Bioinformatics Advance Access published September 14, 2012

Efficient methods for identifying mutated driver pathways in cancer

Junfei Zhao [1,*], Shihua Zhang [1,*†], Ling-Yun Wu [1] and Xiang-Sun Zhang [1]

[1]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: The first step for clinical diagnostics, prognostics, and targeted therapeutics of cancer is to comprehensively understand its molecular mechanisms. Large-scale cancer genomics projects are providing a large volume of data about genomic, epigenomic, and gene expression aberrations in multiple cancer types. One of the remaining challenges is to identify driver mutations, driver genes and driver pathways promoting cancer proliferation and filter out the unfunctional and passenger ones.

Results: In this study, we propose two methods to solve the so-called Maximum Weight Submatrix problem which is designed to *de novo* identify mutated driver pathways from mutation data in cancer. The first one is an exact method which can be helpful for assessing other approximate or/and heuristic algorithms. The second one is a stochastic and flexible method which can be employed to incorporate other types of information to improve the first method. Particularly, we propose an integrative model to combine mutation and expression data. We first apply our methods onto simulated data to show their efficiency. We further apply the proposed methods onto several real biological data sets such as the mutation profiles of 74 head and neck squamous cell carcinomas samples, 90 glioblastoma tumor samples and 313 ovarian carcinoma samples. The gene expression profiles were also considered for the later two data. The results show that our integrative model can identify more biologically relevant gene sets. We have implemented all these methods and made a package called MDPFinder (**M**utated **D**river **P**athway **Finder**) which can be easily used for other researchers.

Availability: A matlab package of MDPFinder is available at http://zhangroup.aporc.org/ShiHuaZhang

Contact: zsh@amss.ac.cn

1 INTRODUCTION

Cancer is a complex disease which has been one of the most serious threats to human health. People have realized that cancer is related to genome aberrations which include gene mutations, copy number alterations and so on (Hanahan *et al*., 2000). Through these aberrations, the cancer cells can acquire the ability of infinite proliferation while normal cells don't due to the self-correction mechanism. Another dreadful feature of cancer cells is that some of them can spread to other tissues through blood circulation or lymphatic system (Fidler *et al*., 2003). This largely reduces the effectiveness of surgery to treat cancer.

Generally, the genome aberrations in cancer cells can be divided into two types: one type is neutral to cancer proliferation; the other can promote the cancer cell to proliferate infinitely and diffuse (Greenman *et al*., 2007). We usually call the former type of mutations as "passenger mutation", the latter as "driver mutation". Undoubtedly, finding out the driver mutation, driver gene as well as driver pathway is a key to understand the molecular mechanisms of cancer progression which further aid in designing effective drugs to treat cancer (Overdevest *et al*., 2009; Swanton *et al*., 2009).

With the development of high-throughput sequencing technologies, a huge number of mutation profiles of samples for many cancer types are available now (TCGA, 2008, 2011; Stransky *et al*., 2011; Chapman *et al*., 2011). Designing effective bioinformatics tools to mine useful information from these data is a challenging task. In gene level, much effort has been devoted to detect the genes with significantly higher mutation rate across samples than background mutation rate (Beroukhim *et al*., 2007; Getz *et al*., 2007). Several studies have detected some important gene mutations in cancer progression, but they can't capture the heterogeneity of genome aberrations. Many studies found that there is little overlap between the gene mutations of two samples even if they come from the same patient (Ding *et al*., 2008; Jones *et al*., 2008).

It is well-known that different gene mutations may target the same pathway (Hahn *et al*., 2002; Vogelstein *et al*., 2004). Therefore, it is necessary to shift the point of view from gene level to pathway level, which is helpful to capture the heterogeneous patterns in cancer. There have been several studies to discover the mutation patterns in pathway level (Boca, 2010; Effoni, 2011). Most of them are based on known information about pathways and try to find out which ones are significantly perturbed. However, this kind of methods has one obvious limitation: they only consider those known pathways. Taking into account the incompleteness of knowledge about pathways, it is indispensable to develop new algorithms to discover mutated driver pathways or gene sets without relying on prior knowledge.

Given the huge number of genes in the whole genome, it seems to be an unsolvable problem to enumerate and test all the candidates due to the enormous number of possible gene combinations. Several recent studies about combinatorial patterns of mutations in cancer shed light on how to solve this problem

*The first two authors contributed equally to the paper

†To whom correspondence should be addressed

(Vogelstein *et al*., 2004; Yeang *et al*., 2008). They found that driver pathways often cover a large number of samples. More importantly, usually a single mutation is enough to perturb one pathway. In other words, the mutation of genes in one pathway usually exhibit mutual exclusivity. There have been several studies using these rules to expand the existing pathway-based methods and identify completely new gene sets (Ciriello *et al*., 2012; Miller *et al*., 2011; David *et al*., 2011). Ciriello *et al*. (2012) proposed the method MEMo to detect modules that obey the mutual exclusivity rule within a gene functional network constructed based on prior knowledge. In contrast, Miller *et al*. (2011) proposed a method that identifies functional modules without any information other than patterns of recurrent and mutually exclusive aberrations. More recently, Vandin *et al*. (2012) introduced a novel scoring function by combining two measures (i.e., "coverage" and "exclusivity") to identify the mutated driver pathway using the mutation data. The maximization of this scoring function is defined as the Maximum Weight Submatrix problem which was solved by a stochastic search method. However, there is still no exact algorithm for solving this problem. Moreover, considering the noise of the mutation data, it's interesting to incorporate more information into this framework to improve this model.

In this study, we propose two methods to solve the so-called Maximum Weight Submatrix problem (Vandin *et al*., 2012) which is designed to *de novo* identify mutated driver pathways from mutation data in cancer. The first one is based on a Binary Linear Programming model which is an exact method. This method can be employed for assessing other approximate or/and heuristic algorithms. The second one is based on the Genetic Algorithm which is a stochastic and flexible method. It can be employed to optimize other scoring functions or incorporate other types of information to improve the first method. We have integrated the gene expression data to achieve this. To test the efficiency of our methods, we first apply them onto simulated data and compare them with another method. We further apply our methods onto five biological datasets. The results show that our integrative model can identify more biologically relevant gene sets than the one without expression data.

2 MATERIALS AND METHODS

2.1 A brief introduction

Two important characteristics on the expected patterns of somatic mutations have been employed to understand the somatic mutational process of cancer in recent years. Particularly, Vandin *et al*. (2012) introduced a measure to find mutated driver pathways with two criteria (Figure 1). The first one is "high coverage" which means many patients have at least one mutation in this pathway; the second one is "high exclusivity" which means that most patients have no more than one mutation in this pathway. Given the mutation data represented by a binary mutation matrix A with m rows (samples) and n columns (genes), the original Maximum Weight Submatrix problem is defined as finding a submatrix M of size $m \times k$ in the mutation matrix A by maximizing the scoring function: $W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$ where $\Gamma(g) = \{i : A_{ig} = 1\}$ represents the set of patients in which gene g is mutated and $\Gamma(M) = \bigcup_{g \in M} \Gamma(g)$, $|\Gamma(M)|$ measures the coverage of M and $\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$ measures the coverage overlap of M. Although the Markov chain Monte Carlo (MCMC) method proposed by Vandin *et al*. (2012) is a potential powerful procedure to solve this problem, it is a stochastic search technique which may be trapped in a local solution.

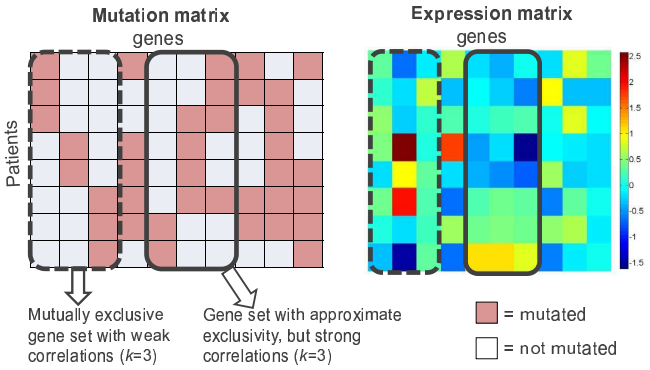

Fig. 1. Illustration of the mutated driver pathway (gene sets) identification problem and how expression profiles improve the identification of mutation patterns with more significant biological relevance. Somatic mutations and expression values in multiple patients are represented in a mutation matrix and expression matrix respectively. The genes in the mutually exclusive gene set (marked by gray dashed box) have very weak expression correlations between each other, while the expression profiles of genes in the second gene set (marked by gray real line box) with approximate exclusivity are strongly correlated with each other.

2.2 Binary Linear Programming (BLP): an exact method

To well understand this problem and assess the stochastic method, we introduce a binary linear programming (BLP) model which can exactly solve this problem using a branch-and-bound algorithm or others. Specifically, we can formulate the original Maximum Weight Submatrix problem into the following optimization problem:

$$\max \quad F(x,y) = 2\sum_{i=1}^{m} y_i - \sum_{j=1}^{n}(x_j \cdot \sum_{i=1}^{m} a_{ij})$$

$$\text{s.t.} \quad \begin{cases} \sum_{j=1}^{n} a_{ij}x_j \geq y_i, & i = 1,\ldots,m \\ \sum_{j=1}^{n} x_j = k, & \\ y_i, x_j \in \{0,1\}, & i = 1,\ldots,m; j = 1,\ldots,n. \end{cases}$$

where x_j is the indicator whether column j of A falls into the submatrix M, so all the columns js with $x_j = 1$ constitute M; y_i is the indicator whether the entries of row i of M are not all zeros. Accordingly, $\sum_{i=1}^{m} y_i$ represents the coverage and $\sum_{j=1}^{n}(x_j \cdot \sum_{i=1}^{m} a_{ij}) - \sum_{i=1}^{m} y_i$ represents the coverage overlap.

Although the problem is NP-hard (Vandin *et al*., 2012), we find that in real application this model can always be solved by a branch-and-bound algorithm efficiently. We use IBM ILOG CPLEX Optimizer to test the effectiveness of this model on simulation data. The experiments are run on a 2.83GHz Core 2 Quad CPU PC. When the gene number of simulation data is smaller than 10000 and the sample number smaller than 500, CPLEX can always get the exact solution in less than 1 second.

2.3 Genetic Algorithm (GA): a stochastic method

As Vandin *et al.* (2012) discussed other criteria can be designed to achieve the similar goal. The BLP model may not be applicable to other new scoring functions. To explore the Maximum Weight Submatrix problem in a more general manner, here we further design a genetic algorithm (GA) (Goldberg *et al*., 1989). The Genetic Algorithm (GA) is an very flexible and powerful technique which can be employed to optimize broad ranges of scoring functions. Moreover, it can be easily extended for integrating other types

of data like gene expression as we will show in the next subsection. The GA method has a natural connection with the current problem in terms of "gene" and "mutation". It simulates the genetic variation in a population and its evolution obeys a random selection mechanism. Moreover, it doesn't need to enumerate all the feasible solutions.

2.3.1 The hypothesis space We use a binary string of 0s and 1s to represent an individual (a feasible solution here) of a population. The length of every string in the hypothesis space is the gene number n. After labeling every gene by $1, 2, \ldots, n$, the value 0 or 1 in the i-th position of an individual characterizes the membership of the i-th gene in the submatrix M. Thus, all of the binary strings with length n and sum k constitute the hypothesis space: $H = \{(x_1, x_2, \ldots, x_n)|x_i \in \{0, 1\}, i = 1, 2, \ldots, n, \sum_i x_i = k\}$.

2.3.2 The fitness function We need to define the fitness function over the hypothesis space H to measure the quality of the candidate solution. In this application, we adopt the rank fitness function. In other words, the fitness of each individual h_i (its corresponding submatrix is M_i) of the population is defined as the rank r_i of the score $W(M_i)$ in the ascending order.

2.3.3 Genetic operators The selection operator, crossover operator and the mutation operator are always problem dependent. In this study, given the rank r_i of an individual h_i based on its score, the selection probability is defined as follows:

$$p_i = \frac{2r_i}{P(P+1)}, \tag{1}$$

where P is the population size. The individual with the highest fitness can be transferred into the next generation with the highest probability.

To ensure every offspring is feasible and reduce the number of iteration, we adopt a crossover operator to inherit properties of its parents. Specifically, the offspring inherits those variables which are common to both parents directly and makes a random selection of variables in the symmetric difference of its parents' genetic makeup.

In the mutation stage, we randomly change one variable value with 1 to 0, and another variable value with 0 to 1. This mutation operator also ensures the feasibility of every offspring. To avoid premature convergence and improve the accuracy of GA algorithm, we employ a local search strategy to improved the search performance. Specifically, we randomly change the value of two variables just as the mutation operator. If such a replacement can improve the current solution, we accept it; when all variables have been tested like this, then we terminate the search.

2.3.4 GA procedure The details of our implementation of GA are described as follows:

Step 0: Given proper parameter settings, i.e., submatrix size k, population size P, mutation rate p_m ($P = n$ and $p_m = 0.1$ were used in this study). Randomly generate the initial population.

Step 1: In every iteration, P couples are selected from current population based on the selection probability p_i and each couple generates an offspring.

Step 2: Each offspring may optionally undergo a mutation with probability p_m.

Step 3: All the parents and offsprings are ranked according to their scoring value, and the best P individuals will make up the next generation, which is used as the current population in the subsequent iteration.

Step 4: Check whether the iteration is trapped in one local solution (e.g., the maximal scoring value does not improve in two consecutive iterations). If so, run local search.

Step 5: Continue in this way until the termination criterion is satisfied (e.g., the current maximal scoring value does not improve in ten consecutive iterations). If so, then terminate the calculation.

2.4 Integrating Mutation and gene Expression data (IME): an integrative model

In real applications, there may be multiple optimal solutions. Moreover, due to the noise in the data or other factors, the most "optimal" ones (the ones with maximal $W(M)$) may not be the best one in biological context. To extract the most biologically significant ones, we try to integrate other types of data to improve this situation. Specifically, we generalize the above model by integrating the gene expression data to improve its performance.

Before describing the simple integrative model here, we note that data integration of many complementary layers is a powerful tool to reduce noise and extract useful information from complexity (Ideker *et al*., 2011). A recent wave of new bioinformatics methods has demonstrated its power (Akavia *et al*., 2010; Zhang *et al*., 2012). For example, Akavia *et al*. (2010) developed a computational framework CONEXIC which integrates chromosomal copy number and gene expression data for detecting aberrations that promote cancer progression. Benefiting from the incorporation of gene expression data, CONEXIC can distinguish the genes within large amplified or deleted regions of a chromosome, while the method based on only the aberration data can not accomplish this.

Our new model is based on such an observation that the genes in the same pathway usually collaborate with each other to execute one function. Therefore, the expression profiles of gene pairs in the same pathway usually have higher correlations than that in different pathways (Qiu *et al*., 2010). Therefore, we can use this characteristic to discriminate the gene sets with the same score. Besides, to filter out the noise in the data and detect more meaningful gene sets, we try to identify such gene sets, whose scores $W(M)$ are close to the optimal solution, but whose member genes have higher correlations with each other (Figure 1).

By combining the gene expression data with the above problem, we define the following new measure:

$$F_{ME} = W(M) + \lambda * R(E_M) \tag{2}$$

where $R(E_M) = \sum_{j_1 \neq j_2} \frac{|pcc(x_{j_1}, x_{j_2})|}{\frac{k(k-1)}{2}}$, E_M is the gene expression submatrix which correspond to the same gene set with the mutation submatrix M, $pcc(\cdot)$ is the Pearson correlation coefficient, and x_{j_1} and x_{j_2} are the expression profiles of gene j_1 and j_2 in E_M respectively. The additional term $R(E_M)$ incorporates information on functional homogeneity to enhance the biological relevance of the identified patterns. Taking into account that $R(E_M)$ is between 0 and 1 and $W(M)$ is integer, setting $\lambda = 1$ we can use F_{ME} to discriminate the gene sets with the same $W(M)$; setting $\lambda \geq 1$ we can detect the gene set with strong correlation and approximate exclusivity. In this study, we report the results with $\lambda = 1$ and $\lambda = 10$. Although the maximization of F_{ME} can be formulated into a binary quadratic programming problem, it is limited by the computational complexity. Here we adopt the GA framework to solve it similarly.

2.5 Biological data

We collected five data sets to assess our methods (Table 1). For the first three data, we only use the mutation data to test BLP and GA. While for the later two data, we use mutation data and expression data together. The first two data sets (LC and GBM1) are obtained from (Vandin *et al*., 2012) directly. We also downloaded another data set about Head and neck squamous cell carcinoma (HNSCC) (Stransky *et al*., 2011).

We obtained Glioblastoma Multiforme (GBM2) and Serous Ovarian Cancer (OC) data from TCGA website (http://tcga-data.nci.nih.gov/tcga/). The data comprise of somatic mutations, copy number aberration and gene expression. Here we only use the data of level 3. After processing the data, we get two types of matrices: mutation matrix A and expression matrix E. A is a binary matrix of size $m \times n$, where m indicates the number of samples, and n indicates the number of genes. Each entry a_{ij} refers to the status of gene j in sample i: $a_{ij} = 1$ if one of the following two conditions holds: (1) The mutation of gene j in sample i is labeled valid somatic (David *et al*., 2011); (2) Gene j is in the statistically significant aberration regions of sample i which is determined by

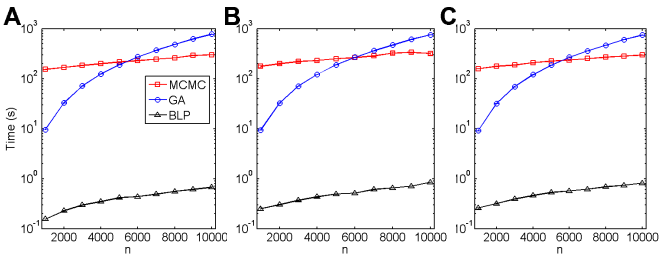

Fig. 2. Comparison of computational time of BLP, GA, and MCMC in terms of gene number n from 1000 to 10000 with different number of patterns: (A) $I = 1$, (B) $I = 5$, and (C) $I = 10$. The y-axis in each plot shows the computational time in seconds. All the markers correspond to the results of an average over 20 realizations.

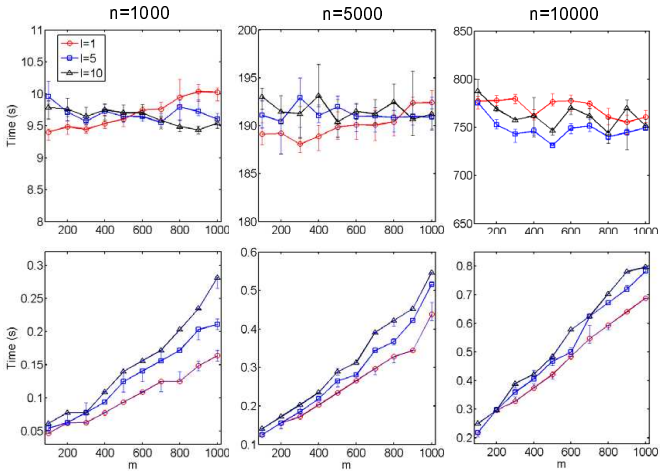

Fig. 3. The time scaling of (A) GA and (B) BLP with sample number m under different gene number $n = 1000$, 5000, 10000 respectively. The curves in each subplot correspond to running time under parameter $I = 1$, 5, 10. All the markers correspond to the results of an average over 20 realizations.

GISTIC (Beroukhim *et al*., 2007). E is a real matrix and each entry e_{ij} is the relative expression of gene j in sample i which is obtained using the method in (Roel *et al*., 2010).

Table 1. Summary of the datasets used in this study and basic information about these datasets. Ct: Cancer type. Spl: number of samples. Ge: number of genes. Am: average number of mutations per sample. Af: average of mutation frequency for all genes. Mf: maximum of mutation frequency of all genes. GE: with or without gene expression data. Ref: references. LC: lung carcinoma. GBM1: glioblastoma multiforme data 1. HNSCC: head and neck squamous cell carcinoma. GBM2: glioblastoma multiforme data 2. OC: ovarian carcinoma.

Ct	Spl	Ge	Am	Af	Mf	GE	Ref
LC	163	356	6.0	2.75	64	no	Vandin *et al*. (2012)
GBM1	84	178	9.6	4.50	43	no	Vandin *et al*. (2012)
HNSCC	74	4920	21.8	1.42	46	no	Stransky *et al*. (2011)
GBM2	90	1126	94.5	1.74	48	yes	TCGA (2008)
OC	313	5385	49.0	2.85	251	yes	TCGA (2011)

3 RESULTS

We first applied our BLP, GA methods onto simulated data to test their performance and compare them with the MCMC method to show their efficiency and characteristics.

3.1 Simulation study

We simulated mutation data starting with gene sets $M_1, M_2, \ldots, M_I$ ($I \geq 1$). Every set has k genes ($k = 10$ has been used in this study). For each patient, we mutate a gene (chosen uniformly at random) in $M_i (i = 1, 2, \ldots, I)$ with probability p_i ($p_i = 1 - i \cdot \Delta$, $\Delta = 0.05$ was used in this study), and if a gene in M_i is mutated, then with probability p_0 we mutate other genes in M_i ($p_0 = 0.04$ was used in this study). Note that p_i and p_0 control the coverage and exclusivity of M_i respectively. The genes not in M_i are mutated at most in three samples. The parameter I controls the complexity of the data structure. When we increase I, the simulation data and the problem get more complicated.

We have compared the time complexity of BLP, GA and MCMC on resolving the original Maximum Weight Submatrix problem through simulation data (Figure 2). We show how the time change with respect to the number of genes under different model complexity. The sample number of simulation data is fixed as 500 which is larger than all our applications. The result of MCMC is obtained based on default parameters. Surprisingly, we can see that our BLP method can get the optimal solution in much shorter time than that of the MCMC. For example, for $n = 10000$ and $I = 5$, the BLP can run in less than 1 second, while MCMC needs more than 300 seconds. We can also observe that the GA is even faster than MCMC with n less than about 5000 which is larger than our all real applications in the following part. In summary, our GA method have competitive efficiency with MCMC, and our exact method BLP can work in a significantly shorter time than others that enables it be more applicable to real data.

From the formulation of BLP model, we can intuitively deduce that when the sample number increases, the constraints number also increases, so the BLP model become more and more complicated and the time required to resolve may rise. We use simulation data to test this conjecture. In Figure 3, we study how the computation time of GA and BLP scales with sample number of simulation data. As the Figure 3 shows, the time of BLP linearly rises with the sample number. However, GA almost remains the same as sample number increases.

Since the BLP model can be solved exactly, so it can always get the optimal solution. As to the GA and MCMC, we found that they both show excellent performance in the data with 'low' complexity (e.g., $I = 1, 2$) (Table 2). However, when we increase the complexity of the data, the GA shows better performance than that of MCMC (Table 2). Note that increasing the number of iterations from 10^6 to 10^7, does not improve the accuracy of MCMC (Supplementary Table S1). This observation clearly shows that our methods are more applicable in real applications.

We also test the effectiveness of GA on detecting the gene set by maximizing the integrative measure F_{ME}. The simulation data is produced according to the following procedures: (1) With predefined m, n and I, produce mutation matrix A using the

Table 2. Accuracy of GA and MCMC with different number of genes n and different number I of embedded patterns.

	$n = 1000$		$n = 5000$		$n = 10000$	
	GA	MCMC	GA	MCMC	GA	MCMC
$I = 1$	100%	100%	100%	100%	100%	100%
$I = 2$	100%	100%	100%	100%	99%	100%
$I = 10$	99%	43%	98%	50%	95%	44%

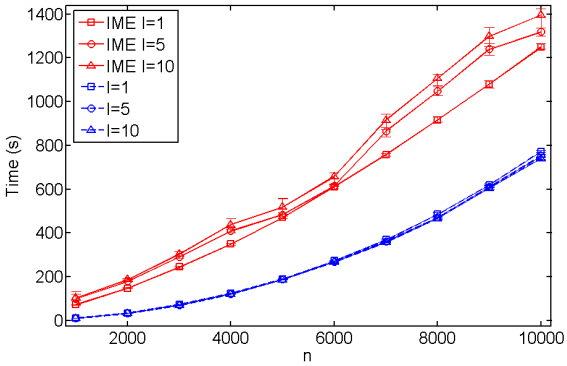

Fig. 4. The time complexity of GA for the integrative model and the original model respectively. The plot shows the scaling of the computer time (in seconds) with respect to the gene number n. The curves correspond to running time under different $I = 1, 5, 10$ and the same sample number $m = 500$.

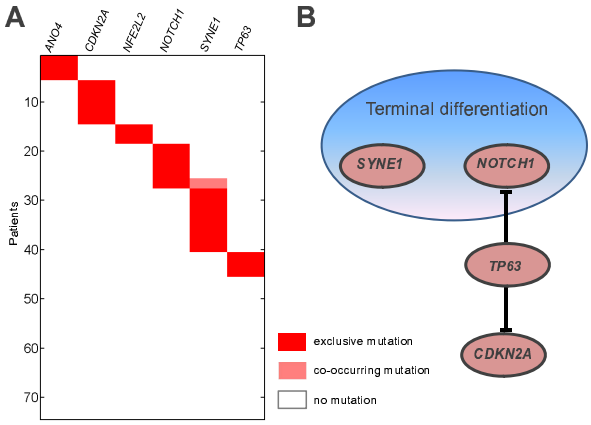

Fig. 5. (A) The submatrix of the "optimal" gene set in the HNSCC data. The legend shows the mutation characteristic between a patient and a gene: (red) exclusive mutation; (soft red) co-occurring mutation; (white) no mutation. It is similar for the other two figures. (B) The known pathway that the identified genes are involved in is terminal differentiation which was reported to be related with HNSCC (Stransky *et al*., 2011). The pathway interactions have been reported in (Stransky *et al*., 2011).

previous methods; (2) Simulate one initial expression matrix E_0, in which each M_I have the exact same expression profile for each gene of it. But we add different noise level according to the Gaussian noise $N(0, (I - i) \cdot \sigma)$ ($\sigma = 0.1$ was used in this study) for the

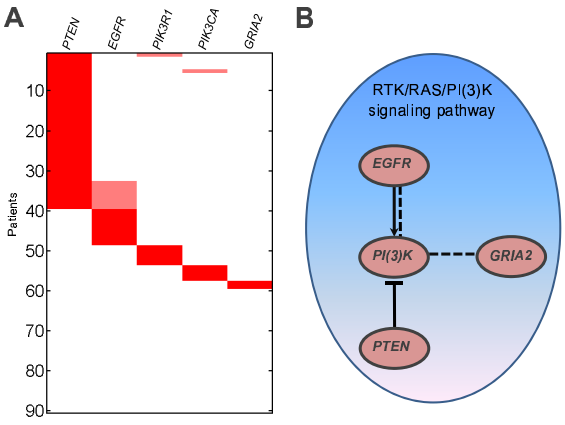

Fig. 6. (A) The submatrix of "optimal" gene set in Glioblastoma data. (B) Four genes in the identified gene patten are involved in the RTK/RAS/PI(3)K signaling pathway which was reported to be related with GBM in TCGA (2008). The pathway interactions marked by real lines are reported in KEGG database, and the dash lines link two genes with significant correlation based on their gene expression profiles (p-value < 0.001). These line types represent the same meaning in next figure.

expression profiles of each M_i to ensure $R(M_1) < R(M_2) < \cdots < R(M_I)$.

Based on these simulation data, we study the accuracy and complexity of GA on maximizing the integrative measure. Similar with the GA for the original model, the integrative model IME optimized by the GA algorithm can get the optimal solution for most cases. Such "optimal" solution may not have the highest mutation score $W(M)$ as we designed for the simulation data. Figure 4 shows how the computation time of GA evolves with gene number n for the integrative model and the original model respectively. The curves clearly show that the GA algorithm for the integrative model costs only about severalfold time of that for the original model. We can also see that the complexity of data doesn't show dramatic impact to the computational time.

In summary, our BLP method is more efficient compared with the MCMC method for solving the original Maximum Weight Submatrix problem. Our GA method have competitive performance compared with the MCMC method and it can also be easily applied for a generalized integrative model to incorporate the gene expression profiles.

3.2 Biological applications

We have applied our methods BLP and GA onto five data sets (Table 1) and compared their performance with MCMC. Note that, when we solve the maximum weight submatrix problem, we represent the genes that are mutated in the same patients as one "metagene" for further analysis. We adopt the permutation test as used by Vandin *et al*. (2012) to assess the significance of the identified gene patterns. We report all the identified patterns with $k = 2$ to 10 (see Supplementary Materials). We also check the second optimal patterns by removing the "optimal" gene set obtained in the whole data set. The information of data resources including the number of samples, the number of genes, the average number of mutations per sample and the maximum number of mutation frequency for each data set have been summarized in Table 1.

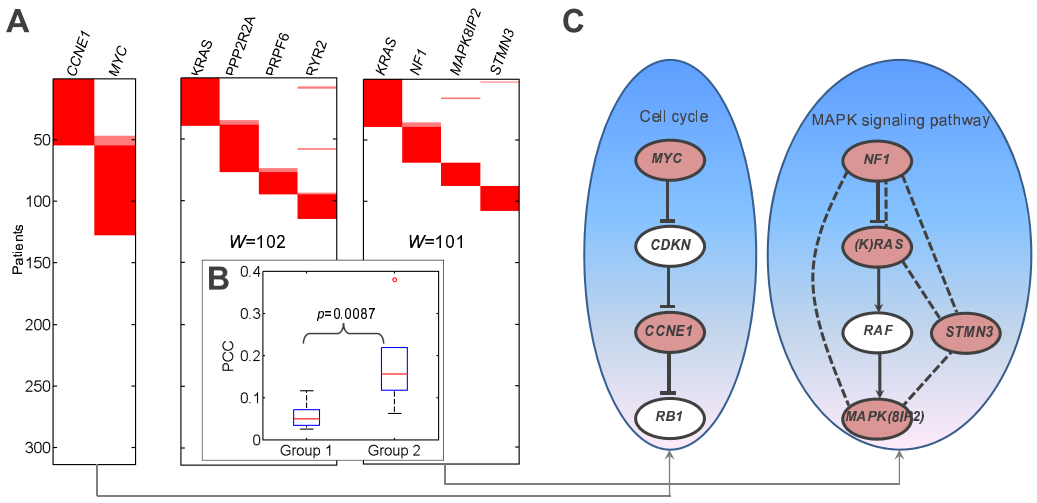

Fig. 7. (A) The submatrix of "optimal" gene sets (*CCNE1,MYC*) in Ovarian carcinoma data with $k = 2$ (the left plot). The right two sub-matrices are the "optimal" gene sets identified by GA and IME respectively by removing the sub-matrix of (*CCNE1,MYC*). Their respective scoring values are $W = 102$ and $W = 101$. But the genes in the later one have significant stronger correlations as illustrated in (B) (p-value = 0.0087, Wilcoxon rank-sum test). (C) The gene sets (*CCNE1, MYC*) and (*KRAS, NF1, MAPK8IP2, STMN3*) are involved in cell cycle and MAPK signaling pathways respectively.

We first apply our BLP and GA methods to the data used by Vandin *et al*. (2012) to assess its performance compared with the MCMC method. The BLP can obtain the exact results in less than 1s, while the GA and MCMC can get them in more than 60s and 5s respectively. This analysis firstly show that the BLP can run in a more efficiently manner than MCMC and GA, and while our GA also has acceptable performance. We found all these three methods got the exact same results. For example, all the three methods can lead to the same gene set (*EGFR, KRAS, STK11*) in Lung adenocarcinoma data set with $k = 3$.

In the following, we further apply our BLP and IME methods onto three datasets that were not used by Vandin *et al*. (2012) and discuss more to show the effectiveness of them. Note that the BLP can efficiently get the exact results like before in less than one second on all these three data sets.

3.2.1 HNSCC It is well-known that HNSCC is a common, morbid, and frequently lethal malignancy (Stransky *et al*., 2011). To uncover its mutational spectrum, Stransky *et al*. (2011) analyzed whole-exome sequencing data from 74 tumor-normal pairs and revealed many genes which have not been implicated in this malignancy in previous studies (Stransky *et al*., 2011). The results imply that the dysregulation of squamous differentiation may play a key driving role in HNSCC carcinogenesis. This mutation dataset covers 74 samples and 4920 genes; on average, there are 130 coding mutations per sample. This mutation matrix is very sparse and only two genes (*TP53* and *TTN*) are mutated in more than 20 samples. They are mutated in 46 and 23 samples respectively.

Considering the prevalence of *TP53* and *TTN* mutation, to identify other pathways not associated with *TP53* and *TTN*, we remove the genes *TP53* and *TTN* from the mutation matrix and ran these three methods on the remaining genes. When $k = 6$, we get unique optimal gene set (*ANO4, CDKN2A, NFE2L2, NOTCH1, SYNE1, TP63*) which is altered in 60.8% (45/74) of the samples with p-value < 0.01. When $k < 6$, the optimal solutions are all subset of these six genes. When $k > 6$, we will identify multiple optimal solutions (see Supplementary Materials).

The analysis result of (Stransky *et al*., 2011) indicates that the mutations of *CDKN2A, NOTCH1, TP63* and *SYNE1* all function in the terminal differentiation in squamous epithelia directly or indirectly (Figure 5). We note that the set of these four genes is one suboptimal solution with $k = 4$.

3.2.2 Glioblastoma The glioblastoma dataset obtained from TCGA (2008) contains DNA copy number alteration and gene expression profiles in 206 glioblastomas samples, and nucleotide sequence aberrations in 91 of the 206 samples. After processing these three types of data (see Materials and Methods), we built a mutation matrix and an expression matrix which cover 90 samples and 1126 genes.

We firstly detect the mutation pattern only depending on the mutation matrix. When $k = 2$, we get two optimal gene sets: the first (*CDKN2A, TP53*) is the part of *p53* signaling pathway; the other is *CDKN2B* and one metagene comprising *CDK4* and *TSPAN31*. After analysis of expression data, we found that the correlation between *CDK4* and *CDKN2B* is stronger than that between *TSPAN31* and *CDKN2B*. So, we have reason to believe that *CDK4* is the one needing more attention. In fact, the pair (*CDK4, CDKN2B*) is the part of *RB* signaling pathway while there is no direct evidence supporting the relation between *TSPAN31* and *CDKN2B*. This example shows one potential advantage of combining expression data with the original model: It can discriminate the genes within identical mutation profile and detect the one with the most relevant functional relationship. We note that, when $k = 3$, the optimal solution is the gene pair (*CDK4, CDKN2B*) together with *RB1*. These two pathways have also been reported by Vandin *et al*. (2012) using another data set (GBM1).

We removed the above five genes and apply the methods to detect the additional gene set. On the remaining genes, when $k = 5$, we identify the gene set (*PTEN, EGFR, PIK3R1, PIK3CA, GRIA2*)

which is mutated in 59 samples ($p < 0.01$). Other than *GRIA2*, the rest four genes are all the part of *RTK/RAS/PI(3)K* signaling pathway which is significantly altered in glioblastoma (Figure 6). Moreover, previous studies have shown that *GRIA2* play important roles in glioma cells (Beretta *et al*., 2009; Maas *et al*., 2001).

3.2.3 Ovarian carcinoma The ovarian carcinoma dataset is obtained from a recent study (TCGA, 2011) which has analysed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number alteration in 489 high-grade serous ovarian adenocarcinomas and the DNA sequences of exons from coding genes in 316 of these tumors. After processing the data, we get a mutation matrix and an expression matrix which cover 313 samples and 6108 genes.

The mutation distribution among genes is very non-uniform. *TP53* is mutated in the majority (251/313) of samples and all the other genes are mutated in less than 25% of samples. In addition, analysis of gene *TTN* indicates that the mutations of *TTN* are likely artifacts (TCGA, 2011). Therefore, we remove *TP53* and *TTN* and ran the methods onto the remaining genes. When $k = 2$, we identify gene pair (*CCNE1, MYC*) which is approximately exclusively mutated in 135 samples ($p < 0.01$) (Figure 7A). *CCNE1* and *MYC* are two important genes engaged in cell cycle progression (Figure 7C). When $k = 3$, the optimal gene set includes this pair and *NINJ2*.

We remove the above genes and apply the methods onto the remaining genes. When $k \leq 3$, none of the optimal solutions are significant ($p < 0.01$). When $k = 4$, the optimal solution is (*KRAS, PPP2R2A, PRPF6, RYR2*) by the BLP model ($p < 0.01$), while using our integrative model ($\lambda = 10$) we get gene set (*KRAS, MAPK8IP2, NF1, STMN3*), which is one of the suboptimal solutions of the original model and is detected as the optimal solution by our integrative model due to the stronger correlations among genes (Figure 7B). *KRAS*, *NF1* and *MAPK8IP2* are all part of *MAPK* signaling pathway (Figure 7C) which regulates cell proliferation and differentiation. The abnormal expression of *STMN3* is associated with malignant progression of multiple cancer types (Fang *et al*., 2009; Singer *et al*., 2009). However, there is no much evidence to support that the genes detected by the original method based only on mutation data show distinct functional relationship. This example shows another advantage of our integrative model. It can detect the gene set which has sub-optimal score but more relevant function relationship. Notably, there may be multiple sub-optimal solutions with the same score. Thus, it is necessary to integrate gene expression data to distinct and identify the underlying key patterns.

4 DISCUSSION AND CONCLUSION

Discovering mutated driver patterns in cancer is an important problem in computational biology. In this paper, we have studied the *de novo* discovery of mutated pathways problem in cancer which has recently been explored by Vandin *et al*. (2012). We first proposed a binary linear programming (BLP) model to exactly solve the so-called Maximum Weight Submatrix problem. The exact method is necessary to evaluate the performance of other approximate or stochastic algorithms. We further suggested a stochastic algorithm – genetic algorithm (GA) which have natural connection with the literal description like "gene" and "mutation". The BLP and GA both show promising performance compared with the original MCMC method. Our study demonstrates that the MCMC encounters serious problems in complex situations with multiple high weight sets of genes for extracting the 'optimal' one.

We should note that this study focus on the mutation at the gene level instead of point mutation (single nucleotide) level. But many known driver genes (e.g., *P53*) have hundreds of point mutations, among which some of drivers and some are passengers. With the development of biological technologies, more and more point mutation data can be available. Our method could be expanded to a more detailed mutation data for finding more sophisticated driver genes and pathways.

We also considered to incorporate the gene expression data into the above model to improve its performance. The new integrative model can be helpful in two aspects. Firstly, the integrative model can be employed to distinguish the genes which have identical mutation profiles. For example, in the Glioblastoma data (GBM2), *CDK4* and *TSPAN31* are in the same copy number abberation region and have the same mutation profiles. Our integrative model can well identify the *CDK4* which has been reported to be related with GBM while *TSPAN31* not. Secondly, some significant biologically relevant gene set with strong correlations among their genes may be not the "optimal" one with the scoring function W. Our integrative model can identify such gene set well with the "optimal" integrative score. For example, in the Ovarian carcinoma data, our integrative model identified one different gene set with the original model. Our analysis demonstrates that these genes have significant biological function connections between each other than that of the original model.

ACKNOWLEDGEMENT

This project was supported by the National Natural Science Foundation of China, No. 11001256 and 11131009, Innovation Project of Chinese Academy of Sciences (CAS), kjcx-yw-s7, the 'Special Presidential Prize' - Scientific Research Foundation of the CAS, the Special Foundation of President of AMSS at CAS for 'Chen Jing-Run' Future Star Program and the Foundation for Members of Youth Innovation Promotion Association, CAS.

REFERENCES

Akavia, U.D., *et al*. (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005-1017.

Beretta, F., *et al*. (2009) The GluR2 subunit inhibits proliferation by inactivating Src-MAPK signalling and induces apoptosis by means of caspase 3/6-dependent activation in glioma cells. *Eur J Neurosci*, **30**, 25-34.

Beroukhim, R., *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA*, **104**, 20007-20012.

Boca, S.M. (2010) Patient-oriented gene set analysis for cancer mutation data. *Genome Biol*, *11*, R112.

Chapman, M.A., *et al*. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature*, **471**, 467-472.

Ciriello, G., *et al*. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398-406.

David, L., *et al*. (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.*, **71**, 4550-4561.

Ding, L., *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069-1075.

Efroni, S. (2011) Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE*, *6*, e14437.

Fang, F., *et al*. (2009) Expression of cyclophilin B is associated with malignant progression and regulation of genes implicated in the pathogenesis of breast cancer. *Am J Pathol*, **174**, 297-308.

Fidler, I.J., *et al*. (2003) The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer*, **3**, 453-458.

Getz, G., *et al*. (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500-1500.

Goldberg, D.E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*. Addison Wesley.

Greenman, C., *et al*. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153-158.

Hahn, W.C., *et al*. (2002) Modelling the molecular circuitry of cancer. *Nat Rev Cancer*, **2**, 331-341.

Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.

Ideker, T., *et al*. (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, **144**, 860-863.

Jones, S., *et al*. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801-1806.

Miller, C.A., *et al*. (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics*, **4**, 34.

Overdevest, J.B., *et al*. (2009) Utilizing the molecular gateway: the path to personalized cancer management. *Clin Chem*, **55**, 684-697.

Qiu, Y.Q., *et al*. (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics*, **11**, 26.

Roel, G.W., *et al*. (2010) Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98-110.

Singer, S., *et al*. (2009) Coordinated expression of stathmin family members by far upstream sequence element-binding protein-1 increases motility in non-small cell lung cancer. *Cancer Res.*, **69**, 2234-2243.

Maas, S., Patt, S., Schrey, M., Rich, A. (2001) Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc Natl Acad Sci USA.*, **98**, 14687-14692.

Stransky, N., *et al*. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science*, **333**, 1157-1160.

Swanton, C., *et al*. (2009) Molecular classification of solid tumours: towards pathway-driven therapeutics. *Br J Cancer*, **100**, 1517-1522.

The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.

The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609-615.

Vandin, F., *et al*. (2011) *De novo* discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375-785.

Vogelstein, B., *et al*. (2004) Cancer genes and the pathways they control. *Nat Med*, **10**, 789-799.

Yeang, C.H., *et al*. (2008) Combinatorial patterns of somatic gene mutations in cancer. *FASEB J*, **22**, 2605-2622.

Zhang, S., *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, doi: 10.1093/nar/gks725.