Bioinformatics, 31(20), 2015, 3306–3314
doi: 10.1093/bioinformatics/btv360
Advance Access Publication Date: 16 June 2015
Original Paper

Systems biology

TENET: topological feature-based target characterization in signalling networks

Huey Eng Chua[1,]*, Sourav S. Bhowmick[1,]*, Lisa Tucker-Kellogg[2] and C. Forbes Dewey Jr[3]

[1]School of Computer Engineering, Nanyang Technological University, [2]Duke-NUS Graduate Medical School, National University of Singapore, Singapore and [3]Biological Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 27, 2014; revised on May 13, 2015; accepted on June 5, 2015

Abstract

Motivation: Target characterization for a biochemical network is a heuristic evaluation process that produces a characterization model that may aid in predicting the suitability of each molecule for drug targeting. These approaches are typically used in drug research to identify novel potential targets using insights from known targets. Traditional approaches that characterize targets based on their molecular characteristics and biological function require extensive experimental study of each protein and are infeasible for evaluating larger networks with poorly understood proteins. Moreover, they fail to exploit network connectivity information which is now available from systems biology methods. Adopting a network-based approach by characterizing targets using network features provides greater insights that complement these traditional techniques. To this end, we present TENET (Target charactErization using NEtwork Topology), a network-based approach that characterizes known targets in signalling networks using topological features.

Results: TENET first computes a set of topological features and then leverages a support vector machine-based approach to identify predictive topological features that characterizes known targets. A characterization model is generated and it specifies which topological features are important for discriminating the targets and how these features should be combined to quantify the likelihood of a node being a target. We empirically study the performance of TENET from a wide variety of aspects, using several signalling networks from BioModels with real-world curated outcomes. Results demonstrate its effectiveness and superiority in comparison to state-of-the-art approaches.

Availability and implementation: Our software is available freely for non-commercial purposes from: https://sites.google.com/site/cosbyntu/softwares/tenet

Contact: hechua@ntu.edu.sg or assourav@ntu.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Complex intra- and inter-cellular signalling drives various biological processes, such as growth, proliferation and apoptosis within systems. In systems biology, these molecular interactions are typically modelled as signalling networks (Klamt *et al.*, 2009) that provide a holistic view of the various interactions between different molecular players in the system. As signalling networks become an increasingly acceptable way for representing biological systems, various *network-based* computational techniques have been developed to analyze these networks with the goal of answering biological needs, such as target characterization (Chua *et al.*, 2014) and target discovery (Yang *et al.*, 2008). In this article, we focus on the target characterization problem for signalling networks.

Target characterization identifies characteristics (e.g. topological features) that distinguishes *targets* (i.e. nodes) from other nodes in the network. These characteristics can be summarized as models which we refer to as *characterization models*. Traditionally, targets are characterized based on their molecular characteristics [e.g. structure and binding sites of targets (Maira *et al.*, 2008)] and biological function [e.g. regulation of apoptosis (Yan *et al.*, 2013)]. These traditional approaches focus primarily on the target 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, the degree centrality of a target may be leveraged to assess potential toxicity of targets as high degree nodes tend to be involved in essential protein–protein interactions (He *et al.*, 2006) and are potentially toxic as a result. In particular, *network-based* target characterization techniques can exploit such topological features for superior characterization of targets.

Recently, there have been increasing efforts toward devising network-based target characterization techniques (Hwang *et al.*, 2008; Zhang *et al.*, 2010; McDermott *et al.*, 2012). These methods focus on using topological features to characterize targets of protein–protein interaction (PPI) networks. Specifically, McDermott *et al.* (2012) performed characterization of targets in *protein co-abundance networks* [The *protein co-abundance networks* are essentially protein–protein interaction (PPI) networks constructed by identifying highly differentially regulated proteins from proteomics data using specific filters.] using several topological features such as degree centrality. Although this study suggests that multiple topological features can be combined for superior target characterization, it did not explore how these topological features should be combined towards this goal. In contrast, Hwang *et al.* (2008) concluded that *bridging centrality* is useful in identifying targets in PPI networks. However, the complexity and diversity of biological networks make target characterization using a single feature challenging as in some networks the chosen feature may perform poorly. Indeed, Chua *et al.* (2014) showed that bridging centrality performs well in the MAPK-PI3K network, but not in the glucose metabolism network. Zhang *et al.* (2010) proposed the use of machine learning techniques such as support vector machines (SVM) and logistic regression for characterizing known targets in a manually curated human PPI network using 15 topological features. In contrast to McDermott *et al.* (2012), their goal was to identify topological characteristics of drug targets in general, instead of for specific diseases. However, characterizing targets in general assumes that targets of different diseases share similar target characteristics, which may not always be true. Indeed, as we shall see in Section 3, known targets in signalling networks tend to be characterized by different sets of topological features. Consequently, target characterization based on individual disease-specific networks may yield better characterization that is specific to the disease.

A common thread running through the aforementioned target characterization techniques is their focus on PPI networks. Surprisingly, similar systematic study in curated signalling networks has been lacking in the literature. Compared to signalling networks, PPI networks may contain many false-positive PPI in the sense that although these proteins can truly physically bind they may never do so inside cells due to different localization or they are not simultaneously expressed. Furthermore, PPI networks are static. That is, the edges in PPI networks are undirected; there is neither flow of information nor mass between nodes. Hence, they lack of knowledge of the underlying mechanism (i.e. actual signal flow) causing the disease. As network quality directly affects the results of network-based target characterization, the aforementioned limitations of PPI networks may adversely impact the search for superior characteristics of targets. Signalling networks, however, model the dynamic interaction of the biological systems and present an attractive alternative to PPI networks.

In our recent work (Chua *et al.*, 2014), we took the first step to demonstrate how signalling networks can be effectively leveraged to identify topological features that are *discriminative* of targets using the Wilcoxon test. However, similar to McDermott *et al.* (2012), this work does not shed any insight on a *predictive model* to combine these features for identifying potential targets. In this article, we address this limitation by presenting TENET (Target charactErization using NEtwork Topology), a network-based approach that characterizes known targets in signalling networks using topological features. Specifically, we use an SVM-based approach to identify the set of topological features (referred to as *predictive topological features*) that characterizes known targets and to generate a *characterization model* using these features. The model specifies which topological features are important for discriminating the targets and how these features should be combined to produce a quantitative *score* that identifies the likelihood of a node being a target. In particular, TENET uses *feature selection* to select *predictive topological features* and *weighted misclassification cost* (WMC) to handle SVM training issues such as noisy labels and imbalanced data. Our empirical study on four real-world curated signalling networks demonstrates the effectiveness and superiority of TENET.

2 Materials and methods

2.1 Terminology

A biological signalling network can be modelled as a directed hypergraph $G = (V, E)$ (Klamt *et al.*, 2009) where the nodes V represent molecules (e.g. proteins) and the *hyperedges* E represent biochemical reactions and processes. A hyperedge connects one node set U to another W, where $U, W \subseteq V$. For instance, in the activation of ERK, the set U in the hyperedge consists of ERK and its kinase, phosphorylated MEK whereas W contains the phosphorylated ERK (ERKPP). Analysis of directed hypergraphs is generally more complex than graphs and many graph algorithms cannot be used directly on hypergraphs. Hence, they are often transformed into graphs containing simple edges for analysis. Methods (e.g. bipartite and substrate graph representation) exist for such transformation (Klamt *et al.*, 2009). In this article, we use the bipartite graph representation as it retains the original information of the hypergraph (Klamt *et al.*, 2009). Signalling networks generally contain characteristics such as feedback and feedforward loops, which are common in complex regulatory control (Kwon *et al.*, 2008). These loops in turn give rise to graph characteristics, such as strongly connected components (SCC).

The activity of nodes in the signalling network is generally governed by complex interconnectivity of various nodes in the same network. We refer to a node as a *candidate target* if when perturbed, it modulates the activity of a specific node (referred to as *output node*). An *output 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. For instance, in the MAPK-PI3K network (Hatakeyama *et al.*,

2003) that is often implicated in cancer, ERKPP can be considered as an output node due to its role in proliferation. Given a signalling network $G = (V, E)$ and an output 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 $T_x \subseteq V_x$.

Network-based analysis can be applied to signalling networks to study the characteristics and properties of these networks. In this article, we examine a total of 16 topological features that are summarized in Table 1. These features are selected based on their role in measuring relative importance of a node in a signalling network. The formal definitions as well as motivation for selecting these features are given in Chua *et al.* (2014) (also detailed in Supplementary Material S1.1).

2.2 Topological feature-based target characterization

Intuitively, the goal of topological feature-based target characterization is to use a set of *predictive topological features* to characterize known targets in a network. Hence, the *topological feature-based target characterization problem* can be formulated as a supervised learning problem. In a supervised learning problem, a training set $\{\langle x_i, f(x_i)\rangle\}$ is given where $f(x_i)$ is the predictor of x_i and the goal is to learn some target function $f : X \rightarrow Y$ which can be applied to predict unseen data w. The problem can be subdivided into two categories: regression when the predictor yields a continuous outcome and classification when the outcome is discrete. A regression problem can be converted into a binary classification problem by specifying a threshold h and assigning x_i with f greater than h to one class and the remaining to the other class. We advocate that the *topological feature-based target characterization problem* is best represented as a regression problem. In this problem, we are interested in finding out how likely one node is a target relative to another node based on a set of predictive topological features. This is different from the target classification problem where we want to find out the class membership of a node. Note that the regression problem can be converted into a classification problem by specifying a threshold h and assigning nodes having target function greater than h to the target class and the rest to the non-target class.

Although we examine 16 topological features, as we shall see later, not all features are relevant to a given signalling network. In fact, incorporating irrelevant features may adversely impact the performance of the prediction model. Hence, it is important to learn a set of predictive topological features that best characterizes targets (referred to as *topological feature selection*) for a given network. Formally, it is defined as follows.

Definition 1: *Given a signalling network $G = (V, E)$ and an output node $x \in V$*, let *$T_x \subseteq V$ and $\mathcal{X}_{\text{all}}$ denote the set of known targets in G relevant to x, and the set of topological features of G, respectively. Then, the goal of 'topological feature selection' is to find a set of 'predictive topological features' $\mathcal{F} \subseteq \mathcal{X}_{\text{all}}$ that maximizes the prediction accuracy for $f(\xi(u, \mathcal{F}))$ subject to the following conditions:*

$$\begin{cases} f(\xi(u, \mathcal{F})) = 1 & \text{when } u \in T_x, \\ f(\xi(u, \mathcal{F})) = 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then the *topological feature-based target characterization problem* is formally defined as follows.

Definition 2: *Given a signalling network $G = (V, E)$, an output node $x \in V$, T_x, and $\mathcal{X}_{\text{all}}$, let $\mathcal{F}$ denote the set of predictive topological features. Then, for a threshold h, the goal of the 'topological feature-based target characterization problem' is to identify a set of predictive topological features $\mathcal{F} \subseteq \mathcal{X}_{\text{all}}$ using topological feature selection and learn a 'characterization model' $g(\xi(u, \mathcal{F}))$ subject to the conditions*

$$\begin{cases} g(\xi(u, \mathcal{F})) \in \Re, \\ g(\xi(u, \mathcal{F})) \geq h & \text{when } u \in T_x, \\ g(\xi(u, \mathcal{F})) < h & \text{otherwise,} \end{cases} \tag{2}$$

that maximizes the target prediction for $g(\xi(u, \mathcal{F}))$.

Table 1. Topological features

Symbol	Description
θ_u	Degree centrality of node u. The in, out and total degree centralities are denoted as $\theta_{\text{in}(u)}$, $\theta_{\text{out}(u)}$ and $\theta_{total(u)}$, respectively
α_u	Eigenvector centrality of node u
β_u	Closeness centrality of node u
γ_u	Eccentricity centrality of node u
δ_u	Betweenness centrality of node u
π_u	Bridging coefficient of node u
ζ_u	Bridging centrality of node u
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
μ_u	Proximity prestige of node u
ω_u	Target downstream effect of node u

Figure 1 depicts a pictorial overview of the topological feature-based target characterization problem. For example, given the MAPK-PI3K signalling network, its associated output node ERKPP, the set of known targets in this network and the topological features in Table 1, the goal of this problem is to produce the followings: (i) identify the set of predictive topological features $\mathcal{F} = \{\delta, \pi, \theta_{in}, \theta_{out}\}$ and (ii) learn a characterization model $g(\xi(\text{ERKPP}, \mathcal{F}))$. Note that in Definition 2, there is no need to explicitly specify a threshold h if we are only interested in obtaining the relative rankings of the nodes. The threshold is required if we want to assign class labels (e.g. target class) to the nodes.

2.3 SVM-based target characterization

We employ *support vector classification* (SVC) to select predictive topological features and *support vector regression* (SVR) to generate the characterization model. The SVC and SVR are typically formulated as constrained optimization problems and solved using the *Lagrangian multiplier method*. In general, SVM models contain multiple parameters, such as the cost parameter C and parameters related to the kernel function, that affect the learning and performance of the models (Chapelle *et al.*, 2002). We follow the method in Hsu *et al.* (2003) for training the SVM. The feature values are scaled linearly to the range of [0, 1] for each signalling network to avoid features with larger ranges dominating those with smaller ranges. We use stratified (The training data were sampled from the original data such that the ratio of the targets to non-targets is similar to that of the original data.) cross-validation (Supplementary Material S1.3) and grid-search (Hsu *et al.*, 2003) on the training data to identify the best values of the model parameters. Note that cross-validation helps us to avoid the issue of overfitting the data whereas stratification enables us to keep the percentage of targets in the different folds similar to the original dataset. The best parameter is the one that yields the best average prediction accuracy for the cross-validation process. Wherever possible (In our study, we set a lower bound of

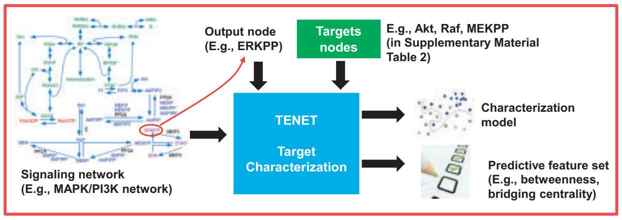

Fig. 1. Target characterization problem

one target in all our test sets.), we use a 10-fold stratified cross-validation as larger fold numbers reduce pessimistic bias and 10-folds generally give good performances (Kohavi *et al.*, 1995).

Several non-trivial issues, namely, irrelevant or redundant features, noisy labels and imbalanced dataset, need to be addressed in training the SVM model for characterizing targets. In particular, we use feature selection to select for appropriate features to be used in the SVM model and cost-sensitive learning to handle the issue of noisy labels and imbalanced dataset. We examine three feature selection approaches, namely, backward stepwise elimination (BSE) (Marill *et al.*, 1963), Wilcoxon-ROC based elimination (WRE) and WRE-BSE. BSE is *classifier-aware* whereas WRE is *classifier-independent*. WRE-BSE which performs WRE followed by BSE is a hybrid approach. Note that compared to classifier-independent methods, classifier-aware methods interact with the classifiers and such interaction can lead to better classification results (Saeys *et al.*, 2007). However, they are typically computationally expensive and run the risk of model-overfitting. Cost sensitive learning is an algorithmic approach that chooses an appropriate strategy specific to the classifier to overcome the bias introduced by imbalanced data and the noise caused by uncertainty in labelling. We use WMC, an approach that proportionates the misclassification cost of the training data according to class. In particular, we use a variable C_i as the cost parameter C:

$$C_i = \begin{cases} C^+ & \text{if } y_i = +1 \\ C^- & \text{if } y_i = -1 \end{cases} \tag{3}$$

subject to the constraints $C^+ + C^- = 1$, $C^+ > 0$ and $C^- > 0$ where y_i is the class predictor and C^+ and C^- denote the misclassification cost of the target and non-target classes, respectively.

2.4 The Tenet algorithm

Given a signalling network $G = (V, E)$, an output node $x \in V$, a known target set $T_x \subseteq V$, a set of topological features $\mathcal{X}_{\text{all}}$ and a step-size of the misclassification cost s, Tenet identifies the set of predictive structural features and a characterization model that best characterizes these known targets. Note that $\mathcal{X}_{\text{all}}$ and s are optional inputs and are set to default values ($\mathcal{X}_{\text{all}}$ is set to the 16 topological features given in Table 1 whereas s is set to 0.1.) if they are not given. The known targets T_x can be extracted by following the curation process described in Chua *et al.* (2014) (Supplementary Material S1.2). The Tenet algorithm comprised three phases, namely, the *pruning phase*, the *feature extraction phase* and the *model training phase*. First, the *pruning phase* identifies relevant nodes (denoted as $V_{\text{candidate}}$) that shall be used for training the SVM. Then, the *feature extraction phase* extracts all the topological features (denoted as $\mathcal{X}_{\text{all}}$) of each candidate node and stores them in a $|V_{\text{candidate}}| \times |\mathcal{X}_{\text{all}}|$ matrix H. Finally, in the *model training phase*, Tenet learns the optimal set of predictive topological features $\mathcal{F}$ and the best model parameters of the characterization model $\mathcal{M}$. We shall now describe these phases in turn. The formal algorithm is given in Supplementary Material S1.4.

2.4.1 Phase 1: Pruning

In this phase, Tenet prunes nodes that do not have paths leading to the output node x. This phase yields a set of potential candidate nodes $V_{\text{candidate}} \subset V$ and is used to reduce the subsequent computation. In the pruning process, the given network G is first preprocessed into a bipartite graph and then converted into a *directed acyclic graph* (DAG), a graph with consistent topological ordering, to facilitate indexing of nodes. Note that the node indices shall be used subsequently to perform reachability check to identify the nodes to be pruned. We adopt the method in Engelfiet *et al.* (1990) for bipartite graph conversion. In order to convert the bipartite graph into its DAG representation, we adopt the approach in Tarjan *et al.* (1972) to identify SCCs and replace each SCC with a representative node (referred to as *meta node*). Then, we adopt the indexing approach of Chen *et al.* (2005) to index the DAG. This indexing approach performs depth-first traversal to assign each node v a *preorder index* (when v is first visited) and a *postorder index* (when all descendent nodes of v are visited). Finally, an index-based reachability algorithm is used to determine whether there exists a path from each node v to the output node x (denoted as $v \rightarrow x$). Given a node v and x, let w be the descendent of v that is not in the *spanning tree* (referred to as *non-spanning tree node*) and *v.preorder* and *v.postorder* denote the preorder and postorder indexes of v, respectively. A path $v \rightarrow x$ exists if any of the following conditions are satisfied (Chen *et al.*, 2005):

1. $v.preorder \leq x.preorder$ and $v.postorder \geq x.postorder$.
2. $w.preorder \leq x.preorder$ and $w.postorder \geq x.postorder$.

Note that the pruning step is beneficial in improving execution time for larger sparsely connected networks and for output node that are positioned further upstream. For instance, in the MAPK-PI3K network, no nodes are pruned when we select ERKPP (downstream) as the output node whereas 17 nodes (47.2%) are pruned when activated Ras (RasGTP) (upstream) is selected.

2.4.2 Phase 2: Feature extraction

In this phase, for all nodes in $V_{candidate}$, Tenet extracts all the topological features in Table 1 for characterizing the known targets.

2.4.3 Phase 3: Model training

Given a matrix of topological feature values H, a target set T_x and a step-size of the misclassification cost s, this phase identifies a set of predictive topological features $\mathcal{F}$ and the best parameters for configuring the characterization model $\mathcal{M}$. First, the misclassification cost of the target class C^+ is initialized to a default value of 0.5. Then, feature selection is used to obtain the predictive topological feature set $\mathcal{F}$. We iterate over three different feature selection approaches (BSE, WRE and WRE-BSE). Next, the step-size s is used to step through the range of misclassification cost (0–1). In each iteration, the misclassification cost of the target class C^+ is incremented according to the number of iterations completed, before the SVM training is performed to obtain the parameter settings of the characterization model $\mathcal{M}$ with the best accuracy.

The BSE approach is a well-known greedy approach that progressively removes features from the naïve SVM model (built using all topological features) and trains a new best model after each feature removal. The elimination process stops when removal of additional features results in a worse average accuracy of the validation set prediction. In contrast, the WRE approach performs two statistical tests, namely, one-tailed Wilcoxon Rank-Sum (referred to as Wilcoxon) and receiver operating characteristics (referred to as ROC). The results are used to eliminate features that do not discriminate between targets and non-targets in a significant manner (based on Wilcoxon) and that do not classify targets well (based on ROC). Note that we perform two 1-tailed Wilcoxon tests and for each test; *P*-values smaller than 0.05 are considered significant. Hence, we take the difference of the *P*-values for both test hypotheses (referred to as *P*-value difference) and remove features with *P*-value difference less than 0.9. For the ROC analysis, features with AUC less than 0.7 (Hosmer Jr *et al.*, 2004) are considered poor performers and are removed. The best characterization model is found by training the SVM using the remaining features. The WRE-BSE approach first performs WRE followed by BSE.

The worst-case time complexity of Tenet is $O((|V|+|E|)^2 + O(\mathcal{G}(\mathcal{X}_{\text{all}})) + O(\mathcal{T}(\cdot)))$ where $\mathcal{G}(\mathcal{X}_{\text{all}})$ is the worst-case time complexity for extracting the features and $O(\mathcal{T}(\cdot))$ is the worst-case time complexity of the feature selection method used. Note that in this article, $\mathcal{G}(\mathcal{X}_{\text{all}}) = O(|V|^3)$ whereas $O(\mathcal{T}(\cdot)) = O(|\mathcal{X}_{\text{all}}|^2 \times k \times |V|^3)$ (for BSE) where k is the number of iterations required for the grid-search. Proofs are given in Supplementary Material S1.5.

3 Results and discussion

Tenet is implemented using Java. We shall now present the experiments conducted to study the performance of Tenet and report some of the results here (additional results are given in Supplementary Material). The experiments are performed on a computer system using a 64-bit operating system with 8 GB RAM and a dual core processor running at 3.60 GHz. We characterize four signalling networks (referred to as *individual networks*) in *BioModels* (I_1–I_4 in Table 2) and a *combined network* that is generated by iteratively performing a union of the nodes and edges in individual networks. The resulting combined network is a graph consisting of four disconnected (The node and edge sets of the individual networks are disjoint.) subgraphs, each representing one individual network. For the combined network, we use each of the signalling network as the test set in turn (C_1–C_4 in Table 2) and examine the effects of generating characterization models from individual networks and from the combined network. Pruning in Tenet is performed on each individual network within the combined network. Supplementary Material S1.3 describes the generation of the training and test data. Note that in all our experiments, we use the linear kernel as it yielded the same accuracy as other kernels but is faster to train (Supplementary Material S1.7.1). We study different variants of Tenet (Table 3) by varying the SVM training approach.

3.1 Performance metrics

We evaluate the performance of Tenet based on prediction *accuracy* [The accuracy for the validation and test sets are denoted as $\phi_X(\text{val})$ and $\phi_X(\text{test})$, respectively, where X indicates the method used for training the SVM model. Average prediction accuracy is denoted as $\overline{\phi}$] (ϕ), *sensitivity* (TPR), *specificity* (TNR) and *precision* (PPV) of the generated characterization models using the same training and test set. The definitions are as follows: $\phi = \frac{\text{tp}+\text{tn}}{\text{tp}+\text{tn}+\text{fp}+\text{fn}}$, $\text{TPR} = \frac{\text{tp}}{\text{tp}+\text{fn}}$, $\text{TNR} = \frac{\text{tn}}{\text{fp}+\text{tn}}$ and $\text{PPV} = \frac{\text{tp}}{\text{tp}+\text{fp}}$ where tp, tn, fp and fn denote true positive, true negative, false positive and false negative prediction, respectively. Note that PPV is set to 0 when the classifier did not make any positive prediction. We include an additional metric *feature reduction factor* (FRF) to compare the performance of the feature selection methods. Formally, $\text{FRF} = 1 - \frac{|\mathcal{F}|}{\mathcal{X}_{\text{all}}}$ where $\mathcal{X}_{\text{all}}$ is the entire set of features considered in the study. The performance of different characterization models is compared using an *integrated performance score* (This score can be modified according to the needs of the application.) $\mathcal{P} = \sum_{\text{m} \in \text{M}} \text{val}_{\text{m}}$ where $M = \{\overline{\phi}(\text{val}), \phi(\text{test}), \text{TPR}, \text{TNR}, \text{PPV}\}$ and val_m is the value of metric m. Note that a larger score indicates better performance.

3.2 Feature selection

First, we examine the performance of different feature selection approaches (Tenet-B, Tenet-R and Tenet-H) and compare it with Tenet-*naïve* for different signalling networks. Note that in this set of experiments, we study the effect of the feature selection approaches in isolation. The effect of incorporating WMC into the SVM shall be investigated later. Table 4 reports the predictive feature sets for each network using different approaches. In total, 24 experiments were conducted as there are three feature selection methods and eight networks (I_1–I_4 and C_1–C_4). Amongst these 24 experiments, 25% of the predictive feature sets consist of only one feature whereas the remaining had multiple features (ranging from 4 to 15 features). This supports our previous observation (Chua *et al.*, 2014) that *multiple features result in better prediction of known targets*. Observe that in Table 4, bridging centrality is not always in the predictive feature set (e.g. I_2). Figure 2 plots the performances of different feature selection approaches. We can make several observations. First, no single approach performs consistently well on all performance metrics. In fact, network topology plays an important role in feature selection. For instance, I_4 has extremely high density of edges (ratio of edges to nodes) compared to other networks. The connectivity features of such networks become less informative and other features such as target downstream effect become more important. Hence, the most appropriate feature selection approach is dependent on the signalling network. However, we note that for larger sized networks, a larger number of features are informative (regardless of feature selection approach). This is perhaps because larger networks provide greater richness of context and diversity of structure in the sub-networks. As network sizes are growing and network analysis demands applicability to larger networks, future methods might benefit particularly from the use of multiple features. Second, feature selection generally led to an improvement in prediction accuracy (87.5% for validation dataset and 50% in test dataset) over the naïve approach. An exception is C_4 in which feature selection resulted in poorer performance. In C_4, the characterization model is generated using I_1, I_2 and I_3 as training data whereas I_4 is used as the test data. The characteristics of the known targets in the training data may be quite different from that of the test data. Indeed, from Table 4, we observe that bridging coefficient π is included in the predictive topological feature set of C_4, but not in I_4. Including redundant features may lead to poorer performance. Third, the models generally have high specificity due to imbalanced dataset. Fourth, Tenet-R has the best runtime performance, followed by Tenet-H and Tenet-B. The poorer performance of Tenet-B is due to the interaction of the feature selection approach with the classifier (classifier-aware approach) which is different from Tenet-R where the feature selection approach is a wrapper layer that sits on top of the classifier. Finally, the size of the networks used for training affects the runtime performance. In general, larger size networks require longer runtime. In the Supplementary Material S1.7.6, we report Tenet's performance on the human cancer signalling network containing more than 2500 nodes.

Table 2. Dataset

Network notation	I_1	I_2	I_3	I_4	C_1	C_2	C_3	C_4
Dataset (BioModel ID)	`MAPK-PI3K` (0000000146)	`glucose`-stimulated `insulin` secretion (0000000239)	`endomesoderm` gene regulatory (0000000235)	`glucose` metabolism (0000000244)	All networks			
Output node(s)	`ERKPP`	$\texttt{ATP}_{mitochondrial}$	`Protein_E_ Endo16`	`acetate`	{`ERKPP`, $\texttt{ATP}_{mitochondrial}$, `Protein_E_Endo16`, `acetate`}			
No. of nodes in dataset	36	59	622	47	764	764	764	764
No. of hyperedges in dataset	34	45	778	109	966	966	966	966
No. (%) of targets in dataset	9 (25%)	6 (10.2%)	206 (33.1%)	16 (34%)	237 (31%)	237 (31%)	237 (31%)	237 (31%)
Cross validation	8-fold	5-fold	10-fold	10-fold	10-fold	10-fold	10-fold	10-fold
Test set	Supplementary Material and Table 5				`MAPK-PI3K`	`glucose`-stimulated `insulin` secretion	`endomesoderm` gene regulatory	`glucose` metabolism
No. (%) of targets in test set	1 (25%)	1 (10%)	21 (34.4%)	2 (40%)	9 (25%)	6 (10.2%)	206 (33.1%)	16 (34%)

Table 3. TENET variant and WMC weight ratios used in experiment

Variant	BSE	WRE	WRE-BSE	WMC
TENET-naïve				
TENET-B	√			
TENET-R		√		
TENET-H			√	
TENET-W				√
TENET-WB	√			√
TENET-WR		√		√
TENET-WH			√	√

Weights ratio ID	C^+	C^-
1	0.1	0.9
2	0.2	0.8
3	0.3	0.7
4	0.4	0.6
5	0.5	0.5
6	0.6	0.4
7	0.7	0.3
8	0.8	0.2
9	0.9	0.1

√ indicates the approach(es) used in the variant.

Table 4. Features selected by various feature selection approaches

Data	TENET-B	TENET-R	TENET-H
I_1	δ, π, θ_{in}, θ_{out}	δ, ζ, β, ϑ, θ_{out}, μ, κ_{undir}	δ, ζ, β, ϑ
I_2	θ_{in}	δ, π, β, κ_{undir}, κ_{cyc}, α, θ_{in}, κ_{in}, μ, κ_{mid}, θ_{out}, θ_{total}	π, β, κ_{cyc}, κ_{undir}
I_3	δ, ζ, π, β, κ_{cyc}, ϑ, α, κ_{in}, κ_{mid}, μ, κ_{out}, θ_{out}, ω, θ_{total}, κ_{undir}	δ, ζ, ϑ, α, κ_{mid}, θ_{out}, θ_{total}, ω, κ_{undir}	δ, ζ, ϑ, α, θ_{out}, θ_{total}, κ_{undir}
I_4	ζ, β, κ_{cyc}, ϑ, α, κ_{in}, κ_{mid}, μ, ω, κ_{out}, θ_{out}, θ_{total}, κ_{undir}	ω	ω
C_1	δ, ζ, π, β, κ_{cyc}, ϑ, α, θ_{in}, κ_{mid}, θ_{out}, μ, ω, κ_{undir}	δ, ζ, π, β, ϑ, α, κ_{mid}, κ_{undir}, θ_{out}	ζ, π, ϑ, α, θ_{out}, κ_{undir}
C_2	δ, ζ, π, β, κ_{cyc}, ϑ, α, θ_{in}, κ_{mid}, θ_{out}, κ_{undir}	δ, ζ, α, κ_{mid}, θ_{out}, ω, θ_{total}, κ_{undir}	δ, ζ, α, κ_{mid}, ω, κ_{undir}
C_3	θ_{in}	ζ	ζ
C_4	ζ, π, β, κ_{cyc}, ϑ, α, κ_{in}, θ_{in}, ω, κ_{out}, θ_{out}, θ_{total}, κ_{undir}	δ, ζ, π, ϑ, α, κ_{mid}, θ_{out}, ω, θ_{total}, κ_{undir}	ζ, π, ϑ, α, κ_{undir}, ω, θ_{out}, θ_{total}, κ_{mid}

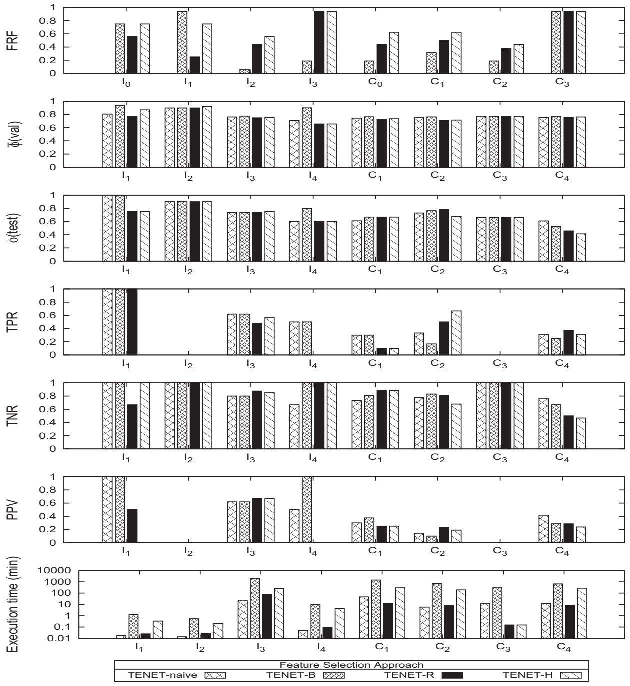

Fig. 2. Performance of different feature selection approaches

3.3 Effect of varying WMC

Intuitively, when we vary the WMC, we expect that as the target misclassification cost C^+ increases, the prediction accuracy, sensitivity, specificity and precision would display a negative skewed, increasing, decreasing and positive skewed distribution, respectively. This is because a large C^+ eventually results in a model that is likely biased towards classifying data as targets. We noted the following when the WMC was varied. First, amongst the individual networks, only I_3 (Fig. 3) displays the expected trends. This could be due to the extreme small target size (1 or 2) in the test set that resulted in extreme fluctuations in the performance metrics and deviation from the expected trends. Hence, the target size of the test set can have significant impact on the observed results. Second, the performance of the combined networks C_1, C_2 and C_4 (Supplementary Material S1.7.2) resembles that of I_3, possibly due to the large size of I_3 dominating over other networks used for training. This implies large training networks can have undue influence on the characterization model. Third, sensitivity generally improves whereas specificity generally deteriorates when the target misclassification cost is set higher than the non-target misclassification cost ($C^+ > C^-$). The choice of an appropriate model depends on the application. Fourth, the prediction accuracy tends to display a skewed distribution where accuracy initially increases (or remains constant) with increasing C^+, and then decreases with increasing C^+. Fifth, individual networks and combined networks behave differently. In individual networks, prediction accuracy, sensitivity and precision generally improve when C^+ is set larger than C^-. However, in combined networks, sensitivity improves whereas other performance criteria deteriorate when C^+ is set larger than C^-. Hence, there is no single universal best value of C^+ and the choice of C^+ depends on the network.

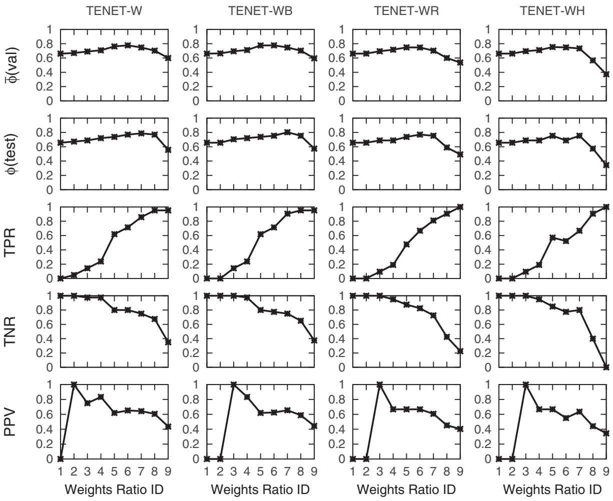

Fig. 3. Performance of TENET variants incorporating feature selection approach and WMC for the `endomesoderm` gene regulatory network

3.4 Best TENET variant

We identify the best TENET variant (Table 5) using the integrated performance score $\mathcal{P}$. We note the following. First, the best TENET variant is network dependent. Second, variants incorporating both WMC and feature selection generally perform well. Specifically, setting C^+ greater than C^- led to better results. Third, TENET variants based on individual networks (I_1 to I_4) outperform that based on combined networks (C_1–C_4). The poorer performance of the combined networks may be due to insufficient number of training networks, inappropriate or insufficient features used for training or that signalling networks by nature have distinct characteristics and it is just not possible to have a generalized model. Finally, the predictive topological features differ across networks (Tables 4 and 5). Hence, as we mentioned in Section 1, a single set of predictive topological features may not effectively characterize known targets in all signalling networks. When we compare the results with that in our previous work, we note that the set of predictive topological features is different from the discriminative topological features (DTF) identified in Chua *et al.* (2014) although there was an overlap of at least 50% of the features (We consider only I_1 to I_3 and exclude I_4 from this comparison as no DTF was found at P-value less than 0.05). The difference is due to the different approach used to identify the features. The characterization models (We use SVM with WMC and WRE to generate the characterization models.) generated by these DTFS also yielded poorer average ROC (0.873) than that generated using TENET (0.913) (Approach DIFFER in Fig. 4).

Table 5. Summary of best TENET variant for different networks

	I_1	I_2	I_3	I_4	C_1	C_2	C_3	C_4
Best approaches	TENET-B[a], TENET-WB ($C^+ = 0.1$, 0.2, 0.3, 0.4)	TENET-WH ($C^+ = 0.9$[a])	TENET-WB ($C^+ = 0.7$[a])	TENET-WB ($C^+ = 0.2$, 0.3, 0.4, 0.6, 0.8[a])	TENET-WH ($C^+ = 0.6$[a])	TENET-R[a]	TENET-WR ($C^+ = 0.8$[a]), TENET-WH ($C^+ = 0.8$)	TENET-naïve[a]
$\mathcal{P}$	4.935	4.109	3.86	4.9	3.08	3.022	3.268	2.917
$\overline{\phi}$(val) [$\Delta\overline{\phi}$(val)]	0.935 [0.16]	0.82 [−0.087]	0.747 [−0.02]	0.9 [0.268]	0.734 [−0.013]	0.711 [−0.052]	0.561 [−0.274]	0.757 [0]
ϕ(test) [$\Delta\phi$(test)]	1 [0]	0.9 [0]	0.803 [0.088]	1 [0.667]	0.694 [0.136]	0.78 [0.070]	0.724 [0.097]	0.609 [0]
TPR [ΔTPR]	1 [0]	1 [∞[b]]	0.905 [0.462]	1 [1]	0.4 [0.333]	0.5 [0.502]	0.602 [∞[b]]	0.313 [0]
TNR [ΔTNR]	1 [0]	0.889 [−0.111]	0.75 [−0.063]	1 [0.499]	0.808 [0.105]	0.811 [0.048]	0.788 [−0.212]	0.767 [0]
PPV [ΔPPV]	1 [0]	0.5 [∞[b]]	0.655 [0.058]	1 [1]	0.444 [0.48]	0.231 [0.615]	0.593 [∞[b]]	0.471 [0]

Note: C^+ values are provided in bracket besides approaches using WMC. $\Delta_x = \frac{x_{\text{best}} - x_{\text{naïve}}}{x_{\text{naïve}}}$ where x_{best} and $x_{\text{naïve}}$ are the values of performance metric x of the best TENET variant and TENET-naïve, respectively.

[a]Best models selected for generating the characterization model.

[b]Instances where $x_{\text{naïve}} = 0$.

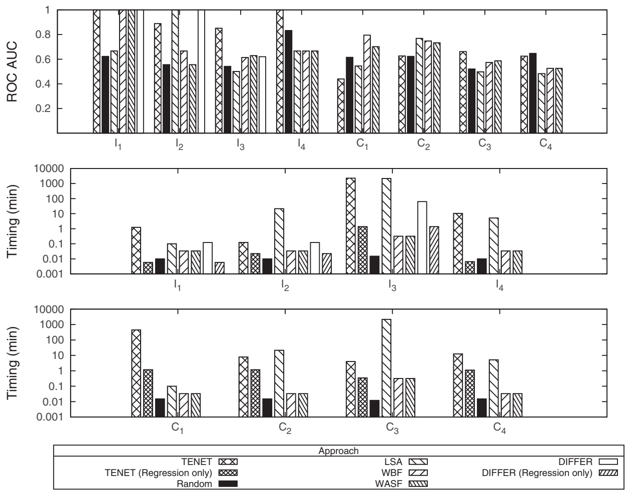

Fig. 4. Performance of different prioritization approaches

3.5 Comparison with state-of-the-art approaches

Recall that state-of-the-art techniques such as McDermott *et al.* (2012), Zhang *et al.* (2010) and Hwang *et al.* (2008) focus on PPI networks instead of signalling networks. To the best of our knowledge, there does not exist any target characterization technique for signalling networks. However, one way to investigate the performance of TENET is to examine how well the characterization model generated by it *prioritizes* known targets. Intuitively, *target prioritization* aims to *rank* the nodes according to their potential of being a target based on some *importance measures* (e.g. gene expression level; Chen *et al.*, 2011). A more detailed exposure to the target prioritization problem as well as how TENET is used to prioritize known targets is given in Supplementary Material S1.6.

For our study, we compare TENET with several *network-aware* target prioritization approaches, namely, random prioritization, LSA (Gustafson *et al.*, 1996) and *NetworkPrioritizer* (Kacprowski *et al.*, 2013). Comparison with *network-unaware* techniques as well as PPI network-based techniques is reported in the Supplementary Material S1.7.3 and S1.7.4, respectively.

In random prioritization, the nodes were randomly assigned a rank in the range $[1–|V|]$ where $|V|$ is the number of nodes in the network and we assume that no ranking ties are present. LSA was performed using *Copasi* (Sahle *et al.*, 2006) 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 as we do not have prior knowledge of the best weights or features to consider. For TENET, we use the characterization model to generate prioritization ranks of known targets. Specifically, we apply the SVM models to obtain these ranks. The SVM type is set to ϵ-SVR [In ϵ-SVR, the error function is an ϵ-insensitive loss function and error smaller than ϵ is ignored (Chang *et al.*, 2011).] with default ϵ value (1×10^{-3}) and the SVM parameters are set according to the best models for each network (Table 5 and Supplementary Material S1.7). Note that the nodes are ranked in decreasing order of the regression score and higher ranked nodes are more likely to be targets.

The experimental results reveal that the *normalized* ranks *(The* normalized rank *of a node* u *for a particular approach* x *is defined as* $\Psi_{\mathrm{norm}(x):u} = \frac{\Psi_{x:u}}{\max_{i \in V}(\Psi_{x}:i)}$.) of a given node vary widely using different approaches (Supplementary Material S1.7.5). Hence, an approach that performs better for one particular network can perform poorly in another. We further perform ROC analysis based on the rankings of the nodes in the test set for each network. From Figure 4, we observe that TENET outperforms other approaches in terms of the quality of the prioritization results, particularly for individual networks, and is comparable in terms of runtime performance when SVM training is performed offline [TENET (Regression only)].

4 Conclusions

We propose TENET, an SVM-based approach that characterizes known targets in signalling networks using topological features by identifying a set of predictive topological features and using them to generate a characterization model. TENET uses feature selection to remove redundant features, thereby improving prediction accuracy of the characterization models and WMC to improve other performance criteria (e.g. sensitivity). Our empirical study reveals that the characterization models generated by TENET outperform state-of-the-art approaches in prioritizing signalling and PPI networks. In summary, the contribution of this work is a machine learning-based framework that affords flexibility in characterizing signalling networks of different sizes and with different number of known targets. Although TENET is evaluated on a small (Manual target curation, a time-intensive process, is needed to identify known targets of signalling networks for validating our experimental results.) number of signalling networks, it can easily incorporate additional signalling networks without any modification to the framework. As part of future work, we intend to explore how the characterization models learnt by TENET can be leveraged for target prioritization of signalling networks with *unknown* targets.

Funding

This work was supported in part by a Singapore MOE AcRF Tier 1 Grant RGC 1/13 to SSB and by a Duke-NUS block grant to LTK.

Conflict of Interest: none declared.

References

Chang,C.-C. *et al.* (2011) Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.

Chapelle,O. *et al.* (2002) Choosing multiple parameters for support vector machines. *Mach. Learn.*, **46**, 131–159.

Chen,L. *et al.* (2005) Stack-based algorithms for pattern matching on DAGs. In: Böhm,K. *et al.* (eds) *VLDB*. VLDB Endowment, Trondheim, Norway, pp. 493–504.

Chen,Y.-A. *et al.* (2011) Targetmine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**, e17844.

Chua,H. *et al.* (2014) One feature doesn't fit all: characterizing topological features of targets in signaling networks. In: Baldi,P. *et al.* (eds) *ACM BCB*. ACM New York, Newport Beach, CA, USA, pp. 426–435.

Engelfiet,J. *et al.* (1990) A comparison of boundary graph grammars and context-free hypergraph grammars. *Inf. Comput.*, **84**, 163–206.

Gustafson,P. *et al.* (1996) Local sensitivity analysis. *Bayesian Stat.*, **5**, 197–210.

Hatakeyama,M. *et al.* (2003) 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.

He,X. *et al.* (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**, e88.

Hosmer,D.,Jr *et al.* (2004) *Applied Logistic Regression*. 2nd edn. John Wiley & Sons, New York.

Hsu,C.-W. *et al.* (2003) *A Practical Guide to Support Vector Classification*. Technical report. Department of Computer Science, National Taiwan University.

Hwang,W.-C. *et al.* (2008) Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin. Pharmacol. Ther.*, **84**, 563–572.

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

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

Kohavi,R. *et al.* (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, **14**, 1137–1145.

Kwon,Y.-K. *et al.* (2008) Coherent coupling of feedback loops: a design principle of cell signaling networks. *Bioinformatics*, **24**, 1926–1932.

Maira,S.-M. *et al.* (2008) Identification and characterization of nvp-bez235, a new orally available dual phosphatidylinositol 3-kinase/mammalian target of rapamycin inhibitor with potent in vivo antitumor activity. *Mol. Cancer Ther.*, **7**, 1851–1863.

Marill,T. *et al.* (1963) On the effectiveness of receptors in recognition systems. *IEEE Trans. Inf. Theory*, **9**, 11–17.

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**, 28.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Sahle,S. *et al.* (2006) Simulation of biochemical networks using copasi: a complex pathway simulator. In: Perrone,L. *et al.* (eds) *WSC*. Winter Simulation Conference, Huntington Beach, CA, USA, pp. 1698–1706.

Tarjan,R. (1972) Depth-first search and linear graph algorithms. *SIAM J. Sci. Comput.*, **1**, 146–160.

Yan,X. *et al.* (2013) The identification of novel targets of mir-16 and characterization of their biological functions in cancer cells. *Mol. Cancer*, **12**, 92.

Yang,K. *et al.* (2008) Finding multiple target optimal intervention in disease-related molecular network. *Mol. Syst. Biol.*, **4**, 228.

Zhang,J. *et al.* (2010) Novel biological network features discovery for in silico identification of drug targets. In: Veinot,T. (ed.) *IHI*. ACM New York, Arlington, VA, USA, pp. 144–152.