.
[4] M. Berlingerio et al. NetSimile: a scalable approach to size-independent network similarity. In *ASONAM*, 2013.
[5] H. Chua et al. PANI:A Novel Algorithm for Fast Discovery of Putative Target Nodes in Signaling Networks. In *ACM BCB*, 2011.
[6] H. Chua et al. TENET:Topological Feature-based Target Characterization in Signaling Networks. *Bioinformatics*, 31(20):3306-3314, 2015.
[7] H. Chua, et al. TAPESTRY: A Network Similarity Ranking-Based Approach For Prioritizing Nodes In Signaling Networks. http://www.ntu.edu.sg/home/assourav/TechReports/TAPESTRY-TR.pdf, Technical Report, 2015.
[8] H. Chua, et al. TAPESTRY: Network-centric Target Prioritization Nodes In Disease-related Signaling Networks. In *ACM BCB*, 2016.
[9] D. Cieslak et al. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. In *ICDM*, 2008.
[10] J. Engelfiet, et al. A comparison of boundary graph grammars and context-free hypergraph grammars. *Inform. Comput.*, 84(2):163-206, 1990.
[11] J. Hartung. A note on combining dependent tests of significance. *Biom. J.*, 41(7):849-855, 1999.
[12] V. Janjić et al. Biological function through network topology: a survey of the human diseasome. *Brief. Funct. Genomics*, els037, 2012.
[13] E. Keogh et al. Derivative Dynamic Time Warping. In *SDM*, 2001.
[14] S. Klamt et al. Hypergraphs and cellular networks. *PLoS Comput. Biol.*, 5(5):e1000385, 2009.
[15] L. Meng, et al. Local versus global biological network alignment. *Bioinformatics* 32(20): 3155-3164, 2016.
[16] NCI drug dictionary. *http://www.cancer.gov/drugdictionary/*, National Cancer Institute, accessed in April 2015.
[17] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177-e183, 2007.
[18] J. O. Ramsay et al. Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc.: Series B*, Statistical Methodology, 69(5), 2007.
[19] J. Raymond et al. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7):521-533, 2002.
[20] A. Ritz et al. Signaling hypergraphs. *Trends Biotechnol.*, 32(7):356-362, 2014.
[21] P. Roberts et al. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, 26(22):3291-3310, 2007.
[22] M. Vanneman et al. Combining immunotherapy and targeted therapies in cancer treatment. *Nat. Rev. Cancer*, 12(4), 2012.