TENET: A Machine Learning-based System for Target Characterization in Signaling Networks

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

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

C. Forbes Dewey, Jr.
Biological Engineering Department
Massachusetts Institute of Technology
Cambridge, MA, USA
Email: cfdewey@mit.edu

***Abstract*—*Target characterization* of a biological network identifies characteristics that distinguish *targets* (nodes that can serve as molecular targets of drugs) from other nodes. In this demonstration, we present TENET (Target charactErization using NEtwork Topology), a software that facilitates topological features-based characterization of known targets in signaling networks modelling dynamic interactions within biological systems. TENET is based on a support vector machine (SVM)-based approach and generates a *characterization model*. These models specify topological features that can discriminate known targets and how these features are combined to quantify the likelihood of a node being a target. Hence, TENET can be used for prioritizing targets and for identifying novel candidate targets that share similar characteristics with known targets. The interactive user interface that TENET provides facilitates users' study and understanding of topological characteristics of targets in signaling networks.**

***Index Terms*—Signaling network, target characterization, support vector machine, topological features, visualization.**

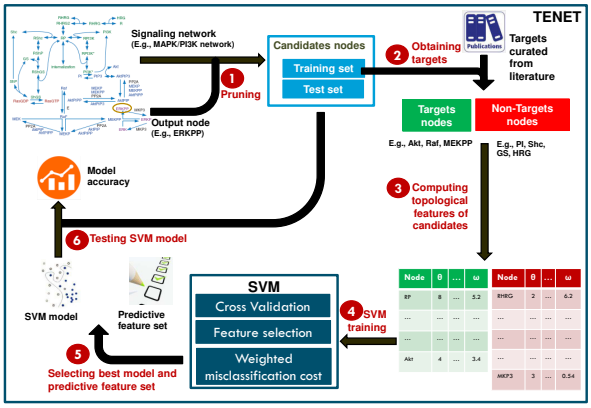

Fig. 1. TENET consists of 6 steps: (1) pruning irrelevant nodes, (2) obtaining curated targets, (3) computing topological features of nodes, (4) training the SVM, (5) selecting the best model and predictive feature set, and (6) testing the SVM model.

I. INTRODUCTION

Cells use sophisticated communication between proteins in order to perform a variety of cellular functions such as growth, survival, proliferation and development. As signaling proteins rarely operate in isolation through linear pathways, cell signaling can be viewed as a large and complex network. Specifically, the network view emerges due to 'cross-talks' between signaling pathways. Understanding signal flow in the network is paramount as alterations of cellular signaling events, such as those that arise by gene mutations or epigenetic changes, can result in various diseases such as cancer. Consequently, in recent times various computational techniques have been developed to analyze signaling networks in order to gain insights in a variety of biological problems such as *target characterization* and drug target discovery. In this paper, we demonstrate a machine learning-based system to *characterize targets* in a disease-related signaling network.

Target characterization identifies characteristics (*e.g.*, topological features) that distinguishes *targets* (*i.e.,* nodes) from other nodes in a biological network. We refer to a node as a *candidate target* if when perturbed, it modulates the activity of an *output node*, which is a molecule that is of interest due to its physiological role in a disease [4]. The characteristics of targets in a biological network can be summarized as models which we refer to as *characterization models*. The knowledge of target characteristics is useful in drug design pertaining to these targets and in the identification of novel targets that share similar characteristics with known targets.

Traditionally, targets are characterized based on their molecular characteristics (*e.g.*, structure and binding sites of target) and biological functions (*e.g.*, regulation of apoptosis). These traditional approaches focus primarily on targets alone and are oblivious to the presence of other interacting molecules in the system. However, understanding how a target interacts with other molecules in a biological system may provide valuable and holistic insights for superior target characterization. For example, degree centrality of targets may be leveraged to assess potential toxicity of targets since high degree nodes tend to be involved in essential protein-protein interactions and are potentially toxic as a result. Hence, *network-based* target characterization techniques can exploit such topological features for superior characterization of targets.

State-of-the-art network-centric approaches for target characterization typically focus on PPI networks [6], [9], [11] instead of signaling networks. However, PPI networks may contain many false-positive interactions in the sense that although these proteins can truly physically bind they may

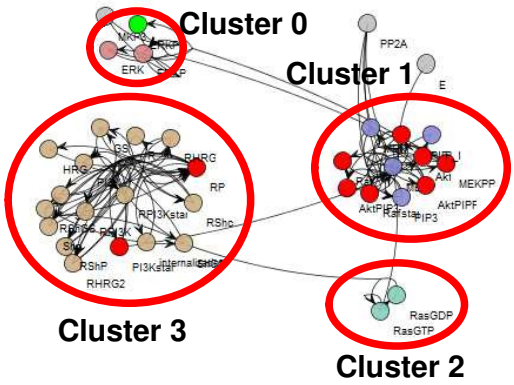

Fig. 2. [Best viewed in colour] Cluster view showing that majority of known targets are in Cluster 1. Target nodes are in red.

never do so inside cells due to different localization or they are not simultaneously expressed. Furthermore, unlike signaling networks, the edges in PPI networks are undirected and lack of knowledge of the underlying mechanism (*i.e.*, actual signal flow) causing the disease, which may adversely impact the search for superior characteristics of targets. Hence, in this demonstration, we present TENET (**T**arget charact**E**rization using **NE**twork **T**opology) [4], a target characterization system that automatically *characterizes* known targets in a signaling network based on topological features. Given a disease-related signaling network $G = (V, E)$, an output node $x \in V$, a set of known targets $T_x \subseteq V$ and a set of topological features, TENET first performs pruning to remove nodes that do not have paths leading to x, hence, reducing irrelevant computation in subsequent steps. Then, it performs feature extraction and utilizes support vector machine (SVM) to train the models. The result is an SVM model and a set of *predictive feature* set characterizing T_x. Fig. 1 illustrates the TENET framework. A practical usage of TENET is *target prioritization* [4].

In order to facilitate understanding of the target characterization results in the context of the signaling network, TENET provides several user-friendly visualization features. First, TENET superimposes known targets onto the graphical representation of the network for users to visualize the location of the targets within the network. For instance, the cluster view (Fig. 2) which displays node clusters by colours, allows users to quickly identify important node clusters in the network (*e.g.*, those containing majority of known targets). Second, the interface design (Fig. 3) is user-friendly but yet configurable to the needs of different users. For instance, novice users can use the default settings to perform target characterization whereas expert users can modify the options provided in the panel (*e.g.*, list of known targets, list of topological features). Third, TENET consolidates target-related information in a single platform. The *Target Information* panel lists the drugs hitting the target. Hyperlinks relating to these drugs, literature evidence associating the drugs to the target and related clinical trials give users access to detailed information easily.

The rest of the paper is organized as follows. We discuss related work in Section II. Sections III and IV present the system overview and demonstration objectives, respectively.

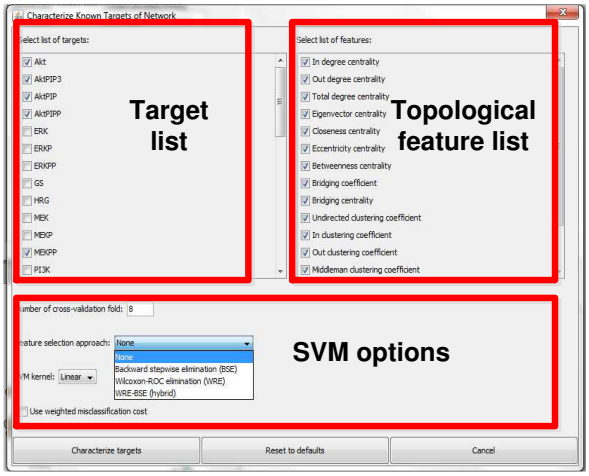

Fig. 3. Target characterization configuration setting.

II. Related Work and Novelty

McDermott *et al.* characterized targets in PPI networks using several topological features such as degree centrality [9]. Their study suggested that multiple features can be combined to improve target characterization. However, it stopped short of examining how these features should be combined. In contrast, Hwang *et al.* proposed a single topological features (*bridging centrality*) for identifying targets in PPI networks [6]. However, given the fact that biological networks are complex and diverse, the chosen feature may not characterize targets effectively [4]. Zhang *et al.* characterized known targets of a manually curated human PPI network using several machine learning techniques including SVM [11]. In particular, their study focused on identifying topological characteristics of targets in general, instead of for specific diseases. Hence, there is an implicit assumption that targets of different diseases share similar target characteristics, which is not necessarily true [4].

We take the first step in [3] to show that signaling networks can be a viable alternative to PPI networks for characterizing targets. However, this work suffers from the same limitation as [9]. TENET extends [3] to address the aforementioned limitations by performing target characterization based on individual disease-specific networks using multiple topological features. The characterization model generated by the SVM-based approach identifies important topological features characterizing the targets and provides a basis for combining these features. Moreover, it provides interactive features such as cluster mode visualization to assist users in analyzing the network. In our seminal work [4] on TENET, we detailed the algorithm for generating characterization models and how it can be effectively used to prioritize nodes in a signaling network. In fact, it outperforms state-of-the-art approaches such as *NetworkPrioritizer* [7]. This paper makes the following additional contributions: (1) we describe the TENET system architecture (Section III) in detail and (2) we demonstrate the interactive features that TENET provide for target characterization and visualization.

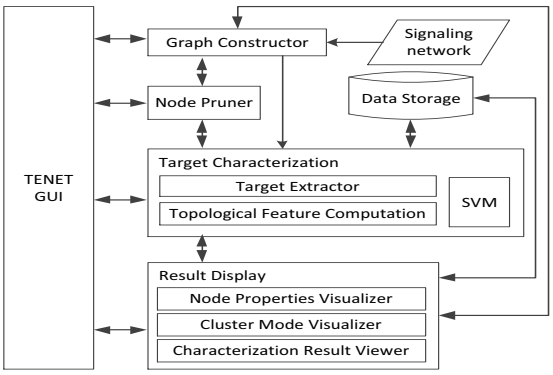

Fig. 4. Architecture of TENET.

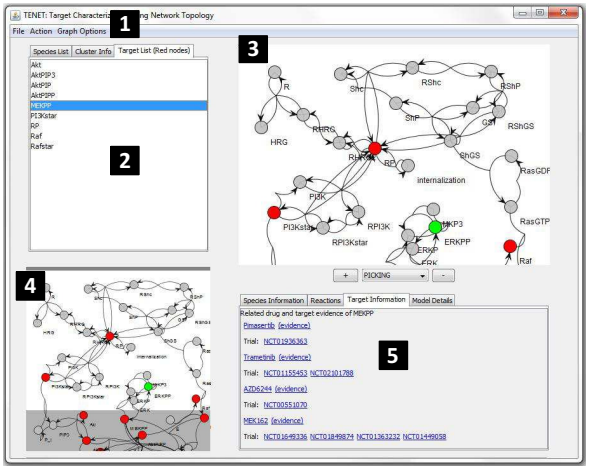

Fig. 5. [Best viewed in color] Graphical user interface of TENET.

III. SYSTEM OVERVIEW

TENET is implemented in Java and uses the *PostgreSQL* database for storing topological features, curated targets and characterization models. Fig. 4 shows the architecture of TENET which consists of the following modules.

The TENET GUI Module. The graphical user interface (GUI) of TENET (Fig. 5) coordinates the various modules of TENET and consists of five components. We shall walkthrough the steps for performing target characterization to illustrate the role of each component of the GUI. First, a user selects a signaling network for analysis using the toolbar (Fig. 5, Component 1). The *Graph Constructor* constructs the graphical representation (referred to as *model graph*) of the signaling network using the Java JUNG library. Note that curated targets are highlighted in a prominent colour to distinguish them from the remaining nodes. This facilitates the user in performing visual analysis of the targets such as finding the relative position of these targets within the network compared to specific molecule(s). The *Result Visualizer* then displays the network nodes in the *Node Panel* (Fig. 5, Component 2) and the model graph in the *Satellite Graph View* (Fig. 5, Component 3) and *Interactive Graph View Panels* (Fig. 5, Component 4). The *Satellite Graph View Panel* displays the overall layout of the graph whereas the *Interactive Graph View Panel* displays the active portion of the graph (*e.g.*, zoomed in view). Next, the user selects the output node from the *Node Panel* and initiates the target characterization process using the toolbar. Details of the characterization model are displayed in the *Information Panel* (Fig. 5, Component 5). Note that the *Information Panel* also displays node properties such as annotations of the selected node.

The Graph Constructor Module. Signaling networks can be modelled as directed hypergraphs [8]. The nodes represent molecules whereas the hyperedges represent biochemical reactions and processes. The *Graph Constructor* creates different graphical representations of the user-selected signaling network. In particular, it generates three types of graphical representations, namely, hypergraph, directed acyclic graph (DAG) and bipartite graph representations. The hypergraph representation models the input signaling network and is used for display; The DAG representation is used to analyze reachability information of node pairs in the *Node Pruner* to remove *redundant*[1] nodes from further computation; The bipartite graph representation as discussed in [8] is a simpler but equivalent structural representation of the hypergraph and is used for computing the topological features by the *Target Characterization* module.

The Node Pruner Module. The *Node Pruner* determines whether a node is redundant by adopting a *reachability rule*-based approach [2]. The *reachability rule* considers a node as redundant if there exists no path from the node to the output node. Briefly, this module indexes the DAG representation of the signaling network using depth-first traversal. Then, the indices are compared and redundant nodes are removed from further computation.

The Target Characterization Module. TENET stores a list of known targets (*Data Storage*) curated from literature and clinical trials repositories (*e.g.,* [1]) that are classified based on organism and disease. The user has to specify the relevant organism and disease of the given signaling network (via TENET GUI) in order to perform target characterization. This restriction ensures that the *Target Extractor* module retrieves appropriate targets relevant to the network for characterization from the *Data Storage*. TENET provides flexibility to users by allowing configuration of certain settings for target characterization (Fig. 3), such as the number of cross-validation folds and heuristics for the SVM training. As described in [4], this module first extracts a set of topological features (some are shown in Fig. 3) of the candidate targets (*i.e.,* nodes that are not redundant) 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

[1]In TENET, we are interested in finding the defining characteristics of known targets relevant to a particular disease. The output node of the network is associated to the manifestation of the disease. Hence, nodes that do not modulate the output node are considered redundant.

characterization model based on these features. 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. The SVM module provides three feature selection heuristics, namely, backward stepwise elimination (BSE), Wilcoxon-ROC based elimination (WRE) and WRE-BSE. These feature selection approaches help to remove redundant features. In contrast, the *weighted misclassification cost* (WMC) heuristic addresses the issue of noisy labels and imbalanced data set. WMC proportionates the misclassification cost of the training data according to class such that misclassification of known targets incur greater cost.

The results of target characterization is a characterization model and its details are presented to the user in the *Information Panel* of the GUI (Fig. 5, Component 5). Users can modify the configuration settings to study the effects of selecting different heuristics, structural feature sets and curated target set on the prediction accuracies.

The Result Display Module. The *Result Display* handles the visualization aspect of TENET and consists of the following submodules. First, the *Cluster Mode Visualizer* handles the modular layout display of the model graph. It assesses the cluster information from the *Graph Constructor* and assigns nodes in the same cluster to the same color (Fig. 2). Note that the user can select the cluster view via a viewing option provided in the toolbar (TENET GUI). Second, the *Node Properties Visualizer* displays node properties such as node annotations. The visualizer also handles target-related information and provides hyperlinks to interface with external databases (*e.g.*, DrugBank [10]). Finally, the *Characterization Result Viewer* displays the results of target characterization in the *Information Panel* (Fig. 5, Component 5). Briefly, the results summarize the configuration setting used to perform target characterization, list the optimal parameters and the list of predictive topological features used for constructing the final SVM model.

IV. Demonstration Objectives

The TENET prototype is ready for demonstration and is available freely for non-commercial purposes from https://sites.google.com/site/cosbyntu/softwares/tenet. We shall pre-install TENET on our laptop and upload data files of several signaling networks (*e.g.*, MAPK−PI3K network [5]) from the *BioModels* database (https://www.ebi.ac.uk/biomodels-main/). We shall use these datasets to demonstrate the process of target characterization and the use of the TENET GUI to perform analysis of known targets and the network.

Interactive Characterization of Targets. The main objective of the demonstration is for users to experience the process of performing target characterization. Users can select from a variety of signaling networks with preloaded data files (including the list of known targets) for characterization. The GUI design provides configurable options that allow users to perform different characterization. Hence, the TENET framework supports comparative study of various selection heuristics (Section III, *Target Characterization Module*) used in target characterization. During the demonstration, the audience will be able to experience the time taken by TENET to format the display of the signaling networks and to perform target characterization of networks of different sizes. They can also load their own models as long as they have a list of targets relevant to those models. We shall explain the steps required to include new signaling networks for characterization using TENET during the demonstration.

Interactive Study of Network and Known Targets. We shall demonstrate the use of the cluster view option (Fig. 2) to study the distribution of known targets within the network and to locate important clusters (those containing many known targets). This feature is useful for applications such as drug discovery since nodes in important clusters that do not correspond to known targets may be potential drug candidates. Users can also find out about drugs and clinical trials associated to known targets using hyperlinks to external online databases provided at the *Information Panel*.

V. Conclusions

TENET [4] facilitates target characterization in signaling network using machine learning and enables users to answer questions pertaining to targets such as what are the defining topological features predictive of targets and how likely it is for a node to be a target relative to other nodes. The use of a relational database (PostgreSQL) as the backend enables TENET to scale up to larger signaling networks. In summary, this demonstration will showcase various features of TENET, the world's first target characterization system for signaling networks.

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

References

[1] NIH. ClinicalTrials.gov. *http://www.clinicaltrials.gov*.
[2] Chen, L. *et al.* (2005). Stack-based algorithms for pattern matching on DAGs. In *VLDB*, 2005.
[3] Chua, H. *et al.* (2014). One feature doesn't fit all: characterizing topological features in signaling networks. In *BCB*, 2014.
[4] Chua, H. *et al.* (2015). TENET: topological feature-based target characterization in signaling networks. *Bioinformatics*, **31**(20), 3306-3314.
[5] 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.
[6] Hwang, W.-C. *et al.* (2008). Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharma. Ther.*, **84**(5), 563-572.
[7] Kacprowski, T. *et al.* (2013). NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, **29**(11), 1471-1473.
[8] Klamt, S. et al. Hypergraphs and cellular networks. *PLoS Comput Biol*, 2009.
[9] McDermott, J. *et al.* (2012). Topological analysis of protein co-abundance networks identifies novel host targets important for hcv infection and pathogenesis. *BMC Syst. Biol.*, **6**(1), 28.
[10] Wishart, D.S. *et al.* (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36:D901-D906.
[11] Zhang, J. *et al.* (2010). Novel biological network features discovery for in silico identification of drug targets. In *IHI*, 144-152.