WMAXC: A Weighted Maximum Clique Method for Identifying Condition-Specific Sub-Network

Bayarbaatar Amgalan, Hyunju Lee*

School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, South Korea

Abstract

Sub-networks can expose complex patterns in an entire bio-molecular network by extracting interactions that depend on temporal or condition-specific contexts. When genes interact with each other during cellular processes, they may form differential co-expression patterns with other genes across different cell states. The identification of condition-specific sub-networks is of great importance in investigating how a living cell adapts to environmental changes. In this work, we propose the weighted MAXimum clique (WMAXC) method to identify a condition-specific sub-network. WMAXC first proposes scoring functions that jointly measure condition-specific changes to both individual genes and gene-gene co-expressions. It then employs a weaker formula of a general maximum clique problem and relates the maximum scored clique of a weighted graph to the optimization of a quadratic objective function under sparsity constraints. We combine a continuous genetic algorithm and a projection procedure to obtain a single optimal sub-network that maximizes the objective function (scoring function) over the standard simplex (sparsity constraints). We applied the WMAXC method to both simulated data and real data sets of ovarian and prostate cancer. Compared with previous methods, WMAXC selected a large fraction of cancer-related genes, which were enriched in cancer-related pathways. The results demonstrated that our method efficiently captured a subset of genes relevant under the investigated condition.

Citation: Amgalan B, Lee H (2014) WMAXC: A Weighted Maximum Clique Method for Identifying Condition-Specific Sub-Network. PLoS ONE 9(8): e104993. doi:10.1371/journal.pone.0104993

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received March 24, 2014; **Accepted** July 7, 2014; **Published** August 22, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Datasets from The Cancer Genome Atlas (TCGA): URL: http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp under Ovarian serous cystadenocarcinoma.

Funding: This work was supported by the Systems biology infrastructure establishment grant provided by Gwangju Institute of Science & Technology in 2014, and by the MSIP (The Ministry of Science, ICT and Future Planning), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0503-13-1019). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: hyunjulee@gist.ac.kr

Introduction

A central problem in network biology is the identification of genes and pathways involved in the same biological processes or physiological conditions. The details of control mechanisms in biological processes can be understood by analyzing interacting neighbors and local patterns. Network structures often have been used to describe these complex bio-molecular pathways and functional modules by representing a whole set of interactions as overlapping sub-networks, each associated with a specific condition [1,2].

Many methods have been developed to construct bio-molecular networks by comparing multiple sets of microarray data under different conditions. Because expressions of different genes in a series of biological conditions influence each other, correlations between genes have been widely used to analyze microarray gene-expression measurements. Waaijenborg and Zwinderman [3] developed a penalized canonical correlation analysis method to extract a subset of variables that capture the common features among genes by maximizing a canonical correlation between expression of genes. Witten and Tibshirani [4] presented some extension formulas to the sparse canonical correlation analysis as a supervised method, which resulted in the identification of linear combinations of sets of variables that are correlated and associated with its outcome.

To fully understand the complex biological processes, the effective integration of diverse sets of data and knowledge is required. Protein-DNA interaction and gene expression data were combined for regulatory network identification [5]. Integrating protein-protein interaction (PPI) data with gene expression data has also been attempted for the identification of biologically meaningful and cancer-related networks in cancer studies [6–9]. The integration increases the accuracy in identifying genes jointly regulated in the same condition. Guo et al. [10] developed an edge-based scoring function for gene-gene co-expression by using both gene expression and PPI data. However, in this method, PPI information is used to define existence of edges in the bio-molecular network so that only gene pairs that are included in the existing PPI network are considered in the scoring function. Lai et al. [11] extended the traditional F-statistic to obtain an expected conditional F-statistic (ECF-statistic), which measures the connectivity between genes. The ECF-statistic was used in a COSINE method [12] to measure gene-gene co-expression from gene expression data. For the use of the PPI information in COSINE, only the number of interactions in a selected sub-network was considered in their scoring function to calculate its sub-network adjust score, instead of using each interaction in the PPI network.

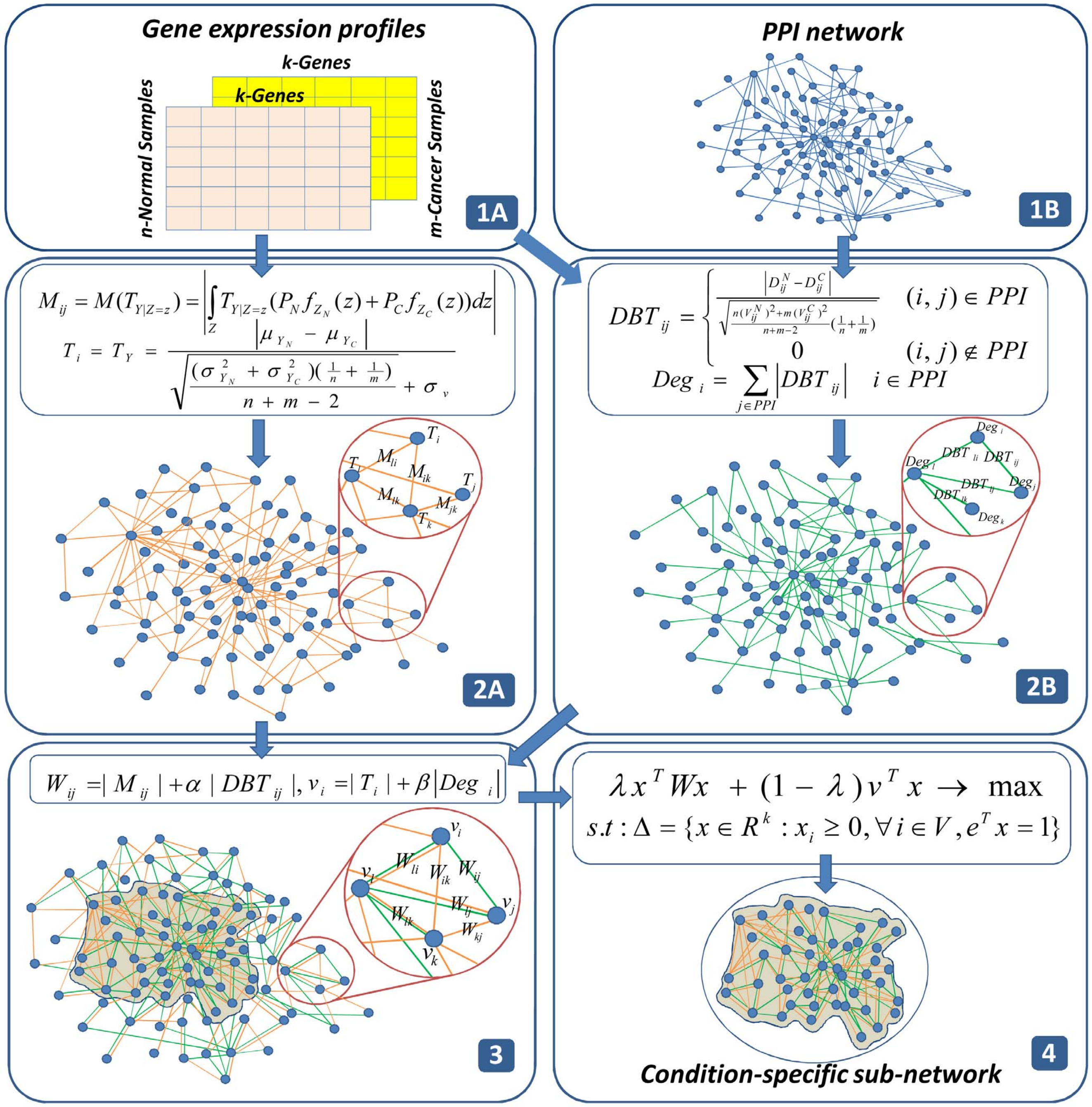
Gene expression profiles
k-Genes
k-Genes
n-Normal Samples
m-Cancer Samples
1A
PPI network
1B
$M_{ij} = M(T_{Y|Z=z}) = \left| \int_Z T_{Y|Z=z}(P_N f_{Z_N}(z) + P_C f_{Z_C}(z))dz \right|$
$T_i = T_Y = \frac{|\mu_{Y_N} - \mu_{Y_C}|}{\sqrt{\frac{(\sigma^2_{Y_N} + \sigma^2_{Y_C})(\frac{1}{n} + \frac{1}{m})}{n+m-2}} + \sigma_v}$
T_i
T_l
T_j
T_k
M_{li}
M_{ik}
M_{lk}
M_{jk}
2A
$DBT_{ij} = \begin{cases} \frac{|D^N_{ij} - D^C_{ij}|}{\sqrt{\frac{n(V^N_{ij})^2 + m(V^C_{ij})^2}{n+m-2}(\frac{1}{n}+\frac{1}{m})}} & (i,j) \in PPI \\ 0 & (i,j) \notin PPI \end{cases}$
$Deg_i = \sum_{j \in PPI} |DBT_{ij}| \quad i \in PPI$
Deg_i
Deg_l
Deg_j
Deg_k
DBT_{li}
DBT_{ij}
DBT_{lj}
DBT_{lk}
2B
$W_{ij} = |M_{ij}| + \alpha |DBT_{ij}|, v_i = |T_i| + \beta |Deg_i|$
v_i
v_l
v_j
v_k
W_{li}
W_{ik}
W_{ij}
W_{lj}
W_{lk}
W_{kj}
3
$\lambda x^T W x + (1-\lambda) v^T x \rightarrow \max$
$s.t: \Delta = \{x \in R^k : x_i \geq 0, \forall i \in V, e^T x = 1\}$
Condition-specific sub-network
4

Figure 1. Workflow of WMAXC. (1) Gene expression data consisting of normal and cancer samples (1A) and the PPI network (1B) are used as inputs. (2) We begin by constructing two responsive networks under the investigated condition: In (2A), we use two statistic measurements to construct a bio-molecular network. For each gene, T_i is used to measure activity of gene i (a node score) and for each pair of genes M_{ij} is used to measure connectivity relationship between gene i and gene j (an edge score). In (2B), for each interaction in PPI network, DBT_{ij} is used to measure activity of interaction behavior between gene i and gene j (an edge contribution score from PPI) and for each gene, Deg_i is used to measure the weighted degree of gene i (a node contribution score from PPI) under the condition. (3) We then combine the two responsive networks to construct the background network by assigning node and edge scores to a set of genes. Orange edges represent gene-gene co-expression estimated from only gene expression data and green edges represent activity of interactions in the PPI network. In the process of combining two networks, new edges are included to (2A) although they are not in the existing PPI network. (4) Finally, we solve the constrained optimization problem to obtain the single optimal sub-network.

doi:10.1371/journal.pone.0104993.g001

A mixed integer linear programming model [13] and an integer linear programming approach [14] were developed to identify differentially expressed pathways using both data sets. However, these two integer linear programming approaches have used expression values of individual genes without accounting for correlation or co-expression information, which might be less informative.

Finding a sub-network that maximizes the score of differential expressions of genes and differential co-expression of gene pairs can be formulated as a combinatorial optimization problem. In practice, bio-molecular networks are often large in scale [15]. Hence, it is impossible to exactly solve a large combinatorial optimization problem within a reasonable time. For examples, in the COSINE method [12], a genetic algorithm was used to find a binary vector, in which 1 or 0 represents presence or absence of a gene in the sub-network, and its space complexity is exponential as 2^p, where p is the total number of genes. In the edge-based method [10], a searching procedure based on simulated annealing was used to find the sub-network. It iteratively tests whether the addition or removal of an edge will increase their sub-network adjust score during the annealing process, and its space complexity is also exponential as 2^q, where q is the number of possible edges (edges in the PPI network). Ideker et al. [6] also used the simulated annealing approach; however, it tended to produce a large sub-network that was often difficult to interpret. Although Rajagopalan and Agarwal [16] and Nacu et al. [17] offered several improvements, they were also based on heuristic techniques with combinatorial selection and required the estimation of additional parameters. Various graph theory-based approaches, such as sequential greedy heuristics [18,19], have been presented. These heuristics generate a maximal scoring sub-network through the repeated addition of a vertex into a partial sub-network, or by the repeated deletion of a vertex from a sub-network. In addition to heuristics, several local optimization-based approaches have been presented to extract condition specific sub-networks. A method developed by Wang and Xia [15] was inspired by a KKT condition and was used to iteratively find a local minimum from a predetermined initial solution. A large number of local solutions can be found for the non-convex problem. Depending on the selection of the initial solution, many of them might not be significant under the investigated condition, which may give rise to false positives.

In our work, we reformulate the sub-network identification problem as a constrained optimization problem for continuous variables. It is an approximation of the general combinatorial problem, based on the theorem posed by Motzkin and Straus [20]. To construct a background network under investigated condition, we used two statistical measurements to represent activity of each genes and interaction behaviors of each pair of genes; a modified T-statistic as the differential expression of each gene and a conditional expectation of the modified T-statistic as the differential gene-gene co-expression. In our first experiment, we used the two measurements to construct the background network that represents the weight parameters of the optimization problem. We then employed a continuous genetic algorithm, which has the advantage of being capable of jumping out of local solutions, and used a projection procedure that maximizes our objective function under a sparsity constraint. In the second experiment, we reconstructed the background network by integrating PPI information with the gene expression profiles for each gene pair, and obtained a more robust estimation of the neighbors of each gene in the network. We first tested the performance of our method using simulated data sets, and then applied the method to analyze human ovarian and prostate cancer data sets. The results demonstrated that our method efficiently captures relevant interaction behaviors under the investigated conditions. An overview of the workflow is presented in Figure 1, and more detailed descriptions are given in the Methods section.

Methods

We first describe scoring functions to measure gene expression differences and gene-gene correlations for given two conditions, which generate weight values of nodes and edges in a background network. We then introduce an optimization model to identify the maximal scoring sub-network. Finally, the proposed model is extended to include PPI interactions.

Scoring function of WMAXC

The entire background network is represented as a graph $\mathbb{G}=(\mathbb{V},\mathbb{E})$, where a set of nodes $\mathbb{V}$ represents genes, and a set of edges $\mathbb{E}$ represents the connectivity relationships among these genes. Let $\mathbb{V}$ be a set of k genes. For each gene, n and m denote the numbers of samples in two different conditions, such as normal and cancer. We then have a gene expression data set, given as two matrices with sizes of $n\times k$ and $m\times k$. A modified T-statistic is used as a scoring function to measure the differential expression of each gene. For each node in $\mathbb{V}$, the differential expression value of the corresponding gene Y is computed as

$$T_Y=\frac{\mu_{Y_N}-\mu_{Y_C}}{\sqrt{\frac{(\sigma^2_{Y_N}+\sigma^2_{Y_C})(\frac{1}{n}+\frac{1}{m})}{n+m-2}+\sigma_v}}, \qquad (1)$$

where μ_{Y_N}, μ_{Y_C} and σ_{Y_N}, σ_{Y_C} are the sample means and standard deviations of gene Y in the normal and cancer conditions, respectively, and σ_v is a constant chosen to minimize the coefficient of variation of T-statistic [21] (see more descriptions and Figure S1 in File S1).

For each edge in $\mathbb{E}$, the conditional expectation of the modified T-statistic is used to measure differential gene-gene co-expressions across two conditions. To measure the gene-gene correlations for a pair of genes, we assume that samples of genes are jointly normal distributed in a particular condition, such as normal (N) or cancer (C). By the assumption, a bivariate normal distribution of two genes Y and Z in the normal condition is

$$(Y_N,Z_N)\sim\mathcal{N}\left((\mu_{Y_N},\mu_{Z_N}),\begin{pmatrix}\sigma^2_{Y_N} & \rho_N\sigma_{Y_N}\sigma_{Z_N}\\ \rho_N\sigma_{Y_N}\sigma_{Z_N} & \sigma^2_{Z_N}\end{pmatrix}\right).$$

The conditional distribution of Y_N given $Z_N=z$ is

$$Y_N|Z_N=z\sim\mathcal{N}\left(\mu_{Y_N}+\rho_N(z-\mu_{Z_N})\frac{\sigma_{Y_N}}{\sigma_{Z_N}},\sigma^2_{Y_N}(1-\rho^2_N)\right). \qquad (2)$$

Similarly, the conditional distribution of Y_C given $Z_C=z$ in the cancer condition is

$$Y_C|Z_C=z\sim\mathcal{N}\left(\mu_{Y_C}+\rho_C(z-\mu_{Z_C})\frac{\sigma_{Y_C}}{\sigma_{Z_C}},\sigma^2_{Y_C}(1-\rho^2_C)\right), \qquad (3)$$

where ρ_N and ρ_C are the sample correlations of a pair of genes (Y,Z) in the normal and cancer conditions, respectively. By replacing mean and variance in Equation (1) with the corresponding conditional means and variances from Equations (2) and (3), we obtain a conditional T-statistic as follows.

$$T_{Y|Z=z} = \frac{\mu_{Y_N} + \rho_N(z-\mu_{Z_N})\frac{\sigma_{Y_N}}{\sigma_{Z_N}} - \mu_{Y_C} - \rho_C(z-\mu_{Z_C})\frac{\sigma_{Y_C}}{\sigma_{Z_C}}}{\sqrt{\frac{[\sigma^2_{Y_N}(1-\rho^2_N)+\sigma^2_{Y_C}(1-\rho^2_C)](\frac{1}{n}+\frac{1}{m})}{n+m-2}}+\sigma_v}. \quad (4)$$

For gene Z, let $f_{Z_N}(z) \sim \mathcal{N}(\mu_{Z_N}, \sigma^2_{Z_N})$ and $f_{Z_C}(z) \sim \mathcal{N}(\mu_{Z_C}, \sigma^2_{Z_C})$ be the probability density function of the normal and cancer conditions, respectively. Then, the probability density function of Z is $f_Z(Z) = p_N f_{Z_N}(z) + p_C f_{Z_C}(z)$, where $p_N = \frac{n}{n+m}$ and $p_C = \frac{m}{n+m}$ are the probabilities that a sample is selected from the normal or cancer conditions. By calculating the expectation over all samples of gene Z, we obtain the connectivity relationship between gene Y and gene Z as follows.

$$\begin{aligned} M(T_{Y|Z=z}) &= \int_Z T_{Y|Z} f_Z(z)dz \\ &= \int_Z (p_N T_{Y|Z=z} f_{Z_N}(z) + p_C T_{Y|Z=z} f_{Z_C}(z))dz \end{aligned} \quad (5)$$

As we described above, $T_{Y|Z}$, $f_{Z_N}(z)$ and $f_{Z_C}(z)$ are functions of z so that the integration in Equation (5) can be numerically computed over all samples in the normal and cancer conditions.

In summary, the modified T-statistic is used to measure the differential expression of each gene, and the conditional expectation $M(T_{Y|Z=z})$ is used to measure differential co-expression of each pair of genes across two conditions. Note that we calculated the differential co-expression patterns $M(T_{Y|Z})$ for all pairs of nodes in the background network because the co-expression of two genes might be significant although each gene may not be differentially expressed. In this work, we will use a matrix notation M_{ij} for $M(T_{Y|Z})$ and a vector notation T_i for T_Y, and the matrix M is symmetric, in which its entries represent weights of undirected edges and the entries in the vector T represent weights of nodes in the background network.

Optimization model

A remarkable connection between the maximum clique problem and a certain standard quadratic programming problem was established [20] by providing an alternative proof of a slightly weaker version of the fundamental theorem [22].

Let $\bar{\mathbb{G}} = (\bar{\mathbb{V}}, \bar{\mathbb{E}})$ be an unweighted and undirected graph, and Δ denotes the standard simplex in the k-dimensional Euclidean space $\mathbb{R}^k$, and $\Delta_{\mathbb{S}} = \{x \in \Delta : x_i = 0 \text{ if } i \notin \mathbb{S}\}$ is the face of Δ corresponding to a subset $\mathbb{S} \subseteq \bar{\mathbb{V}}$. A characteristic vector $x^{\mathbb{S}}$ denotes the vector in Δ defined by $x^{\mathbb{S}}_i = 1/\|\mathbb{S}\|$ if $i \in \mathbb{S}$ and $x^{\mathbb{S}}_i = 0$, otherwise. The maximum clique problem can then be formulated as the following quadratic programming problem.

$$\begin{aligned} &f(x) = x^T A_{\bar{\mathbb{G}}} x \to \max, \\ &\Delta = \{x \in \mathbb{R}^k : x_i \geq 0, \ \forall i \in \bar{\mathbb{V}}, \ e^T x = 1\}, \end{aligned} \quad (6)$$

where e^T denotes the transpose of the vector e consisting of unit entries, $A_{\bar{\mathbb{G}}} = (a_{ij})_{i,j \in \bar{\mathbb{V}}}$ is the adjacency matrix (binary matrix) of $\bar{\mathbb{G}}$, and x^* is a global solution of f on Δ. Motzkin et al. [20] proved that the clique number of $\bar{\mathbb{G}}$ is related to $f(x^*)$ in the following formula.

$$\omega(\bar{\mathbb{G}}) = \frac{1}{1-f(x^*)} \geq \frac{1}{1-f(x)}, \ \forall x \in \Delta, \quad (7)$$

where $\omega(\bar{\mathbb{G}})$ is the size of the maximum clique in $\bar{\mathbb{G}}$. Essentially, they proved that a subset of nodes $\mathbb{S}$ is a maximum clique of $\bar{\mathbb{G}}$ if and only if its characteristic vector $x^{\mathbb{S}}$ is a global solution of f on Δ (for a comprehensive review, see [23]).

Based on the theoretical validation, we reformulate the sub-network identification problem as a continuous optimization problem that is an approximation of the general combinatorial problem and a generalization of the problem in Equation (6). The proposed method relates the densest part (a maximum scored clique) of a weighted graph to the optimization of a quadratic function under sparsity constraints. A weaker formula of the maximum clique problem, the optimization problem for identifying a condition-specific sub-network from a bio-molecular network, can be formulated as follows.

$$\begin{aligned} &\lambda x^T W x + (1-\imath\lambda) v^T x \to max, \\ &\Delta = \{x \in \mathbb{R}^k : x_i \geq 0, \ \forall i \in \mathbb{V}, \ e^T x = 1\}, \end{aligned} \quad (8)$$

where v is a vector in which v_i represents a node score measuring differential expression for gene i, W is a symmetric matrix in which W_{ij} represents an edge score measuring the connectivity strength between gene i and gene j, and λ is a positive parameter to balance and to integrate the two terms of the objective function in Equation (8). A k-dimensional non-negative vector $x = (x_1, x_2, ..., x_k)$, determined by solving our optimization problem, represents the contribution to each gene belonging to the condition specific sub-network. Particularly, x_i indicates whether its corresponding node is contained in a selected sub-network ($x_i > 0$) or not ($x_i = 0$). Since we maximize the interconnectivity of sub-network, a gene i with both a high node confidence score in v and high confidence scores in W_{ij} should be selected in the sub-network and its corresponding x_i should be assigned to have a high contribution score. Therefore, the subset of variables corresponding to the nonzero elements in the optimal solution x^* forms the maximum scored sub-network in the background network. Moreover, the genes that have higher contribution scores are more likely to be related to the phenotype (cancer) being analyzed.

Sub-network identification from gene expression and PPI network data

We extend our model to incorporate the assumption that the significance score of one gene or its interactions in a network depends not only on its own gene expression profile but also on the profiles of its neighbors in the PPI network. Some interactions in the PPI network are activated under the investigated condition while others are not activated. If two genes interact with each other under a particular condition, the expression distance between them might be significantly changed across two conditions (a normal condition and a cancer condition). Based on this assumption, we propose a scoring function called the distance-based T-score to measure the change in gene expression distances across two conditions for each pair of genes in the PPI network. This function is used to test the significance score of each interaction in the PPI network.

As described in the section of scoring function of WMAXC, the gene expression profile data are given as two matrices with sizes of $n \times k$ and $m \times k$. Let $(g^1_{1i}, g^1_{2i}, g^1_{3i}, ..., g^1_{ni})^T$ and $(g^2_{1i}, g^2_{2i}, s^2_{3i}, ..., g^2_{mi})^T$ be the samples of gene i in normal and cancer conditions, respectively, and $\mathbb{PPI}$ denote a set of pair indexes of genes with interactions in the PPI network. For a pair $(i,j)\in\mathbb{PPI}$, let $D^N_{ij}=\frac{1}{n}\sqrt{\sum_{l=1}^{n}(g^1_{li}-g^1_{lj})^2}$ and $D^C_{ij}=\frac{1}{m}\sqrt{\sum_{l=1}^{m}(g^2_{li}-g^2_{lj})^2}$ be the normalized distances between gene i and gene j in the normal and cancer conditions, respectively. Then, the distance-based T-score, DBT_{ij}, can be formulated as follows.

$$DBT_{ij}=\begin{cases}\dfrac{D^N_{ij}-D^C_{ij}}{\sqrt{\dfrac{n*(V^N_{ij})^2+m*(V^C_{ij})^2}{n+m-2}\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}}, & (i,j)\in\mathbb{PPI}\\ 0, & (i,j)\notin\mathbb{PPI},\end{cases} \tag{9}$$

where V^N_{ij} and V^C_{ij} are the geometric averages of the standard deviations of samples of gene i and gene j in the normal and cancer conditions, respectively (see more descriptions and Figure S2 in File S1). One of the advantages of using DBT score is that less relevant interactions in the PPI network under the investigated condition can be thinned out. Since both the conditional expectation of the modified T-statistic and the distance-based T-score are estimated from the same population, we reconstruct our background network by integrating the two types of information. To quantify absolute changes of expression under the investigated condition, we transformed T_i, M_{ij}, Deg_i and DBT_{ij} in their absolute values. The weight parameters of the objective function in Equation (8) are then calculated as follows.

$$W_{ij}=|M_{ij}|+\alpha*|DBT_{ij}|,\quad v_i=|T_i|+\beta*Deg_i, \tag{10}$$

where $Deg_i=\sum_j|DBT_{ij}|$ is a weighted degree for gene i in the PPI network, $i,j=1,2,...,k$ are gene indexes, α and β are positive scaling parameters defining the contribution of PPI information to the condition-specific network. We set $\alpha=\frac{\max_{i,j}\{M_{ij}\}}{\max_{i,j}\{DBT_{ij}\}}$ and $\beta=\frac{\max_i\{T_i\}}{\max_i\{Deg_i\}}$, where α ensures that the maximum of entries in the M matrix is equal to the maximum of entries in the DBT matrix and similar normalization is applied to β.

The PPI network is very sparse; the percentage of interacting protein pairs is only 0.0888% for the ovarian cancer data and 0.1280% for the prostate cancer data. Moreover, some interactions in the existing PPI network might not be significantly activated under the investigated condition. In this case, they might have very small values in DBT_{ij} (near to zero). Therefore, the DBT score matrix becomes more sparse, and only the entries corresponding to significant interactions in the PPI network have observable contributions to the matrix W and the vector v. However, the weight matrix M and the weight vector T are not sparse, because until now we used all differential expression and co-expression values without considering the significance of them. Therefore, genes with non-significant changes and correlations in expression values contribute their weights to W and v. In other words, all non zero entries in W and v can influence the behavior (slope) of the continuous objective function even though their values are not high. The influence in the objective function may lead the solution to a local maximum. To avoid the local convergence, we define two hard thresholds as follows.

$$T_i=\begin{cases}|T_i|, & |T_i|<\mu_v\\ 0, & otherwise,\end{cases}\qquad M_{ij}=\begin{cases}|M_{ij}|, & |M_{ij}|<\mu_W\\ 0, & otherwise,\end{cases} \tag{11}$$

where μ_v is the mean value of T, computed as $\mu_v=\frac{1}{k}\sum_{i=1}^{k}|T_i|$, and μ_W is the maximum of $\mu_{W_1},\mu_{W_2},...,\mu_{W_k}$, and for each i, $\mu_{W_i}=\frac{1}{k}\sum_{j=1}^{k}|M_{ij}|$.

Estimation of weight parameter λ

A reasonable balance between the quadratic and linear terms of the objective function in Equation (8) is essential for the stability and robustness of the optimization result. Unfortunately, it is difficult to define an objective criterion for biological relevance. Regarding optimization, the term with greater value should be more informative in our task. To achieve the stability of the solution, we compared the magnitudes of edge and node scores in terms of statistical distributions of score values. Because of the restrictions imposed by the sparsity constraint, the sub-network scores are not affected by their size. For both score terms, sub-network scores of different sizes behave similarly, and the majorities of edge and node score values fall in the region around their means. Therefore, the edge and node score values of a randomly selected sub-network have more chances to fall around the mean (distribution plots are given in Figures S3 and S4 in File S1), which allows the possibility of estimating the magnitudes that make the solution more stable. To estimate the magnitudes of the two terms, we used the following procedure: 1) Randomly sample a large number of sub-networks by selecting points on Δ. Each point $x^i\in\Delta$ presents a sub-network, and the edge and node scores of the sub-network are calculated as $x^{i^T}Wx^i$ and v^Tx^i, respectively; 2) Compute the means μ_W and μ_v, standard deviations s_W and s_v of two score terms; 3) The magnitudes of both edge and node score terms are defined as follows: $M_W=\frac{\mu_W}{s_W}$, $M_v=\frac{\mu_v}{s_v}$; and then 4) We set $\lambda=\frac{M_W}{M_W+M_v}$.

Searching for condition specific sub-network

If the matrix $-W$ is positive definite, the objective function in Equation (8) is concave (concave maximization is equivalent to convex minimization), and any local solution can also be the global maximum. Unfortunately, in the WMAXC method, the weight matrix W is generally indefinite. There are a large number of local maxima, each representing a densely connected subgraph. Because of the high complexity of the problem, it is a common practice to solve it using metaheuristics, such as evolutionary algorithms. This study implements a combination of a continuous genetic algorithm [24] and a projection procedure [25–27] to avoid the local maxima and reduce computational costs.

The continuous genetic algorithm is a parallel search procedure commonly used in a high-dimensional global optimization problem. For constrained optimizations, depending on the shape of the constraint and the dimension of the problem, the implementation of a genetic algorithm should be adapted to a particular problem. Because the standard simplex $\Delta\subset\mathbb{R}^k$ is a subspace of the l_∞ ball $[0,1]^k\subset\mathbb{R}^k$ (see Figure S5 in File S1), we first apply the continuous genetic algorithm to maximize the objective function in Equation (8) over $[0,1]^k$ and find a single optimal solution. Let $\hat{x}$ be the global maximum of Equation (8) over $[0,1]^k$. We then project the solution $\hat{x}$ onto Δ in the Euclidean space $\mathbb{R}^k$ using the algorithm [26] to obtain a sparser solution [27]. Since projection of any point onto convex set is unique (see the projection theorem and Figure S6 in File S1) and the Δ is convex, the problem of finding the Euclidean projection of a vector $\hat{x}\in\mathbb{R}^k$ onto Δ can be described by the following convex optimization problem.

$$x^* = \arg\min_{x \in \Delta} \|x - \hat{x}\|^2, \tag{12}$$

where x is a k-dimensional vector used as an optimization variable, and $x^* = (x_1^*, x_2^*, \ldots, x_k^*)$ is the projection of $\hat{x}$ onto Δ (see File S1). In view of the Lagrangian duality, the optimal solutions of the primal and dual problems are equal to a saddle point. In our task, the optimization variable of the dual problem is simply a scalar. Hence, it is more efficient to solve the dual problem instead of directly solving the primal problem in Equation (12). By considering the well-known result of the projection onto standard simplex [25], the optimal projection x^* is expressed as follows.

$$x_i^* = \begin{cases} \hat{x}_i - \theta, & \hat{x}_i > \theta \\ 0, & otherwise, \end{cases} \tag{13}$$

where the dual optimal point θ^* is given by the root of the following equation:

$$D'(\theta) = \sum_{i=1}^{n} \max\{0, \hat{x}_i - \theta\} - 1 = 0. \tag{14}$$

The root θ^* of Equation (14) can be numerically found using the algorithm from [26] and is used in Equation (13). The solution of Equation (13) represents the optimal solution of the problem in Equation (8). After the procedure is completed, the nodes corresponding to non-zero entries in the solution vector are selected to represent the subset of genes forming the condition-specific sub-network.

Data sets

- Protein-protein interaction data: We used the Human Protein Reference Database released in 2010 [28]. There were 39,240 binary protein-protein interactions involving 9,617 genes. After excluding self-interactions, 37,080 interactions remained.
- Ovarian cancer data: Gene expression data was collected from the TCGA project [29]. It contains 17,814 genes with the expression profiles of 587 cancer samples and 62 normal samples. For the cancer samples, we only considered 332 samples without missing values. For the normal samples, we imputed the missing values in the samples using the Weighted K-Nearest Neighbors method, based on available observations in the samples. For the analysis of ovarian cancer, we considered only the genes that were included in the PPI network, which consists of 8,721 genes and 33,771 interactions.
- Prostate cancer data: Gene expression data was collected from [30] with gene expression omnibus (GEO) accession number GSE3933. It contains the gene expression profiles of 71 cancer samples and 41 normal samples. The gene expression data was initially used in COSINE [12]. In order to compare our results with the results of COSINE using the same data, we used the data set prepared by the authors of COSINE. This data set consists of 5,335 genes with 18,234 interactions in the PPI network.

Results

Simulation studies

We showed the performance of WMAXC on simulated data by comparing it with COSINE [12], because COSINE was initially compared to several other methods, including jActiveModules [6], an edge-based method [10], and a local method [15]. We constructed five simulation data sets from multivariate normal distributions: four case data sets and one reference data set.

Each data set consisted of 1,000 variables (genes) for 50 samples. For each gene i, we draw the mean μ_i and the standard deviation σ_i from the uniform distribution on the observed range of normal data $[-0.5, 0.5]$, and then estimated the correlation coefficient ρ_{ij} for each pair of genes (i,j). Let $\mu = (\mu_1, \mu_2, \ldots, \mu_{1000})$ denote the mean vector, and $\Sigma = \{\sigma_{ij}\}_{i,j=1,\ldots,1000}$ denote the covariance matrix, where $\sigma_{ij} = \rho_{ij} * \sigma_i * \sigma_j$ is the covariance between gene i and gene j.

We first simulated the reference data from joint normal distribution $N(\mu, \Sigma)$, and the reference data was compared to each of the four case data sets. In each of four case data sets, 200 of 1,000 genes were selected as significant genes. For the particular genes, some shifts in the mean expressions and higher correlations among the genes were assigned to represent significant alterations in expression values. Let $\mathbb{I}$ denote the set of indexes for the significant genes.

$$\hat{\mu}_i = \begin{cases} \mu_i + k, & i \in \mathbb{I} \\ \mu_i, & i \notin \mathbb{I}, \end{cases} \hat{\rho}_i = \begin{cases} \rho_{ij} + r, & i,j \in \mathbb{I} \\ \rho_{ij}, & i,j \notin \mathbb{I}, \end{cases} \tag{15}$$

In each of four case data sets, the expression data was simulated from jointly normal distribution $N(\hat{\mu}, \hat{\Sigma})$, where $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{1000})$ is the mean vector, and $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}_{i,j=1,\ldots,1000}$ is the covariance matrix and its entries were calculated as $\hat{\sigma}_{ij} = \hat{\rho}_{ij} * \sigma_i * \sigma_j$. For case data 1, $k = 0.6$ and $r = 0.65$ were used in Equation (15), $k = 0.3$ and $r = 0.65$ for case data 2, and $k = 0.6$ and $r = 0.35$ for case data 3, and $k = 0.3$ and $r = 0.35$ for case data 4. Compared to the results of COSINE, WMAXC provided higher accuracy in sub-network size, recall, precision, and F-measure (Table 1). In the simulated data, the sub-network sizes identified by COSINE were smaller than desired, whereas WMAXC identified approximately 200 genes. Regarding the running time, WMAXC was much faster than COSINE; in the analysis on simulated data, it took 10 minutes to convergence, while COSINE took around 30 hours.

Comparison with other methods on real data sets

We applied the WMAXC method on gene expression profiles of two real data sets to identify a cancer type specific sub-network. Then, the performance of WMAXC was compared to other methods, COSINE [12] and BMRF [5].

Comparison on ovarian cancer. To evaluate whether the genes in the identified sub-network are related to ovarian cancer, we used a set of 379 experimentally verified ovarian cancer-related genes from the Dragon Database of Genes [31] as reference genes. Among the 379 genes, 315 were included in the 8,721 genes of our data set. For WMAXC, the initialization parameters of the genetic algorithm were set as follows: the number of iterations = 60,000, mutation rate = 1/(k+1) and crossover rate = 0.5, where k is the number of optimization variables (the number of genes). We first applied our method using only the gene expression profile data, and then integrated the PPI network with gene expression profiles

Table 1. Comparison on simulated data with COSINE.

methods	WMAXC				COSINE			
Case data	Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
sub-network size	204	230	219	237	137	178	126	128
λ	0.157	0.215	0.087	0.223	0.61	0.85	0.11	0.39
Recall	1	0.97	1	0.975	0.66	0.685	0.61	0.605
Precision	0.9803	0.8434	0.9132	0.8227	0.9635	0.7696	0.9682	0.9453
F-measure	0.9901	0.9023	0.9546	0.8924	0.7833	0.7248	0.7484	0.7378

Recall, precision and F-measure are defined as follows: Recall (R) $= \frac{TP}{TP+FN}$, Precision (P) $= \frac{TP}{TP+FP}$ and F-measure $= \frac{2PR}{P+R}$, where TP, FP, and FN represent true positive, false positive, and false negative, respectively.
doi:10.1371/journal.pone.0104993.t001

using the Distance Based *T*-score as described in the Methods section. Performances of the two approaches are shown in Table 2. The fold enrichment of the genes selected among the ovarian cancer-related genes had increased from 1.828 to 2.454 when the PPI network was integrated with the gene expressions, compared to using only the gene expressions. A list of 100 genes with the highest contribution scores to the condition-specific network is given in Table S1 in File S1. Using a CPU with 3.40 GHz and 32 GB RAM, it took 26 hours to search for the maximal scoring sub-network. To test whether the genetic algorithm reached the convergence, we quantified the variations in objective function values for the population in each iteration. For both ovarian and prostate cancer data, the variation was almost zero after running 60,000 iterations, and the minimum values of the objective function became stable in the last 10,000 iterations. This result suggests that the solution had reached convergence.

For the COSINE method, both gene expression profile and PPI data were used. After five different λ's were tried, the λ giving the highest adjusted score of the scoring function was used. As a result, the sub-network with a size of 806 was selected. The fold enrichment was 1.237, and it took around 56 hours with 1,000 iterations. For the BMRF method, both gene expression profile and PPI data were used. In addition, a set of hub genes related to the investigated condition was required as an input. Hence, we collected a set of 209 hub genes from KEGG, consisting of genes included in the ovarian cancer-related pathways, such as ubiquitin, coagulation, and hedgehog signaling pathways. BMRF extracted the sub-network of the size of 916 genes, and the fold enrichment was around 1.75. However, its accuracy was depending on the choice of the set of hub genes; for a set of randomly selected genes, the fold enrichment was decreased to 1.041. Overall, the comparison demonstrated that WMAXC outperformed both the COSINE and BMRF methods on the real data set.

Comparison on prostate cancer. For the method evaluation, 703 genes related to prostate cancer from the Dragon Database of Genes [32] were used as reference genes. Among the 703 genes, 400 were included in our dataset. $\lambda = 0.041$ was used for the WMAXC method, and the initialization parameters of the genetic algorithm were the same as those used in the analysis of the ovarian cancer data. WMAXC extracted a sub-network of the size 539 and a fold enrichment of 2.35. A list of 100 genes with the highest contribution scores to the condition specific network is given in Table S2 in File S1. COSINE selected a relatively smaller network with the size of 243 and a fold enrichment of 1.262 [12]. With a set of hub genes included in the mark signaling pathway from KEGG, BMRF selected the sub-network with the size of 601 genes and a fold enrichment of 2.086. However, for a set of randomly selected genes, the fold enrichment was decreased to 0.98. Performances of methods are summarized in Table S3 in File S1, confirming that WMAXC outperformed the other two methods.

Analysis on real data

Although only 57 out of 643 genes forming the condition-specific sub-network were included in DDOC as shown in Table 2, some ovarian cancer-related genes might not be included in DDOC. Hence, we manually checked whether 20 candidate genes with the highest contribution scores were ovarian cancer-related genes. Among them, 16 genes were known to be related to ovarian cancer by DDOC or the manual literature search (Table S4 in File S1). The remaining four genes are SELL, UBAP2L, TFEB and DPPA4. Although there were no evidences of their involvements in ovarian cancer development, these four genes were highly co-expressed with other ovarian cancer and cancer

Table 2. Performance on ovarian cancer data.

Methods	COSINE	BMRF	WMAXC1	WMAXC2
λ	0.871	-	0.173	0.2715
Selected genes	806	916	567	643
Recovered interactions	275	635	483	2015
Recovered genes	36	58	38	57
Fold enrichment	1.237	1.753	1.828	2.454

WMAXC1 represents the results obtained using only gene expression profile data, whereas the WMAXC2 results were obtained by integrating gene expression profiles and PPI network data. 'Fold enrichment' was used to evaluate the performance of the methods and was calculated as $\frac{\text{`}Recovered\ genes\text{'} * \text{`}All\ genes\text{'}}{\text{`}Selected\ genes\text{'} * \text{`}Reference\ genes\text{'}}$, where 'Selected genes' is the number of selected genes by the method, 'Reference genes' is the number of reference genes from the Ovarian Cancer Dragon Database of genes, 'Recovered genes' is recovered genes by the method among the reference genes, and 'All genes' represents all genes in the entire network. 'Recovered interactions' represents the number of interactions recovered from the PPI network.
doi:10.1371/journal.pone.0104993.t002

related genes in the condition specific network. As shown in Figure 2, they directly shared significant co-expression patterns with 159 neighbors. Surprisingly, 59.1% (94/159) of neighbors of these four genes are ovarian cancer-related genes and 35.8% (57/159) are other cancer-related genes. A list of these genes with their literature evidences is shown in Table S5 in File S1. For only 5% (8/159) of neighbor genes, we cannot find literature evidences showing their relevance to cancer. We further investigated that these four genes were actively involved in the other cancer types and biological phenomenon. L-selectin, SELL, is a member of a family of adhesion receptors that play important roles in lymphocyte-endothelial cell interactions. Resto et al. [33] investigated adhesive interactions between lymphocytes and head and neck cancer cells (HNSCC cells) under shear stress, and the interactions can be mediated by L-selectin. Kuiper et al. [34] investigated that upregulation of the transcription factor TFEB in some particular chromosomal position may play an important role in the regulation of renal cancer progression. Maldonado-Saldivia et al. [35] provided an evidence that DPPA4 is downregulated during fetal germ line progression and this process might be required to facilitate appropriate germ line differentiation. Moreover, it may provide an implication in the development of germ cell cancer in human. Although there was no much evidence of their involvements in ovarian cancer, our results suggest that these genes might be closely related to ovarian cancer progression.

By applying the WMAXC method to the ovarian cancer data, we identified a biologically meaningful sub-network involved in many ovarian cancer related pathways. Measured using Database for Annotation, Visualization and Integrated Discovery (DAVID) [36], 60 pathways were significantly enriched by the KEGG pathway, including the ErbB signaling pathway, the Notch signaling pathway, and the TGF-β signaling pathway (Table S6 in File S1). The epidermal growth factor receptor (EGFR) is a member of the ErbB family of tyrosine kinase receptors. Overexpression of EGFR and its downstream targets are associated with resistance to chemotherapy for ovarian cancer [37]. Ovarian cancer cells, in which Notch3 was frequently amplified and overexpressed, are dependent on the Notch3 signaling pathway for cellular survival and growth. Notch3 expression is also associated with chemo resistance in ovarian high-grade serous carcinoma [38]. The TGF-β signaling pathway is activated in ovarian cancer, and the inhibition of this pathway by a small molecule is a promising strategy in the treatment of ovarian cancer [39].

The sub-network identified using the prostate cancer data included 62 significantly enriched KEGG pathways, including prostate cancer, neurotrophinn signaling, MAPK signaling, Wnt signaling, TGF-β signaling, chemokine signaling pathways, and the regulation of actin cytoskeleton (Table S7 in File S1). For instance, it has been demonstrated that the progression of prostate cancer is affected by changes in the expression of auctocrine neurotrophins [40]. MAPK signaling is shown to be activated in prostate cancer, especially in later stages of the disease [41–43], and it was recently suggested that the MAPK signaling pathway may be a target for prostate cancer therapy, if it is inhibited simultaneously with other pathways, such as PI3K/AKT signaling [44]. The upregulation of some Wnt pathway members was observed in ERG-positive prostate cancers, and it has been shown that knockdown of the ERG gene in VCaP prostate cancer cells causes an activation of cell adhesion and expression changes in Wnt signaling. These findings were validated by gene expression data from both clinical prostate cancer samples and from ERG over-expressing non-transformed prostate epithelial cells [45]. Several studies have shown that changes in the levels of TGF-β pathway components are related to prostate cancer progression and cellular responses [46], [47], and [48]. In addition, chemokine signaling pathways and the regulation of actin cytoskeleton have been studied and experimentally validated to be associated with prostate cancer [15,49].

In summary, the analysis of both simulated and real data provides evidence that the WMAXC method can yield new insights that contribute to a better understanding of diseases.

Discussion

Our main goal was to design an algorithm that reveals a subset of genes closely related to a particular disease. Based on an optimization framework, we proposed an effective method, WMAXC, for identifying a condition-specific sub-network under a particular condition. WMAXC has the following advantages: (1) It extracts the global optimal sub-network that exhibits significant alterations across two phenotypes; WMAXC considers the weighted contributions of both expression difference for each single gene and the differential correlation of each pair of genes. (2) WMAXC effectively integrates diverse sets of data and knowledge to construct the background network under a particular condition. (3) An optimization formulation with strong theoretical validation is used to represent a continuous version of the general combinatorial problem for identifying a condition-specific sub-network. (4) WMAXC considers all nodes and edges at the same time to search a single optimal sub-network. (5) Genetic algorithm and a projection procedure are combined to approximate the

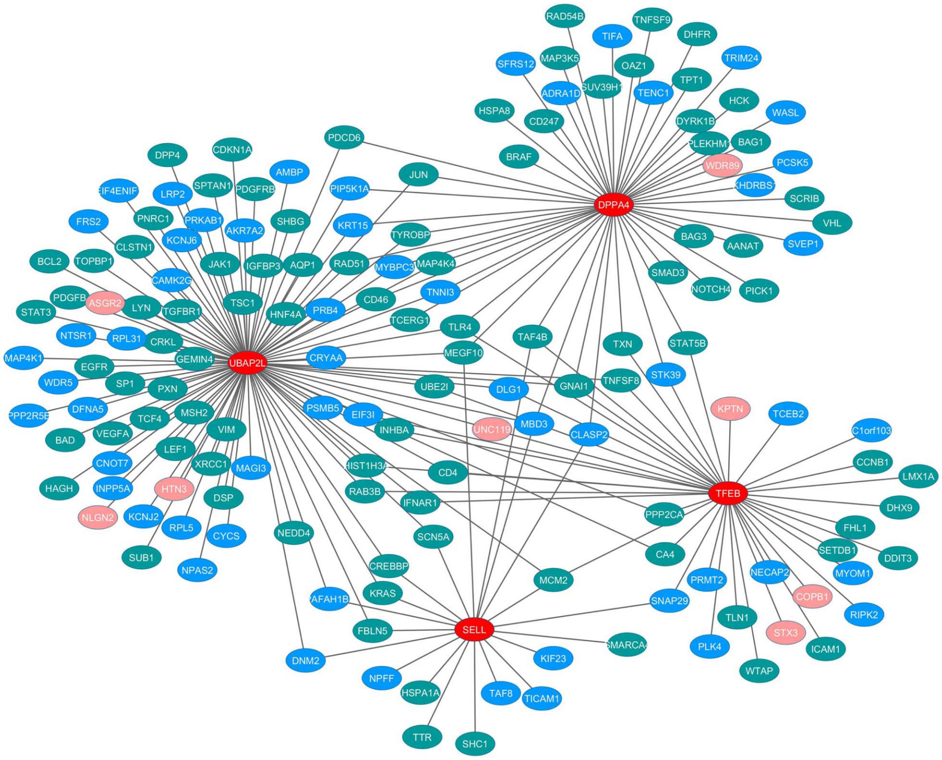

Figure 2. The four candidate genes for ovarian cancer and their neighbor genes in the condition specific network. The four candidate ovarian cancer-related genes are colored in red, ovarian cancer-related genes in green, cancer-related genes in blue and the remaining genes in pink. Edges represent significant co-expressions between genes in the given ovarian cancer.
doi:10.1371/journal.pone.0104993.g002

global solution to our problem. (6) A weight parameter λ is chosen to make the solution stable to the problem, and it is also adaptive to the specific dataset being analyzed.

WMAXC integrates gene expressions and the PPI network, and the positive scaling parameters α and β in Equation (10) are contribution factors of the PPI network in identifying disease-specific genes. We expect that for larger α and β values, the accuracy will increase with a high-quality PPI network, while false positive genes might be included with a low-quality PPI network. In the Results section, we used a high-quality PPI network, and $\alpha = \frac{\max_{i,j}\{M_{ij}\}}{\max_{i,j}\{DBT_{ij}\}} = 26.336$ and $\beta = \frac{\max_i\{T_i\}}{\max_i\{Deg_i\}} = 31.242$ provide a stable performance, and the fold enrichment is relatively high as around 2.45. To show how the performances of the method are affected depending on PPI data quality and scaling parameters, we simulated three PPI data sets with different qualities by randomly removing a fraction of existing edges from the PPI network and randomly adding the same numbers of edges into the PPI network. Then, the simulated PPI data sets were used with three different sets of scaling parameters in our model and corresponding results for each scaling parameter are given in Table 3. For relatively high-quality data sets such as the original PPI and data set-1, the fold enrichment was increased with larger values of α and β. On the other hand, for the low-quality PPI data sets such as data set-3, the fold enrichment was decreased with larger values of α and β.

WMAXC is flexible. It can be simply adapted to directed graphs or even to the integration of gene expression and pathways. For example, instead of using a PPI network, the union of regulatory pathways can be used to represent directed interaction and to compute the DBT score for a pair of genes. In this case, the weight matrix W is non-symmetric, and only slight modifications are required to construct a bio-molecular network from the gene expression profile. The solution to the constrained optimization problem can be approximated by combining the genetic algorithm and the projection procedure.

Table 3. Performances on simulated PPI data with different scaling parameters.

Parameters	Original PPI network	PPI data set-1	PPI data set-2	PPI data set-3
$\alpha = 5,\ \beta = 6$	2.2704	1.7681	2.0443	1.7569
$\alpha = 26.336,\ \beta = 31.224$	2.454	2.3092	2.2176	1.615
$\alpha = 50,\ \beta = 60$	2.4592	2.3843	2.1657	1.5638

For the PPI data set-1, set-2 and set-3, 30%, 50% and 70% of edges from the original data are randomly removed and then the same number of edges are randomly added, respectively. Performances are measured using the fold enrichment, which is described in Table 2.
doi:10.1371/journal.pone.0104993.t003

Supporting Information

File S1 Supplementary material. The combined supporting information file contains multiple supporting Figures, Tables and Descriptions of some fundamental concepts used in the work. (DOCX)

Author Contributions

Conceived and designed the experiments: BA HL. Performed the experiments: BA. Analyzed the data: BA HL. Contributed reagents/materials/analysis tools: HL. Wrote the paper: BA HL.

References

1. Wu X, Jiang R, Zhang M, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4: 189.
2. Barabási A, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12: 56–68.
3. Waaijenborg S, Zwinderman A (2009) Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. The American Journal of Pathology 10: 315.
4. Witten D, Tibshirani R (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8: 28.
5. Chen L, Xuan J, Riggins R, Wang E Y Hoffman, Clarke R (2010) Multilevel support vector regression analysis to identify condition-specific regulatory networks. Bioinformatics 26: 1416–1422.
6. Ideker T, Ozier O, Schwikowski B, Siegel A (2003) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18: s233–s240.
7. Chuang H, Lee E, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Molecular Systems biology 3.
8. Chen L, Xuan J, Riggins R, Wang E Y Hoffman, Clarke R (2012) Identifying protein interaction subnetworks by a bagging markov random field-based method. Nucleic Acids Research 41: e42.
9. Taylor I, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology 27: 199–204.
10. Guo Z, Li Y, Gong X, Yao C, Ma W, et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. Bioinformatics 23: 2121–2128.
11. Lai Y, Wu B, Chen L, Zhao H (2004) A statistical method for identifying differential gene-gene coexpression patterns. Bioinformatics 20: 3146–3155.
12. Ma H, Schadt E, Kaplan L, Zhao H (2011) Condition-specific sub-network identification using a global optimization method. Bioinformatics 27: 1290–1298.
13. Qiu YQ, Zhang S, Zhang XS, Chen L Identifying differentially expressed pathways via a mixed integer linear programming model. IET Systems Biology 3: 475–486.
14. Backes C, Rurainski A, Klau W, et al. (2011) An integer linear programming approach for finding deregulated subgraphs in regulatory networks. Nucleic Acids Research 40: e43.
15. Wang Q, Symes A, Kane C, Freeman A, Nariculam J, et al. (2010) A novel role for wnt/ca^{2+} signaling in actin cytoskeleton remodeling and cell motility in prostate cancer. PLoS ONE: e10456.
16. Rajagopalan D, Agarwal P (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics 21: 788–793.
17. Nacu S, Critchley-Thorne R, Lee P, Holmes S (2007) Gene expression network analysis and applications to immunology. Bioinformatics 23: 850–858.
18. Breitling R, Amtmann A, Herzyk P (2004) Graph-based iterative group analysis enhances microarray interpretation. BMC Bioinformatics 5.
19. Ulitsky I, Karp R, R S (2008) Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In Proceedings of Research in Computational Molecular Biology 4955: 347–359.
20. Motzkin S, Straus G (1965) Maxima for graphs and a new proof of a theorem of turan. Canadian Journal of Mathematics 17: 533–540.
21. Tusher V, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98: 5116–5121.
22. Turan P (1941) On an extremal problem in graph theory. Mat Fiz Lapok 48: 435–452.
23. Du D, Pardalos P (1999) Handbook of combinatorial optimization.
24. Haupt R, Haupt S (2004) Chapter 3: The continuous genetic algorithm. Book, Practical Genetic Algorithms Second Edition: 51–66.
25. Michelot C (1986) A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. Journal of Optimization Theory and Applications 50: 195–200.
26. Songsiri J (2011) Projection onto an l_1-norm ball with application to identification of sparse au-toregressive models. Asian Symposium on Automatic Control Vietnam.
27. Kyrillidis A, Becker S, Cevher V, Koch C (2013) Sparse projections onto the simplex. 30th International Conference on Machine Learning (ICML) 28: 235–243.
28. Prasad T, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human protein reference database. Nucleic Acids Research 37: D767–D772.
29. Atlas TCG The results published here are in whole or part based upon data generated by the cancer genome atlas pilot project established by the nci and nhgri. information about tcga and the investigators and institutions who constitute the tcga research network can be found at http://cancergenome.nih.gov. Data accessed 20 July 2011 .
30. Lapointe J, Li C, Higgins J, van de Rijn M, Bair E, et al. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS 101: 811–816.
31. Kaur M, Radovanovic A, Essack M, Schaefer U, Maqungo M, et al. (2008) Ddpc: Database for exploration of functional context of genes implicated in ovarian cancer. Nucleic Acids Research 37: D820–D823.
32. Maqungo M, Kaur M, Kwofie S, Radovanovic A, Schaefer U, et al. (2011) Ddpc: Dragon database of genes associated with prostate cancer. Nucleic Acids Research 39: D980–D985.
33. Resto VD, Burdick M, Dagia N, McCammon S, Fennewald S (2008) L-selectin-mediated lymphocyte-cancer cell interactions under low fluid shear conditions. The Journal of Biological Chemistry 283: 15816–15824.
34. Kuiper R, Schepens M, Thijssen J, van Asseldonk M, van den Berg E, et al. (2003) Upregulation of the transcription factor tfeb in t(6;11) (p21;q13)-positive renal cell carcinomas due to promoter substitution. Human Molecular Genetics 12: 1661–1669.
35. Maldonado-Saldivia J, van den Bergen J, Krouskos M, Gilchrist M, Lee C, et al. (2007) Dppa2 and dppa4 are closely linked sap motif genes restricted to pluripotent cells and the germ line. Stem Cells 25: 19–28.
36. Huang D, Sherman B, Lempicki R (2008) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature protocols 4: 44–57.
37. de Graeff P, Crijns A, ten Hoor K, Klip H, Hollema H, et al. (2008) The erbb signalling pathway: protein expression and prognostic value in epithelial ovarian cancer. British Journal of Cancer 99: 341–349.
38. Park J, Chen X, Trop C, Davidson B, Shih IM, et al. Notch3 over expression is related to the recurrence of ovarian cancer and confers resistance to carboplatin .
39. Yamamura S, Matsumura N, Mandai M, Huang Z, Oura T, et al. (2012) The activated transforming growth factor-beta signaling pathway in peritoneal metastases is a potential therapeutic target in ovarian cancer. International Journal of Cancer 130: 20–28.
40. Satoh F, Mimata H, Nomura T, Fujita Y, Shin T, et al. (2001) Autocrine expression of neurotrophins and their receptors in prostate cancer. International Journal of Urology 8: s28–s34.
41. Kinkade C, Castillo-Martin M, Puzio-Kuter A, Yan J, Foster T, et al. (2008) Targeting akt/mtor and erk mapk signaling inhibits hormone-refractory prostate cancer in a preclinical mouse model. The Journal of Clinical Investigation 118: 3051–3064.
42. Gioeli D, Mandell J, Petroni G, Frierson HJ, Weber M (1999) Activation of mitogen-activated protein kinase associated with prostate cancer progression. Cancer Research 59: 279–284.
43. Abreu-Martin M, Chari A, Palladino A, Craft N, Sawyers C (1999) Mitogen-activated protein kinase kinase kinase 1 activates androgen receptor-dependent transcription and apoptosis in prostate cancer. Molecular and Cellular Biology 19: 5143–5154.
44. da Silva H, Amaral E, Nolasco E, de Victo N, Atique R, et al. (2013) Dissecting major signaling pathways throughout the development of prostate cancer. Prostate Cancer 2013: 23 pages.
45. Gupta S, Iljin K, Sara H, Mpindi J, Mirtti T, et al. (2007) Fzd4 as a mediator of erg oncogene- induced wnt signaling and epithelial-to-mesenchymal transition in human prostate cancer cells. Cancer Research 70: 6735–6745.
46. Wikstrm P, Stattin P, Franck-Lissbrant I, Damber JE, Bergh A (1998) Transforming growth factor beta1 is associated with angiogenesis, metastasis, and poor clinical outcome in prostate cancer. The Prostate 37: 19–29.
47. Adler H, McCurdy M, Kattan M, Timme T, et al. (1999) Elevated levels of circulating interleukin-6 and transforming growth factor-β1 in patients with metastatic prostatic carcinoma. The Journal of Urology 161: 182–187.
48. Shariat S, Shalev M, Menesses-Diaz A, Kim I, Kattan M, et al. (2001) Preoperative plasma levels of transforming growth factor beta1 (tgf-β1) strongly predict progression in patients undergoing radical prostatectomy. Journal of Clinical Oncology 19: 2856–2864.
49. Wu YM, Robinson D, Kung HJ (2004) Signal pathways in up-regulation of chemokines by tyrosine kinase mer/nyk in prostate cancer cells. Cancer Research 64: 7311–7320.