Gabr *et al. EURASIP Journal on Bioinformatics and Systems Biology* (2015) 2015:10
DOI 10.1186/s13637-015-0031-8

EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

# Computing interaction probabilities in signaling networks

Haitham Gabr[1*], Juan Carlos Rivera-Mulia[2], David M. Gilbert[2] and Tamer Kahveci[1]

**Abstract**

Biological networks inherently have uncertain topologies. This arises from many factors. For instance, interactions between molecules may or may not take place under varying conditions. Genetic or epigenetic mutations may also alter biological processes like transcription or translation. This uncertainty is often modeled by associating each interaction with a probability value. Studying biological networks under this probabilistic model has already been shown to yield accurate and insightful analysis of interaction data. However, the problem of assigning accurate probability values to interactions remains unresolved. In this paper, we present a novel method for computing interaction probabilities in signaling networks based on transcription levels of genes. The transcription levels define the signal reachability probability between membrane receptors and transcription factors. Our method computes the interaction probabilities that minimize the gap between the observed and the computed signal reachability probabilities. We evaluate our method on four signaling networks from the Kyoto Encyclopedia of Genes and Genomes (KEGG). For each network, we compute its edge probabilities using the gene expression profiles for seven major leukemia subtypes. We use these values to analyze how the stress induced by different leukemia subtypes affects signaling interactions.

**Keywords:** Biological networks, Signaling, Interaction probability, Reachability, Leukemia

## 1 Introduction

Biological networks describe how different molecules, such as proteins, interact with each other to carry out various cellular functions. Studying biological networks gives us deep insight into cellular mechanics and allows us to understand how biological processes are governed. Discovering signaling pathways [1], mapping transcription regulation [2], and identifying the reasons behind and the consequences of various disorders [3, 4] are only a few examples to many applications which are possible through studying biological networks.

Biological networks are often modeled as graphs, where each node denotes a molecule and each edge denotes an interaction. One of the critical factors that affects our analysis of biological networks is that their topologies are often uncertain. This uncertainty follows from the fact that key biological processes governing these interactions, like DNA replication, gene transcription, and epigenetic mutations, are themselves inherently uncertain events. For instance, in higher eukaryotes, DNA replication can start at different chromosome locations with different probabilities [5]. Also, different biological processes like replication timing, gene expression, and transcription regulation vary across different cell types [6–9], and also from healthy cases to different disorders [10, 11]. Probabilistic networks model this uncertainty in a mathematically sound manner. Briefly, a probabilistic biological network associates each edge of the underlying network with a probability value indicating the chance that the corresponding interaction takes place.

Taking the edge probabilities into account is extremely important in studying biological networks as they improve the accuracy of analysis of these networks and can lead to biologically significant observations that are impossible to achieve otherwise. Signaling pathway detection [1], network topology characterization [12], signal reachability [13], node centrality, and network stability [14] are just a few examples to the applications that have already been benefiting from this knowledge. Therefore, having

*Correspondence: hgabr@cise.ufl.edu
[1] Department of Computer & Information Science & Engineering, University of Florida, Gainesville, Florida, USA
Full list of author information is available at the end of the article

accurate knowledge of edge probabilities is of utmost importance.

In the literature, interaction probabilities are computed in several ways. MINT [15] and STRING [16], for instance, provide a confidence value for each interaction. Confidence value of an interaction represents the level of certainty in observing that interaction. This way of probability assignment compensates for the level of noise in the experiment used to observe the interaction. However, it does not account for the inherent stochasticity of the interaction events. Sharan et al. [17] addressed this problem by utilizing features like the volume of evidence present for the interaction, gene expression correlation, and network topology to learn the edge probabilities. This strategy however accounts only for the correlation between the interacting gene products, ignoring their relations with the rest of the network. *Thus, new methods which can compute edge probabilities by taking the entire network into consideration are direly needed.*

**Contributions** In this paper, we present a novel method for computing edge probabilities for a given signaling network topology. We use end-to-end signal reachability probabilities between pairs of genes to guide our computation. While it is hard to observe the probability of each individual interaction, target reachability values are much easier to observe experimentally. Moreover, they can be observed in different cell types or under different disorders, paving way for computing phenotype-specific edge probabilities.

Correlation between the transcription levels of genes has been widely used as the primary evidence for signaling and regulation [18–22]. Here, we also use gene expression correlation as the guide for signal reachability between gene pairs. For each pair of source (i.e., membrane receptor) and target (reporter, i.e., transcription factor) genes, we compute the normalized Pearson correlation value between their gene expression levels as the *empirical* signal reachability between that pair of genes. Our method computes the probability values for all the edges so that the resulting *computed* signal reachability probabilities for all source-target pairs are as close as possible to the input empirical reachability values. Given a network with $n$ edges, reachability probability can be expressed as an $n$th degree function of $n$ variables [23]. Optimizing this function in an exact manner requires solving a system of $n$ simultaneous derivative equations. The key challenge arises from the fact that computing the function itself has an exponential time complexity, equivalent to computing all combinations of $n$ objects. This makes exact optimization impossible even for medium-sized networks. To address this challenge, we develop a two-phase strategy. The first phase is global optimization using a genetic algorithm, where we search the entire space of possible edge probability assignments to obtain a good initial probability assignment. The second phase is local optimization using hill climbing technique. Here, we optimize the initial solution we found in the first phase by gradually improving the edge probabilities one edge at a time, until no further improvement is possible. More specifically, instead of optimizing all $n$ variables simultaneously, we seek to optimize the value of one variable at a time. That is, at each step, we consider only one edge probability for optimization, fixing the probability values of all other edges. We show that our method produces a result that is very close to the objective. Our experiments demonstrate that our method can compute edge probabilities with high accuracy. They also show that these probability values help in identifying specific genes and interactions that characterize major leukemia subtypes.

The rest of this paper is organized as follows. Section 2 describes the method in detail. Section 3 discusses our results. Section 4 concludes the paper.

## 2 Method

In this section, we explain our method for computing edge probability values of a given probabilistic signaling network. Our method consists of two phases: global optimization and local optimization. Section 2.1 describes the key notation needed to understand our method and formally defines the problem. Sections 2.2 and 2.3 discuss the global and local optimization phases respectively.

### 2.1 Preliminaries

Throughout the rest of this paper, we denote a probabilistic signaling network as a graph $G = (V, E, P)$, where $V$ denotes the set of nodes (i.e., genes), $E$ denotes the set of directed edges (i.e., interactions), and $P : E \rightarrow \mathbb{R} \cap [0, 1]$ denotes the function that returns the existence probability of each edge in $E$. We also define the two sets $S \subseteq V$ and $T \subseteq V$ as the sets of source nodes (i.e., receptor genes) and target nodes (i.e., reporter genes). We define the $|S| \times |T|$ matrix $C$ as the gene coexpression matrix, such that $C[s, t]$ is the absolute value of the Pearson correlation coefficient between the expressions of genes $s$ and $t$, for all $s \in S$ and $t \in T$. Given a probability function $P : E \rightarrow \mathbb{R} \cap [0, 1]$, we define the $|S| \times |T|$ matrix $R_P$ as the signal reachability matrix, such that $R_P[s, t]$ is the probability of a signal propagating successfully from $s$ to $t$, for all $s \in S$ and $t \in t$ using $P$. In these definitions, $C$ represents the *empirical* reachability probability between receptor and reporter genes based on their transcriptional activities. This is motivated by evidence of a strong link between gene coexpression and signaling and regulation [18–22]. On the other hand, $R_P$ denotes the *computed* reachability probability between the same gene pairs, having $P$ as the probability function. Thus, the Euclidean L2 norm $\|C - R_P\|_2$ is the error introduced by the

function $P$. Since we are only interested in the magnitude of the difference between $C$ and $R_P$, we used L2 norm to disregard the sign of this difference. Following from this observation, next we mathematically define the problem considered in this paper.

**Problem definition.** Given $V$, $E$, $S$, $T$, and $C$, find the function $P : E \rightarrow \mathbb{R} \cap [0,1]$ such that $\|C - R_P\|_2$ is minimum.

Notice that the problem above differs from the classical reachability problem. In the reachability problem, $P$ is known and the goal is to find $R_P$ [23]. On the other hand, in the problem considered in this paper, $P$ is not known. In fact, the goal is to compute $P$ with the guidance of $C$. That said, in order to understand our method in this paper, it is essential to know the original reachability problem well. In the following, we take a brief detour to summarize the PReach method that solves the reachability problem. For further details, we refer the reader to Gabr et al. [23].

Let $U = \{1, \ldots, n\}$, where $n = |E|$. Let $X$ and $Y$ be two sets of $n$ variables, where $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$. Let $\Theta$ be a subset of $U$. Let $S_1, \ldots, S_k$ be $k$ different subsets of $\Theta$. Let $x_{S_i} = \prod_{j \in S_i} x_j$ and $y_{S_i} = \prod_{j \in S_i} y_j$, where $i \in \{1, \ldots, k\}$. Let $x^*$ and $y^*$ be two free variables. Let $a_1, \ldots, a_k, b$ and $c$ be real numbers. PReach defines an *xy-polynomial* over $\Theta$ as $F = \sum_{i=1}^{k} a_i x_{S_i} y_{\Theta \setminus S_i} + bx^* + cy^*$. Except for the free variables, each term in the above summation contains each of the indices $j \in \Theta$ either as a product term $x_j$ or $y_j$.

PReach associates every edge $e_j \in E$ with a variable $x_j \in X$ and a variable $y_j \in Y$, where $j \in U$. In this notation, $x_j$ and $y_j$ represent the cases where $e_j$ is present and absent, respectively. In the above summation, each of the non-free terms $a_i x_{S_i} y_{\Theta \setminus S_i}$ corresponds to a combination where $e_j$ is present $\forall j \in S_i$ and absent $\forall j \in \Theta \setminus S_i$, and $a_i$ is the probability of observing this specific combination. The free variable $x^*$ represents the case where $T$ is reachable from $S$, and $b$ designates its probability. Inversely, the free variable $y^*$ represents the case where $T$ is unreachable from $S$, and $c$ designates its probability.

Let $p_i = P(e_i)$ and $q_i = 1 - p_i, \forall e_i \in E$. PReach starts by associating every edge $e_i \in E$ with a binomial $p_i x_i + q_i y_i$. It then proceeds by multiplying these binomials together into a growing $xy$-polynomial. After each multiplication step, PReach checks the polynomial for the non-free terms that can be *collapsed* into one of the two free terms as follows. For any of the non-free terms $a_i x_{S_i} y_{\Theta \setminus S_i}$, if the edges associated with $S_i$ form at least one path from $S$ to $T$, it replaces those terms with $a_i x^*$. Inversely, if the edges associated with $\Theta \setminus S_i$ form at least one cut between $S$ and $T$, it replaces those terms with $a_i y^*$. Any further multiplication of a new term $p_i x_i$ with $bx^*$ results in $bp_i x^*$. Similarly, $(p_i x_i)(cy^*) = cp_i y^*$, $(q_i y_i)(bx^*) = bq_i x^*$, and $(q_i y_i)(cy^*) = cq_i y^*$. The reachability problem is a computationally hard problem; it belongs to the #P-complete class [24]. However, thanks to the repeated application of the collapsing operation, PReach tries to avoid exponential growth of the size of the $xy$-polynomial.

### 2.2 Phase 1: global optimization

We are now ready to describe the method developed in this paper. The first phase of our method is a genetic algorithm to find a population of probability functions as an initial candidate solution. Note that this is a best-faith solution that will be further optimized in the second phase of our method.

We represent a candidate solution as a vector with $|E|$ entries and denote it with $\psi$, where the $i$th entry $\psi[i]$ is the probability assigned to edge $e_i \in E$. Let us denote the computed reachability matrix obtained using the solution $\psi$ as $R_\psi$. That is, $\forall s \in S$, $\forall t \in T$, $R_\psi[s,t]$ is the signal reachability probability between $s$ and $t$, computed based on edge probabilities in $\psi$ (see Section 2.1). We define the fitness $F_\psi$ of a candidate solution $\psi$ as $1 - \frac{\|C - R_\psi\|_2}{|S \times T|}$. In this formulation, the fitness $F_\psi$ takes a value in the $[0,1]$ interval. A larger value indicates a better solution. In the extreme case when the empirical and computed reachability probabilities are identical (i.e., $C = R_\psi$), $F_\psi$ is equal to 1, indicating that the solution is 100 % accurate.

Our genetic algorithm consists of four steps: initialization, crossover, mutation, and selection. We elaborate on these steps next.

1. **Initialization** We start by generating a set $\Psi$ of random candidate solutions. These solutions serve as the seed population of solutions. We generate each seed candidate solution $\psi \in \Psi$ by assigning a random number between 0 and 1 to each entry in $\psi$. We then compute the fitness values $F_\psi, \forall \psi \in \Psi$. In our experiments, we set the population size to $|\Psi| = 50$, thus generate 50 random seeds.
2. **Crossover** This step improves the solutions in the set $\Psi$ by combining pairs of existing solutions, also known as the crossover operation. To do that, We define a *gap* value $g_\psi$ for every $\psi \in \Psi$ as $g_\psi = \sum_{i=1}^{|S|} \sum_{j=1}^{|T|} C[i,j] - R_\psi[i,j]$. The value of $g_\psi$ shows how much the reachability $R_\psi$, computed based on the solution $\psi$, deviates from the target $C$ in total. A positive gap value indicates that the solution $\psi$ underestimates the probability of some of the edges. Inversely, a negative gap value indicates that the solution $\psi$ overestimates the probability of some of the edges. We then randomly select two solutions $\psi_1$ and $\psi_2$ from $\Psi$ using biased sampling, where the chance of selecting a sample $\psi_i$ is directly proportional to its fitness $F_{\psi_i}$. We use $\psi_1$ and $\psi_2$ to generate a new candidate solution as follows. For each entry $i \in \{1, .., |E|\}$, we choose either entry $\psi_1[i]$

or $\psi_2[i]$ based on which is more likely to produce a candidate solution with a higher fitness. There are three possible scenarios: if both $g_{\psi_1}$ and $g_{\psi_2}$ are positive, both $\psi_1[i]$ and $\psi_2[i]$ are possibly underestimated, so we choose the higher. Inversely, if both $g_{\psi_1}$ and $g_{\psi_2}$ are negative, both $\psi_1[i]$ and $\psi_2[i]$ are possibly overestimated, so we choose the lower. If one of $g_{\psi_1}$ and $g_{\psi_2}$ is positive and the other is negative, then we randomly select between $\psi_1[i]$ and $\psi_2[i]$, where the chance of each is proportional to the fitness of its corresponding solution. We expect this strategy to produce a new solution that is better than both $\psi_1$ and $\psi_2$, as we reduce the gap value while constructing it. We repeat the crossover step 50 times (i.e., $|\Psi|$ times) and include the resulting solutions to $\Psi$.

3. **Mutation** In this step, our genetic algorithm aims to avoid local minima by adding a small amount of random diversity to the existing set of solutions. More specifically, for each solution $\psi \in \Psi$, we iterate over all entries $\psi[i]$. For each entry $\psi[i]$, we perform a Bernoulli trial with probability of 0.01. If the trial yields success, the entry value is replaced with a new value drawn uniformly at random from the range [0, 1].
4. **Selection** After crossover and mutation, the size of $\Psi$ doubles to 100. This step ensures the set of solutions in $\Psi$ does not grow. To do this, from the 100 solutions in $\Psi$, we select five which have the highest fitness values. Additionally, we randomly select another 45 solutions from the remaining 95, where every solution has a chance of selection that is proportional to its fitness. We remove the non-selected 50 solutions from $\Psi$.

We repeat the crossover, mutation, and selection steps for a large number of iterations, updating the population $\Psi$ each time. The number of iterations needed for convergence depends on the size and properties of the target network and is a matter of trial and error. We then select the solution which has the highest fitness in the final population as the output of this phase. We use it as an input to our next local optimization phase.

### 2.3 Phase 2: local optimization

At the end of the first phase, we have a solution $\psi$ that has the highest fitness value in the entire population $\Psi$. Although $\psi$ is expected to yield small errors in signal reachability, it is not necessarily optimal. In this phase, we develop a hill climbing algorithm, which gradually alters the probability assignment of each edge in the solution, one edge at a time. At each step, it ensures that the probability assignment $\psi[e]$ of the edge $e$ being altered becomes optimal (i.e., yields the highest possible fitness value) given the probability assignment of all other edges. We continue altering the solution until no edge probability value can be altered without increasing $\|C - R_\psi\|_2$. In the following, we describe in detail how at a given step we alter one probability assignment $\psi[e]$, given all other values in $\psi$.

**Optimizing a single edge probability** Assume that for only one edge $e \in E$, the probability $p_e$ of this edge is unknown. Also, assume that the probability values of all the remaining edges in $E - \{e\}$ are known. Here, we compute the value of $p_e$ that guarantees to minimize the reachability error $\|C - R_\psi\|_2$. For this purpose, we develop a new method which is built on the PReach method [23].

Unlike PReach, our method allows one of the edge probabilities $p_e$ to be a variable. This additional unknown alters the form of the $xy$-polynomial constructed by PReach (see Section 2.1) as the unknown $p_e$ can get multiplied by all the terms of the original $xy$-polynomial. This new variable can increase the polynomial size dramatically, depending on the combination of the terms in the polynomial. We avoid this problem through a simple observation that the final $xy$-polynomial is independent of the order in which we multiply individual edge binomials. Following from this observation, we defer the multiplication of the edge binomial corresponding to $e$ until all other edge binomials are multiplied. Thus, until before the edge binomial of $e$ is multiplied, our method yields the same intermediate $xy$-polynomial as PReach. After multiplying the final binomial by the intermediate $xy$-polynomial, the coefficient of $x^*$ in the final $xy$-polynomial has the form $\alpha + \beta p_e$, where $\alpha$ and $\beta$ are real numbers. We mathematically deduce the values of $\alpha$ and $\beta$ in the following theorem.

**Theorem 1.** *Assume the binomials of all edges except that of e has been already multiplied into the xy-polynomial* $F = \sum_{i=1}^{l} a_i x_{S_i} y_{\Theta \setminus S_i} + bx^* + cy^*$, *where the terms* $a_i x_{S_i} y_{\Theta \setminus S_i}$ $\forall i \in \{1, \ldots, l\}$ *are l terms that are not yet collapsed into either* $x^*$ *or* $y^*$. *The reachability probability, which is the final coefficient of* $x^*$ *after multiplying the binomial of e, is* $\alpha + \beta p_e$, *where* $\alpha = b$ *and* $\beta = \sum_{i=1}^{l} a_i$.

*Proof.* Multiplying the $e$ binomial $(p_e x_e + (1 - p_e) y_e)$ by $F$, the final $xy$-polynomial $F_{\text{final}}$ has the following form.

$$
\begin{aligned}
F_{\text{final}} = & \\
& (p_e x_e + (1 - p_e) y_e) \left( \sum_{i=1}^{l} a_i x_{S_i} y_{\Theta \setminus S_i} + bx^* + cy^* \right) \\
& = p_e x_e \sum_{i=1}^{l} a_i x_{S_i} y_{\Theta \setminus S_i} + (1 - p_e) y_e \sum_{i=1}^{l} a_i x_{S_i} y_{\Theta \setminus S_i} \\
& \quad + p_e b x_e x^* + (1 - p_e) b y_e x^* \\
& \quad + p_e c x_e y^* + (1 - p_e) c y_e y^*
\end{aligned}
$$

Since $e$ is the last edge to multiply in the network, it is guaranteed that $x_e x_{S_i}$ and $y_e y_{\Theta \setminus S_i}$ will collapse to $x*$ and $y^*$, respectively, for all $i \in \{1, \ldots, l\}$ [23]. Also, we already know that $x_e x^* = x^*$ and $y_e y^* = y^*$ for any edge $e$. Thus, we have

$$\begin{aligned} F_{final} = & \\ & p_e \sum_{i=1}^{l} a_i x^* + (1-p_e) \sum_{i=1}^{l} a_i y^* \\ & + p_e b x^* + (1-p_e) b x^* + p_e c y^* + (1-p_e) c y^* \\ & = \left( p_e \sum_{i=1}^{l} a_i + p_e b + (1-p_e) b \right) x^* \\ & + \left( (1-p_e) \sum_{i=1}^{l} a_i + p_e c + (1-p_e) c \right) y^* \\ & = \left( p_e \sum_{i=1}^{l} a_i + b \right) x^* + \left( (1-p_e) \sum_{i=1}^{l} a_i + c \right) y^* \end{aligned}$$

The reachability probability is the final coefficient of $x^*$. Therefore its value is $p_e \sum_{i=1}^{l} a_i + b$. i.e., $\alpha = b$ and $\beta = \sum_{i=1}^{l} a_i$. This means that after multiplying the binomials of all the edges except $e$, $\alpha$ is the coefficient of $x^*$, and $\beta$ is the sum of the coefficients of all non-free terms. □

Notice that the coefficient of $x^*$ (i.e., $\alpha + \beta p_e$) is also a polynomial of first degree in $p_e$. Using this observation, we solve for the objective of this paper (which is to minimize $\| C - R_\psi \|_2$) by solving for $p_e$ as follows. We first compute $R_\psi[s,t] = \alpha_{st} + \beta_{st} p_e, \forall (s,t) \in S \times T$. We then derive the optimal value for $p_e$ as:

$$\begin{aligned} \text{Minimize} \quad & \| C - R_\psi \|_2 \\ & = \sum_{(s,t) \in S \times T} \left( C[s,t] - R_\psi[s,t] \right)^2 \\ & = \sum_{(s,t) \in S \times T} (C[s,t] - \alpha_{st} - \beta_{st} p_e)^2 \end{aligned}$$

To solve the minimization function above, we equate its derivative to zero.

$$\begin{aligned} & \frac{d}{dp_e} \sum_{(s,t) \in S \times T} (C[s,t] - \alpha_{st} - \beta_{st} p_e)^2 = 0 \\ & \therefore \sum_{(s,t) \in S \times T} -2\beta_{st} (C[s,t] - \alpha_{st} - \beta_{st} p_e) = 0 \\ & \therefore \sum_{(s,t) \in S \times T} -2\beta_{st} (C[s,t] - \alpha_{st}) + p_e \sum_{(s,t) \in S \times T} 2\beta_{st}^2 \\ & = 0 \end{aligned}$$

Thus, we get

$$p_e = \frac{\sum_{(s,t) \in S \times T} \beta_{st} (C[s,t] - \alpha_{st})}{\sum_{(s,t) \in S \times T} \beta_{st}^2}$$

The formula above constitutes the optimal value for $p_e$ given all other probability values. However, there is no guarantee that optimal value of $p_e$ falls within the proper probability range of $[0,1]$. This is because the derivation above gives the optimal result across all real numbers. However, by taking the second derivative of the objective function, one can easily see that the objective function is continuous, convex, and has only one solution to $\frac{d}{dp_e} = 0$. This implies that the closer the value of $p_e$ is to its unconstrained optimal value, the smaller the error is in the objective function. Therefore, if the optimal value of $p_e$ is above 1, we replace it with 1. If it is below 0, we replace it with 0. This way, we find the best possible value for $p_e$ in $[0,1]$.

## 3 Experimental results

In this section, we experimentally evaluate our method on four major signaling networks from the Kyoto Encyclopedia of Genes and Genomes (KEGG), including cell cycle, programmed cell death, and immune response regulation pathways (see Table 1 for dataset details). We use the gene expression samples for the leukemia subtypes from Zhang et al. [10]. This dataset contains gene expression values for 413 patients, each having one of seven different leukemia subtypes (six B-ALL subtypes plus T-ALL). We use this dataset as it provides a large number of samples for a wide spectrum of leukemia subtypes. We perform a comprehensive comparative analysis of how interaction probabilities vary across these leukemia subtypes (Section 3.2). Using the interaction probability values we find, we also compute gene centrality values for all network and leukemia subtype combinations (Section 3.3). We finally extract the genes which behave differently in specific combinations of network and leukemia subtype, and analyze the significance of these genes in these combinations (Section 3.4).

**Table 1** Networks used in our experiments, their sizes (nodes and edges), running time of our method to compute their interaction probabilities, and the quality of the resulting probabilities. For every network, time is the average running time over the seven leukemia subtypes in seconds, and quality is the average result quality over the seven leukemia subtypes

| Network | Nodes | Edges | Time (s) | Quality (%) |
|---|---|---|---|---|
| Apoptosis | 48 | 58 | 77.78 | 95.37 |
| Cell cycle | 66 | 79 | 274.78 | 96.08 |
| Complement and coagulation | 57 | 67 | 126.57 | 96.88 |
| Chemokine | 51 | 62 | 302.86 | 95.4 |

### 3.1 Comparison with logistic regression

In this section, we present comparative analysis of the results of our method against the logistic regression method presented by Sharan et al. [17]. Throughout this section, we refer to our method as PReach, and to the logistic regression method as LogReg. LogReg learns interaction probabilities through three features: available evidence, interactor small-world properties, and their gene expression. The latter is different across the leukemia subtypes, therefore LogReg produces different probability values for each subtype. In addition to the four networks described in Table 1, we also use four more networks to obtain more conclusive results. The additional four networks are ErbB, Wnt, NF-kappaB, and p53 from KEGG.

First, we compare the edge probability values produced by both methods. For each network, we run both methods once for each leukemia subtype. For every pair of network and subtype, we compute the average edge probability for both methods. We then compute the log of the ratio between them for comparison. Figure 1a shows the results. We observe from the figure that PReach produces higher probabilities on average for all pairs of network and subtypes. The biggest gap between PReach and LogReg occurs in Wnt and complement and coagulation cascades (ccc).

Next, we compare the quality of the results of PReach vs LogReg. There exists no ground truth to compare the edge probability values against. However, we compare the outputs of the two methods with respect to two measures. First, we inspect how spread is the output probabilities across the $[0,1]$ spectrum. To do this, we divide the $[0,1]$ range into a set of ten bins $B = \{[0,0.1),[0.1,0.2),\ldots,[0.9,1]\}$, and count the number of times a probability appears in each bin. For every pair of network and subtype, we compute the entropy of the method as $-\sum_{b\in B} p(b)logp(b)$, where $p(b)$ is the number of values in the bin $b$ divided by the total number of values. These entropy values are higher if the values are more spread across the $[0,1]$ spectrum, and lower if they are crowded in a less fraction of the spectrum. Figure 1b shows the results. We observe that the entropy in PReach is higher than LogReg in almost all cases. This means that PReach output probabilities are more spread across the $[0,1]$ spectrum, while those of LogReg tend to be more discrete. More detailed inspection reveals that this happens because LogReg assigns similar probability values to most of the interactions most of the time (results not shown due to space limit). Thus, it fails to provide fine-grained distinction between the likelihoods of interactions, while our method successfully provides such distinction.

To further investigate the results quality of both methods, we inspect how much each method differentiates leukemia subtypes with respect to their edge probability values. For every pair of network and subtype, we arrange the edge probability values produced by each method in a vector with the same ordering. Next, for a given network, we compute the Euclidean distance between the vectors of each pair of leukemia subtypes. Then for every subtype, we compute the average distance between its vector and those of all other subtypes. We compare this average distance when computed using PReach versus LogReg. Figure 1c shows the results. We observe that the distances computed based on PReach are higher in the vast majority of network-subtype pairs. This means that our method can differentiate leukemia subtypes while LogReg fails to do that.

### 3.2 Interaction probability in leukemia

In this experiment, we explore the differences on interaction probabilities of signaling networks in distinct leukemia subtypes. Our aim is to identify specific gene interaction differences between distinct leukemia subtypes. To achieve this, we use our method to compute interaction probability values for the KEGG signaling networks. For each network, we run our method seven times, once for each leukemia subtype. Before conducting detailed analysis, however, we first need to validate that our method is computationally feasible, that it scales to networks under consideration. To evaluate its performance, we measure the time our method takes to compute the probabilities of all the interactions for every network in every leukemia subtype. Also, based on our original optimization target (see Section 2.1), we need to know how accurate our results are (i.e., how close the computed reachability probability $R_\psi$ is to the input $C$). To do this, we measure the quality of the resulting interaction probabilities as $1-\frac{\|C-R_\psi\|}{|S\times T|}$. The closer this value is to 100 %, the better the quality is.

Table 1 shows the size of each network along with its average time and quality over the seven leukemia subtypes. Our results demonstrate that our method easily scales to the networks under consideration. It computes the interaction probabilities in about 5 min or less for all the networks we tested. More importantly, our method is highly accurate. The computed reachability values deviate from the empirical reachability values by less than 5 % for all the networks. These results are highly encouraging as they show that our method is both accurate and has practical running time. Thus it can be applied on real datasets to compute interaction probabilities.

Next, we analyze the differences in interaction probabilities across leukemia subtypes. For each network, we represent each leukemia subtype by a vector of the edge probabilities computed for it. We then compute a hierarchical clustering of these vectors. Figure 2 shows the results. From the figure, we observe that the probability value of some interactions vary significantly across

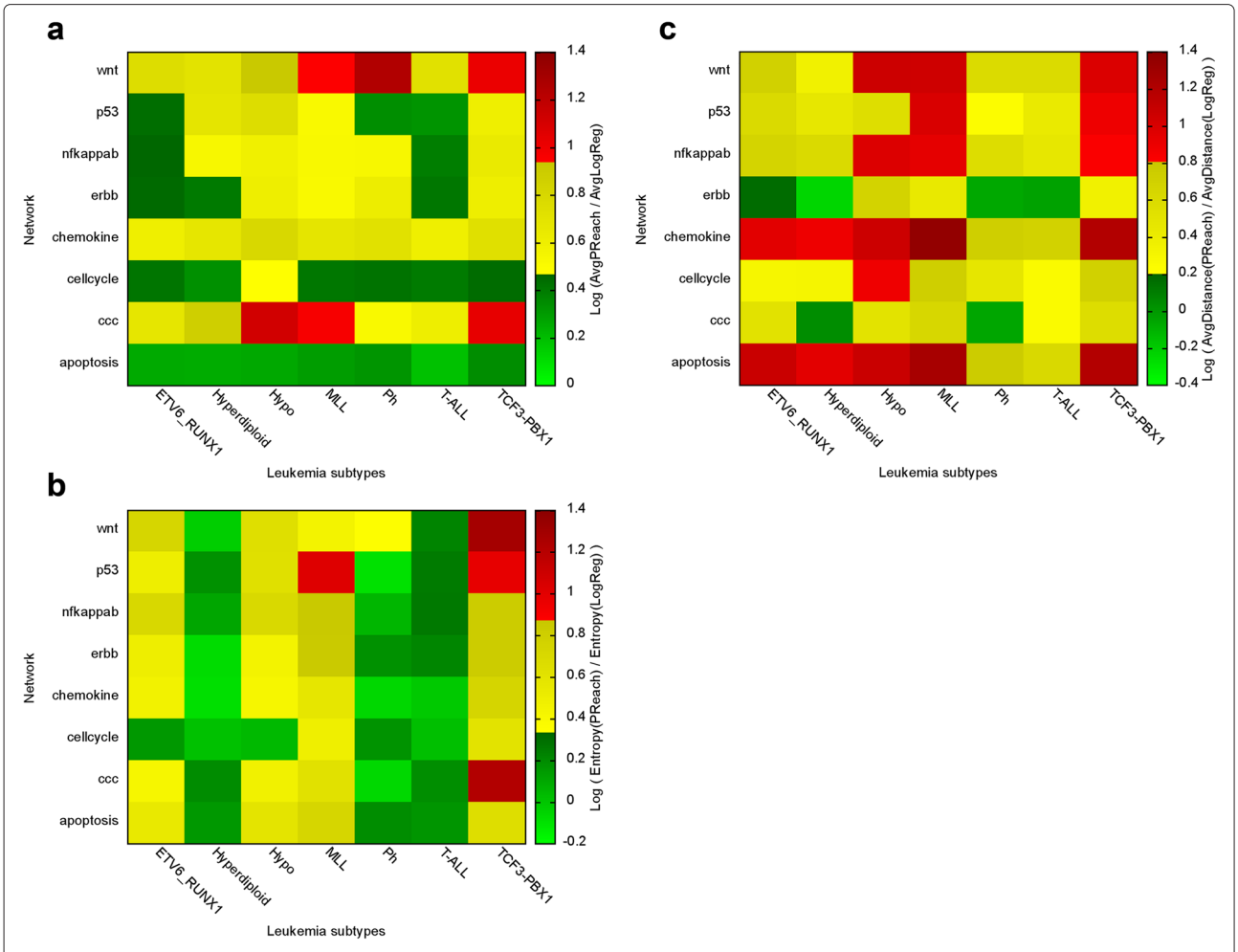


**Fig. 1** Comparison of PReach vs LogReg for computing edge probabilities in different leukemia subtypes. *X*-axis: leukemia subtypes. *Y*-axis: networks. All values are computed for both methods and compared using logarithm of the ratio (i.e., more than zero means PReach is higher). **a** Average edge probability. **b** Entropy of edge probability distribution. **c** Average distance between the edge probabilities of a each subtype and other subtypes

different leukemia subtypes. For instance, CASP3 is the target in three different apoptosis interactions whose probability in a subtype is at least 2 standard deviations away from their mean values among other subtypes. These interactions are (CASP10 → CASP3) in hyperdiploid, (CASP12 → CASP3) in T-ALL, and (BIRC8 → CASP3) in Ph (circled in Fig. 2a). Similarly, CHEK1 is the source in two different cell cycle interactions whose probability in a subtype is at least 2 standard deviations away from their mean values among other subtypes. These interactions are (CHEK1 → CDC25A) in T-ALL, and (CHEK1 → TP53) in MLL (circled in Fig. 2b). CASP3 is already linked to B-cell lymphoma [25], lung [26], skin [27], breast [28], and other cancers. Defects in apoptosis signaling and cell cycle pathways play an essential role in leukemogenesis. CASP3 is an effector caspase that has been associated with B-cell lymphoma [25], lung [26], skin [27], breast [28], and other cancers. Moreover, regulation of CASP3 activation has been linked to the prognosis and remission in B-ALL [29]. Notably, in the three leukemia subtypes with different apoptotic signals targeting CASP3, the programmed cell death is inhibited; interactions of CASP3 with its activators are weaker in hyperdiploid and T-ALL (CASP10 → CASP3 and CASP12 → CASP3, respectively) while the interaction with its inhibitor is increased in B-ALL with Philadelphia chromosome (BIRC8 → CASP3). CHEK1 is a cell cycle checkpoint response protein that is linked to oral squamous cell carcinoma [30] and colorectal cancer [31]. Recently, increased levels of CHEK1 have been associated to B-ALL and T-ALL [32]. Our observation makes both CASP3 and CHEK1 strong candidates for investigation in their respective subtypes of leukemia.

Additionally, the hierarchy of the leukemia subtypes gives an insight about which subtypes have similar signaling behavior. T-ALL and TCF3-PBX1 are closest to each other in apoptosis, complement and coagulation, and

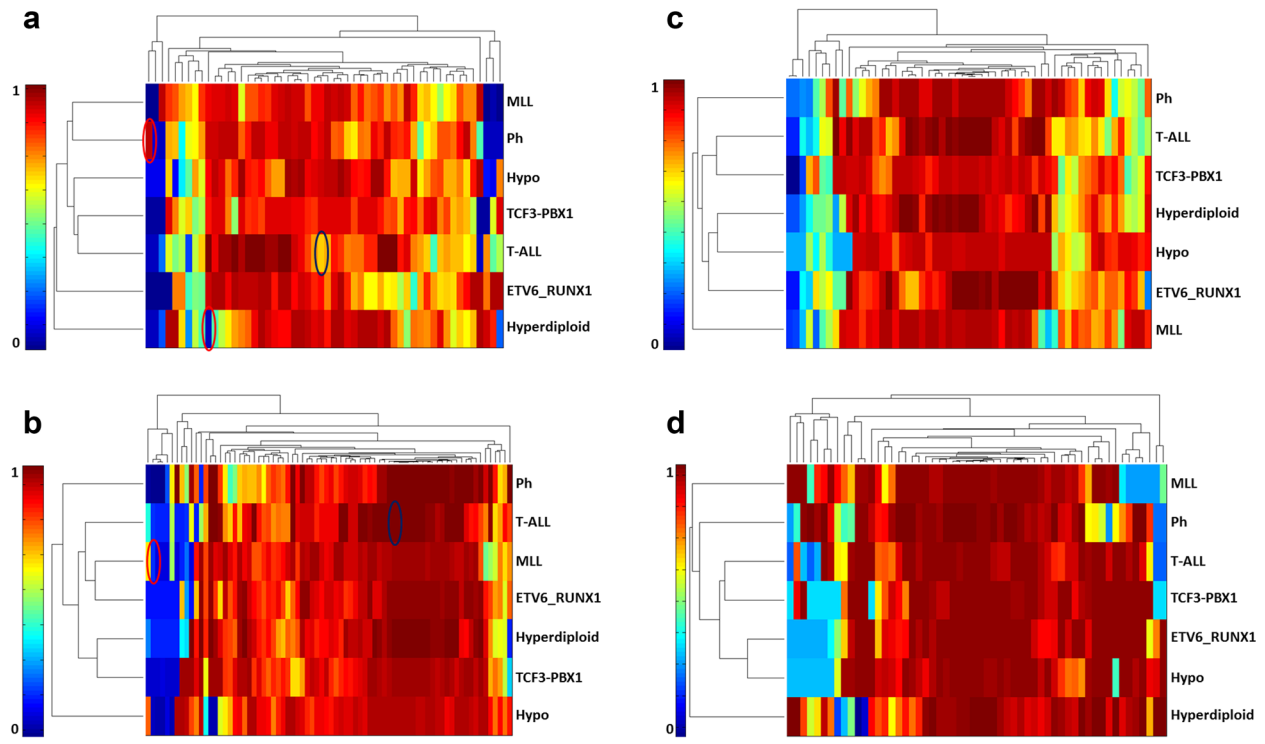


**Fig. 2** Interaction probability values of the four KEGG networks in seven leukemia subtypes. **a** Apoptosis, **b** cell cycle, **c** complement and coagulation, and (**d**) chemokine. Rows represent leukemia subtypes, and columns represent network interactions. Both rows and columns are hierarchically clustered

chemokine, noticeably more distant in cell cycle. Hyperdiploid is very similar to TCF3-PBX1 in cell cycle, but more distant from it in the other three networks. In fact, hyperdiploid is the most distant from all other subtypes in both apoptosis and chemokine. This information can guide us to build on the existing knowledge about signaling behavior in a certain subtype, using appropriate experiments, to develop new findings about other subtypes with similar behavior.

### 3.3 Gene centrality for leukemia subtypes

Next, we use the interaction probability values we computed in Section 3.2 to compute centrality values of the genes in each network. Briefly, we compute the centrality of a gene as its contribution to signal reachability probability between all pairs of genes (see Gabr et al. [14] for details). We compute centrality values for the genes in each of the four networks for each of the seven leukemia subtypes. For each network, we represent each leukemia subtype by a vector of the node centrality values computed for it. We then compute a hierarchical clustering of these vectors. Figure 3 presents the results.

Interestingly, the figure shows variation of centrality value across different leukemia subtypes for only a small number of genes. Notice that this variation is not as diverse as that of interaction probability values (see Fig. 2). In apoptosis, BID is the top outstanding gene in hyperdiploid, with centrality values 2.8 standard deviations higher than their mean centrality in other subtypes (circled in Fig. 3a). We notice loss of centrality for other key regulators of apoptosis like RIPK1 and CASP7 in ETV6_RUNX1 and Ph, respectively (circled in Fig. 3a). Similarly, cell cycle regulators like CDK1 and PLK1 have an increased centrality in hypo, with centrality values 2.6 standard deviations higher than their mean centrality in other subtypes (circled in Fig. 3b). BID remains as key regulator in hyperdiploid but its centrality is lost in other samples, suggesting a disruption of the programmed cell death regulation in most of the leukemia subtypes. RIPK1 and CASP7 are linked to colorectal cancer [33]. CDK1 and PLK1 induce cell cycle progression and have been associated with distinct types of cancer [34–36] including leukemia and lymphoma [37–39]. Our results suggest that these genes are interesting targets for studying in the scope of their respective leukemia subtypes.

### 3.4 Enrichment analysis of outstanding genes

Following from the previous results, we want to know which network plays a key role in a certain leukemia subtype. In other words, we want to know which network's

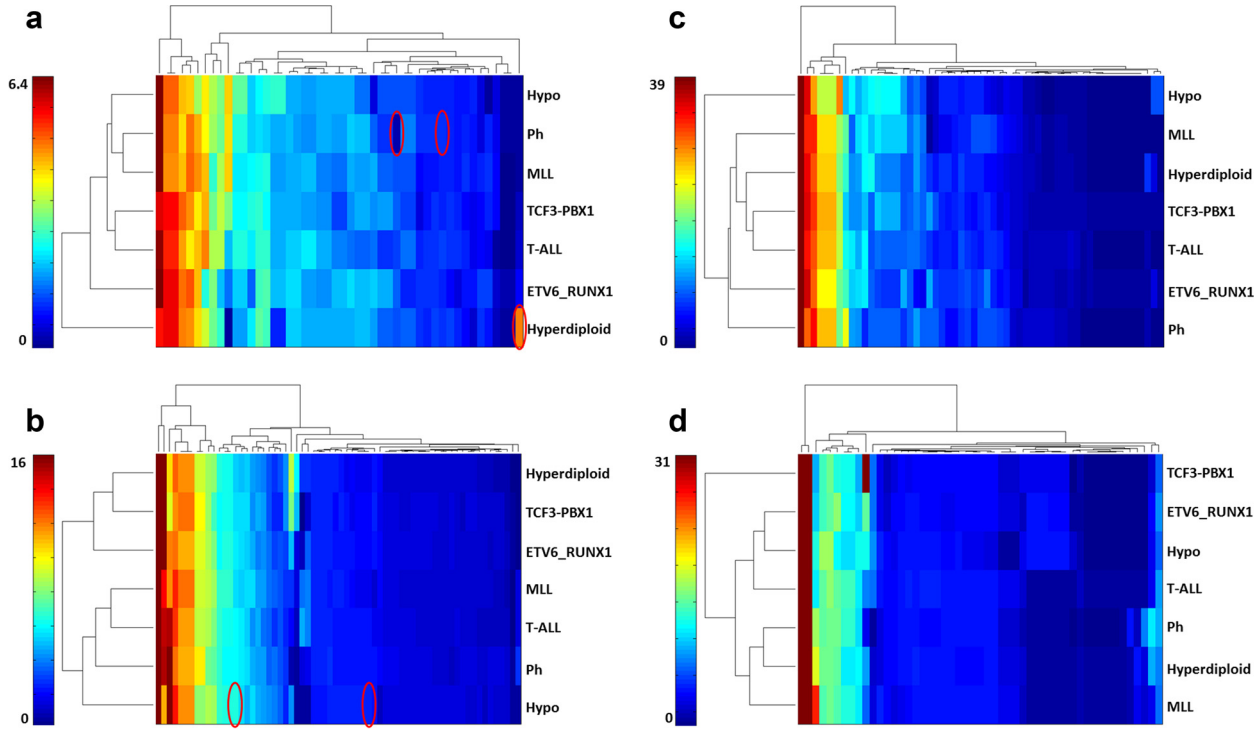


**Fig. 3** Gene centrality values of the four KEGG networks in seven leukemia subtypes. **a** Apoptosis, **b** cell cycle, **c** complement and coagulation, and (**d**) chemokine. *Rows* represent leukemia subtypes, and *columns* represent genes. Both *rows* and *columns* are hierarchically clustered

outstanding set of genes is highly enriched in a specific leukemia subtype. To achieve this, we first extract the set $L$ of outstanding genes for every network in every subtype. For every edge $e = (u, v)$ in a given network, we compute the mean $\mu_e$ and the standard deviation $\sigma_e$ of its probability values in all leukemia subtypes. Then for every subtype, for every edge $e$, we check if the probability of $e$ in this subtype was at least $2\sigma_e$ away from $\mu_e$. If it is, we add $u$ and $v$ to $L$. We then perform gene set enrichment analysis (GSEA) [40] on $L$ for every network in every

**Table 2** Highly enriched gene sets in specific combinations of signaling networks and leukemia subtypes, with the nominal *p* values produced by GSEA for these sets in their respective combinations

| Subtype | Network | *p* value | Gene set |
|---|---|---|---|
| Hyperdiploid | Apoptosis | 0.0083 | NFKB1, RELA, BCL2, PPP3CA, PPP3CB, PPP3CC, IL3RA, TNF, BAD, PPP3R1, AKT3, AKT1, AKT2, CHUK, TNFRSF1A, CASP7, DFFA, IKBKB, IKBKG, CASP3, IL3, CASP10, CSF2RB |
| ETV6_RUNX1 | Cell cycle | 0.0151 | TFDP1, TFDP2, E2F1, RBP3, E2F2, E2F3, CCND3, RBL1, PRB1, RBL2, CCNE1, CCNE2, CDK4, CDK6, CCND1, CCND2, CCNA1, CDK2, CCNA2 |
| T-ALL | Apoptosis | 0.0162 | BIRC2, BIRC3, XIAP, BIRC7, CASP7, NFKB1, IL1RAP, TRADD, FASLG, RELA, CASP3, DFFA, FADD, CASP8, IL1R1, FAS |
| TCF3-PBX1 | Apoptosis | 0.0167 | PRKACA, PRKACB, PRKACG, PRKAR1A, PRKAR1B, IL1RAP, FADD, PRKAR2A, PRKAR2B, PRKX, BAD, IL1A, IL1B, IL1R1, TNFRSF10B, TNFRSF10C, TNFRSF10D |
| Hypo | Apoptosis | 0.0484 | CAPN1, CAPN2, IRAK3, IRAK1, IRAK4, MAP3K14, BCL2, TP53, NGF, NTRK1 |
| Ph | Cell cycle | 0.088 | BUB1, BUB3, CDKN2A, CDK4, CDK6, CCND1, GADD45A, GADD45B CCND2, CCND3, RB1, CDC45L, MCM7, MCM2, CDC25A, CDKN1A, MCM6, MCM5, MCM4, MCM3, CCNL1, LAT, CCNE1, CCNE2, CDK2, ORC3L, ORC5L, ORC4L, ORC2L, ORC1L, ORC6L, TP53, GADD45G, CDC2, CCNA2, CCNA1, CDKN1B, CDKN1C |

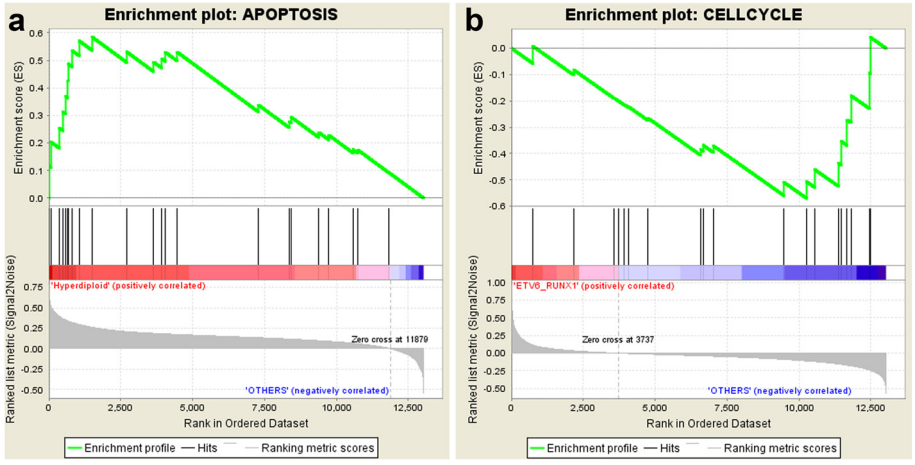


**Fig. 4** Gene set enrichment results for the highest two enriched gene sets in their respective leukemia subtypes. **a** Apoptosis in hyperdiploid and (**b**) cell cycle in ETV6_RUNX1

leukemia subtype. For every pair of network and subtype, we set the phenotype $A$ as the subtype samples, and phenotype $B$ as all samples from other subtypes. We then run GSEA on the network's outstanding gene set $L$ to measure its differential significance from $A$ to $B$. We consider gene sets whose $p$ value is below 0.1 as highly enriched. Table 2 lists these gene sets and their $p$ values in their respective leukemia subtypes. Figure 4 shows the gene set enrichment plots for the two highest enriched gene sets.

We observe from Table 2 that apoptosis and cell cycle signaling networks are dominant in all gene sets that are highly enriched. This implies a fundamental role for these two networks in the listed subtypes. It also implies that these subtypes are either caused by or leading to a perturbation in their respective gene sets. Another noteworthy observation is that, although all the highly enriched gene sets belong to only two networks, there is little overlap between them. In apoptosis for instance, PPP3 genes are dominant in hyperdiploid, while BIRC genes are dominant in T-ALL, and PRKA genes are dominant in TCF3-PBX1. Additionally, from Fig. 4, we observe that, although apoptosis and cell cycle have the highest enriched gene sets for hyperdiploid and ETV6_RUNX1, respectively, their relations to their respective leukemia subtypes are not the same. All genes in the apoptosis set in hyperdiploid exhibit higher expression than in other subtypes, which implies up-regulation of these genes in hyperdiploid. On the other hand, most of the genes in the cell cycle set in ETV6_RUNX1 have lower expression than in other subtypes, which indicates down-regulation of these genes in ETV6_RUNX1.

## 4 Conclusions

In this paper, we presented a novel method for computing edge probability in signaling networks. Our method uses gene coexpression as input and computes the edge probabilities so that reachability between edge terminals is as close as possible to their empirical values obtained from gene transcription levels. We used our method to compute edge probabilities for four KEGG signaling networks, using gene expression data for seven leukemia subtypes. We also used the computed edge probabilities to compute a centrality value for every gene in every leukemia subtype. We analyzed the interactions and genes with outstanding probability and centrality in specific subtypes. We also analyzed similarities and differences among these subtypes based on their edge probabilities. We performed gene set enrichment analysis on the set of edges with outstanding probabilities in each subtype to study the significance of the results. Our analysis provided evidence that links specific gene sets to specific leukemia subtypes, which makes them strong candidates for investigation in the scope of their respective subtypes.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

HG and TK analyzed the problem and designed the methods and experiments. HG implemented the methods and experiments. JCRM and DMG analyzed the observations and did the results' discussion. HG and JCRM authored the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

This work was supported by NSF Grant CCF-1251599 to TK, and NIH Grant GM083337 to DMG.

### Author details

[1]Department of Computer & Information Science & Engineering, University of Florida, Gainesville, Florida, USA. [2]Department of Biological Science, Florida State University, Tallahassee, Florida, USA.

Received: 12 August 2015 Accepted: 30 October 2015
Published online: 11 November 2015

**References**


1. J Scott, T Ideker, RM Karp, R Sharan, Efficient algorithms for detecting signaling pathways in protein interaction networks. J. Comput. Biol. **13**(2), 133–144 (2006)
2. TI Lee, NJ Rinaldi, F Robert, DT Odom, Z Bar-Joseph, G Gerber, NM Hannett, CT Harbison, CM Thompson, I Simon, J Zeitlinger, EG Jennings, HL Murray, DB Gordon, B Ren, JJ Wyrick, JB Tagne, TL Volkert, E Fraenkel, DK Gifford, RA Young, Transcriptional regulatory networks in saccharomyces cerevisiae. Science. **298**(5594), 799–804 (2002)
3. A-L Barabási, N Gulbahce, J Loscalzo, Network medicine: a network-based approach to human disease. Nat Rev. Genet. **12**(1), 56–68 (2011)
4. D-Y Cho, Y-A Kim, TM Przytycka, Network biology approach to complex diseases. PLoS Comput. Biol. **8**(12), 1002820 (2012)
5. T Ryba, I Hiratani, T Sasaki, D Battaglia, M Kulik, J Zhang, S Dalton, DM Gilbert, Replication timing: a fingerprint for cell identity and pluripotency. PLoS Comput. Biol. **7**(10), 1002225 (2011)
6. I Hiratani, A Leskovar, DM Gilbert, Differentiation-induced replication-timing changes are restricted to at-rich/long interspersed nuclear element (line)-rich isochores. Proc. Nat. Acad. Sci. USA. **101**(48), 16861–16866 (2004)
7. I Hiratani, T Ryba, M Itoh, T Yokochi, M Schwaiger, C-W Chang, Y Lyou, TM Townes, D Schübeler, DM Gilbert, Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. **6**(10), 245 (2008)
8. I Hiratani, T Ryba, M Itoh, J Rathjen, M Kulik, B Papp, E Fussner, DP Bazett-Jones, K Plath, S Dalton, PD Rathjen, DM Gilbert, Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res. **20**(2), 155–169 (2010)
9. RE Thurman, N Day, WS Noble, JA Stamatoyannopoulos, Identification of higher-order functional domains in the human encode regions. Genome Res. **17**(6), 917–927 (2007)
10. J Zhang, L Ding, L Holmfeldt, G Wu, SL Heatley, D Payne-Turner, J Easton, X Chen, J Wang, M Rusch, C Lu, SC Chen, L Wei, JR Collins-Underwood, J Ma, KG Roberts, SB Pounds, A Ulyanov, J Becksfort, P Gupta, R Huether, RW Kriwacki, M Parker, DJ McGoldrick, D Zhao, D Alford, S Espy, KC Bobba, G Song, D Pei, *et al*, The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. **481**(7380), 157–163 (2012). doi:10.1038/nature10725
11. T Ryba, D Battaglia, BH Chang, JW Shirley, Q Buckley, BD Pope, M Devidas, BJ Druker, DM Gilbert, Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res. **22**(10), 1833–1844 (2012)
12. A Todor, A Dobra, T Kahveci, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB). **10**(4), 970–983 (2013)
13. A Todor, H Gabr, A Dobra, T Kahveci, Large scale analysis of signal reachability. Bioinforma. **30**(12), 96–104 (2014). doi:10.1093/bioinformatics/btu262
14. H Gabr, T Kahveci, in *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference On*. Characterization of probabilistic signaling networks through signal propagation, (2014), pp. 1–2. doi:10.1109/ICCABS.2014.6863909
15. A Zanzoni, L Montecchi-Palazzi, M Quondam, G Ausiello, M Helmer-Citterich, G Cesareni, Mint: a molecular interaction database. FEBS letters. **513**(1), 135–140 (2002)
16. A Franceschini, D Szklarczyk, S Frankild, M Kuhn, M Simonovic, A Roth, J Lin, P Minguez, P Bork, C von Mering, String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research. **41**(D1), 808–815 (2013)
17. R Sharan, Conserved patterns of protein interaction in multiple species. PNAS (2002)
18. K-C Li, Genome-wide coexpression dynamics: theory and application. Proc. Natl. Acad. Sci. **99**(26), 16875–16880 (2002)
19. S Horvath, J Dong, Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. **4**(8), 1000117 (2008)
20. P Dhaeseleer, S Liang, R Somogyi, Genetic network inference: from co-expression clustering to reverse engineering. Bioinforma. **16**(8), 707–726 (2000)
21. BA Novak, AN Jain, Pathway recognition and augmentation by computational analysis of microarray expression data. Bioinforma. **22**(2), 233–241 (2006)
22. DJ Allocco, IS Kohane, AJ Butte, Quantifying the relationship between co-expression, co-regulation and gene function. BMC bioinforma. **5**(1), 18 (2004)
23. H Gabr, A Todor, H Zandi, A Dobra, T Kahveci, in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. Preach: Reachability in probabilistic signaling networks (ACM, New York, NY, USA, 2013), p. 3
24. H Gabr, A Todor, A Dobra, T Kahveci, Reachability analysis in probabilistic biological networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics. **12**(1), 53–66 (2015)
25. DF Dukers, JJ Oudejans, W Vos, RL ten Berge, CJ Meijer, Apoptosis in b-cell lymphomas and reactive lymphoid tissues always involves activation of caspase 3 as determined by a new in situ detection method. J. Pathol. **196**(3), 307–315 (2002)
26. T Takata, F Tanaka, T Yamada, K Yanagihara, Y Otake, Y Kawano, T Nakagawa, R Miyahara, H Oyanagi, K Inui, H Wada, Clinical significance of caspase-3 expression in pathologic-stage i, nonsmall-cell lung cancer. Int. J. Cancer. **96**(S1), 54–60 (2001)
27. F Cousin, S Baldassini, D Bourchany, A Claudy, J Kanitakis, Expression of the pro-apoptotic caspase 3/cpp32 in cutaneous basal and squamous cell carcinomas. J Cutan. Pathol. **27**(5), 235–241 (2000)
28. N ODonovan, J Crown, H Stunell, AD Hill, E McDermott, N OHiggins, MJ Duffy, Caspase 3 in breast cancer. Clinical Cancer Res. **9**(2), 738–742 (2003)
29. LH Meyer, L Karawajew, M Schrappe, W-D Ludwig, K-M Debatin, K Stahnke, Cytochrome c–related caspase-3 activation determines treatment response and relapse in childhood precursor b-cell all. Blood. **107**(11), 4524–4531 (2006)
30. RA Parikh, LJ Appleman, JE Bauman, M Sankunny, DW Lewis, A Vlad, SM Gollin, Upregulation of the atr-chek1 pathway in oral squamous cell carcinomas. Genes Chromosomes Cancer. **53**(1), 25–37 (2014)
31. H Gali-Muhtasib, D Kuester, C Mawrin, K Bajbouj, A Diestel, M Ocker, C Habold, C Foltzer-Jourdainne, P Schoenfeld, B Peters, M Diab-Assaf, U Pommrich, W Itani, H Lippert, A Roessner, R Schneider-Stock, Thymoquinone triggers inactivation of the stress response pathway sensor chek1 and contributes to apoptosis in colorectal cancer cells. Cancer Res. **68**(14), 5609–5618 (2008)
32. Y Dodurga, Y Oymak, C Gündüz, NL Satıroglu-Tufan, C Vergin, Ç,etingül, ÇB Avci, N Topçuoğlu, Leukemogenesis as a new approach to investigate the correlation between up regulated gene 4/upregulator of cell proliferation (urg4/urgcp) and signal transduction genes in leukemia. Mol. Biol. Reports. **40**(4), 3043–3048 (2013)
33. YS Chae, JG Kim, SK Sohn, SJ Lee, BW Kang, JH Moon, JY Park, SW Jeon, HI Bae, GS Choi, SH Jun, Ripk1 and casp7 polymorphism as prognostic markers for survival in patients with colorectal cancer after complete resection. J. Cancer Res. Clin. Oncol. **137**(4), 705–713 (2011)
34. A Linton, YY Cheng, K Griggs, MB Kirschner, S Gattani, S Srikaran, S Chuan-Hao Kao, BC McCaughan, S Klebe, N van Zandwijk, G Reid, An rnai-based screen reveals plk1, cdk1 and ndc80 as potential therapeutic targets in malignant pleural mesothelioma. Br. J. Cancer. **110**(2), 510–519 (2013)
35. A Barascu, P Besson, O Le Floch, P Bougnoux, M-L Jourdan, Cdk1-cyclin b1 mediates the inhibition of proliferation induced by omega-3 fatty acids in mda-mb-231 breast cancer cells. Int. J. Biochem. Cell Biol. **38**(2), 196–208 (2006)
36. M Wierer, G Verde, P Pisano, H Molina, J Font-Mateu, L Di Croce, M Beato, Plk1 signaling in breast cancer cells cooperates with estrogen receptor-dependent gene transcription. Cell Reports. **3**(6), 2021–2032 (2013)
37. D Włowiec, P Deviller, D Simonin, C Souchier, R Rimokh, M Benchaib, P-A Bryon, M Ffrench, Cdk1 is a marker of proliferation in human lymphoid cells. Int. J. Cancer. **61**(3), 381–388 (1995)
38. L Liu, M Zhang, P Zou, Expression of plk1 and survivin in diffuse large b-cell lymphoma. Leuk. Lymphoma. **48**(11), 2179–2183 (2007)
39. AG Renner, C Dos Santos, C Recher, C Bailly, L Créancier, A Kruczynski, B Payrastre, S Manenti, Polo-like kinase 1 is overexpressed in acute myeloid leukemia and its inhibition preferentially targets the proliferation of leukemic cells. Blood. **114**(3), 659–662 (2009)
40. A Subramaniana, P Tamayoa, VK Moothaa, S Mukherjeed, BL Eberta, MA Gillettea, A Paulovichg, SL Pomeroyh, TR Goluba, ES Landera, JP Mesirova, Gene set enrichment analysis: A knowledge-based approach for interpreti ng genome-wide expression profiles. Proc. Nat. Acad. Sci. USA. **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102