molecules

MDPI

Article

Multi-Objective Optimization Algorithm to Discover Condition-Specific Modules in Multiple Networks

Xiaoke Ma [1,*], Penggang Sun [1] and Jianbang Zhao [2]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; psun@mail.xidian.edu.cn

[2] College of Information Engineering, Northwest Agriculture and Forestry University, Yangling, Xianyang 712100, China; zhaojianbang@nwsuaf.edu.cn

* Correspondence: xkma@xidian.edu.cn; Tel.: +86-29-8820-2427

Received: 27 October 2017; Accepted: 11 December 2017; Published: 14 December 2017

Abstract: The advances in biological technologies make it possible to generate data for multiple conditions simultaneously. Discovering the condition-specific modules in multiple networks has great merit in understanding the underlying molecular mechanisms of cells. The available algorithms transform the multiple networks into a single objective optimization problem, which is criticized for its low accuracy. To address this issue, a multi-objective genetic algorithm for condition-specific modules in multiple networks (MOGA-CSM) is developed to discover the condition-specific modules. By using the artificial networks, we demonstrate that the MOGA-CSM outperforms state-of-the-art methods in terms of accuracy. Furthermore, MOGA-CSM discovers stage-specific modules in breast cancer networks based on The Cancer Genome Atlas (TCGA) data, and these modules serve as biomarkers to predict stages of breast cancer. The proposed model and algorithm provide an effective way to analyze multiple networks.

Keywords: multiple networks; specific modules; multi-objective optimization; network analysis

1. Introduction

Recent advances in high-throughput biological technologies enable the generation of genome-wide profiles of many patients with various conditions, such as clinical stages, cancer subtypes and time points. Additionally, the network has been proven to be powerful for describing and analyzing the profile data, for which each vertex represents a gene and each edge corresponds to an interaction between a pair of genes. For example, in gene co-expression networks [1], the weight on an edge quantifies the correlation between a pair of genes on the basis of the gene expression profiles. There are various biological networks, such as gene regulation networks [2], signal transduction networks [3], protein–protein interaction (PPI) networks [4], disease networks [5] and gene regulation networks [6–8].

The accumulated biological networks provide an opportunity to explore the mechanisms of cells via mining the graph patterns. Great efforts have been devoted to network analysis, for which the graph patterns shed light on the structure–function relations in biology. For example, Taylor et al. [9] analyzed the PPI network and demonstrated that the genes with large degrees (hub genes) play a critical role in the prognosis of breast cancer. Among these graph patterns, module detection in networks has been extensively studied because this plays an important role in revealing the mechanisms of cells. For example, the dense subgraphs in protein interaction networks are very likely to be protein complexes, which are a cornerstone of many biological processes, and together they form various types of molecular machinery that perform a vast array of biological functions [10]. Furthermore, Ideker et al. [11] showed that the pathways through which genes are differentially expressed between two cohorts of cancer patients serve as biomarkers for predicting cancer metastasis.

Thus, great efforts have been devoted to discovering modules in networks [12–16]. These algorithms mainly differ in their characterization of the module structure and their strategy of module discovery.

Although these methods are promising in discovering modules in networks, they solely focus on identifying a module in a network. In fact, each gene has multiple attributes, indicating that an interaction cannot fully characterize the relation between a pair of genes. For example, proteins possess multiple features, such as physical and co-localization features [17]. To this end, the interactome of proteins for some organisms is up to several distinct network layers accounting for different genetic and physical interactions, each layer containing thousands of protein–protein relationships [18]. The cancer deleterious is dynamic, implying that multiple networks are required to model the progression of diseases, for which each network corresponds to a specific stage.

Fortunately, many algorithms have been developed to extract modules in multiple networks [14–22]. For instance, Ma et al. [20,21] designed the *M-Module* algorithm to discover common modules within multiple networks, which can trace the dynamics of pathways associated with cancer progression. Kelly et al. [16] extracted the conserved modules in multiple networks for various species, which can infer homologous proteins across species. These results demonstrate that discovering graph patterns within multiple networks is promising.

Although great efforts have been devoted to common module detection, few attempts have been made to extract the condition-specific modules in multiple networks, because it is difficult to characterize the specific modules. To accurately depict the specific modules, we must balance the specificity and modularity of modules. Currently, the available algorithms handle this issue by separating the specificity and modularity. Specifically, for each condition, a specific network is constructed for which the edge weight quantifies the specificity of the corresponding edge across all the conditions (details are presented in the next section). Then, module search algorithms for the constructed network, such as WGCNA [23], are employed to obtain the modules. The advantage of this strategy is simplicity, as any module search algorithm can be directly applied. However, it is difficult to achieve a good trade-off between the specificity and modularity because these are independent; this is the major motivation of the present study.

To overcome this problem, an efficient heuristic algorithm is proposed for the specific modules in multiple networks (SMMN), which discovers the condition-specific modules by considering multiple networks without collapsing networks [24]. However, the SMMN algorithm transforms the problem into a single objective optimization, which cannot fully characterize the condition-specific modules in multiple networks. However, it has been shown that intelligent algorithms, such as genetic algorithms (GAs) and particle swarm optimization (PSO), provide an effective strategy to address the optimization problems. For example, Kowk et al. showed that PSO algorithms are effective and efficient in image processing [25,26], industry applications [27] and graph clustering [28]. Knowles et al. [29] demonstrated that the multiobjective optimization is promising in bioinformatics. Inspired by the intelligent algorithms, we present a multi-objective genetic algorithm for condition-specific modules (MOGA-CSM) for condition-specific modules in multiple networks. We demonstrate that the MOGA-CSM outperforms state-of-the-art methods by using artificial and real-world multiple networks.

The rest of the paper is organized as follows: Section 2 proposes the mathematical model and algorithm. The related materials are presented in Section 3. The experimental results are provided in Section 4. The conclusion is discussed in Section 5.

2. Methods

In this section, we first discuss the mathematical model for the condition-specific modules and then describe the MOGA-CSM. It is shown that the traditional algorithms cannot effectively characterize the specificity of modules within multiple networks [24] (Figure 1a,b). The ultimate goal is to develop a multi-objective GA for this issue (Figure 1c,d).

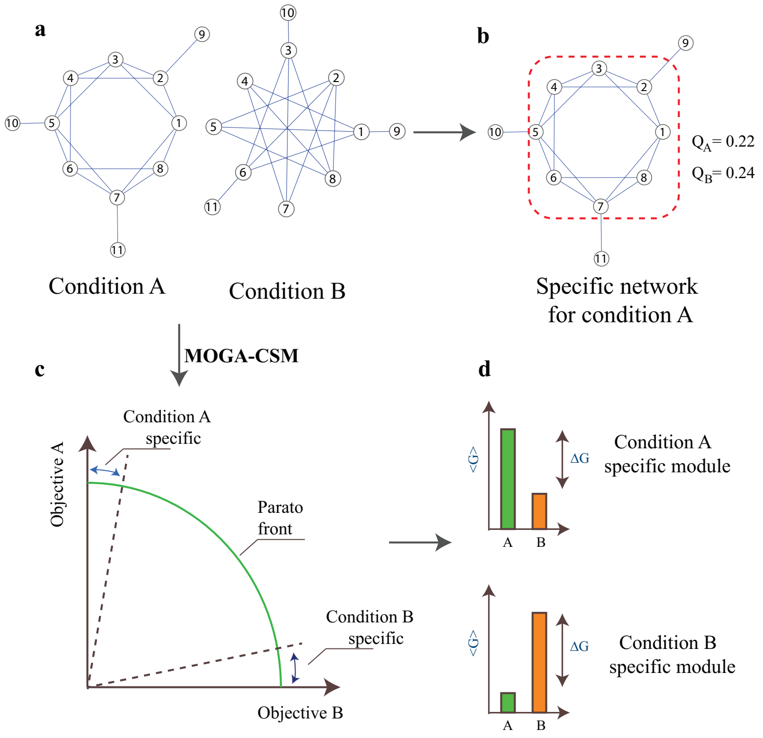


Figure 1. A schematic example of the limitations of the available approaches for condition-specific modules and the overview of the multi-objective genetic algorithm for condition-specific modules (MOGA-CSM). (**a**,**b**) Limitation of the current algorithms: (**a**) Two networks under conditions A and B; (**b**) The condition A-specific networks, for which the module is surrounded by the red dashed line, are specifically obtained by the WGCNA algorithm. The connectivity of the condition A-specific module obtained by the current algorithm in network B is even stronger than that in network A, which contradicts intuition; (**c**,**d**) Overview of the proposed algorithm: (**c**) the MOGA-CSM transforms the condition-specific module detection into a multi-objective optimization problem, for which the specific modules can be obtained by maximizing the connectivity of modules and minimizing the connectivity of modules in other networks; and (**d**) the connectivity of specific modules obtained by MOGA-CSM.

2.1. Multi-Objective Mathematical Model

Let $\{1, 2, \ldots, M\}$ be a finite set of conditions, and let the attached subscript m be the value of the variable under condition m. The multiple network $\mathcal{G}$ is defined as a sequence of networks $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$, where G_m is the network at condition m with a vertex set V and an edge set E_m. The adjacency matrix for $\mathcal{G}$ is defined as $\mathcal{W} = (w_{ijm})_{n\times n\times M}$, where n is the number of genes in $\mathcal{G}$ (i.e., $n = |V|$) and w_{ijm} is the weight on the edge connecting the ith and jth gene in G_m.

Given network $G_m = (V, E_m)$, the module detection aims at obtaining a hard partitioning of V, that is, $\{C_{1m}, C_{2m}, \ldots, C_{km}\}$ (denoted by $\{C_{im}\}_{i=1}^{k}$, such that $C_{im} \cap C_{jm} = \emptyset$ if $i \neq j$ and $V = \sum_i C_{im}$), where k is the number of modules. Given the partitioning $\{C_{im}\}_{i=1}^{k}$ of G_m, an $n \times k$ index matrix X is constructed to represent the memberships of genes such that columns correspond to modules and rows correspond to genes. Element x_{ij} = 1 if the ith gene belongs to module C_{jm} and is 0 otherwise. The connectivity of module C_{tm} in network G_ι is quantified by the modularity Q [30]. According to [24], the overall function of the condition-specific modules for the condition m is defined as

$$\begin{array}{ll} \max\limits_{X} & F(X) \\ & x_{ij} \in \{0,1\} \\ \text{s.t.} & \sum\limits_{j=1}^{k} x_{ij} = 1 \\ & \sum\limits_{i=1}^{n} x_{ij} \geq 1 \end{array} \tag{1}$$

where $F(X) = (F_1(X), \ldots, F_m(X), \ldots, F_M(X))$ are the multi-objective functions, for which $F_i(X) = Q_i(\{C_{tm}\}_{t=1}^{k})$ for $i \neq m$, and $F_m(X) = 1 - Q_i(\{C_{tm}\}_{t=1}^{k})$. Because this is an NP-hard problem, we employ a heuristic algorithm to obtain the solution for Equation (1).

Differently from [24] using a single objective optimization problem, we present a GA to directly address the multi-objective optimization problem in Equation (1). We first introduce the *Pareto front* for solutions. Given two solutions $X^{[1]}$ and $X^{[2]}$ to the multi-objective optimization problem in Equation (1), $X^{[1]}$ is dominated by $X^{[2]}$, denoted by $X^{[1]} \prec X^{[2]}$, if and only if

$$\forall i : F_i(X^{[1]}) \leq F_i(X^{[2]}) \wedge \exists i, \text{s.t.} F_i(X^{[1]}) < F_i(X^{[2]})$$

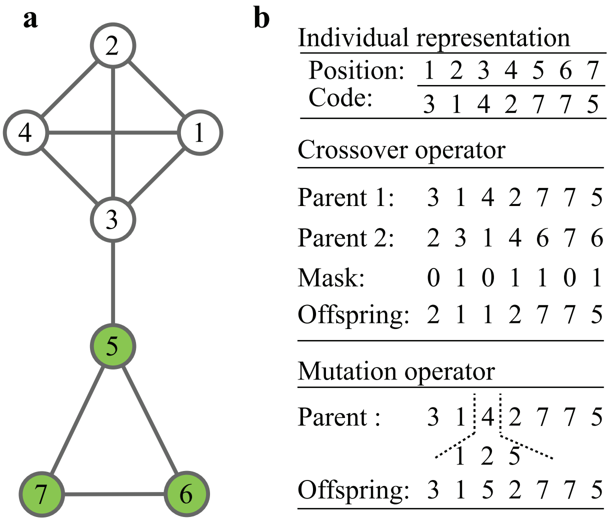
Instead, a nondominated solution is one for which an improvement in one objective requires a degradation of the other(s). The set of these nondominated solutions is called the *Pareto front*.

2.2. The MOGA-CSM

GAs are a class of adaptive search methods inspired by natural evolution [31], which evolves a population of individuals using the operators of *crossover* and *mutation*. Each individual represents a candidate solution to the problem in Equation (1). The *fitness value* of an individual quantifies how good it is with respect to the other solutions in the population. The crossover operator generates an individual by combining two individuals in the population, while the mutation operator randomly alters the individual. GAs, for example, the *nondominated sorting genetic algorithm (NSGA-II)*, have been successfully applied to multi-objective optimization problems (MOGA) [32]. Recently, Gu et al. [33] proposed an innovative semi-active storey isolation system by utilizing the NSGA-II based on the dynamic crowding distance (DCD), which significantly improved the performance. GAs have been widely applied to network clustering [34].

Individual representation: The locus-based adjacency representation is adopted [35–37]. In this graph-based representation, an individual is denoted by $P = (g_1, \ldots, g_n)$, where g_i is one of the neighbors of node i such that (i, g_i) is an edge belonging to one of the modules of the graph. The schematic example of representation for a graph (Figure 2a) is illustrated in Figure 2b, where the two modules are encoded.

To decode the module within an individual, the disjoint set algorithm [38] is employed, in which the modules correspond to a set of disjoint dynamic sets, where each set is represented by a rooted tree. The root is defined as the representative, and the rest node i of the tree points only to its parent $parent(i)$. The level of node i is defined as the length of the shortest path connecting i to the root, denoted by $level(i)$. At the beginning, the decode procedure initializes each vertex as a set (step 1); that is, the parent of i is itself and the level is 0. Then, for each edge (i, g_i), it tracks the roots of the tree(s) of i and g_i, denoted by r_1 and r_2 (step 2). If the node i and g_i belong to various trees, that is, $r_1 \leq r_2$, it merges the two trees as a new tree (step 3). Otherwise, it updates the levels of genes within the tree (step 4).



Individual representation

Position:	1	2	3	4	5	6	7
Code:	3	1	4	2	7	7	5

Crossover operator

Parent 1:	3	1	4	2	7	7	5
Parent 2:	2	3	1	4	6	7	6
Mask:	0	1	0	1	1	0	1
Offspring:	2	1	1	2	7	7	5

Mutation operator

Parent :	3	1	4	2	7	7	5
		1	2	5			
Offspring:	3	1	5	2	7	7	5

Figure 2. Illustration of procedure of the multi-objective genetic algorithm for condition-specific modules (MOGA-CSM): (**a**) A network with seven vertices partitioned into two modules, $\{1,2,3,4\}$ and $\{5,6,7\}$; (**b**) Procedure of MOGA-CSM: top panel corresponds to a locus-based representation of the network on the left; middle panel contains the example of uniform crossover; bottom panel represents the mutation.

Algorithm 1 Decoding Procedure

```
Input:
    $P$: an individual of the population.
Output:
    $X_P$: the module structure of $P$.
1: For each node $i \in V$, set $parent(i) = i$ and $level(i) = 0$.
2: For each edge $(i, g_i)$, find the roots of $i$ and $g_i$, denoted by $r_1$ and $r_2$, respectively.
3: If $r_1 = r_2$, update $level(r_2) = level(r_1) + 1$; else goto step 4.
4: If $level(r_1) > level(r_2)$, update $parent(r_2) = r_1$; else update $parent(r_1) = r_2$.
5: return Disjoint sets.
```

Crossover operator: Given two parents in population $P^{[1]} = (g_1^{[1]}, \dots, g_n^{[1]})$ and $P^{[2]} = (g_1^{[2]}, \dots, g_n^{[2]})$, the child $S = (g_1, \dots, g_n)$ is generated by randomly selecting each component from one of the parents; that is, g_i is either $g_i^{[1]}$ or $g_i^{[2]}$. This procedure is fulfilled by a random binary mask: when mask is 0, $g_i = g_i^{[1]}$; otherwise $g_i = g_i^{[2]}$ (Figure 2c). The advantage of the crossover is to maintain node connections in the child individual.

Mutation operator: Given an individual $P^{[1]} = (g_1, \dots, g_n)$, the mutation operator randomly changes the value of g_i. To guarantee the connections, only the neighbors of node i are candidates for replacing g_i (Figure 2d).

The pseudo-code of MOGA-CSM is presented in Algorithm 2. Given the multiple networks $\mathcal{G} = \{G_1, G_2, \dots, G_M\}$ and the condition m, MOGA-CSM generates a population of random individuals. Specifically, given an individual $P = (g_1, \dots, g_i)$, we randomly select one vertex from the neighbors of vertex i and assign it g_i. After the population is generated, it decodes the individuals of the population to produce the partitioning and evaluates the objective values. The individuals are ranked according to the Pareto dominance. The crossover and mutation operators are employed to create the new population. Finally, the solutions in the Pareto front are returned, where each of them corresponds to a trade-off among multiple functions. Therefore, a criterion is required to select one solution with respect to another. We choose the solution with the maximum modularity for G_m on the basis of the fact that the Pareto front has already selected the nondominated solutions that best satisfy all the functions.

Algorithm 2 The MOGA-CSM

```
Input:
    $\mathcal{G}$: the involved multiple networks.
    $m$: the specific condition.
Output:
    $\{C_{tm}\}_{t=1}^{k}$: the condition-specific modules.
1: Create a population of random individuals for $G_m$.
2: Decode each individual $P$ of the population using the decoding procedure.
3: Obtain the rank of each individual according to nondomination rank.
4: Generate new offspring using the crossover and mutation operators.
5: Combine the parents and offspring into a new pool and rank them.
6: Select the individuals with lower rank for the next generation.
7: If the termination criterion is not satisfied, goto step 1; otherwise, goto step 8.
8: return $\{C_{tm}\}_{t=1}^{k}$ with the maximum modularity.
```

2.3. Algorithm Analysis

In terms of space complexity, the space for the adjacency matrix of multiple networks is $O(n^2M)$. For each network, the space for the population is $O(np)$, where p is the size of the population. The space complexity of the indicator matrix for modules is $O(nkM)$, where k is the number of modules. Because $k \ll n$, the total space complexity of the proposed algorithm is $O(n^2M)$.

In terms of time complexity, the MOGA-CSM makes use of *NSGA-II* to rank the non-dominance [34], which requires time $O(tp(\log p)^{h-1})$, where t is the number of generations, p is the size of the population, and M is the number of objective functions. Because MOGA-CSM optimizes M networks, the time complexity is $O(tp(\log p)^{M-1})$. For each generation, the crossover needs $O(n)$ time, and mutation requires $O(1)$ time. Furthermore, the decoding procedure requires $O(n \log n)$ time [38]. Thus, the total time complexity of MOGA-CSM is $O(tp(\log p)^{M-1} n \log n)$.

3. Materials

3.1. Statistical Significance of Specific Modules

The statistical significance of specific modules was computed on the basis of the null score distribution of specific modules generated using randomized networks. Each network was completely randomized 100 times by degree-preserved edge shuffling. To construct the null distribution for specific module scores, we performed the MOGA-CSM on the randomized networks. Using the null distribution, the empirical p-value of a specific module was calculated as the probability of the module having the observed score or greater by chance; p-values were corrected for multiple testing using the method of Benjamini–Hochberg [39]. An adjusted p-value of 0.05 was considered as significant.

3.2. Features for Support Vector Machine on Specific Modules

Given a module C, we normalize the expression level of each gene across all samples using the z-score transformation [11], denoted by Exp_{ij} for the ith gene and jth patient. For each sample j, the activity score of the kth module is defined as the average gene expression of all genes within the module, that is,

$$e_C = \sum_{i \in C} Exp_{ij} / \sqrt{|C|} \tag{2}$$

where $|C|$ is the number of genes in C. For each patient sample, a feature vector is constructed by all modules.

3.3. Artificial Networks

The artificial network is introduced in [30]. In each network, 128 nodes are grouped into 4 clusters of equal size. Every node has an average degree of 16 and shares k_{out} edges connecting nodes outside of the module to which it belongs. As parameter k_{out} increases from 1 to 8, the detection of clusters in the networks becomes increasingly difficult.

3.4. Breast Cancer Networks

The gene expression data for breast cancer was downloaded from the TCGA Data Portal, where the clinical stage information for patients is also available. The RPKM (Reads Per Kilobase per Million mapped reads) values are used. There are 715 samples across four stages (stage I: 119; stage II: 407; stage III: 189).

For each stage, we construct a gene co-expression network, where the edge weight is defined as the absolute value of the Pearson correlation of the gene expression profiles of a pair of genes. To remove indirect correlation due to a third gene, we use the first-order partial Pearson correlation coefficient (PCIT package [40]). The breast cancer networks contain 6643 genes and about 2.6 million edges.

4. Results

To fully test the performance of the proposed algorithm, we compared MOGA-CSM with the available algorithms. We note that the current approaches differ greatly on their strategy of how to extract the modules from the constructed condition-specific networks. Thus, we adopted three well-known algorithms, including the WGCNA [23], spectral clustering (SPEC) [41] and nonnegative matrix factorization (NMF) algorithms [42]. The reason that these algorithms were selected was that they achieve excellent performance in detecting modules in networks.

Two types of networks, both artificial and real biological networks, were employed for a comparison among various algorithms. The artificial networks were adopted to test the accuracy of the MOGA-CSM, and the breast cancer networks were used to determine the the applicability of the proposed algorithm in discovering condition-specific modules in real networks with a strong background. The parameters for the MOGA-CSM were set as follows: $crossoverrate = 0.8$ and $mutationrate = 0.2$. The reason was that in general, a high crossover rate and low mutation rate are suggested in GAs. Furthermore, we set $elitereproduction = 10\%$ of the population size, and the number of generations as 500 (how the parameters affect the performance is discussed in the following section).

4.1. Benchmarking Performance of the Artificial Networks

In the artificial networks, we constructed two networks by combining a network with a known module structure (Materials) and a size-matched random network. Therefore, the modules in the benchmark network were specific modules, because the random network was not expected to exhibit a modular structure. To quantify the performance of the algorithms, the modularity Q was used.

Prior to giving the performance of the algorithms, we first investigate the effects of the parameters for MOGA-CSM on the artificial networks. The results are shown in Figure 3A, where the crossover rate ranges from 0.1 to 0.8 with a gap 0.1 and the mutation rate ranges from 0.2 to 0.8 with a gap 0.2. It can be observed that they do not present high variation.

We compare the WGCNA, SPEC, NMF and MOGA-CSM algorithms on the artificial networks in terms of accuracy, as shown in Figure 3B. From this, we conclude that the performance of the algorithms decreases dramatically as k_{out} increases from 1 to 8, because the module structure becomes fuzzy as k_{out} increases. For example, the Q value is 0.7 when k_{out} = 1, and it is 0.22 when k_{out} = 8. Furthermore, the proposed algorithm has a similar performance to NMF when $k_{out} \leq 4$, while it outperforms NMF if $k_{out} > 4$. Furthermore, both MOGA-CSM and NMF are superior to the SPEC and WGCNA algorithms. The SPEC algorithm is inferior to others, indicating that the spectral features are insufficient to characterize the specific modules.

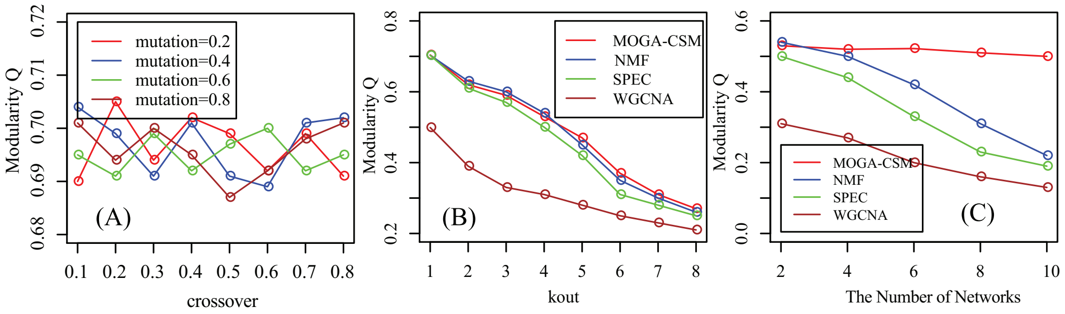


Figure 3. Performance of the compared algorithms on artificial multiple networks. (**A**) Parameter effect: modularity for different combinations of crossover and mutation rates for the artificial networks; (**B**) Performance as a function of the amount of noise in simulated networks, where modularity *Q* is used as the performance measure. Shown here are average *Q* values of 50 runs of each method at each noise level; (**C**) Performance as a function of the number of networks in the artificial networks, where *Q* is used as the performance measure.

As shown in Section 2, the available methods are sensitive to the number of networks. Therefore, we investigated whether the proposed algorithm is also sensitive to the number of networks. We increased the number of random networks from 2 to 10 and tested the performance of various algorithms, as shown in Figure 3C. From this, we conclude that the performance of NMF, SPEC, and WGCNA decreases dramatically as the number of networks increases. However, the MOGA-CSM is not sensitive to the number of networks. The results demonstrate that the proposed algorithm is more accurate and robust than state-of-the-art approaches in discovering condition-specific modules.

4.2. Benchmarking Performance of the Breast Cancer Networks

On the basis of the clinical stages for breast cancer, we constructed a gene co-expression network for each stage (Materials). By applying the MOGA-CSM to the breast cancer networks, we obtained 27 (stage I), 5 (stage II) and 9 (stage III) specific modules (Figure 4A).

The homeostasis has been proven to be a critical complex for breast cancer diagnosis and therapy [43]. Remarkably, the MOGA-CSM obtained a stage I-specific module that was significantly enriched by homeostasis (p-value = 1.6×10^{-2}, corrected by BH test), as shown in the top panel of Figure 4C. There were six genes (*ANTXR2*, *FHL1*, *AVPR2*, *PLEKHM3*, *PKD2*, and *CNRIP1*), for which genes PKD2, FHL1 and AVPR2 had the function homeostasis. To check whether the module was stage I-specific, we calculated the connectivity of the modules in all three networks, as shown in the bottom panel of Figure 4C. The density of the module in the stage I network was 0.4, while the density in the stage II and III networks was 0.13. These results met our expectation, because the connectivity was strong in the stage I network and weak in others.

Then, we checked the functions of the genes within the stage-specific modules, as shown in Figure 4B. We found that the genes within stage I-specific modules were more likely to be enriched by the signaling pathways (red bars), such as the tumor necrosis factor-mediated signaling pathway (p-value = 2.7×10^{-2}, corrected by BH test), the receptor guanylyl cyclase signaling pathway (p-value = 4.5×10^{-2}, corrected by BH test), and the endothelial growth factor receptor signaling pathway (p-value = 4.7×10^{-2}, corrected by BH test). These signaling pathways are critical for breast cancer [44].

To check the specificity of the modules obtained by MOGA-CSM, we compared the distribution of densities of modules for each network. The results are shown in Figure 4D, where it is indicated that the connectivity of the specific modules can capture the specificity, because the modules are well connected in the corresponding network and weak in others. For example, the density of stage I-specific modules is more significant than that of modules in networks at stage II and III (stage I vs. stage II: p-value = 4.6×10^{-6}; stage I vs. stage III: p-value = 3.1×10^{-5}, Student's t-test).

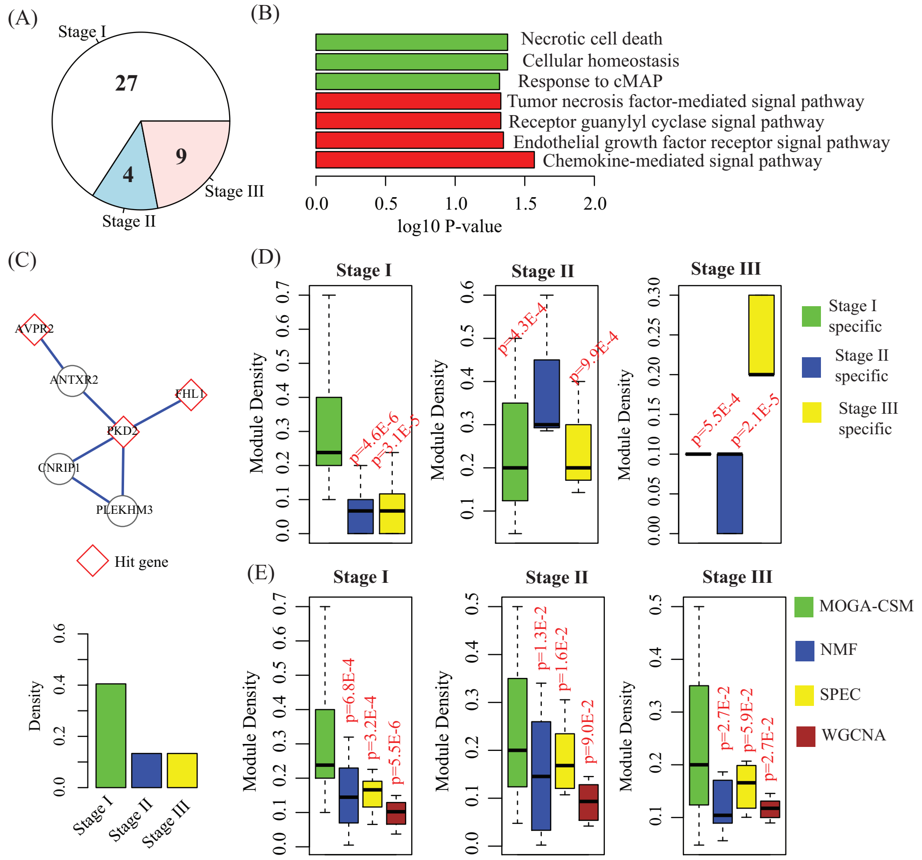


Figure 4. Performance of the compared algorithms on the TCGA breast cancer networks. (**A**) Piechart for the number of stage-specific modules obtained by the multi-objective genetic algorithm for condition-specific modules (MOGA-CSM); (**B**) the barplot for the functions of genes within specific modules, where the red color indicates the genes within stage I-specific modules and green indicates the genes within stage II- and III-specific modules; (**C**) a schematic example of a stage I-specific module obtained by the MOGA-CSM, where the top panel is the topological structure of the module and the bottom panel contains the density of the module in each network; (**D**) distribution of density of stage-specific modules in each network; (**E**) comparison of various algorithms in terms of distribution of density of specific modules obtained by algorithms in each network.

Finally, we compared the WGCNA, SPEC, NMF and MOGA-CSM algorithms in terms of discovering the condition-specific modules by applying them to the breast cancer networks. We compared the distribution of densities of the modules obtained for each algorithm, which are shown in Figure 4E. From these, it is easy to conclude that the MOGA-CSM is significantly better than the others, as the density of the modules obtained by MOGA-CSM was much higher than that of others. For instance, in the stage I network, the means of the density of the modules were 0.22 (MOGA-CSM), 0.17 (NMF), 0.19 (SPEC), and 0.12 (WGCNA) (MOGA-CSM vs. NMF: p-value = 6.8×10^{-4}; MOGA-CSM vs. SPEC: p-value = 3.2×10^{-4}; MOGA-CSM vs. WGCNA: p-value = 5.5×10^{-6}, Student's t-test). These results imply that the proposed algorithm is more accurate than state-of-the-art approaches for the specific module detection in biological networks.

4.3. Stage-Specific Modules Serve as Biomarkers to Predict Breast Cancer Stages

Taylor et al. [9] showed that the hub genes are predictive for breast cancer diagnosis. Ideker et al. [11] demonstrated that the modules can serve as biomarkers to predict metastasis of breast cancer. Thus, we hypothesized that the stage-specific modules could also be used to predict the stages of breast cancer.

For a baseline comparison, we compared the classification accuracy by using the following feature sets: stage-specific modules generated by SPEC, NMF, WGCNA and MOGA-CSM. We trained the support vector machine (SVM) classifier to perform multi-class classification. The SVM employs the accuracy (percentage of patients that are corrected classified) to measure performance. The results of TCGA breast cancer data by using five-fold cross-validation are presented in Figure 5A. The stage-specific modules obtained by SMMN were more discriminative than the others. Specifically, the MOGA-CSM had a significantly higher accuracy than WGCNA (73.3% vs. 69.9%). The WGCNA algorithm had a similar performance to NMF, and they outperformed the other methods. These results demonstrate that the stage-specific modules obtained by MOGA-CSM capture the specificity of pathways for breast cancer progression.

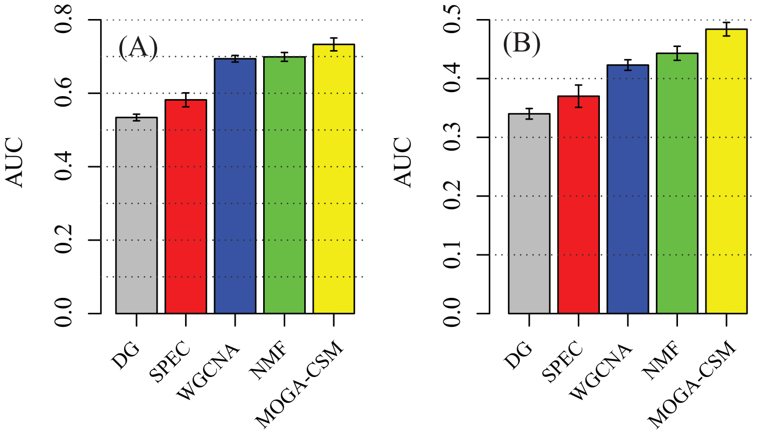


Figure 5. Subtype-specific methylation modules improve the accuracy of breast cancer stage classification using 50 independent five-fold cross-validations. (**A**) Classification accuracy of breast cancer stages using different feature sets, including the stage-specific modules obtained by various algorithms. Accuracy is defined as the number of patient samples correctly classified. The *Y*-axis is the accuracy and the error bar is for the standard deviation; (**B**) External validation by training on TCGA data and testing on external data.

To further validate the performance of various algorithms, we evaluated the performance of the SVM classifiers by using external data (GSE5874). We trained the SVM classifier on the TCGA data and tested on an external microarray dataset. The consistent results indicated that the performance was not due to the hidden confounding factors in the TCGA dataset (Figure 5A). The accuracy of MOGA-CSM was 48.4%, while the accuracy was 44.3%, 42.3%, 37.1%, and 34.3% for NMF, WGCNA, SPEC and differentially expressed genes (DG), respectively. The results show that the proposed algorithm is better than the available approaches in terms of discovering specific modules in multiple networks.

4.4. Benchmarking Performance of Cancer Co-Methylation Networks

To fully explore the performance of the proposed algorithms, we compared these algorithms by using the cancer subtype long non-coding RNA co-methylation networks [45]. There are four subtypes of breast cancer, Luminal A, Luminal B, Her2 and Basal-like. For each subtype, there is a corresponding long non-coding RNA gene co-methylation network.

The results are shown in Figure 6, where the distributions of the graph density of the obtained subtype-specific modules are presented. From these panels, it is easy to assert that the proposed method significantly outperforms the others, because the modules are more specific in terms of density for almost all the subtypes. Therefore, we assert that the proposed method is promising in discovering the condition-specific modules.

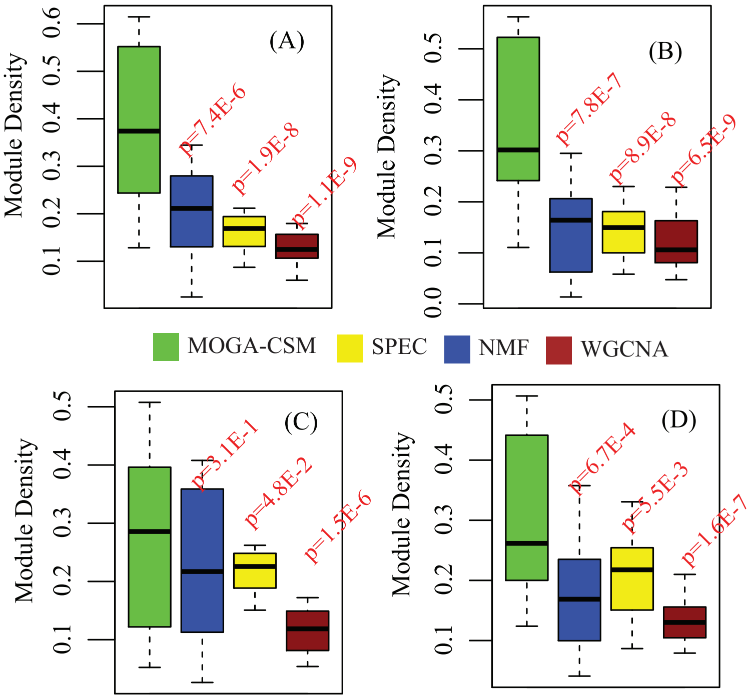


Figure 6. Performance of the compared algorithms on the multiple subtype co-methylation networks for long noncoding RNA networks. Distribution of density of subtype-specific modules in each subtype: (**A**) Luminal A; (**B**) Luminal B; (**C**) Her2 and (**D**) Basal-like.

5. Discussion and Conclusions

Recent technology has enabled the possibility of generating multiple genomic profiling of biological samples for different stages or time points. However, the systematic integrative analysis of multiple-stage (or time-point) data associated with disease progression or cell differentiation for discovering biological relevant patterns is currently lacking. The accumulated multiple networks provide an opportunity to explore the underlying mechanisms of diseases. Although great efforts have been devoted to multiple networks analysis, few attempts have been made to extract the specific modules in multiple networks.

The available algorithms first construct a specific network by compressing the multiple networks. Then, they discover modules in the constructed networks. The strategy is criticized for its low accuracy, because of the separation of specificity and modularity. To overcome this problem, we characterize the specific modules on the basis of the topology of multiple networks rather than the constructed network, which provides a better characterization of modules. Then, a multi-objective optimization model is developed for specific module detection in multiple networks. Finally, a multi-objective optimization algorithm is designed to obtain specific modules. The results demonstrate that the proposed algorithm is better than the current approaches. We wish to point out several unique insights: (i) the integrative analysis of multiple networks without collapsing them is promising, which is overlooked by the available methods; and (ii) the biologically inspired computational approaches, such as GAs, provide an efficient tool to extract the graph patterns in multiple networks.

For further research, we see ample opportunities to improve on the basic concept of the MOGA-CSM. First, although this study uses breast cancer as a proof-of-principle, it is flexible to any diseases, as it is a generalized framework for any cohort of patients with various conditions. Second, data integration might further expand the applicability of the proposed model and algorithm.

Acknowledgments: This work was supported by the NSFC (Grant Nos. 61772394 and 61502363), the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2017JM6030) and the Fundamental Research Funds for the Central Universities (Grant No. Z109021508).

Author Contributions: X.M. and P.S. conceived and designed the approach; X.M., S.P. and J.Z. analyzed the data; X.M. and J.Z. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stuart, J.M.; Segal, E.; Koller, D.; Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **2003**, *302*, 249–255.
2. Vaquerizas, J.M.; Kummerfeld, S.K.; Teichmann, S.A.; Luscombe, N.M. A census of human transcription factors: Function, expression and evolution. *Nat. Biotechnol.* **2009**, *10*, 252–263.
3. Anastas, J.N.; Moon, R.T. WNT signaling pathways as therapeutic targets in cancer. *Nat. Rev. Cancer* **2013**, *13*, 11–26.
4. Schwikowski, B.; Uetz, P.; Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **2010**, *18*, 1257–1261.
5. Menche, J.; Sharma, A.; Kitsak, M.; Ghiassian, S.D.; Vidal, M.; Loscalzo, J.; Barabási, A.-L. Uncovering disease-disease relationships through the incomplete interactome. *Science* **2015**, *347*, 841–841.
6. Zeng, X.; Lin, W.; Guo, M.; Zou, Q. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* **2017**, *13*, e1005420, doi:10.1371/journal.pcbi.1005420.
7. Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A.E.; Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* **2017**, doi:10.1093/bioinformatics/btx622.
8. Li, P.; Guo, M.; Wang, C.; Liu, X.; Zou, Q. An overview of SNP interactions in genome-wide association studies. *Brief. Funct. Genom.* **2015**, *14*, 143–155.
9. Taylor, I.W.; Linding, R.; Warde-Farley, D.; Liu, Y.; Pesquita, C.; Faria, D.; Bull, S.; Pawson, T.; Morris, Q.; Wrana, J.L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **2009**, *27*, 199–204.
10. Hartwell, L.H.; Hopfield, J.J.; Leibler, S.; Murray, A.W. From molecular to modular cell biology. *Nature* **1999**, *402*, C47–C52.
11. Chuang, H.Y.; Lee, E.; Liu, Y.; Lee, D.; Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **2007**, *3*, 140, doi:10.1038/msb4100180.
12. Cancer Genome Atlas Research Network; Albert Einstein College of Medicine; Analytical Biological Services; Barretos Cancer Hospital; Baylor College of Medicine; Beckman Research Institute of City of Hope; Buck Institute for Research on Aging; Canada's Michael Smith Genome Sciences Centre; Harvard Medical School; Helen F. Graham Cancer Center & Research Institute at Christiana Care Health Services; et al. Integrated genomic and molecular characterization of cervical cancer. *Nature* **2017**, *543*, 378–384.
13. Zeng, X.; Zhang, X.; Zou, Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* **2016**, *17*, 193–203.
14. Yang, Y.; Han, L.; Yuan, Y.; Li, J.; Hei, N.; Liang, H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **2014**, *5*, 3231, doi:10.1038/ncomms4231.
15. De Lichtenberg, U.; Jensen, L.J.; Brunak, S.; Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **2005**, *307*, 724–727.
16. Kelley, B.P.; Sharan, R.; Karp, R.M.; Sittler, T.; Root, D.E.; Stockwell, B.R.; Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 11394–11399.
17. Schlecht, U.; Miranda, M.; Suresh S.; Davis, R.W.; St Onge, R.P. Multiplex assay for condition-dependent changes in protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 9213–9218.
18. Stark, C.; Breitkreutz, B.J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. Biogrid: A general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34*, D535–D539.

19. Ma, X.; Tang, W.; Wang, P.; Guo, X.; Gao, L. Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, doi:10.1109/TCBB.2016.2625791.
20. Ma, X.; Gao, L.; Tan, K. Modeling disease progression using dynamics of module connectivity. *Bioinformatics* **2014**, *30*, 2343–2350.
21. Ma, X.; Gao, L.; Karamanlidis, G.; Gao, P.; Lee, C.F.; Garcia-Menendez, L.; Tian, R.; Tan, K. Revealing module dynamics in heart diseases by analyzing multiple differential networks. *PLoS Comput. Biol.* **2015**, *11*, e1004332, doi:10.1371/journal.pcbi.1004332.
22. Gevaert, O.; Tibshirani, R.; Plevritis, S.K. Pancancer analysis of DNA methylation-driven genes using MethyMix. *Genome Biol.* **2015**, *16*, 17, doi:10.1186/s13059-014-0579-8.
23. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559, doi:10.1186/1471-2105-9-559.
24. Ma, X.; Sun, P.; Qin, G. Identifying condition-specific modules by clustering multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, doi:10.1109/TCBB.2017.2761339.
25. Kwok, N.W.; Shi, H.Y.; Ha, Q.; Jia, X. Simultaneous image color correction and enhancement using particle swarm optimization. *Eng. Appl. Artif. Intell.* **2013**, *26*, 2356–2371.
26. Mai, C.; Nguyen, M.T.T.; Kwok, N.W. A modified snsharp masking method using particle swarm optimization. In Proceedings of the 2011 4th International Conference on Image and Signal Processing, Shanghai, China, 15–17 October 2011; Volume 2, pp. 646–650.
27. Fang, G.; Kwok, N.M.; Ha, Q. Automatic fuzzy membership function tuning using the particle swarm optimization. In Proceedings of the 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, China, 19–20 December 2008; Volume 2, pp. 324–328.
28. Folino, F.; Pizzuti, C. An evolutionary multiobjective approach for community discovery in dynamic networks. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1838–1852.
29. Handl, J.; Kell, D.B.; Knowles, J. Multiobjective optimization in Bioinformatics and Computational Biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2007**, *4*, 279–292.
30. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *77*, 026113, doi:10.1103/PhysRevE.69.026113.
31. Holland, J.H. *Adaption in Natural and Artificial Systems*; University of Michigan Press: Ann Harbor, MI, USA, 1975.
32. Sriniva, N.; Deb, K. Multi-objective optimization using nondominated sorting in genetic algorithms. *Evol. Comput.* **1994**, *2*, 221–248.
33. Gu, X.; Li, J.; Li, Y.; Alamdari, M.M. Semi-active storey isolation system employing MRE isolator with parameter identification based on NSGA-II with DCD. *Earthq. Struct.* **2011**, *11*, 1101–1121.
34. Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*; Wiley: Chichester, UK, 2001.
35. Pizzuti, C. A multi-objective genetic algorithm to find communities in complex networks. *IEEE Trans. Evol. Comput.* **2012**, *3*, 418–430.
36. Handle, J.; Knowles, J. An evolutionary genetic algorithm to multi-objective clustering. *IEEE Trans. Evol. Comput.* **2007**, *11*, 56–76.
37. Park, Y.; Song, M. A genetic algorithm for clustering problems. In Proceedings of the 3rd International Conference on Genetic Algorithms, Madison, WI, USA, 22–25 July 1998; pp. 2–9.
38. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2007.
39. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methadol.* **1995**, *57*, 289–300.
40. Reverter, A.; Chan, E.K. Combining partial correlation and an information theory approach to the reverse engineering of gene co-expression networks. *Bioinformatics* **2008**, *24*, 2491–2497.
41. Newman, M.E.J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **2006**, *74*, 036104, doi:10.1103/PhysRevE.74.036104.
42. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.
43. Rodina, A.; Wang, T.; Yan, P.; Gomes, E.D.; Dunphy, M.P.; Pillarsetty, N.; Koren, J.; Gerecitano, J.F.; Taldone, T.; Zong, H.; et al. The epichaperome is an integrated chaperome network that facilitates tumour survival. *Nature* **2016**, *538*, 397–401.

44. Wang, X.; Abraham, S.; McKenzie, J.A.G.; Jeffs, N.; Swire, M.; Tripathi, V.B.; Luhmann, U.F.O.; Lange, C.A.K.; Zhai, Z.; Arthur, H.M.; et al. LRG1 promotes angiogenesis by modulating endothelial TGF-β signalling. *Nature* **2013**, *499*, 306–311.
45. Ma, X.; Yu, L.; Wang, P.; Yang, X. Discovering DNA methylation patterns for long non-coding RNAs associated with cancer subtypes. *Comput. Biol. Chem.* **2017**, *69*, 164–170.