Toward Accurate Reconstruction of Functional Protein Networks: Supporting Information

Nir Yosef et al.

Analyzing the apoptosis data set

The APT protein set was manually assembled and contains 77 proteins known to act as the core machinery of the apoptotic pathways in Humans. Among those are key regulators of apoptosis such as different Caspases (Kumar, 2007; Bao & Shi, 2007), the Bcl-2 family of proteins (Youle & Strasser, 2008), death receptors and their ligands (Locksley et al, 2001), and several caspase substrates. We consider a subgroup of six caspase substrates as the anchor points of the apoptosis core machinery. These substrates are directly responsible for different events that lead to the collapse of the cell such as chromatin condensation, DNA fragmentation, disruption of the cell's cytoskeleton etc.

The APT set, the anchor points and the subnetwork models generated by the different methods are available as Supplemental Material. In the following we examine the biological significance of the obtained subnetworks from two perspectives: biological case analysis as well as large scale systematic validation.

Biological case analysis. Caspase3 activation by the extrinsic pathway involves the extracellular binding of death ligands to different death receptors. Adaptor proteins (such as FADD) are then recruited to the death receptors, which in turn recruit initiator caspases (Bao & Shi, 2007) such as caspase8, that lead to the activation of the effector caspases: caspase3 and caspase7 (Kumar, 2007). SI Figures 1a–c present the ways by which the three main death ligands (FASL, TNF and TRAIL (TNFSF10)) are linked to the APT node in the different models. The global approach correctly channeled the ligands through FADD (SI Figure 1c). Conversely, the shortest pathways to the APT node, found individually for each ligand (*i.e.,* the local approach) do not go through the canonical FADD or caspase8, which seems less plausible (SI Figure 1a). The Charikar-0.25

algorithm gave an intermediate solution connecting two ligands out of the three through FADD (SI Figure 1b). Interestingly, this case resembles the adversary example in SI Figure 5. The global approach seeks the best solution for all three ligands concomitantly, and chooses to connect them through a single adaptor, whereas the local approach seeks the best solution independently for each ligand, resulting in less plausible pathways.

Another example is the Bcl-2 family of proteins, one of the key regulators of the apoptotic machinery. In the models produced by the global approach and the Charikar-0.25 algorithm, the Bcl-2 family members were grouped together in a single connected component, reflecting their well-studied biological function during apoptosis (Youle & Strasser, 2008) (light blue nodes in SI Figures 1f and 1e, respectively). The local approach on the other hand divided these proteins into six connected components of sizes 1 to 3, scattered al over the network (SI Figure 1d).

Functional coherence based on manual annotation. For a more systematic comparison, we classified the proteins in the APT set into 14 functional groups (see legend of SI Figure 1). We then measured the *functional coherence* of a subnetwork model as the tendency of the different functional groups to be clustered together (Supporting Methods). As in the gene ontology based measure (Figure 2), the results in SI Figure 2 show a clear increase in accuracy as we go from the local to the global extreme. We also experimented with a less refined partition (using 8 groups) and achieved similar results (data not shown). The partition into the functional groups is available as Supplemental Material.

Biological case analysis of the TLM system

Comparative analysis

Inspection of specific pathways and proteins within the Steiner tree (global) and the Charikar-α models shows the superiority of the latter methods, and in particular of the Charikar-0.25. A good example of these differences is provided by Est1/Est2/Est3, three well-characterized telomeric proteins (Zakian, 1996) that are all included in the telomerase machinery group we assembled. Est encodes a subunit of telomerase, the nucleoprotein complex in charge of synthesizing telomeric DNA repeat (Counter et al, 1997). Est1/Est3 are accessory proteins that regulate the recruitment of telomerase to the telomere (Evans & Lundblad, 1997). Est2 appears unconnected to the rest of the network in the Steiner tree model, the Charikar-0.5 model and the Charikar-0 model, but is

connected to to other nodes in the graph in the Charikar-0.25 model. Similarly, the Steiner tree model has Est1 unconnected, and Est3 only connected to a single protein, Fyv13. In contrast, both Est1 and Est3 are well connected in all the Charikar-α models. Two additional well-characterized telomeric proteins, Stn1 and Cdc13 (Grandin et al, 1997) appear as unconnected or sparsely connected in the Steiner tree model, whereas they are hubs connecting telomeres to replicative and checkpoint functions in the Charikar-α models.

Another manifestation of the superiority of the Charikar-α models is given by the checkpoint proteins Rad53 and Tel1. These two protein kinases are the yeast orthologs of mammalian Chk2 and Atm, respectively, and have been shown to play an important role in the response to DNA damage such as double-strand breaks (Branzei & Foiani, 2006). Surprisingly, it was recently found that they are also important in preventing attempts by the cells to repair the natural chromosomal ends as if they were double-strand breaks (Viscardi et al, 2005). Rad53 is connected in the Charikar-0.25 and 0.5 models to the telomere protein Cdc13, and to the DNA checkpoint proteins Mec1 (the yeast ortholog of mammalian ATR), Dun1, and nucleolin/Nsr1. In contrast, the Steiner tree model retained only the checkpoint links to Mec1 and Dun1 (the Charikar-0 model shows only links to Cdc13 and Dun1). In the case of Tel1, the Charikar-0 and 0.25 models link Tel1 to the telomere through the single-stranded DNA binding protein Rfa1 and the Mismatch repair Msh2/Exo1 nucleases. In contrast, both the Steiner tree model and the Charikar-0.5 model trace the connection between Tel1 and the telomeres through a long pathway that also requires additional proteins, such as the MRX complex and the Sgs1 helicase (Steiner tree) or the Rsc transcription mediator (Charikar-0.5).

Biological insights from the TLM system – the RSC case

The Charikar 0.25 model predicts that mutations in the essential gene RSC8, encoding a subunit of the RSC chromatin remodeling complex (Cairns et al, 1996), should exhibit long telomere phenotypes. Using a constitutively destabilized (DAmP) allele of the gene citepSchuldiner-05, we have experimentally confirmed this Prediction (SI Figure 3). The Charikar 0.25 model placed Rsc8p as the major hub connected to Rsc2p, Npl6p and Htl1p, other components of the RSC chromatin remodeling complex (SI Figure 3) (Wilson et al, 2006). Surprisingly, the model suggests Sin3p and Sap30p as the upstream proteins connected to Rsc8p. Together with Rpd3, Sin3p and Sap30p form a histone deacetylase complex involved in silencing at telomeres (Sun & Hampsey, 1999). Recent studies have implicated RSC, together with additional chromatin remodeling enzymes, in

promoter-dependent transcription initiation (Carey et al, 2006). Upstream to Sin3p, the Charikar 0.25 model placed Opi1p and Dep1p. These two genes encode transcriptional regulators of phospholipid biosynthetic genes(Lamping et al, 1994; Sreenivas & Carman, 2003). This network suggest a new telomere length regulation pathway. The Opi1p and Dep1p could affect Sin3 activity (either directly, or, most likely, through a signal transduction pathway involved phospholipids). The SIN3 histone deacetylase would then cooperate with RSC to control transcription of telomere-length regulating genes.

Supporting Methods

Large scale performance measures

Predictive score. Measures the ability to recover new unannotated proteins that are relevant to a system of interest using cross validation. In each iteration we hide the annotation of a subset of the terminal proteins and measure how well the resulting model managed to recover them. The recovery success is measured using the Jaccard measure (Yosef et al, 2006) accompanied by a hypergeometric p-value. For the TLM data set we use 2-fold cross validation, dividing the proteins in two groups according to their experimental source ((Askree et al, 2004) and (Gatbonton et al, 2006), proteins that were reported in both are randomly distributed among the groups). For the APT data set we use 2-fold cross validation based on random partitioning of the proteins and repeat this procedure 100 times. Notably, we repeated this experiment with various cross validation levels ranging from leaving out a single protein to 20% of the proteins. We obtained qualitatively similar results to the ones in Figure 2 in all cases except for the leave one out extreme where the global approach performed better than the local one (data not shown). In both data sets, the p-values from the different cross validation repeats are combined into a single p-value using Fisher's method (Fisher, 1948). We report the mean jaccard value of the subnetwork models that had a significant p-values ($p \leq 0.05$).

Functional coherence. We compute for each terminal protein the union of all paths connecting it to the root node, and call this path-collection a pathway. For each such pathway P we derive a functional enrichment score based on the gene ontology (GO) annotations (Harris et al, 2004) of its members (discarding the ends of the pathway). The enrichment of each GO class within P is computed by a hypergeometric p-value. This p-value is compared to 100 p-values obtained

by randomly selecting connected components of the same size, yielding an empirical p-value. The reported functional coherence scores are the fraction of pathways with at least 3 inner nodes that were significantly coherent (empirical p-value lower than 0.05).

We use an additional measure for the apoptosis models. Specifically, we divide the terminal proteins into several finely tuned functional groups (SI Figure 1). For each pair of vertices we calculate the similarity of their neighborhoods using the Jaccard measure as in (Goldberg & Roth, 2003). The distribution of similarity values for pairs within the same functional group is then compared to the distribution of pairs that span different groups using the Wilcoxon rank-sum test. The reported functional coherence scores (SI Figure 2) are the mean jaccard values. As before, we only report significant cases ($p \leq 0.05$). Notably the results are qualitatively similar to those obtained with the gene ontology based measure.

Monochromaticity. The monochromaticity of a pathway P reflects the coherence of the effect of the TLM proteins it contains on telomere length. It is the larger between the fraction of short-effect TLM proteins in P and the fraction of long-effect TLM proteins in P. This measure is motivated by the observation that proteins on the same TLM-related pathway tend to have a similar effect on the telomere length (Shachar et al, 2008). We report the mean monochromaticity across all pathways.

Phenotype vs. location (PvL). The PvL measure quantifies the correspondence between the effect of a protein and its location in the subnetwork. The knockout of TLM genes may have effects of varying magnitude on telomere length, leading to slight, moderate or large alterations (short or long) (Askree et al, 2004; Gatbonton et al, 2006). We quantify the correspondence between the location of a protein in the resulting network model and the magnitude of its phenotype, by considering the extent to which TLM proteins that occupy internal nodes have a stronger effect on telomere length than leaf nodes (Shachar et al, 2008). To rule out the putative effect of the distance of TLM network proteins from the telomere binding proteins, we also compute a partial correlation index that factors out the distance (Shachar et al, 2008).

Model validation based on new TLM data

We define the success rate of a model with respect to experimental data of both positive and negative cases (genes that exhibited/ did not exhibit defects in telomere length) as the mean of the

true positive rate and the true negative rate.

To predict a phenotype (short/long) for non-terminal proteins we define the *effect* of a path from a TLM node to the *TELOMERE* node according to the phenotype of the participating TLM proteins. The effect can be either *long* if most proteins have a long effect on telomere length, *short* in the opposite case, or *undecided* if it contains equal numbers of shortening and elongating proteins. We then classify a given unannotated protein to the *long* class if all the pathways to which it belongs are either *long* or *undecided* and at least one of them is *long*. Similarly, we classify a protein as *short* if all the pathways in which it is involved are either *short* or *undecided* and at least one of them is *short*. If none of these conditions hold, we leave this protein as *undecided*. The reported accuracy measure is the percentage of correctly predicted phenotypes out of all phenotypes predicted as either *short* or *long*.

The Charikar-α algorithm

For convenience, we describe the algorithm for directed graphs, as an undirected edge can be replaced by two oppositely directed edges. The algorithm does not operate directly on the input graph but rather on its transitive closure, namely, it is assumed that between every pair of nodes u and v there exists an edge (u, v) with a cost equal to the weight of the shortest path between them.

Given such a transitive closure G, a root node r, and a set of terminals X, the algorithm finds a low cost subgraph connecting the terminals to the root, where the relative importance of the overall cost and the costs of the specific terminal-root paths is determined by α. We define the *density* of a subgraph H as $d(H) = \frac{C(H)}{k(H)^{\alpha+0.5}}$ where $C(H)$ is the sum of edge weights in H and $k(H)$ is the number of terminals in H. The algorithm operates in an iterative manner; at each iteration, a 2-level tree (*i.e.,* a tree in which the distance of a leaf from the root is at most 2) with a low density is identified, and the set of nodes covered by this tree is removed from X. Upon termination, the algorithm returns the union of all these identified trees. (See Algorithm 1 for pseudocode of the algorithm.)

Implementation details

We applied the Charikar-α algorithm to the TLM data set with three different α values $(0, 0.25, 0.5)$. Our theoretical analysis (Table 1) indicates that the best approximation ratio guarantee with respect to the combined goal is attained at $\alpha = 0.25$. The two other α values lead to different solutions than those obtained by the extreme approaches and, notably, provide optimal approximation guar-

antees for different choices of the normalizing factor c (it can be shown that for $c = \frac{OPT_L}{\sqrt{k} \cdot OPT_G}$ the best α is 0, and for $c = \frac{\sqrt{k} \cdot OPT_L}{OPT_G}$ the best α is 0.5, where k is the number of terminal nodes).

For the local optimization approach we computed the union of all the shortest paths from the terminals to the root. For the Steiner tree construction (the global approach) we used the minimum spanning tree heuristic (Gilbert & Pollak, 1968) which guarantees a factor 2 approximation ratio. Notably, most subsequent approximation algorithms for the problem, including the tightest approximation algorithm known (Gilbert & Pollak, 2005), rely on the minimum spanning tree heuristic as a first step (Gropl et al, 2001). Hence, these algorithms are amenable to the worst-case scenario depicted in SI Figure 5, and do not provide a better tradeoff with the local approach then the minimum spanning tree heuristic.

In both the shortest paths, Steiner tree and intermediary (Charikar-α) algorithms there might exist few solutions with the same best score. To avoid an arbitrary choice among equally-good solutions, we adjusted our implementations to record multiple solutions. This was done by 100 random shuffling of the order by which the algorithms process their input, and taking the union of all solutions obtained.

Approximation ratios of the Charikar-α algorithm

In the following we denote $\beta = \alpha + 0.5$. Let X denote the set of terminals, k denote the number of terminals, and r denote the root. We assume that the input graph is directed and that between every pair of nodes u and v there exists an edge (u, v) with a cost $C_{u,v}$ equal to that of a shortest path between u and v in the original graph. We further assume, without loss of generality, that all terminals are at the leaves of any Steiner tree (otherwise, connect the real terminal to a dummy terminal and add a zero cost edge). For a subtree T, we denote its total cost by $C(T)$, the number of terminals it spans by $k(T)$, and its density by $d(T) = C(T)/[k(T)^{\beta}]$.

Let T be a 2-level tree rooted at a node r. T can be decomposed into a set of 2-level trees by separately considering each of the nodes adjacent to the root. Let v be such an adjacent node. Denote by T_v the tree rooted in r and containing v and the set of terminals connected to v. Denote the latter by X_v, and let $k_v = |X_v|$.

Theorem 1. *Charikar-α approximates F_L to within a factor of $O(k^{\frac{1}{2}+\alpha})$.*

Proof. Let T denote the tree returned by the Charikar-α algorithm. For a node v adjacent to r, let $OPT(r, X_v)$ be the sum of shortest paths from r to X_v. From the construction of T it follows

that for all $t \in X_v$, $\frac{C(T_v)}{k_v^\beta} \leq \frac{C_{r,t}}{1^\beta} = C_{r,t}$. Thus:

$$\begin{aligned} C(T_v) &\leq k_v^\beta \cdot min_{t \in X_v} C_{r,t} \\ &\leq k_v^\beta \cdot \frac{OPT(r, X_v)}{k_v} \\ &\leq k_v^{\beta-1} \cdot OPT(r, X_v) \end{aligned}$$

Let W_v be the sum of weights of shortest paths from the r to X_v in T_v. Using the above bound we have:

$$W_v \leq k_v \cdot C(T_v) \leq k_v^\beta \cdot OPT(r, X_v)$$

Finally, let W and OPT_L be the sums of weights of shortest paths from r to X in T and in the original graph, respectively. Denote by $N(r)$ the set of nodes adjacent to r in T. It follows that

$$\begin{aligned} W &= \sum_{v \in N(r)} W_v \\ &\leq \sum_{v \in N(r)} k_v^\beta \cdot OPT(r, X_v) \\ &\leq k^\beta \cdot OPT_L = k^{0.5+\alpha} \cdot OPT_L \end{aligned}$$

□

Theorem 2. *Charikar-α approximates F_G to within a factor of $O(k^{1-\alpha})$.*

Proof. We prove the theorem by extending the lemmas and theorems given in (Charikar et al, 1999) to our definition of density. For a given graph, we denote by $ST(r, X)$ the Steiner tree problem with a root r and a terminal set X.

Definition 1. *An $f(k)$ partial approximation procedure for the Steiner tree problem $ST(r, X)$ is a procedure which constructs a tree T' rooted at r and spanning k' terminals in X such that: $d(T') = \frac{C(T')}{k'^\beta} \leq f(k) \cdot \frac{C(T_{OPT})}{k^\beta}$ where T_{OPT} is the optimal solution.*

Let $\Psi(r, X)$ denote a partial approximation procedure. This procedure can be applied iteratively such that in each iteration the terminals which were already linked to the root are removed from X. Let us denote the iterative procedure as $\Phi(r, X)$.

Lemma 3. *Given $\Psi(r, X)$ and an $f(k)$ partial approximation procedure for $ST(r, X)$, where $\frac{f(x)}{x}$ is a decreasing function of x, the algorithm $\Phi(r, X)$ gives a $k^{1-\beta} \int_0^k \frac{f(x)}{x} dx$ approximation for*

$ST(r, X)$.

Proof. Let us denote the tree returned by $\Psi(r, X)$ as T_1, and the number of terminals in T_1 as k_1. We prove the claim by induction on k, the size of the terminal set. The base case $k = 1$ follows as $C(T_1) \leq f(1) \cdot C(T_{OPT}) \leq \int_0^1 \frac{f(x)}{x} dx$ (by the decreasing property of $\frac{f(x)}{x}$). Suppose the claim is true for all values less than k. Since $d(T_1) = \frac{C(T_1)}{k_1^\beta} \leq f(k) \cdot \frac{C(T_{OPT})}{k^\beta}$, we have:

$$\begin{aligned} c(T_1) &\leq k_1^\beta \cdot k^{1-\beta} \cdot \frac{f(k)}{k} \cdot C(T_{OPT}) \\ &\leq k^{1-\beta} \cdot C(T_{OPT}) \cdot \int_{k-k_1^\beta}^{k} \frac{f(x)}{x} dx \end{aligned}$$

If $k_1 = k$ then the claim clearly follows. Suppose $k_1 < k$, and let X_1 be the set of terminals in T_1. Let T_2 be the tree returned by the call to $\Phi(r, X \setminus X_1)$. By the inductive hypothesis, we have $c(T_2) \leq (k - k_1)^{1-\beta} \cdot \int_0^{k-k_1} \frac{f(x)}{x} dx$. Taking the union of T_1 and T_2 we obtain:

$$\begin{aligned} &c(T_1) + c(T_2) \leq \\ &\quad k^{1-\beta} \cdot C(T_{OPT}) \cdot \int_{k-k_1^\beta}^{k} \frac{f(x)}{x} dx \\ &\quad +(k - k_1)^{1-\beta} \cdot \int_0^{k-k_1} \frac{f(x)}{x} dx \\ &\quad \leq k^{1-\beta} \cdot C(T_{OPT}) \cdot \int_0^{k} \frac{f(x)}{x} dx \end{aligned}$$

□

Lemma 4. *The trees T_{BEST} chosen during the* while *loop of the Charikar-α algorithm are $2k^{\frac{1}{2}}$ partial approximation for the $ST(r, X)$ problem.*

Proof. Let $T_{OPT}^{(2)}$ be the optimal 2-level tree that solves $ST(r, X)$ (recall that we are working with the transitive closure of the graph). Consider an outgoing edge (r, v) in $T_{OPT}^{(2)}$, such that the subtree T_v (rooted at v) is the minimum density subtree of $T_{OPT}^{(2)}$. Notice that T_v is exactly a 1-level tree. Let v' be the child of r with minimal density, namely $d(T_{v'}) = \frac{w(r,v)+C(T_{v'})}{k_{v'}^\beta}$ is minimum. It is easy to see that for $0.5 \leq \beta \leq 1$ we get $d(T_{v'}) \leq d(T_v) \leq d(T_{OPT}^{(2)})$.

It follows that:

$$d(T_{BEST}) = d(T_{v'}) \leq d(T_{OPT}^{(2)}) = \frac{C(T_{OPT}^{(2)})}{k^\beta}$$

A rooted tree in a transitively closed graph can be transformed into an l-level tree defined on the

same set of nodes, while blowing up the overall cost by no more than $O(lk^{1/l})$ (Zelikovsky, 1997). Using this fact we obtain a partial approximation bound of $2k^{1/2}$:

$$d(T_{BEST}) \leq 2k^{1/2}\frac{C(T_{OPT})}{k^{\beta}}$$

□

Combining lemmas 3 and 4, we get that the approximation ratio of Charikar-α is no more than

$$k^{1-\beta} \cdot \int_0^k \frac{2x^{1/2}}{x}dx = 4k^{1.5-\beta} = 4k^{1-\alpha}$$

which completes the proof of the theorem. □

Theorem 5. *Charikar-α approximates the combined objective $c \cdot F_G + F_L$ to within a factor of $O(k^{max\{1-\alpha,\frac{1}{2}+\alpha\}})$.*

Proof. Let H be the subgraph returned by the Charikar-α algorithm. Let H^* be an optimum solution under the combined objective. Clearly, $c \cdot F_G(H^*) + F_L(H^*) \geq c \cdot OPT_G + OPT_L = 2 \cdot OPT_L$. It follows that the approximation ratio is bounded by $R = \frac{c \cdot F_G(H) + F_L(H)}{2 \cdot OPT_L}$. Plugging into this ratio the bounds proved in Theorems 1 and 5 we conclude:

$$\begin{aligned} R &\leq \frac{c \cdot 4k^{1-\alpha} \cdot OPT_G + k^{0.5+\alpha} \cdot OPT_L}{2 \cdot OPT_L} \\ &= \frac{(4k^{1-\alpha} \cdot + k^{0.5+\alpha}) \cdot OPT_L}{2 \cdot OPT_L} = O(k^{max\{1-\alpha,0.5+\alpha\}}) \end{aligned}$$

□

Figure Legends

SI Figure 1. APT network models. Top: The FADD example. Bottom: Whole network views. The nodes are color coded according to their functional groups. The Bcl-2 family of proteins is colored in light blue. In both panes the results are displayed in the following order (from left to right): shortest paths, Charikar-0.25, and Steiner tree.

Figure 2. Functional coherence of the different apoptosis models. The Jaccard coefficient measures the extent to which the manual partition of the APT proteins into groups is reflected by the model.

SI Figure 3. RSC subunit interactions. Left: the inferred subnetwork. Green frame: subunits of the RSC complex; Red frame: histone deacetylase complex; Blue frame: Opi1 and Dep1 transcriptional regulators. Yellow nodes: essential proteins that showed a TLM phenotype in our new experimental data set. Other nodes are color coded as in Figure 3. Right: Telomere Southern blot of the essential gene RSC8 from the DAmP Yeast Library. DNA was digested with XhoI and probed with telomeric sequences and with unique genomic sequences used as markers (Askree et al, 2004). A red line marks the telomere size of the wild-type strain.

SI Figure 4. Protein phosphatase activity plays a central role in determining telomere length. Nodes are color coded as in Figure 3.

SI Figure 5. The local-global tradeoff. Blue nodes: terminals, orange node: root, gray nodes: other members of the network. A. A toy example in which the optimal Steiner tree T_G is given by the dashed edges and the optimal shortest-paths solution T_L is given by the solid edges (the edge between t_k and r is shared between both solutions). Here, $F_L(T_G) = k(k+1)/2$, $F_L(T_L) = k + (k-1)\epsilon$. Hence, the approximation ratio attained by T_G w.r.t. the local criterion is $\Omega(k)$. In fact, when comparing T_G and T_L w.r.t. the combined function $c \cdot F_G + F_L$ we find that the approximation ratio of the Steiner-tree solution w.r.t. the combined function is $\Omega(k)$ as well. (Moreover, in this example, the minimum spanning tree heuristic for Steiner tree construction (Gilbert & Pollak, 1968) will output T_G. Hence, all approximation algorithms that use this heuristic as a starting point will follow the same approximation bounds.) B. An example in which the optimal Steiner tree T_G is given by the dashed lines and the optimal shortest-paths solution T_L is given

by the solid lines. Here $F_G(T_L) = k$ while $F_G(T_G) = 1 + \epsilon$ and, hence, the approximation ratio attained by T_L w.r.t. the global criterion is $\Omega(k)$. By comparing T_G and T_L w.r.t. the combined criterion we see that the approximation ratio achieved by the shortest-paths solution is $\Omega(k)$.

SI Figure 6. TLM network models. Green nodes: TLM proteins whose mutants have elongated telomeres; blue nodes: TLM proteins whose mutants have short telomeres; beige nodes - TLM protein from the literature whose effect (short or long) on telomere length is not readily available; red nodes: protein products of essential genes (not on the TLM list); grey nodes: protein products of non essential genes and not on the TLM list; purple: the *TELOMERE* anchor node. The paths from the TLM proteins to the telomere in the Steiner tree model tend to be very long while in the Charikar-0.25 and shortest paths the networks are centered around the telomere and the paths are substantially shorter.

SI Figure 7. Performance on the apoptosis data using different length penalties. Presented measures include functional coherence (fraction of functionally coherent pathways) and the predictive score. Length penalty values were determined according to the weight of an edge at the 10^{th} percentile and at the 50^{th} percentile.

SI Algorithm 1. Pseudocode of the Charikar-α algorithm. The procedure $DT(v, k')$ returns a lowest-density 1-level tree in the transitive closure G rooted at v and containing k' terminals. The density of a tree T is $d(T) = C(T)/[k(T)^{\alpha+0.5}]$ where $C(T)$ is the sum of edge weights in T, and $k(T)$ is the number of terminals it contains. We denote by $V(T)$ the vertex set of T.

Figures

Figure 1:

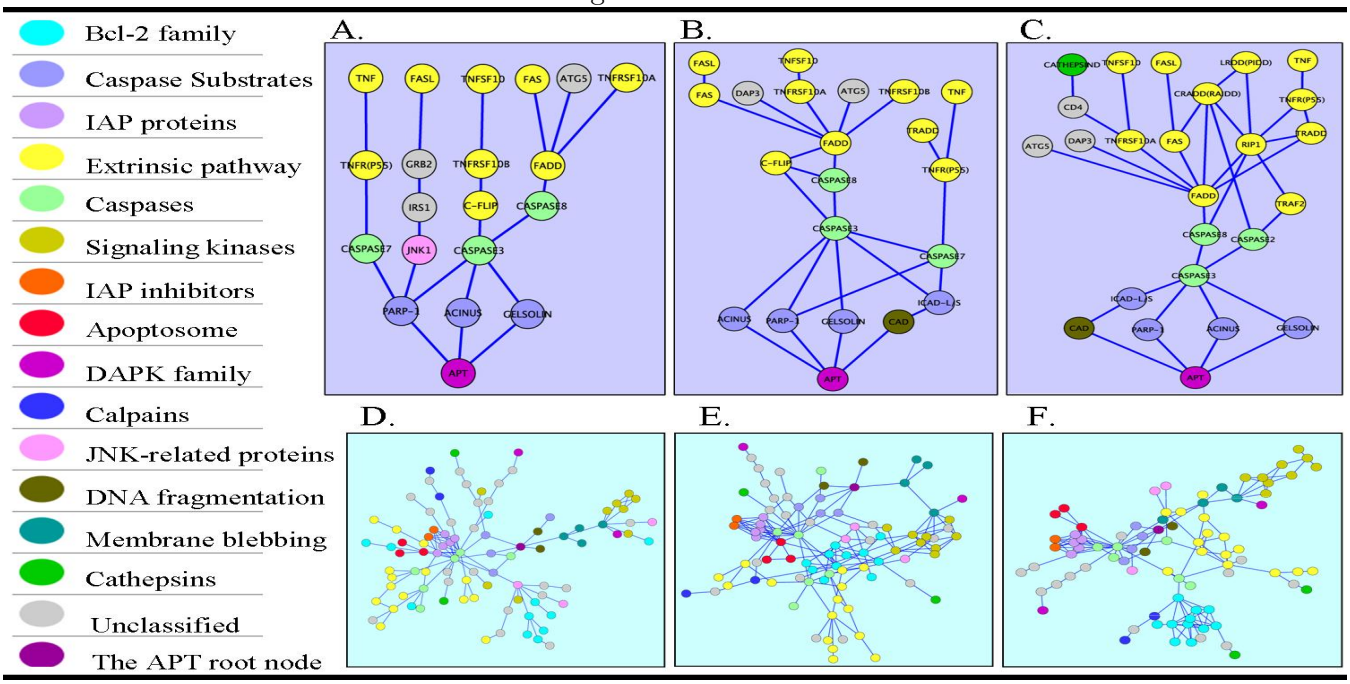


Figure 2:

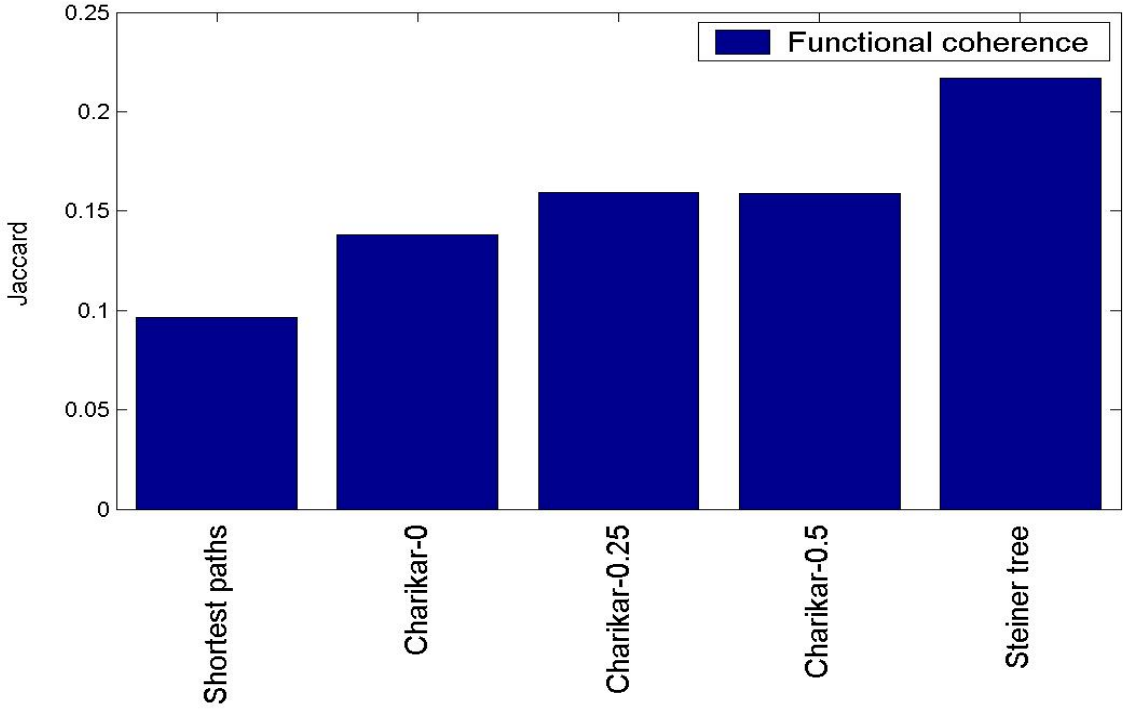

Functional coherence
Jaccard
0.25
0.2
0.15
0.1
0.05
0
Shortest paths
Charikar-0
Charikar-0.25
Charikar-0.5
Steiner tree

Figure 3:

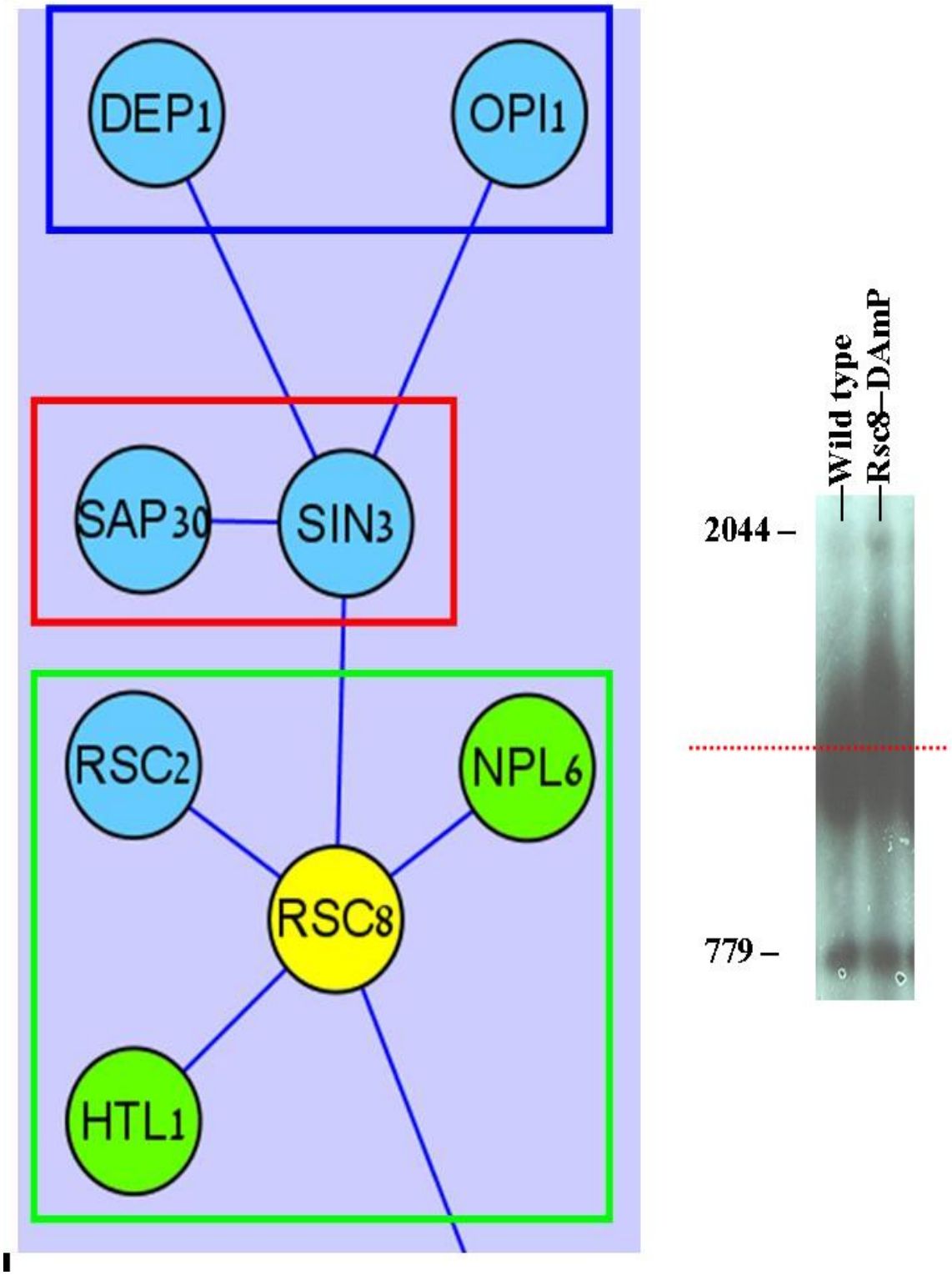

DEP1
OPI1
SAP30
SIN3
RSC2
NPL6
RSC8
HTL1
—Wild type
—Rsc8–DAmP
2044 –
779 –

Figure 4:

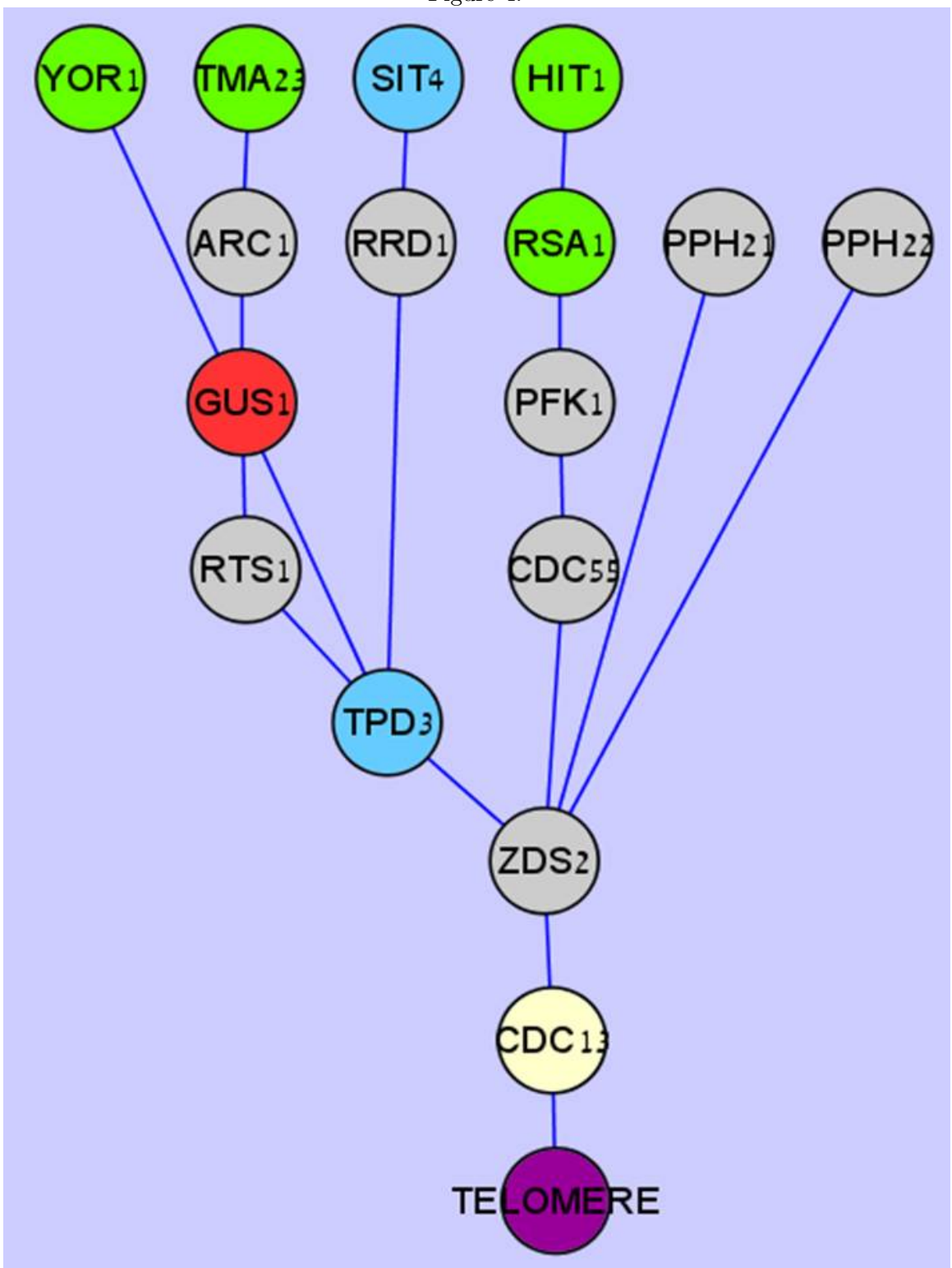

YOR1
TMA23
SIT4
HIT1
ARC1
RRD1
RSA1
PPH21
PPH22
GUS1
PFK1
RTS1
CDC55
TPD3
ZDS2
CDC13
TELOMERE

Figure 5:

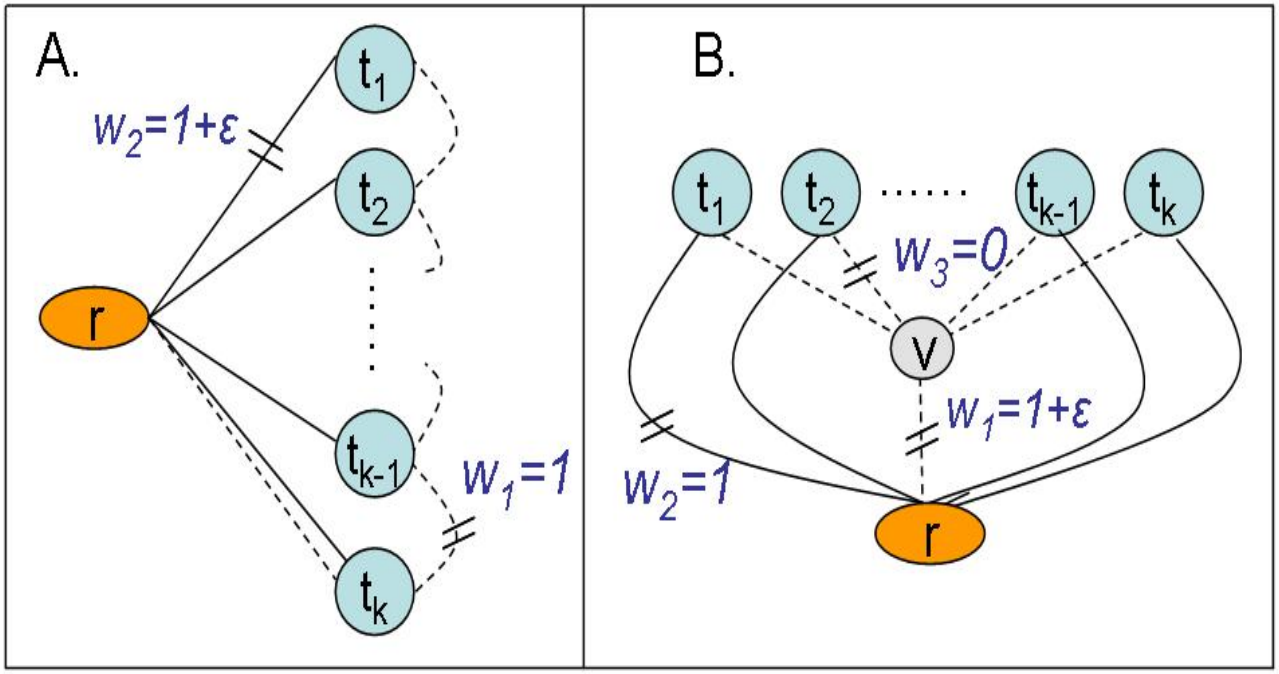

A.
$w_2=1+\varepsilon$
t_1
t_2
r
t_{k-1}
t_k
$w_1=1$
B.
t_1
t_2
.....
t_{k-1}
t_k
$w_3=0$
v
$w_1=1+\varepsilon$
$w_2=1$
r

Figure 6:

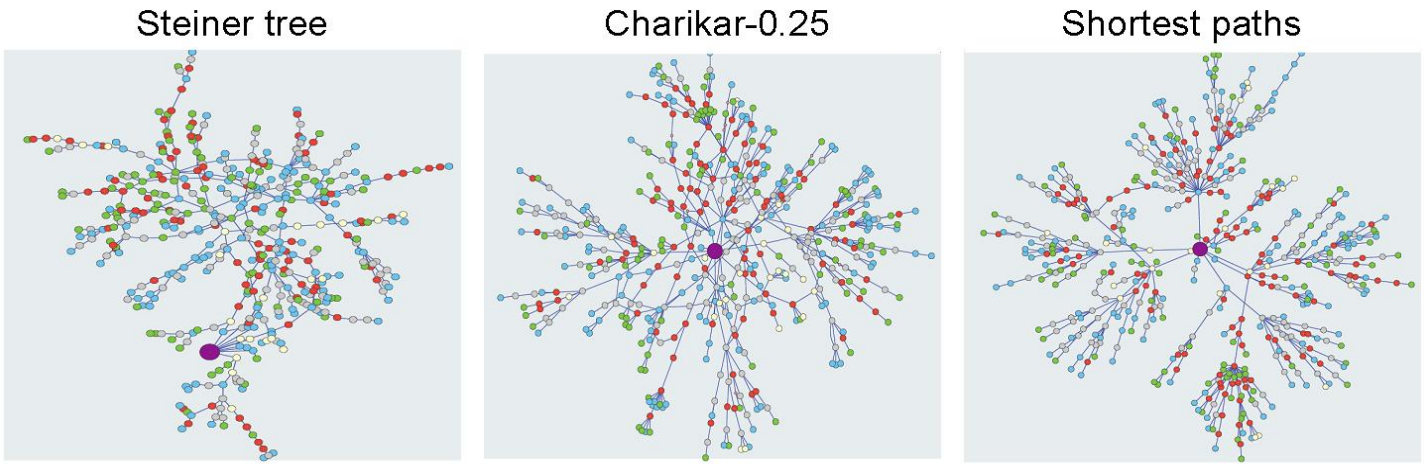
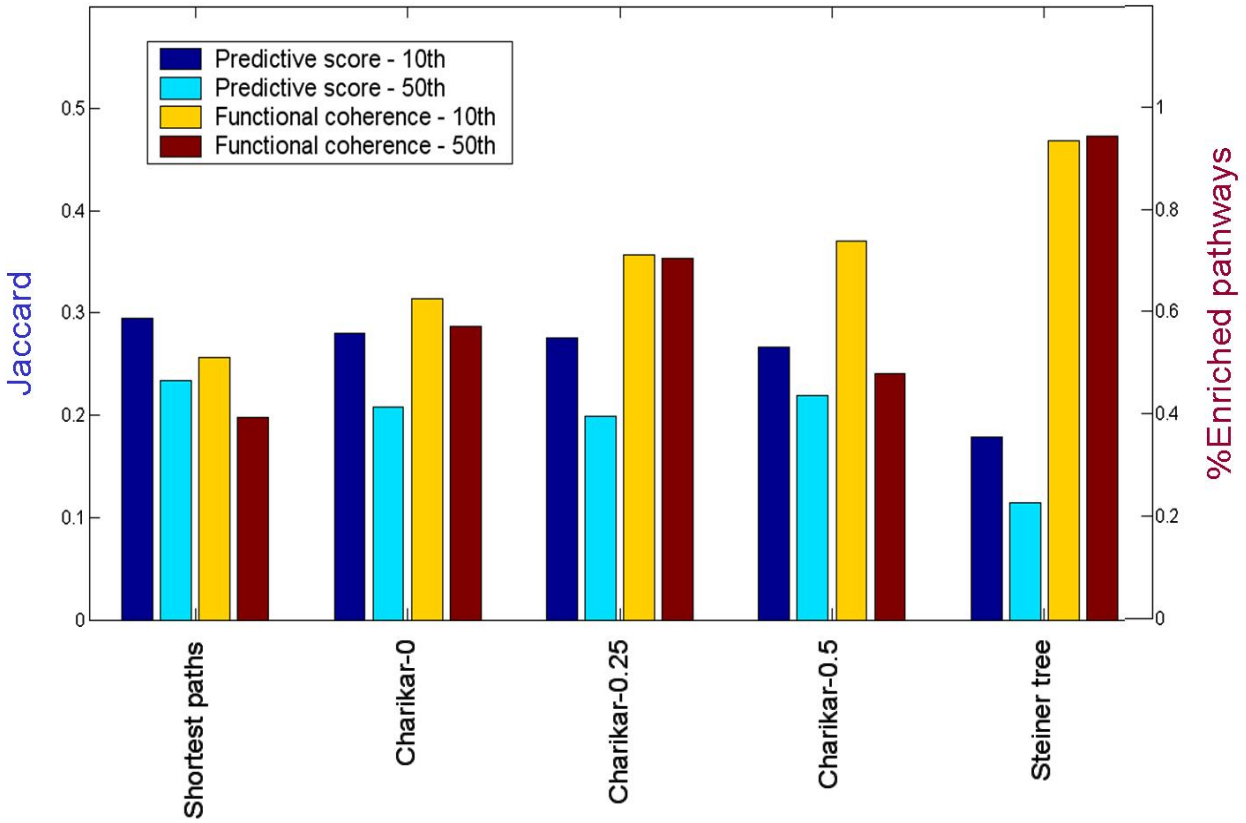


Figure 7:


Predictive score - 10th
Predictive score - 50th
Functional coherence - 10th
Functional coherence - 50th
Jaccard
%Enriched pathways
0
0.1
0.2
0.3
0.4
0.5
0
0.2
0.4
0.6
0.8
1
Shortest paths
Charikar-0
Charikar-0.25
Charikar-0.5
Steiner tree

Algorithm 1

procedure CHARIKAR-$\alpha(G, r, X, \alpha)$
  If not all terminals in X are reachable from r **then return** $\emptyset$
  $T \leftarrow \emptyset$
  $k \leftarrow |X|$
  while $k > 0$ **do**
    $T_{BEST} \leftarrow \emptyset$
    $d_{BEST} \leftarrow \infty$
    for each $v \in V$ and each k', $1 \leq k' \leq k$ **do**
      $T' \leftarrow DT(v, k') \cup \{(r, v)\}$
      If $d(T') < d_{BEST}$ **then**
        $T_{BEST} \leftarrow T', \quad d_{BEST} \leftarrow d(T')$
    end for
    $T \leftarrow T \cup T_{BEST}$
    $k \leftarrow k - |X \cap V(T_{BEST})|$
    $X \leftarrow X \setminus V(T_{BEST})$
  end while
  return T
end procedure

Tables

Table 1: Empirical approximation ratios of the different approaches on the TLM data.

Method	F_G	F_L	$c \cdot F_G + F_L$
Shortest paths	≤ 2.24	**1.00**	≤ 1.62
Charikar-0	≤ 2.22	1.01	$\leq$ **1.61**
Charikar-0.25	≤ 2.21	1.02	$\leq$ **1.61**
Charikar-0.5	≤ 2.22	1.06	≤ 1.63
Steiner-tree	$\leq$ **2.00**	2.75	≤ 2.37

Optimal value for F_G is bounded from below by half the score of the Steiner tree approximation algorithm (denoted OPT_G). The optimal value of F_L is the score of the shortest paths construction (denoted OPT_L). The optimal value for $F_L + c \cdot F_G$ is bounded from below by $2 \cdot OPT_L$. The normalization factor c is bounded from above by $\frac{OPT_L}{0.5 \cdot OPT_G}$. Considering the global optimization function F_G, the Charikar-α variants perform quite similarly to each other, and outperform the shortest paths construction. On the other hand, lower values of α yield a better approximation of the local measure F_L. In the combined objective the Steiner tree solution performs significantly worse than the rest of the algorithms.

Table 2: Empirical approximation ratios of the different approaches on the apoptosis data.

Method	F_G	F_L	$c \cdot F_G + F_L$
Shortest paths	$\leq \mathbf{2.42}$	**1**	≤ 1.71
Charikar-0	≤ 2.54	1.01	≤ 1.77
Charikar-0.25	≤ 2.26	1.03	$\leq \mathbf{1.65}$
Charikar-0.5	≤ 2.57	1.09	≤ 1.83
Steiner-tree	$\leq \mathbf{2.00}$	1.43	≤ 1.71

As in the TLM case, the best approximation for the combined objective is achieved by the Charikar-0.25 algorithm.

References

Askree S, Yehuda T, Smolikov S, et al (2004) A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length. *Proc Natl Acad Sci USA* **101**: 8658–8663.

Bao Q, Shi Y (2007) Apoptosome: a platform for the activation of initiator caspases. *Cell Death Differ* **14**: 56–65.

Branzei D, Foiani M (2006) The Rad53 signal transduction pathway: Replication fork stabilization, DNA repair, and adaptation. *Exp Cell Res* **312**: 2654–2659.

Cairns B, Lorch Y, Li Y, Zhang M, Lacomis L, Erdjument-Bromage H, Tempst P, Du J, Laurent B, Kornberg R (1996) RSC, an essential, abundant chromatin-remodeling complex. *Cell* **87**: 1249–1260.

Carey M, Li B, Workman J (2006) RSC exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation. *Mol Cell* **24**: 481–487.

Charikar M, Chekuri C, Cheung T, Dai Z, Goel A, Guha S, Li M (1999) Approximation Algorithms for Directed Steiner Tree Problems. *Journal of Algorithms* **33**: 73–91.

Counter C, Meyerson M, Eaton E, Weinberg R (1997) The catalytic subunit of yeast telomerase. *Proc Natl Acad Sci USA* **94**: 9202–9207.

Evans S, Lundblad V (1997) Positive and negative regulation of telomerase access to the telomere. *J Cell Sci* **113 Pt 19**: 3357—3364.

Fisher R (1948) Combining independent tests of significance. *Am Stat* **2**: 30.

Gatbonton T, Imbesi M, Nelson M, Akey J, Ruderfer D, Kruglyak L, Simon J, Bedalov A (2006) Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet* **2**: e35.

Gilbert E, Pollak H (1968) Steiner minimal trees. *SIAM J Appl Math* **16**: 1–29.

Gilbert E, Pollak H (2005) Tighter Bounds for Graph Steiner Tree Approximation. *SIAM J Disc Math* **19**: 122 – 134.

Goldberg D, Roth F (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* **100**: 4372–4376.

Grandin N, Reed S, Charbonneau M (1997) Stn1, a new Saccharomyces cerevisiae protein, is implicated in telomere size regulation in association with Cdc13. *Genes Dev* **11**: 512–527.

Gropl C, Hougardy S, Nierhoff T, Promel H (2001) Approximation algorithms for the Steiner tree problem in graphs. *In Steiner Trees in Industry, Cheng X and Du DZ (Eds) Kluwer Academic Publishers* : 235–279.

Harris M, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin G, Blake J, Bult C, Dolan M, Drabkin H, Eppig J, Hill D, Ni L, Ringwald M, et al (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **1**: D258–D261.

Kumar S (2007) Caspase function in programmed cell death. *Cell Death Differ* **14**: 32–43.

Lamping E, Luckl J, Paltauf F, Henry S, Kohlwein S (1994) Isolation and characterization of a mutant of Saccharomyces cerevisiae with pleiotropic deficiencies in transcriptional activation and repression. *Genetics* **137**: 55–65.

Locksley R, Killeen N, Lenardo M (2001) The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* **104**: 487–501.

Shachar R, Ungar L, Kupiec M, Ruppin E, Sharan R (2008) A Systems-level Approach to Mapping the Telomere-length Maintenance Gene Circuitry. *Mol Syst Biol* **4**: 172.

Sreenivas A, Carman G (2003) Phosphorylation of the yeast phospholipid synthesis regulatory protein Opi1p by protein kinase A. *J Biol Chem* **278**: 20673–20680.

Sun Z, Hampsey M (1999) A general requirement for the Sin3-Rpd3 histone deacetylase complex in regulating silencing in Saccharomyces cerevisiae. *Genetics* **152**: 921–932.

Viscardi V, Clerici M, Cartagena-Lirola H, Longhese M (2005) Telomeres and DNA damage checkpoints. *Biochimie* **87**: 613–624.

Wilson B, Erdjument-Bromage H, Tempst P, Cairns B (2006) The RSC chromatin remodeling complex bears an essential fungal-specific protein module with broad functional roles. *Genetics* **172**: 795–809.

Yosef N, Kaufman A, Ruppin E (2006) Inferring functional pathways from multi-perturbation data. *Bioinformatics* **22**: e539–546.

Youle R, Strasser A (2008) The BCL-2 protein family: opposing activities that mediate cell death. *Nat Rev Mol Cell Biol* **9**: 47–59.

Zakian V (1996) Structure, function, and replication of Saccharomyces cerevisiae telomeres. *Annu Rev Genet* **30**: 141–172.

Zelikovsky A (1997) A series of approximation algorithms for the acyclic directed Steiner tree problem. *Algorithmica* **18**: 99 – 110.