TAPESTRY: Network-centric Target Prioritization in Disease-related Signaling Networks

Huey Eng Chua[§,‡] Sourav S Bhowmick[§,‡] Jie Zheng[§,‡] Lisa Tucker-Kellogg[†]
[§]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[‡]Complexity Institute, Nanyang Technological University, Singapore
[†]Duke-NUS Graduate Medical School, National University of Singapore, Singapore
hechua|assourav|zhengjie@ntu.edu.sg, lisa.tucker-kellogg@duke-nus.edu.sg

ABSTRACT

Target prioritization ranks molecules in biological networks according to a score that seeks to identify molecules that fulfill particular roles (*e.g.*, drug targets). We study this problem in the context of partial information (*e.g.*, unknown targets) and present TAPESTRY, a network-based approach that prioritizes candidate targets in a given signaling network with unknown targets by utilizing knowledge (target characteristics) gained from curated targets in another set of signaling networks. We consider both *topological* and *dynamic features* and use a weighted sum approach to examine the relative influence of these two classes of features on the prioritization results. TAPESTRY exploits a knowledge base of *characterization models* and *predictive topological features* of a set of signaling networks (*candidate networks*) with curated targets. Then, given a signaling network G with unknown targets, TAPESTRY identifies a candidate network most *similar* to G and selects its characterization model as *prioritization model* for computing a *topological feature-based rank* of each *candidate* node in G. Next, a *dynamic feature-based rank* is computed for these nodes by leveraging the time-series curves of ODEs associated with the edges in G. Finally, these two ranks are *integrated* and used for prioritizing candidate targets. We experimentally study the performance of TAPESTRY using signaling networks from *BioModels* with real-world curated outcomes. Our results demonstrate its effectiveness and superiority in comparison to state-of-the-art approaches.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.3 [**Life and Medical Sciences**]: Biology & genetics

General Terms

Algorithms

Keywords

Target prioritization, network similarity measure

BCB'16, October 2–5, 2016, Seattle, WA, USA.

http://dx.doi.org/10.1145/2975167.2975178.

1. INTRODUCTION

Systems biology models biological systems as networks of interacting molecules. In particular, signaling networks are useful in capturing signal flows that underly biological processes such as growth and apoptosis. Hence, these networks are increasingly used to analyze various characteristics of biological systems such as *prioritizing targets* for gene annotation and drug discovery [7, 19]. Intuitively, a *target* is typically an endogenous molecule such as a protein, a gene or a nucleic acid sequence that affects the outcome of a disease or a medical condition[1]. The *target prioritization* problem aims to rank candidate targets according to their potential of being a target based on some *importance measures* such as gene expression level [7]. Ideally, for a given network, this problem is best tackled when we have knowledge of known targets, complete topological and dynamic (in the form of concentration or flux measurements) information. However, this is difficult to achieve in practice. Hence, in this paper we study this problem in the context of partial information (*e.g.*, unknown targets).

Several approaches have been recently proposed in the literature for target prioritization. They can be broadly classified into *network-oblivious* and *network-centric* categories. *Network-oblivious* approaches prioritize candidate targets using experimental data (*e.g.*, gene expression [26]) and non-network features (*e.g.*, gene structure and sequence [44], sensitivity coefficient [39]). In contrast, *network-centric* approaches (*e.g., NetworkPrioritizer* [19]) adopt a network view of biological entities and perform *in silico* network-based analysis to prioritize candidate targets. Unlike network-oblivious approaches that treat the biological system as a black box, network-centric approaches consider such system as interactions of molecules. This enables us to capture the complexity of these interactions and their system-level impact during target prioritization. Hence, we advocate that network-centric approaches are more appropriate for prioritizing candidate targets.

Recent network-centric target prioritization approaches include LSA [15], *GeneWanderer* [22] and *NetworkPrioritizer* [19]. *GeneWanderer* is based on a random walk algorithm and uses a score derived from the distance of a gene to known disease genes for prioritization. The reliance on known disease genes limits the usage of *GeneWanderer* to networks associated to well-studied diseases such as cancer. Moreover, the use of a single metric assumes that targets can

[1]In pathogen-related diseases, the target can sometimes be endogenous to the pathogen, instead of the host. In this paper, our focus is on targets related to non-pathogen-related diseases.

be characterized well by one topological feature, which may not necessarily be true [9]. In contrast, LSA and *NetworkPrioritizer* prioritize candidate targets using network dynamics and multiple topological features, respectively. In particular, *NetworkPrioritizer* provides flexibility on the choice of topological features, feature weights, and rank aggregation method for the prioritization task. However, the need to manually configure these settings makes it less palatable to end users as it requires a keen understanding of the underlying network and target characteristics. Note that end users may not necessarily be equipped with such in-depth knowledge of the underlying network.

Moreover, *GeneWanderer* and *NetworkPrioritizer* only leverage topological features of the network ignoring its dynamic characteristics, even though the latter is an important aspect of biological systems and an integral part of comprehending signaling networks. In our recent work [8] (referred to as PANI), we took the first step to investigate the role of dynamics in target prioritization by examining two topological features and one dynamic feature called *profile shape similarity distance* (PSSD[2]) for prioritization in curated signaling networks. We observed that superior prioritization results can be generated when the interplay of dynamic and topological features is exploited, suggesting that network dynamics may also play an important role for this problem.

In this paper, we propose TAPESTRY (**TA**rget **P**rioritization using n**E**twork **STR**ucture and d**Y**namics), a network-centric approach that prioritizes candidate targets in signaling networks with respect to a *disease* node using both topological *and* dynamic features. It is built on top of TENET, a *target characterization*[3] technique [9]. Specifically, TENET deploys a support vector machine (SVM)-based strategy to learn offline the set of *predictive topological features* for characterizing *known* curated targets in a set of publicly-available signaling networks (referred to as *candidate networks*) and generates a set of *characterization models* based on these features. Then given a disease-related signaling network with unknown targets (referred to as *unseen network*) and the set of characterization models generated by TENET, TAPESTRY *prioritizes* its nodes (ranks nodes based on some criteria) with respect to a *disease node* by leveraging a characterization model and network dynamics. Specifically, it selects the "best" characterization model it should adopt as its *prioritization model* from the collection of characterization models of the candidate networks. A *prioritization score* (referred to as *putative target score*) is then derived from the selected model and the dynamics of the unseen network. Subsequently, we use this score to prioritize candidate targets.

Remark. Each of the topological features and dynamic features, that TAPESTRY considers, encapsulates a different possible "reason" why a candidate node might be important and influential in a signaling network (related to the disease). TAPESTRY seeks a context-dependent selector for the feature set that is related to node importance. In particular, the TAPESTRY algorithm searches for the most similar network that has known feature weightings. These known weightings suggest a weighting of features for that context,

[2]Briefly, it identifies the most relevant upstream regulators by assessing the similarity of the concentration-time series profiles of a target and its upstream regulators.

[3]*Target characterization* is the process of defining the characteristics of the targets and is useful in drug design for these targets and in the identification of novel targets that share similar characteristics with known targets.

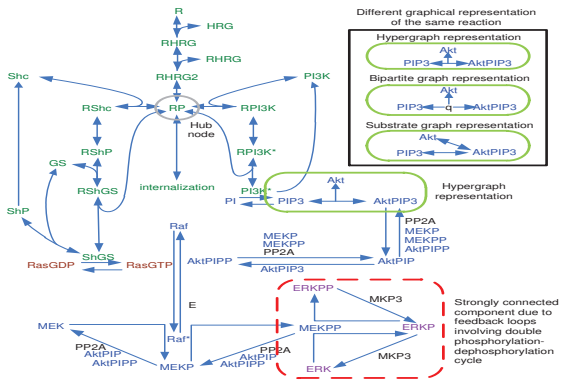


Figure 1: MAPK-PI3K network adapted from [17].

and they exemplify how a node can be influential in that type of network based on that particular feature set.

In summary, this paper makes the following contributions:

- In Section 3, we present a novel *disease node-driven target prioritization* problem that aims to rank candidate targets in a disease-related signaling network based on a specific disease node instead of ranking them independent of any disease.
- In Section 4, we present a novel network-centric target prioritization algorithm called TAPESTRY, that utilizes a machine learning-based target characterization framework (TENET) and statistical analysis to address the disease node-driven target prioritization problem in a signaling network with unknown targets.
- In Section 5, we conduct a detailed empirical study on real signaling networks and drug target data to demonstrate the effectiveness and superiority of TAPESTRY compared to the state-of-the-art generic network-centric target prioritization techniques.

2. PRELIMINARIES

In this section, we first describe the essential graphical representation of signaling networks and ordinary differential equation (ODE) models. Next, we briefly give an overview of TENET [9], a machine learning-based target characterization framework on which the TAPESTRY algorithm is built.

2.1 Graph Model of Signaling Networks

A biological signaling network describes the interaction between molecular species involved in the network. The interaction takes the form of biochemical reaction and processes. An example of such biochemical reaction is the activation of ERK into phosphorylated ERK (ERKPP) by its kinase, phosphorylated MEK in Figure 1. Graphically, this reaction is typically represented as a directed hyperedge connecting one set of nodes (*e.g.*, {ERK and MEK}) to another set (*e.g.*, {ERKPP}) [21, 36]. 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 [21] (*e.g.,* Figure 1). In this paper, we convert signaling network hypergraphs into bipartite graphs by adopting the

method in [13]. We chose the bipartite graph representation as it retains the original structural information of the hypergraphs [21][4]. Note that the transformed bipartite graph is used to compute the topological features.

In these graphs, the edges can be further annotated with dynamic information associated with the biochemical process. This generally takes the form of ordinary differential equation (ODE). The resulting ODE model describes the system's behaviour over time by using mass-action kinetics to model the production and consumption rates of different molecular species [1]. These models are typically constructed by translating prior knowledge of production and consumption rate of different molecular species into differential equations. The determination of these reaction kinetics can be technically challenging. Hence, a large proportion of these kinetics are usually estimated using parameter estimation techniques [34]. Despite this uncertainty, these underdetermined ODE systems can still model real, observable biological behaviour, providing valuable means for quantitative study. Although ODE models are still small in size now, they are expected to grow in the future and become an important and accepted way of representing biological knowledge [1]. In this paper, we use hypergraphs containing ODEs (Table 3) for simulation to obtain *concentration-time series profiles*[5] of nodes. These profiles are used to compute a dynamic feature (PSSD [8]) which is subsequently used for *disease node-driven target prioritization*.

For clarity, we refer to a node as a *candidate target* if, when perturbed, it modulates the activity of a specific *disease node*. A *disease node* is a protein that is involved (or hypothesized to be involved) in some disease-causing process (*e.g.*, phosphorylated ERK in the MAPK-PI3K network [17]). 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 $T_x \subseteq V_x$.

2.2 Target Characterization using TENET

TENET [9] is a network-centric, *in silico* target characterization system. TENET uses signaling networks having known targets from publicly-available signaling network repositories (e.g., *BioModels* [24]) 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. 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.

Specifically, given a signaling network and a disease node, TENET identifies nodes that are likely regulators (nodes positioned upstream) of the disease node based on the interconnections of nodes. Then, it extracts a set of topological features (Table 1) of these candidates from the network and ranks each candidate target based on each topological feature. Next, it partitions the preprocessed data into a training set, a model selection (validation) set and a test set. An SVM-based algorithm is deployed to learn the set of predictive topological features that best characterizes known targets of the network and a characterization model based on these features. The SVM algorithm uses *structural feature selection* and *weighted misclassification cost* to improve the performance of the final characterization model by addressing the issues of irrelevant features, class membership uncertainty and imbalanced data set. It uses a set of known targets curated from literature and clinical trials repositories, such as [25] as the benchmark for learning. The curation process is detailed in [9]. Table 2 shows the characterization models of a set of signaling networks generated by TENET. Details of these networks are given in Table 3.

Table 1: Target features. Topological and dynamic features are denoted as T and D, respectively.

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. $\kappa_{undir(u)}$, $\kappa_{in(u)}$, $\kappa_{out(u)}$, $\kappa_{cyc(u)}$ and $\kappa_{mid(u)}$ denotes undirected, in, out, cycle and middleman clustering coefficients, respectively.	T
μ_u	Proximity prestige of node u.	T
ω_u	Target downstream effect of node u [8].	T
$\Phi_{(u,v)}$	Profile shape similarity distance (PSSD) between nodes u and v [8].	D

3. DISEASE NODE-DRIVEN TARGET PRIORITIZATION PROBLEM

Recall that target prioritization is the process of ranking candidate targets according to some criteria (*e.g.*, sensitivity, gene expression level) so that a target node has higher priority if the disease node is more sensitive to its changes. It is potentially useful in helping to plan experiments since resources are limited and experiments can be costly and time-intensive. This is especially true in drug development [28]. Specifically, our goal is to generate a *putative target score* for *ranking* nodes in a given signaling network G with unknown targets (referred to as *unseen network*) according to their ability in modulating a disease node of G.

DEFINITION 1. *Given an unseen signaling network $G = (V, E)$ and a disease node $x \in V$, the* ***disease node-driven target prioritization problem*** *assigns a* ***target rank*** *r_u for each node $u \in V$. For any two nodes $u, v \in V$, u is more likely to achieve better modulation of x compared to v if $r_u < r_v$.*

State-of-the-art network-centric target prioritization approaches generate target ranks differently. Some approaches (*e.g.*, local sensitivity analysis (LSA) [15]) use a single score (*e.g.*, sensitivity coefficient) to assign target ranks whereas others (*e.g.*, *NetworkPrioritizer* [19]) use an aggregated score (*e.g.*, *Weighted Borda Fuse*). We adopt the latter approach by using an aggregated score referred to as *putative target score*. Specifically, we use 16 topological and one dynamic features[6] as listed in Table 1 to compute this score. These features are selected based on their role in measuring relative importance of a node in a signaling network. The formal

[4]The "dummy" nodes generated due to this transformation are not ranked during target prioritization.

[5]Plots of concentration against time.

[6]Although it is desirable to study a variety of dynamic features, we did not find any suitable ones besides PSSD that we proposed in [8].

Table 2: Characterization models of a set of networks generated by TENET.

	MAPK-PI3K ($\mathbf{I}_1$)	glucose-stimulated insulin secretion ($\mathbf{I}_2$)	endomesoderm gene regulatory ($\mathbf{I}_3$)	glucose metabolism ($\mathbf{I}_4$)
Linear Kernel C	$2^{-0.4}$	2^{10}	$2^{0.8}$	$2^{11.2}$
Target Misclassification Cost C^+	0.5	0.9	0.7	0.8
Feature Selection Approach$^\sharp$	BSE	hybrid	BSE	BSE
Features Selected	δ,π,θ_{in}, θ_{out}	π,β,κ_{cyc}, κ_{undir}	δ,ζ,π,β,κ_{cyc},γ,α,κ_{in},κ_{mid}, μ,κ_{out},θ_{out},ω,θ_{total},κ_{undir}	ζ,β,κ_{cyc},γ,α,κ_{in},κ_{mid}, μ,ω,κ_{out},θ_{total},κ_{undir}

$^\sharp$ BSE=backward stepwise elimination. *Hybrid* involves filtering using statistical test results followed by BSE

Table 3: Signaling networks used in the experiments.

Network notation	$\mathbf{I}_0$	$\mathbf{I}_1$	$\mathbf{I}_2$	$\mathbf{I}_3$	$\mathbf{I}_4$
Data set (BioModel ID)	Ras activation (0000000161)	MAPK-PI3K (0000000146)	glucose-stimulated insulin secretion (0000000239)	endomesoderm gene regulatory (0000000235)	glucose metabolism (0000000244)
Disease node	$\text{RasGTP}_{\text{PM}}$	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%)

definitions as well as motivation for selecting these features are given in [10]. Note that in signaling networks, an aggregated score may be more appropriate as targets are generally characterized by multiple features [9].

DEFINITION 2. *Given an unseen network $G = (V, E)$ and a disease node $x \in V$, the **putative target score** of a node $v_i \in V$ is defined as*

$$\varrho_{v_i} = \begin{cases} \varpi_1 \frac{D_{v_i}}{\text{MAX}_{u \in V}(D_u)} + \varpi_2 \frac{T_{v_i}}{\text{MAX}_{u \in V}(T_u)} & \text{if } \mathcal{R}(v_i, x) \text{ is true,} \\ 0 & \text{otherwise} \end{cases}$$

*where D_{v_i} and T_{v_i} are **dynamic** and **topological feature-based ranks** of v_i, respectively; $\varpi_1 + \varpi_2 = 1$; MAX(.) is the maximum operator and $\mathcal{R}$(u,v) is a boolean function that returns true when there is a path from u to v and false when otherwise.*

The nodes in V are ranked based on decreasing ϱ such that for two nodes $v_i, v_j \in V$, v_i is ranked higher than v_j if $\varrho_{v_i} > \varrho_{v_j}$. Observe that in contrast to state-of-the-art network-centric approaches, our proposed score considers both topological *and* dynamic features. Dynamics play an important role in understanding biological systems [33]. Although topological features are found related to dynamics and modularity in biological networks [31], they are unable to explain the temporal aspects of the networks [33]. Hence topology and dynamics complement each other by providing different perspectives of the biological system. We use a weighted sum approach for computing the score to incorporate the relative influence of topological and dynamic features in prioritizing candidate targets.

Additionally, existing network-centric approaches are generally *disease node-unaware*. That is, they rank *all* nodes in the given network regardless of the fact that some nodes may not influence the activity of the disease node. In contrast, we address this limitation by ensuring that the ranking based on putative target score is *disease node-driven*.

Remark. It is worth mentioning that the target prioritization problem differs from the target characterization problem embodied by TENET in the following key way. The goal of the latter is to identify a set of characteristics (*i.e.*, features such as betweenness centrality) that characterizes the *known* targets in a given signaling network. In contrast, the goal of the target prioritization problem that TAPESTRY seeks to address is to rank nodes (*candidate* targets) of a given signaling network. This input network does not necessarily have known targets.

4. THE TAPESTRY ALGORITHM

The TAPESTRY algorithm is designed to address the disease node-driven target prioritization problem by leveraging the output of TENET. Given an unseen network G and a disease node x, the key idea deployed here is to *select* the characterization model of a signaling network with known targets that *best matches* G as its prioritization model and then use it to rank the nodes (candidate targets) in G with respect to x. Specifically, the ranking exploits topological and dynamic network features of G to obtain the topological features-based (TFB) and dynamic feature-based (DFB) ranks, respectively. Note that the input signaling hypergraph is transformed to a bipartite graph in TENET for computing the topological features. We use the original hypergraph for computing PSSD in TAPESTRY. Algorithm 1 outlines the TAPESTRY algorithm. It comprises of three key phases, namely, the *preprocessing* phase (Line 1), the *prioritization model selection* phase (Line 2), and the *target ranking* phase (Line 3). We shall elaborate on them in turn.

Phase 1: Preprocessing. In a signaling network, certain nodes may not influence the disease node and these nodes can be removed from further processing. Hence, in this phase TAPESTRY traverses G in a depth-first manner and uses a *reachability rule* based on the transitive closure of G to identify potential *candidate* nodes by eliminating nodes in G that cannot reach x. The reader may refer to [8] for details related to this rule.

Phase 2: Prioritization model selection. In this phase, TAPESTRY leverages on TENET to *select* an appropriate *prioritization model* that is most suited for G from the collection of known characterization models of a set of signaling networks $\mathcal{L}$ with known targets. Recall that such collection of characterization models is generated by TENET. Intuitively, it selects the characterization model of the network that is most *similar* to G as its prioritization model for target prioritization. Due to space constraints, we only provide an overview here. Detailed description of the FINDBESTMATCHEDNETWORK procedure is given in [10].

Observe that the key challenge in this phase is to measure the *network similarity distance* $D(G, L)$ between a pair of signaling networks, G and L, based on similarity of their *target features*. Hence, given an unseen network G and two

Algorithm 1 Algorithm TAPESTRY

Require: Unseen network G and disease node x_G; set of candidate networks $\mathcal{L}$, their disease nodes x, and known targets T, relaxation parameter p_r (optional), weights learning parameter t_w (optional) and weight range step-size $step_w$ (optional).
Ensure: $|V_{can}| \times |N|$ matrix of TAPESTRY-prioritized node $\mathcal{P}$.
1: $V_{can} \leftarrow$ filterCandidate(G, x_G)
2: $G_{best} \leftarrow$ findBestMatchedNetwork$(G, x_G, \mathcal{L}, x, T, p_r)$
3: $\mathcal{P} \leftarrow$ prioritizeTargets$(G, x_G, V_{can}, \mathcal{L}, x, G_{best}, t_w, step_w)$

candidate networks $L_i \in \mathcal{L}$ and $L_j \in \mathcal{L}$, we consider G is more *similar* to L_i if the topological and dynamic feature (Table 1) distributions of the candidate targets in G is more *similar* to that of the targets in L_i, across all network features being considered. We use the Wilcoxon Rank-Sum (Wilcoxon) and Kolmogorov-Smirnov (KS) statistical measures[7] for assessing distribution similarity. These tests yield a set of p-values which can be aggregated into a combined p-value using *Fisher's inverse* χ^2 *method* (Fisher) [30] or *Stouffer's method* (Stouffer). A smaller combined p-value implies a closer feature distribution similarity. In TAPESTRY, we rank the networks according to increasing combined p-value and use this rank as the network similarity distance.

Since there are recent efforts to determine similarity between networks by employing network structure-based measures [4, 32], at first glance a keen reader may question the justification behind proposing yet another network similarity measure for target prioritization. Specifically, *graphlet degree distribution* (GDD) [32] and NETSIMILE [4] both use local measures to determine network similarity. A common theme that runs through these approaches is their generality and applicability to other types of networks such as social networks. Hence, why we cannot adopt these techniques to realize this phase? We provide an answer to this question by highlighting the differences between the aforementioned network similarity measure deployed in TAPESTRY and these existing measures.

- First, the network similarity problems are defined differently in existing work. In GDD [32], 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 characteristics. The characteristics are measured using topological *and* dynamic features.
- Second, GDD is applicable only to undirected networks due to the definition of the graphlet patterns. Hence, GDD is not suitable for signaling networks.
- Third, we consider a wider variety of network features inclusive of both *topological* and *dynamic* features. Although GDD uses a large number of graphlet patterns, it does not consider dynamic features. Similarly, NETSIMILE uses only seven topological features[8].
- 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, although such knowledge may yield interesting insights that are unique to specific problems, paving way to solutions that are more effective. In contrast, our similarity measure is "target-aware" and is designed specifically for signaling networks to address the disease node-driven target prioritization problem.

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

[8]The NETSIMILE feature vector consists of degree, clustering coefficient, average number of two-hop neighbours, average clustering coefficient, number of edges in a node's egonet, number of outgoing edges from the egonet and number of neighbours of the egonet.

Algorithm 2 Procedure PRIORITIZETARGETS

Require: Unseen network G, its disease node x_G and candidate targets V_{can}; set of candidate networks $\mathcal{L}$ and their disease nodes x, best matched network G_{best}, weights learning parameter for AUROC $t_{w(\text{AUROC})}$ (optional), weights learning parameter for AUPR $t_{w(\text{AUPR})}$ (optional) and weight range step-size $step_w$ (optional).
Ensure: $|V_{can}| \times |N|$ matrix of TAPESTRY-prioritized node $\mathcal{P}$.
1: $t_{w(\text{AUROC})}, t_{w(\text{AUPR})}, step_w \leftarrow$ init$(t_{w(\text{AUROC})}, t_{w(\text{AUPR})}, step_w)$
2: **for** iteration i=1 to $|\mathcal{L}|$ **do**
3: $Y_{\mathcal{L}[i]} \leftarrow$ filterCandidate$(\mathcal{L}_i, x_{\mathcal{L}[i]})$
4: $\mathcal{S}_{\mathcal{L}[i]} \leftarrow$ getTopologicalRank$(\mathcal{L}_i, \mathcal{L}_i, Y_{\mathcal{L}[i]})$
5: $\mathcal{D}_{\mathcal{L}[i]} \leftarrow$ getDynamicsRank$(\mathcal{L}_i, x_{\mathcal{L}[i]}, Y_{\mathcal{L}[i]})$
6: **for** iteration j=1 to $\frac{10}{step_w}$ **do**
7: $\rho_{\mathcal{L}[i]} \leftarrow$ getPutativeTargetScore$(\mathcal{L}_i, \mathcal{D}_{\mathcal{L}[i]}, \mathcal{S}_{\mathcal{L}[i]}, j \times step_w, 1 - j \times step_w)$
8: $\mathcal{P}_{\mathcal{L}[i]} \leftarrow$ rank$(\rho_{\mathcal{L}[i]})$
9: $\text{ROC}_{\mathcal{L}[i]} \leftarrow$ rocAnalysis$(\mathcal{P}_{\mathcal{L}[i]}, \text{getKnownTargets}(\mathcal{L}_i))$
10: $\text{AUPR}_{\mathcal{L}[i]} \leftarrow$ auprAnalysis$(\mathcal{P}_{\mathcal{L}[i]}, \text{getKnownTargets}(\mathcal{L}_i))$
11: **if** $\text{ROC}_{\mathcal{L}[i]} \geq t_{w(\text{AUROC})}$ **and** $\text{AUPR}_{\mathcal{L}[i]} \geq t_{w(\text{AUPR})}$ **then**
12: $W_{\mathcal{L}[i]} \leftarrow$ appendToWeightList$(j \times step_w)$
13: **end if**
14: **end for**
15: **end for**
16: $\mathcal{S} \leftarrow$ getTopologicalRank(G, G_{best}, V_{can})
17: $\mathcal{D} \leftarrow$ getDynamicsRank(G, x, V_{can})
18: $W_{best} \leftarrow$ getWeightList(W, G_{best})
19: **for** iteration i=1 to $|W_{best}|$ **do**
20: $\rho_i \leftarrow$ getPutativeTargetScore$(G, \mathcal{D}, \mathcal{S}, W_{best[i]}, 1 - W_{best[i]})$
21: $\mathcal{P}_i \leftarrow$ rank(ρ_i)
22: **end for**

Detailed empirical study related to the superiority of our network similarity distance computation technique in comparison to the aforementioned state-of-the-art approaches is reported in [10] and is orthogonal to this work. Certainly, any other superior target-aware signaling network similarity computation technique can be seamlessly integrated to our TAPESTRY framework. Note that the choice of the network similarity computation technique affects our selection of *best matched network* and hence the ranking of the targets. The TAPESTRY framework allows a comparison of different computation technique to identify the best technique for the purpose of target prioritization. Details of such comparison are available in [10].

Based on the aforementioned network similarity distance measure, the *best matched network* $G_{best} \in \mathcal{L}$ of G is selected as follows. Given an unseen network G and its disease node x_G, a set of candidate networks $\mathcal{L}$ with their known targets and characterization models (*e.g.,* Table 2), and an optional *relaxation parameter*[9] p_r, first it learns the type of features (topological or dynamic) and the method for combining p-values (Fisher or Stouffer) that are the most relevant for

[9]An optional parameter for relaxing the criteria for filtering out dissimilar networks.

finding G_{best}. Next, 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 ordered according to decreasing combined p-values. The top-rank network is selected as G_{best}. Subsequently, we use the characterization model of G_{best} as the prioritization model in the next phase.

Phase 3: Candidate target ranking. In this phase (Algorithm 2), TAPESTRY first performs *weight learning* to explore appropriate weights (in the range of 0 to 1) to assign to the topological feature-based (TFB) (Line 4) and dynamic feature-based (DFB) (Line 5) ranks using the candidate networks. In particular, TAPESTRY computes the putative target scores using different weight allocation (Lines 7- 8), then performs ROC and AUPR analysis on the rankings (Lines 9-10). Those weights resulting in AUROC and AUPR greater than or equal to the pre-specified thresholds[10] $t_{w(\text{AUROC})}$ and $t_{w(\text{AUPR})}$ are stored in a weight array W (Line 12).

Next, it performs *target ranking*. The TFB ranks of the unseen network are generated using G_{best} (Line 16) and DFB (PSSD) ranks are obtained by exploiting the time-series curves of ordinary differential equations (ODEs) associated with the edges in G (Line 17). Briefly, PSSD computes the distance between the time-series curve of each node and the disease node using a measure derived from z-normalized *dynamic time warping* (DTW) [20] where a smaller DTW value implies greater similarity between two time-series curves. The relevant list of weight (denoted as W_{best}) for the best matched network is retrieved from W. Finally, for each weight in W_{best}, the putative target score is computed and used for producing a prioritized rank list $\mathcal{P}_i$ (Lines 20-21).

THEOREM 1. *The worst-case time complexity of* TAPESTRY *is* $O((|\mathcal{L}|+1)((|V_{\mathcal{L}[i]}|+|E_{\mathcal{L}[i]}|)^2+\mathcal{G}(\mathcal{X}_{all})+|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|)+\mathcal{H}(G,\mathcal{L}))$ *where* $\mathcal{H}(\cdot)$ *and* $\mathcal{G}(\cdot)$ *are the worst time complexity of the given network similarity ranking approach (Phase 2) and the feature extraction function;* $\mathcal{X}_{all}$ *is the set of features;* $|V_{\mathcal{L}[i]}|$ *and* $|E_{\mathcal{L}[i]}|$ *are number of nodes and edges of* $\mathcal{L}_i$ *(the* i^{th} *network in* $\mathcal{L}$*), respectively; and* $|\varphi_{\mathcal{L}[i]}|$ *is the number of time points in the time series of* $\mathcal{L}_i$.

PROOF. For the FILTERCANDIDATE procedure, the conversion of the input signaling network $G=(V,E)$ to a bipartite graph $G_{bi}=(V_{bi},E_{bi})$ takes $O(|V_{bi}|+|E_{bi}|)$ time. In directed acyclic graph (DAG) conversion, $O(|V_{bi}|+|E_{bi}|)$ time is required for finding strongly connected components (SCC, see Figure 1) using Tarjan's algorithm. In the worst case, the signaling network is a complete directed graph and conversion to a DAG takes $O(|V_{bi}|^2+|E_{bi}|)$ time since $|V_{bi}|<|V_{bi}|^2$. In the indexing of the DAG graph $G_{dag}=(V_{dag},E_{dag})$, the depth-first traversal requires $O(|V_{\text{DAG}}|+|E_{\text{DAG}}|)$ time while computing the set of nodes that can reach x takes $O(|V_{\text{DAG}}|)$ time. Hence, FILTERCANDIDATE algorithm takes $O(|V_{bi}|^2+|E_{bi}|)$ time since $|V_{bi}|=|V|+|E|$, $|E_{bi}|=\sum_{(U,W)\in E}(|U|+|W|)$ and $(|V_{bi}|+|E_{bi}|)\geq(|V_{\text{DAG}}|+|E_{\text{DAG}}|)$.

The time complexity of the FINDBESTMATCHEDNETWORK procedure (denoted as $\mathcal{H}(\cdot)$) depends on the algorithm that is used for network similarity ranking. In TAPESTRY, our proposed similarity computation approach has complexity of $O(|\mathcal{L}|(|\mathcal{X}_{all}|\mathcal{G}(\mathcal{X}_{all})+|\mathcal{A}|(|\mathcal{L}|-1)(|V_{\mathcal{L}[i]}||V_{\mathcal{L}[j]}|)^2))$ time in the worst case, where $\mathcal{X}_{all}$ is the set of network 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 variants [10].

In Algorithm 2, FILTERCANDIDATE; GETPUTATIVETARGETSCORE; RANK; ROCANALYSIS; AUPRANALYSIS; APPENDTOWEIGHTLIST and GETWEIGHTLIST procedures require $O(|V_{bi,\mathcal{L}[i]}|^2+|E_{bi,\mathcal{L}[i]}|)$; $O(|V_{\mathcal{L}[i]}|)$; $O(|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|))$; $O(|V_{\mathcal{L}[i]}|^2)$ [14]; $O(|V_{\mathcal{L}[i]}|^2)$; $O(1)$ and $O(1)$ time, respectively, where $|V_{bi,\mathcal{L}[i]}|$ and $|E_{bi,\mathcal{L}[i]}|$ are the number of nodes and edges of the bipartite graph of $\mathcal{L}_i$ (the i^{th} network in $\mathcal{L}$) and $|V_{\mathcal{L}[i]}|$ is the number of nodes in the $\mathcal{L}_i$. In GETTOPOLOGICALRANK procedure, feature extraction takes $\mathcal{G}(\mathcal{X}_{all})$; generation of the TFB scores takes $O(|\nu|)$ [5] where ν is the number of support vectors (in the worst case, $|\nu|=|V_{\mathcal{L}[i]}|$); and ranking using heapsort takes $O(|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|))$ time [35]. Hence, the time complexity of GETTOPOLOGICALRANK is $\mathcal{G}(\mathcal{X}_{all})+O(|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|))$. In GETDYNAMICSRANK procedure, in the worst case, $O(|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|)$ time is needed for Z-normalization, for performing profile inversion, and for DTW computation using FASTDTW [38], where $|\varphi_{\mathcal{L}[i]}|$ is the number of points in the time-series profile of $\mathcal{L}_i$. The heapsort used for ranking in the RANK procedure takes $O(|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|))$ time. Hence, the time complexity of GETTOPOLOGICALRANK is $O(|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|+|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|))$. Taken together, the time complexity of Algorithm 2 is $O((|\mathcal{L}|+1)(|V_{bi,\mathcal{L}[i]}|^2+|E_{bi,\mathcal{L}[i]}|+|V_{\mathcal{L}[i]}|log(|V_{\mathcal{L}[i]}|)+\mathcal{G}(\mathcal{X}_{all})+|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|))$. This can be further reduced to $O((|\mathcal{L}|+1)((|V_{\mathcal{L}[i]}|+|E_{\mathcal{L}[i]}|)^2+\mathcal{G}(\mathcal{X}_{all})+|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|))$ since the signaling network is a single strongly connected component in the worst case and $O(|E_{bi,\mathcal{L}[i]}|)=O(|V_{bi,\mathcal{L}[i]}|^2)$ and $|V_{bi,\mathcal{L}[i]}|=|V_{\mathcal{L}[i]}|+|E_{\mathcal{L}[i]}|$.

Hence, the TAPESTRY algorithm requires $O((|\mathcal{L}|+1)((|V_{\mathcal{L}[i]}|+|E_{\mathcal{L}[i]}|)^2+\mathcal{G}(\mathcal{X}_{all})+|\varphi_{\mathcal{L}[i]}||V_{\mathcal{L}[i]}|)+\mathcal{H}(G,\mathcal{L}))$ time. □

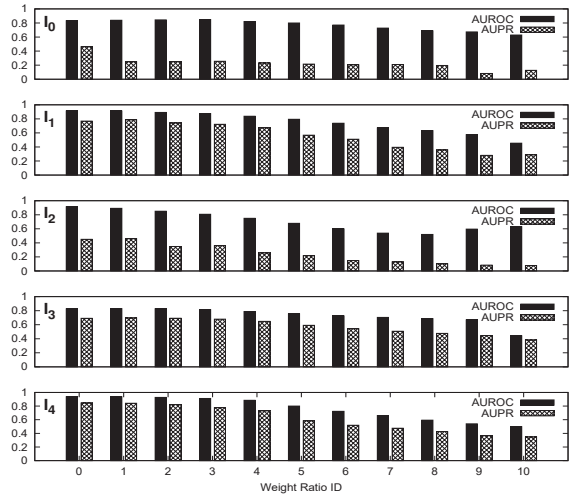
5. PERFORMANCE STUDY

TAPESTRY is implemented using Java. In this section, we investigate the performance of this algorithm. The 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.

5.1 Experimental Setup

Datasets. We use an unseen network (I_0) and four candidate networks (I_1 to I_4) for our experiments as shown in Table 3. Recall that TAPESTRY selects the candidate network with the most similar target features as that of the unseen network as the best matched network. Hence, the pool of networks used as candidate networks affect the final choice of the best matched network and the ranking of the target genes. Clearly, a larger and varied pool of candidate networks (with details of curated targets) allows comparison to be made across a wider spectrum of networks in order to identify a best matched network. Unfortunately, such candidate network pool is currently unavailable in the literature. Consequently, we created a pool of candidate networks by performing *manual* target curation from a large volume of biomedical literature in order to identify known targets of signaling networks for validating our experimental results. We restricted this pool to 5 networks as manual curation is time-intensive. Observe that the largest network studied here comprises of 622 nodes. Although larger signaling networks are desirable, to the best of our knowledge,

[10] AUROC in the range of [0.7, 0.8] and [0.8, 0.9] indicate acceptable and excellent performances, respectively [18]. We set $t_{w(\text{AUROC})}$ and $t_{w(\text{AUPR})}$ to 0.8 to ensure excellent performance of the prioritization model.



Weight Ratio ID	0	1	2	3	4	5	6	7	8	9	10
ϖ_1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ϖ_2	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0

Figure 2: Effect of varying the weights of TFB and DFB on prioritization ranks for candidate networks (I_1 to I_4) and the unseen network (I_0). For I_0, we use the prioritization model of the best matched network (selected using AUROC-based model selection method as described in [10]).

no publicly-available large signaling networks (*e.g.,* human cancer signaling network [12]) contain dynamic information of *all* edges (ODEs).

The curated targets of the unseen network I_0 (`Ras activation` network) are $\texttt{EGFR}_{\texttt{PM}}$, activated `EGFR`, `Ca`, $\texttt{RasGTP}_{\texttt{PM}}$ and $\texttt{RasGDP}_{\texttt{PM}}$ [3, 29]. The curated targets of the candidate networks are given in [10]. Note that in our experimental study **we assume that the targets of I_0 are unknown**. Hence, these targets are used only to validate the *quality* of target prioritization by TAPESTRY and existing network-centric target prioritization approaches.

Network-centric target prioritization approaches. We compare TAPESTRY with several state-of-the-art network-centric target prioritization approaches, namely, random prioritization, local sensitivity analysis (LSA) [15], *NetworkPrioritizer* (WBF and WASF) [19] and PANI [8]. In random prioritization, the nodes were randomly assigned a rank in the range $[1–|V|]$ where $|V|$ is the number of nodes in the signaling network and we assume that no ranking ties are present. LSA was performed using *Copasi* [37] with the following configuration: {task=sensitivities; subtask=time series; function=all variables of the model; and variable=all parameter values}. We consider both *Weighted Borda Fuse* (WBF) and *Weighted AddScore Fuse* (WASF) in *NetworkPrioritizer* and consider all features provided. Note that uniform weights were used for rank aggregation since we do not have prior knowledge of the best weights or features to consider. The weights allocation in PANI are set according to [8].

Performance metrics. In an ideal situation, the performance of the target prioritization tools should be measured as the closeness of the predicted ranks and a set of *benchmark ranks* for a given disease-related signaling network. However, to the best of our knowledge, such benchmark ranks are not publicly available. Hence, we resort to using AUROC or AUPR measure which provides us an idea of whether known targets are ranked higher than non-targets in general. Specifically, in our experiments, we prioritize the candidate targets in the `Ras activation` network (unseen network) using TAPESTRY and other state-of-the-art prioritization tools.

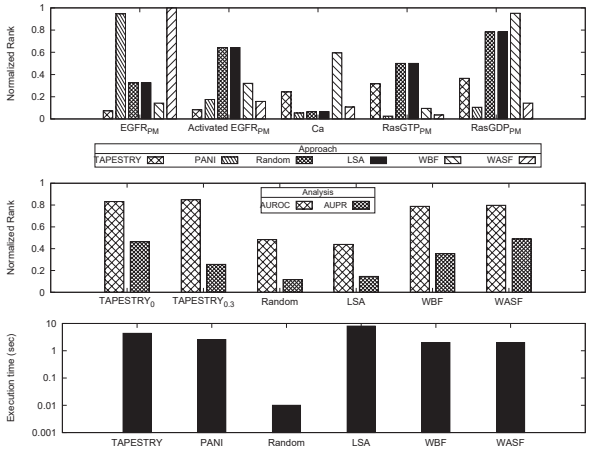


Figure 3: Performance of different target prioritization approaches. Top: normalized ranks of targets (shorter histograms imply higher rank), center: AUROC and AUPR, bottom: runtime performance.

Prioritization Model. We select the characterization model of the `glucose metabolism` network (I_4) as the prioritization model as Phase 2 of TAPESTRY identifies it as best matched network[11] of I_0 (detailed in [10]). Note that it may seem that the `Ras activation` network to be most similar to the `MAPK-PI3K` network (I_1) as both are implicated in cancer and share several common nodes such as `Ras`. However, when considering target feature-based network similarity, this may not necessarily be true as the similarity is affected by the target sets and the topological layouts of the networks instead of disease or functional similarity.

5.2 Experimental Results

Weight settings by TAPESTRY. The TAPESTRY algorithm (Phase 3) uses a weighted sum score of DFB ranks (PSSD) and TFB ranks (topological features) (denoted as ϖ_1 and ϖ_2, respectively) to compute a putative target score for prioritizing candidate targets (Definition 2). Hence, the weights ϖ_1 and ϖ_2 invariably affect the prioritization results. In this set of experiments, we vary these weights to examine the impact of including PSSD as a feature on the prioritization results. The weight parameters ϖ_1 and ϖ_2 are set such that $\varpi_1 + \varpi_2 = 1$ and $\varpi_1, \varpi_2 \in [0 - 1]$. Hence, prioritization is based on PSSD alone when $\varpi_1 = 1$ and on topological features alone when $\varpi_2 = 1$. The range of weights we tested in shown in Figure 2. In the candidate networks (Figure 2, I_1 to I_4), we observe a decreasing trend and setting ϖ_1 less than or equals to 0.3 yielded AUROC greater than or equals to 0.8. Except for I_2, the rest

[11]The exact network similarity ranking of I_1 to I_4 with respect to I_0 is $I_4 \succ I_1 \succ I_2 \succ I_3$ where I_4 is the most similar network to I_0.

Table 4: Ranks and normalized ranks of candidate nodes in I_0 using different approaches.

No.	Node id	Curated target	TAPESTRY		PANI		Random		LSA		WBF		WASF	
			Rank	Norm. rank	Rank	Norm. rank	Rank	Norm. rank	Rank	Norm. rank	Rank	Norm. rank	Rank	Norm. rank
1	RasGTP_Golgi_GM	No	9	0.22	34	0.85	39	0.85	34	0.81	2	0.07	3	0.08
2	EGF_EC	No	3	0.07	18	0.45	26	0.57	42	1	27	1	36	0.95
3	CAPRI$_{cyt}$	No	35	0.85	37	0.93	40	0.87	39	0.93	27	1	30	0.79
4	serca	No	40	0.98	7	0.18	30	0.65	36	0.86	27	1	34	0.89
5	PIP$_{PM}$	No	21	0.51	12	0.3	31	0.67	3	0.07	25	0.93	29	0.76
6	PI$_{PM}$	No	25	0.61	9	0.23	5	0.11	37	0.88	27	1	38	1
7	Shc$_{PM}$	No	5	0.12	40	1	2	0.04	7	0.17	21	0.78	23	0.61
8	CaCAPRI_PM$_{PM}$	No	32	0.78	31	0.78	22	0.48	20	0.48	10	0.37	15	0.39
9	RactCa	No	33	0.80	36	0.9	42	0.91	8	0.19	26	0.96	28	0.74
10	Shc_star$_{PM}$	No	1	0.02	21	0.53	34	0.74	29	0.69	9	0.33	10	0.26
11	EGFR$_{PM}$	Yes	3	0.07	12	0.3	15	0.33	6	0.14	27	1	36	0.95
12	PLC_act$_{PM}$	No	39	0.95	3	0.08	9	0.20	32	0.76	11	0.41	12	0.32
13	RasGTP_pal$_{cyt}$	No	19	0.46	38	0.95	19	0.41	26	0.62	7	0.26	9	0.24
14	PLC$_{PM}$	No	20	0.49	37	0.93	13	0.28	15	0.36	23	0.85	27	0.71
15	PIP2$_{PM}$	No	23	0.56	19	0.48	6	0.13	30	0.71	22	0.81	25	0.66
16	Activated EGFR$_{PM}$	Yes	4	0.10	35	0.88	44	0.96	27	0.64	8	0.30	6	0.16
17	ca_buffer$_{cyt}$	Yes	10	0.24	18	0.45	3	0.07	25	0.60	3	0.11	2	0.05
18	Ract	No	37	0.90	32	0.8	28	0.61	5	0.12	26	0.96	28	0.74
19	Rinh	No	37	0.90	25	0.63	43	0.93	21	0.5	26	0.96	28	0.74
20	RinhCa	No	33	0.80	20	0.5	4	0.09	35	0.83	26	0.96	28	0.74
21	IP3	No	12	0.29	4	0.1	29	0.63	22	0.52	5	0.19	7	0.18
22	RasGDP_Golgi$_{GM}$	No	14	0.34	23	0.58	12	0.26	33	0.79	6	0.22	5	0.13
23	Ca_RasGRP_GM$_{GM}$	No	31	0.76	13	0.33	18	0.39	13	0.31	15	0.56	21	0.55
24	DAG_GM$_{GM}$	No	11	0.27	15	0.38	41	0.89	31	0.74	16	0.59	13	0.34
25	RasGRP_DAG$_{GM}$	No	30	0.73	33	0.83	35	0.76	19	0.45	14	0.52	19	0.5
26	CaCAPRI$_{cyt}$	No	24	0.59	16	0.4	32	0.70	18	0.43	12	0.44	17	0.45
27	DAG$_{PM}$	No	29	0.71	26	0.65	20	0.43	11	0.26	27	1	37	0.97
28	RasGTP_depal$_{cyt}$	No	18	0.44	17	0.43	1	0.02	12	0.29	6	0.22	8	0.21
29	RasGDP_depal$_{cyt}$	No	17	0.41	24	0.6	37	0.80	14	0.33	13	0.48	11	0.29
30	RasGDP_pal$_{cyt}$	No	17	0.41	25	0.63	14	0.30	16	0.38	13	0.48	11	0.29
31	Ca_PLCe$_{cyt}$	No	22	0.54	10	0.25	10	0.22	28	0.67	19	0.70	22	0.58
32	Ras_CaPLCe$_{GM}$	No	26	0.63	34	0.85	33	0.72	42	1	17	0.63	18	0.47
33	PIP2_GM$_{GM}$	No	34	0.83	27	0.68	46	1	42	1	27	1	32	0.84
34	ER$_{erMembrane}$	No	41	1	28	0.7	11	0.24	27	0.64	24	0.89	14	0.37
35	Ca$_{ER}$	No	16	0.39	30	0.75	17	0.37	24	0.57	24	0.89	14	0.37
36	Sos$_{cyt}$	No	6	0.15	6	0.15	8	0.17	17	0.40	27	1	31	0.82
37	Grb2$_{cyt}$	No	7	0.17	5	0.13	27	0.59	23	0.55	27	1	31	0.82
38	PLCe$_{cyt}$	No	38	0.93	1	0.03	45	0.98	41	0.98	27	1	33	0.87
39	buffer$_{cyt}$	No	36	0.88	2	0.05	24	0.52	2	0.05	27	1	35	0.92
40	ca_buffer$_{cyt}$	No	36	0.88	8	0.2	21	0.46	9	0.21	27	1	35	0.92
41	SosGrb2$_{cyt}$	No	2	0.05	11	0.28	38	0.83	27	0.64	25	0.93	20	0.53
42	SGS$_{PM}$	No	8	0.20	22	0.55	25	0.54	38	0.90	7	0.26	16	0.42
43	RasGTP$_{PM}$	Yes	13	0.32	14	0.35	23	0.5	4	0.10	1	0.04	1	0.03
44	RasGDP$_{PM}$	Yes	15	0.37	25	0.63	36	0.78	40	0.95	4	0.15	4	0.11
45	RasGRP$_{cyt}$	No	27	0.66	29	0.73	16	0.35	1	0.02	20	0.74	26	0.68
46	CaRasGRP1$_{cyt}$	No	28	0.68	39	0.98	7	0.15	10	0.24	18	0.67	24	0.63

of the networks had the best AUROC when $\varpi_1 = 0.1$. Similar results is observed for the AUPR. *The results suggest that* TFB *ranks play a more important role in target prioritization compared to* DFB *ranks, although inclusion of* DFB *ranks was observed to improve the prioritization results.* In the unseen network (Figure 2, I_0), there is also a downward trend. In particular, the best AUROC and AUPR are 0.849 ($\varpi_1 = 0.3$) and 0.464 ($\varpi_1 = 0$), respectively. Interestingly, the range of ϖ_1 ([0–0.5]) yielding an AUROC greater than or equals to 0.8 correspond to that of I_4, the best matched network. Although the best weight setting differs for different networks, the aforementioned bound on the weight setting (based on AUROC) serves as a good value for an unseen network. Hence, in our experiments we consider two variants of TAPESTRY by setting $\varpi_1 = 0$ and $\varpi_1 = 0.3$.

Comparison with state-of-the-art. We compare TAPESTRY with the state-of-the-art target prioritization approaches. We denote the TAPESTRY variant with ϖ_1 set to x as TAPESTRY$_x$. We compare the performance in terms of AUROC and AUPR, relevance of top-ranked nodes in TAPESTRY that were missed by other approaches and those that were ranked top in other approaches but missed by TAPESTRY. Figure 3 reports the performance results. We first examine the performance in terms of AUROC and AUPR. Observe that both TAPESTRY$_0$ and TAPESTRY$_{0.3}$ outperform all the other approaches in terms of AUROC and TAPESTRY$_0$ is ranked second in terms of AUPR (Figure 3, center).

When we examined the ranks given to known targets (Figure 3, top), TAPESTRY$_{0.3}$ *is the only approach that ranked all known targets in the top-50% ranked nodes.* The actual ranks are given in Table 4. Other approaches such as WBF and WASF miss certain targets (`EGFR_PM`, `Ca` and `RasGDP_PM`) in their top-50% ranked nodes[12]. We further examine if nodes given high ranks in TAPESTRY but low ranks in other approaches were biologically relevant. In particular, we are interested in nodes that differ significantly in ranks predicted

[12]Note that for the random technique, we expect 50% of the known targets to be predicted correctly. Two targets ((`EGFR_PM` and `Ca`) out of five were found in the top-50% ranked nodes and `RasGTP_PM` is actually found in the 50^{th} percentile. This correlates well with the expected result for the random technique.

Table 5: Nodes that are ranked significantly different by different approaches.

No.	Node id	Ranked high in	Ranked low in
1	RasGTP_Golgi$_{GM}$	TAPESTRY, WASF, WBF	LSA, Random, PANI
2	EGF$_{EC}$	TAPESTRY, Random, PANI	WASF, WBF, LSA
3	serca	PANI	TAPESTRY, WASF, WBF, LSA, Random
4	PI$_{PM}$	Random	TAPESTRY, WASF, WBF, LSA, PANI
5	Shc$_{PM}$	TAPESTRY, WASF, LSA, Random	WBF, PANI
6	Shc_star$_{PM}$	TAPESTRY, WASF, WBF	LSA, Random, PANI
7	EGFR$_{PM}$	TAPESTRY, LSA, Random, PANI	WASF, WBF
8	PLC_act$_{PM}$	WASF, WBF, Random, PANI	TAPESTRY, LSA
9	Activated EGFR$_{PM}$	TAPESTRY	LSA, Random, PANI
10	RinhCa	Random	TAPESTRY, WASF, WBF, LSA, PANI
11	DAG_GM$_{GM}$	TAPESTRY, WASF, WBF, LSA, PANI	Random
12	ER$_{erMembrane}$	WASF, Random, PANI	TAPESTRY, WBF, LSA
13	Sos$_{cyt}$	TAPESTRY, LSA, Random, PANI	WASF, WBF
14	Grb2$_{cyt}$	TAPESTRY, LSA, Random, PANI	WASF, WBF
15	PLCe$_{cyt}$	PANI	TAPESTRY, WASF, WBF, LSA, Random
16	buffer$_{cyt}$	LSA, PANI	TAPESTRY, WASF, WBF, Random
17	ca_buffer$_{cyt}$	LSA, PANI	TAPESTRY, WASF, WBF, Random
18	SosGrb2$_{cyt}$	TAPESTRY, WASF, PANI	WBF, LSA, Random
19	SGS$_{PM}$	TAPESTRY, WASF, WBF, Random, PANI	LSA
20	RasGDP$_{PM}$	TAPESTRY, WASF, WBF, Random, PANI	LSA
21	RasGRP$_{cyt}$	LSA	PANI, WASF, WBF, Random, TAPESTRY

by TAPESTRY as compared to other approaches. We define such nodes as those having differences of *normalized ranks* greater or equal to 0.5. Note that we use the normalized ranks instead of the actual ranks for comparison as each approaches gave a different range of node ranks. Hence for a fair comparison, normalization is performed on the ranks for each approach using the formula:

$$\text{normalized rank} = \frac{maxRank - rank}{maxRank} \quad (1)$$

where $maxRank$ is the maximum rank in the range assigned to the nodes by the approach. The nodes that are given significantly different ranks are given in Table 5.

Twelve nodes are ranked high by TAPESTRY and given low ranks in at least one state-of-the-art approaches (shaded rows in Table 4). Among these 12 nodes, 3 correspond to known curated targets. For the remaining, Shc$_{PM}$, Shc_star$_{PM}$, Sos$_{cyt}$, Grb2$_{cyt}$, SosGrb2$_{cyt}$ and SGS$_{PM}$ (complex of Shc_star$_{PM}$ and SosGrb2$_{cyt}$) are nodes implicated in the complex formation of the Sos/Grb2 complex whose translocation to the plasma membrane is found to trigger the activation of Ras [42]. In addition, Topham and Prescott have also discovered the regulation of Ras activation by DAG kinase ζ through a novel mechanism that controls the local accumulation of DAG [41]. EGF is also known to activate the EGF receptor which then triggers a cascade of activity involving other adaptor proteins such as Grb2 and Shc leading to Ras activation [23]. A study by Apolloni and colleagues on H-ras and K-ras reveals that palmoitolylated H-ras, but not K-ras, traffics to the plasma membrane via the Golgi complex [2]. This suggests that interaction with Golgi is crucial for certain form of Ras and RasGTP_Golgi$_{GM}$ is a potential target for activation of Ras at the plasma membrane.

We performed a similar analysis for nodes (9 nodes in all) ranked low by TAPESTRY and given high ranks in at least one state-of-the-art approaches (shaded rows in Table 5). Among the nodes, serca, RinhCa, buffer$_{cyt}$ and ca_buffer$_{cyt}$ are implicated in calcium induced Ras activation [11]. In addition, PLCg was found to activate Ras on the Golgi apparatus via RasGRP1 [6]. Note, however, that most studies find PLC (including PLCe) as effectors of Ras signaling instead of regulators of Ras activation [16, 40, 43]. We did not find evidence supporting the role of the remaining nodes in the activation of Ras. Hence, *nodes prioritized by* TAPESTRY *are biologically relevant and many of these are missed by at least one state-of-the-art approaches.*

Runtime performance. Lastly, we study the runtime performance of TAPESTRY. We observe that TAPESTRY has moderate runtime compared to other approaches (Figure 3).

6. RELATED WORK

Since network-centric target prioritization using data analytics techniques is still in its infancy, there are only few work in the literature such as LSA [15], *NetworkPrioritizer* [19], and PANI [8] that attempt to address this challenging problem. TAPESTRY differs from LSA [15] and *NetworkPrioritizer* [19] in two key ways. First, TAPESTRY considers both topological *and* dynamic features for target prioritization. In contrast, LSA and *NetworkPrioritizer* only considers network dynamics and topology, respectively. Although the dynamic feature (PSSD) we consider only improve the prioritization results modestly, it highlights the fact that network dynamics may be important for target prioritization and should not be ignored. Second, TAPESTRY makes use of insights about characteristics of known targets in signaling networks whereas LSA and *NetworkPrioritizer* do not. Such insights can lead to better results (*e.g.*, better AUROC).

TAPESTRY differs from our previously proposed target prioritization technique called PANI [8] in the following key ways. First, we select a more comprehensive collection of sixteen topological features (instead of two) in order to enhance the quality of target prioritization. Second, TAPESTRY exploits a novel strategy to choose an appropriate prioritization model (obtained using a machine learning-based technique) that is subsequently leveraged to select weight settings for topological features. Third, the weight allocation of topological features are automatically determined in TAPESTRY using the prioritization model unlike PANI which burdens users with the task of weight selection.

7. CONCLUSIONS & FUTURE WORK

In this paper, we present TAPESTRY, a network feature-based target prioritization approach for signaling networks. It is built on top of a target characterization module called TENET. Given a signaling network G with unknown targets and a disease node, it first preprocesses G to filter candidate nodes that are likely regulators of the disease node. Then, it selects the best prioritization model from TENET for G by using a network similarity-based approach, which ranks the set of candidate networks based on their similarity to G with respect to their target features. Next, the selected

prioritization model is used to generate a TFB rank of the nodes in G which is integrated with their DFB rank using a weighted-sum approach to produce a final prioritization rank for each candidate target. Our experimental results show that TAPESTRY can produce superior quality results.

As part of future work, we would like to investigate a set of disease nodes in order to prioritize targets instead of a single disease node. In our study, the impact of the dynamic feature (PSSD) on prioritization results appears to be moderate compared to topological features. This does not necessarily undermine the importance of dynamic features since we study significantly larger number of topological features and the relative importance of the features can be affected by both the choice of the features as well as the number of features studied. Hence, it would be worthwhile to extend this study to include more dynamic features.

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

8. REFERENCES

[1] B. Aldridge, et al. Physicochemical modelling of cell signalling pathways. *Nat. Cell Biol.*, 8(11), 2006.

[2] A. Apolloni, et al. H-Ras but Not K-Ras Traffics to the Plasma Membrane through the Exocytic Pathway. *Mol. Cell. Biol.*, 20(7):2475-2487, 2000.

[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] P. Bermolen, et al. Support vector regression for link load prediction. *Computer Networks*, 53(2):191-201 2009.

[6] T. Bivona, et al. Phospholipase C gamma activates ras on the golgi apparatus by means of RasGRP1. *Nature*, 424(6949):694-698 2003.

[7] Y.-A. Chen, et al. TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, 6(3):e17844, 2011.

[8] H. Chua, et al. PANI:A Novel Algorithm for Fast Discovery of Putative Target Nodes in Signaling Networks. In *ACM BCB*, 2011.

[9] H. Chua, et al. TENET:Topological Feature-based Target Characterization in Signaling Networks. *Bioinformatics*, 31(20):3306-3314, 2015.

[10] H. Chua, et al. TAPESTRY: A Network Similarity Ranking-Based Approach For Prioritizing Nodes In Signaling Networks. `www.cais.ntu.edu.sg/~assourav/TechReports/TAPESTRY-TR.pdf`, Technical Report, 2015.

[11] P. Cullen, et al. Integration of calcium and Ras signalling. *Nat. Rev. Mol. Cell Biol.*, 3(5): 339-348, 2002.

[12] Q. Cui, et al. A map of human cancer signaling. *Mol. Syst. Biology*, 3: 152, 2007.

[13] J. Engelfiet, et al. A comparison of boundary graph grammars and context-free hypergraph grammars. *Inform. Comput.*, 84(2):163-206, 1990.

[14] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Mach. Learn.*, 31:1-38, 2004.

[15] P. Gustafson, et al. Local sensitivity analysis. *Bayesian statistics*, 5:197-210, 1996.

[16] T. Harden, et al. Phospholipase C isozymes as effectors of ras superfamily GTPases. *J. Lipid Res.*, 50(Suppl):S243-S248, 2009.

[17] M. Hatakeyama, et al. A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. *Biochem. J.*, 373(Pt 2):451-463, 2003.

[18] D. Hosmer Jr., et al. *Applied Logistic Regression. Second Edition*, John Wiley & Sons, 2004.

[19] T. Kacprowski, et al. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471-1473, 2013.

[20] E. Keogh, et al. Derivative Dynamic Time Warping. In *SDM*, 2001.

[21] S. Klamt, et al. Hypergraphs and cellular networks. *PLoS Comput. Biol.*, 5(5):e1000385, 2009.

[22] S. Köhler, et al. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, 82(4), 2008.

[23] R. Mauricio, et al. Controlling epidermal growth factor (EGF)-stimulated ras activation in intact cells by a cell-permeable peptide mimicking phosphorylated EGF receptor. *J. Biol. Chem.*, 271:27456-27461, 1996.

[24] N. Le Novère, et al. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res.*, 34:D689-D691, 2006.

[25] NIH. ClinicalTrials.gov. *http://www.clinicaltrials.gov*.

[26] X. Ma, et al. CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data. *Bioinformatics*, 23(2):215-221, 2007.

[27] T. Milenković, et al. Uncovering biological network function via graphlet degree signatures. *Cancer Inform.*, 6, 2008.

[28] S. Morgan, et al. The cost of drug development: a systematic review. *Health Policy*, 100(1), 2011.

[29] NCI drug dictionary. *http://www.cancer.gov/drugdictionary/*, National Cancer Institute, accessed in April 2015.

[30] W. Piegorsch, et al. Combining information. *Wiley Interdiscip Rev Comput Stat*, 1(3):354-360, 2009.

[31] Y. Pritykin, et al. Simple topological features reflect dynamics and modularity in protein interaction networks. *PLoS Comput. Biol.*, 9(10):e1003243, 2013.

[32] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177-e183, 2007.

[33] T. Przytycka, et al. Toward the dynamic interactome: it's about time. *Brief. Bioinform.*, 11(1), 2010.

[34] 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.

[35] K. Reinhardt. Sorting in-place with a worst case complexity of n log n- 1.3 n+ O (log n) comparisons and ε n log n+ O (1) transports. In *Algorithms and Computation*, 1992.

[36] A. Ritz, et al. Signaling hypergraphs. *Trends in biotechnology*, 32(7):356-362, 2014.

[37] S. Sahle, et al. Simulation of biochemical networks using COPASI: a complex pathway simulator. In *WSC*, 2006.

[38] S. Salvadore, et al. FastDTW: Toward accurate dynamic time warping in linear time and space. In *KDD/TDM*, 2004.

[39] B. Schoeberl, et al. Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. *Sci. Signal.*, 2(77):ra31, 2009.

[40] J. Seifert, et al. Dual activation of phospholipase C-epsilon by Rho and Ras GTPases. *J. Biol. Chem.*, 283(44):29690-29698, 2008.

[41] M. Topham, et al. Diacylglycerol kinase zeta regulates Ras activation by a novel mechanism. *J. Cell Biol*, 152(6):1135-1143, 2001.

[42] F. Walker, et al. Activation of the Ras/Mitogen-Activated Protein Kinase Pathway by Kinase-Defective Epidermal Growth Factor Receptors Results in Cell Survival but Not Proliferation. *Mol. Cell. Biol.*, 18(12):7192-7204, 1998.

[43] M. Wing, et al. PLC-epsilon: a shared effector protein in Ras-, Rho-, and G alpha beta gamma-mediated signaling. *Mol. Interv.*, 3(5):273-280, 2003.

[44] H.-Y. Yuan, et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.*, 34(suppl 2):W635-W641, 2006.