doi:10.4149/neo_2016_007

Identification of driver pathways in cancer based on combinatorial patterns of somatic gene mutations

H. T. LI[1,4], J. ZHANG[1], J. XIA[2,3,*], C. H. ZHENG[1,3,*]

[1]College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China; [2]Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China; [3]Center of Information Support & Assurance Technology, Anhui University, Hefei, Anhui 230601, China; [4]School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 273165, China

**Correspondence: jfxia@ahu.edu.cn, zhengch99@126.com*

Received March 10, 2015 / Accepted August 21, 2015

With the availability of high-throughput technologies, a huge number of biological data (e.g., somatic mutation, DNA methylation and gene expression) in multiple cancers have been generated. A major challenge is to identify functional and vital driver mutation import for the initiation and progression of cancer. In this paper, we introduce a novel method, named Co-occurring mutated metagene Genetic Algorithm (CoGA), to solve the maximum weight submatrix problem, with the aim of distinguishing mutated driver pathways in cancer. The algorithm relies on the combinatorial properties of mutations in the same pathways: high coverage and mutual exclusivity, and the possible properties of mutations in different pathways: co-occurring pattern. We carried out the experiment with glioblastoma multiform (GBM) data. The experimental results show that compared with the original model, our algorithm has more potential to identify driver pathways in cancer with biological significance.

Key words: driver pathway, co-occurring pattern, genetic algorithm, glioblastoma multiform

Cancer is an extremely complex disease which is driven largely by somatic mutations. Finding the pathogenesis of cancer is still difficult. With the development of high-throughput technologies, several large-scale cancer genomics projects, e.g., the Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium (ICGC), which provide huge amounts of biological data, can help researchers to explore a large variety of biological and biomedical problems at the genome-wide scale. To understand the formation and progression of cancer, one of the challenges is to identify functional mutations vital for cancer development. In other words, distinguishing functional "driver mutations" and filtering out the unfunctional and random "passenger mutations", is a key to understand the molecular mechanisms of cancer initiation and progression [1]. Based on driver mutations, driver genes, and driver pathways, researchers have designed efficient treatments for cancer patients [2, 3, 4].

In the past several years, identifying recurrent mutations and recurrently mutated genes are the most commonly used approaches to uncover driver mutations/genes in a large amount of cancer patients. This method has identified several significant driver genes, e.g., *KRAS* in lung cancer, *PIK3CA* in colorectal cancer, and *ERRB2* in glioblastoma, etc. [5, 6, 7]. However, lots of studies discovered that there is little overlap between gene mutations of two cancer cells even if they come from the same patient [8, 9]. Because of the heterogeneity of genome aberrations, this approach has not revealed all of the driver mutations in individual cancer. In other words, even cancer genomes from the same type of cancer, no two genomes exhibit exactly the same complement of somatic aberrations.

One reason for this heterogeneity is that driver mutations target signaling and regulatory pathways which have multiple points of failure [10, 11]. Hence, it is significant to study mutation in pathway level, rather than in gene level. Recent cancer genome sequencing studies have demonstrated that it is easy to capture the heterogeneous phenomenon in cancer cells in pathway level [12, 13, 14, 15]. For example, Ciriello et al. proposed an approach called MEMo (Mutual Exclusivity Modules) to detect oncogene network modules within a constructed network using gene mutation information and a human reference network (including protein interactions and signal transduction pathways) [12].

However, biological interaction networks are far from complete. Meanwhile, a lot of pathway databases may contain noise data [16]. Taking into account these obvious limitations, it is indispensable to use *de novo* discovery of mutated driver pathways without relying on prior information.

Given a very large number of genes in the whole genome, it seems implausible to enumerate and test all the candidates due to the enormous number of possible gene sets. For instance, there are more than 10^{26} sets of seven genes [12, 16, 17, 18, 19]. In recent years, several studies have examined the patterns of somatic mutations of cancers to solve this problem [12, 17, 18, 19]. In these studies, the researchers found that there are two constraints on the majority of combinatorial patterns of aberrations in cancer [20, 21]. First, a driver mutation is generally very rare. Particularly, researchers found that a single driver mutation is frequently enough to perturb one pathway in most cases. In other words, there is a phenomenon of mutual exclusivity between driver mutations, which is called high exclusive. Second, an important cancer pathway should cover a great majority of patients. Thus, the aberrations should be contained by most patients in the pathway, which is called high coverage. Miller et al. proposed a method called RME that distinguishes functional modules without any prior knowledge other than patterns of recurrent and mutually exclusive mutations [19].

Lately, Vandin et al. proposed an approach, called Dendrix (*de novo* driver exclusivity), to discover driver pathways using the somatic mutation and copy number variations data [16]. They devised a novel scoring function using the above two constraints (i.e., exclusivity and coverage). The maximization of this function is defined as the maximum weight submatrix problem. However, this problem is computationally difficult to solve. Vandin et al. used a Markov Chain Monte Carlo (MCMC) method to solve this problem. After that, Zhao et al. proposed two approaches to address the problem based on genetic algorithm (GA) and binary linear programming, respectively [22].

Mutations of genes in the same pathways are mutually exclusive in the great majority of the combinatorial mutational patterns. However, several observations violate this hypothesis [21]. For example, several gene pairs in the Ras and IGF-AKT pathways show co-occurring mutational phenomena: *KRAS* and *PTEN* (endometrium), *BRAF* and *PTEN* (skin), and *NRAS* and *PTEN* (acute lymphoblastic leukemia) [21]. Hence, mutual exclusivity is a fairly strong assumption [16]. Taking into account this situation, in this paper, we propose a novel algorithm, called Co-occurring mutation metagene Genetic Algorithm (CoGA) method, to solve this problem. The proposed CoGA method is based on GA, and it constructs co-occurring mutated genes as a "metagene". The results on the glioblastoma multiform (GBM) data illustrate that our algorithm is able to detect functional driver mutations in pathway with biological significance.

Materials and methods

Construction of glioblastoma multiform mutation matrix. We obtain the glioblastoma multiform data from [16] directly, which comprise of somatic mutations, and copy number variant (CNV) data, respectively. Here, we use the somatic mutation data of Level 2 and copy number aberration data of Level 3. After preprocessing the data, we get mutation matrix A, which is a binary matrix of size $m \times n$, here m indicates the number of patients and n indicates the number of genes. Each entry a_{ij} refers to the status of gene j in patient i: $a_{ij}=1$ if one of the following two conditions is satisfied: (i), the mutation status of gene j in sample i is labeled valid somatic. Here, extremely low frequency somatic mutated genes were removed (removing genes with mutation frequency < 5%). In addition, some known 'artifacts genes' from GBM datasets reported by M. Lawrence (http://1.usa.gov/RBtuz7) [17], for instance *TTN*, were deleted; and (ii), gene j is located in the statistically significant variable regions of sample i, which is determined by Genomic Identification of Significant Targets in Cancer (GISTIC) [5, 23].

Maximum weight submatrix problem. It is difficult to identify driver pathways. Considering this point, Vandin et al. transformed this problem into maximum weight submatrix problem using two constraints, which are "high coverage" and "high exclusivity" [16]. The first one, "high coverage," means that the majority of samples have at least one mutation in driver pathway. The second one, "high exclusivity", means that lots of samples have no more than one mutation in one driver pathway. They reflect these two properties using a mutation matrix and a scoring function. A binary mutation matrix A is constructed by m rows (samples) and n columns (genes). The maximum weight submatrix problem is defined as selecting a submatrix M of size $m \times k$ in the mutation matrix A by maximizing the scoring function:

$$\mathrm{W(M)} = |\Gamma(\mathrm{M})| - \omega(\mathrm{M}) = 2|\Gamma(\mathrm{M})| - \Sigma_{g \in M}|\Gamma(g)| \quad (1)$$

where $\Gamma(g) = \{\mathrm{i}:\mathrm{A}_{ig} = 1\}$ denotes that gene g in ith row (sample) is mutated, $\Gamma(\mathrm{M}) = \mathrm{U}_{g \in M}) \, \Gamma(g)$ indicates the coverage of M, which represents the set of patients where at least one of the genes in M is aberrations, and $\omega(\mathrm{M}) = \Sigma_{g \in M}|\Gamma(g)|-|\Gamma(\mathrm{M})|$ denotes the coverage overlap weight.

Researchers have proved that selecting the largest set of genes is a NP-hard. Hence, no algorithm can obtain a satisfactory result in every case. In view of the situation, some researchers tried to solve this problem with stochastic search methods, for example, Vandin et al. proposed a MCMC method [16]. After that, Zhao et al. used GA to solve this problem, and obtained satisfactory results [22].

However, in some situations, some genes have co-occurring mutations in one pathway in several cancers [21]. As Vandin have pointed out, high exclusivity is a fairly strong assumption, and genes mutate with co-occurring in some cases, such as CBF translocations and kinase mutations in acute myeloid leukemias, and VHL/SETD2/PBRM1 mutations in renal cancer. So we propose a new algorithm, CoGA method, to solve this problem (Figure 1).

CoGA algorithm. The pipeline of the CoGA algorithm is shown in Figure 2. The details of our implementation for the maximum weight submatrix problem are described as follows.

Step 1 Download somatic mutation and CNV data of a cancer from [16].

Step2 Construct mutation matrix. The data preprocessing is described in Construction of GBM mutation matrix section. After that, we will get a binary mutation matrix A.

Step 3 Construct co-occurrence mutation metagene. As we will describe in details below, we construct an adjacency matrix C, which is called "co-occurring mutation adjacency matrix". The co-occurring mutation adjacency matrix encodes the connection strength between each pair of genes i and j as

$$c_{ij} = \begin{cases} 1 \text{ } \mathit{if}\ p < 0.01\ (\mathit{right - side\ Fisher's\ exact\ test\ between\ gene\ i\ and\ j)\ and} \\ \qquad \mathit{number\ of\ genes\ lar\ ger\ then}\ 40\%\ \mathit{sample} \\ \mathit{0} \qquad\qquad\qquad \mathit{otherwise} \end{cases}$$

The $m \times n$ mutation matrix A = $[a_{ij}]$ is transformed into an $n(n\text{-}1)/2 \times n(n\text{-}1)/2$ co-occurring mutation adjacency matrix C = $[c_{ij}]$, which is a symmetric matrix with binary entries.

If the average of several genes' clustering coefficient is 1, these genes will be constructed as a metagene. The clustering coefficient is

$$CC_1(\mathrm{v}) = \frac{2|E(G_1(v))|}{\deg(v)\ (\deg(v) - 1)} \tag{2}$$

where v represent a gene, $G_1(v)$is the neighborhood of gene v, deg(v) is the degree of gene v.

Step 4 Integrate these genes to create a new column into mutation matrix. The weight of 'close' between gene i and gene j is

$$w_{ij} = \frac{1}{-\log\,[\min\,(P)]} \times [-\log\,(p_{ij})] \tag{3}$$

where p_{ij} is p-value of right-side Fisher's exact test between gene i and gene j, min(P) is the minimum of these p-value between each pair of gene i and gene j. In one sample, if the mutation of metagene number n' is larger than half of metagene number n, we multiply n' by the average impact factor of metagene W as integration value of this patient. The average impact factor is

$$W = \Sigma_{i \neq j} \frac{w_{ij}}{\frac{n(n-1)}{2}} \tag{4}$$

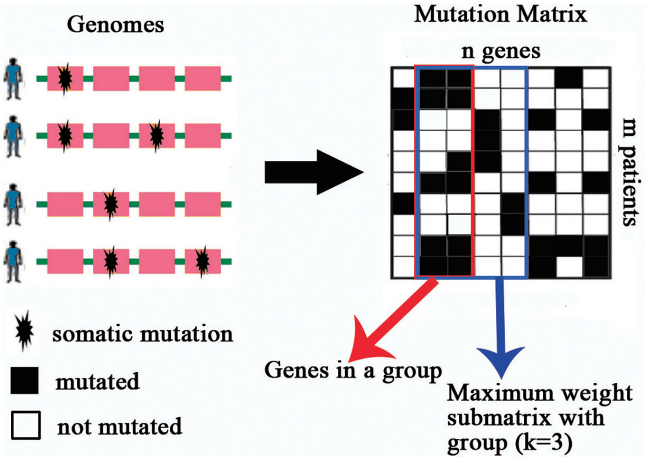

Figure 1. Somatic mutations in samples (patients) are represented in a mutation matrix. The second and third genes show highly co-occurring mutation levels across the 10 samples. According to this, we combine these two genes as a "metagene". Gene sets are identified as exclusive submatrices or high weight submatrices.

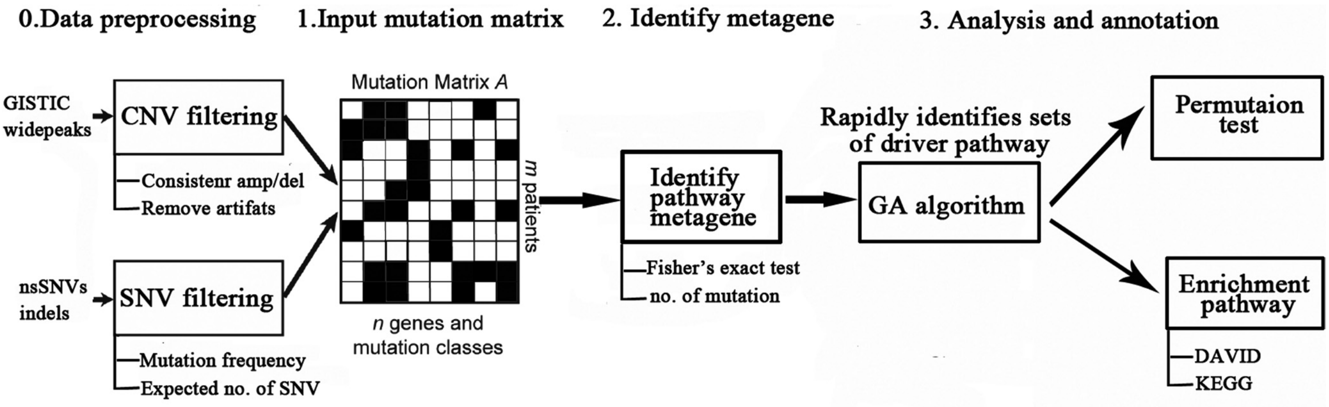

Figure 2. Illustration of CoGA method pipeline. CoGA analyzes integrated mutation data from a variety of sources including somatic mutations and copy number aberrations. We construct co-occurring mutation as a metagene. After processing, co-occurring mutation columns integrate into "metagene column". And then "metagene column" integrate into the matrix. With GA method, we find the maximum weight submatrix which is mutated driver pathway.

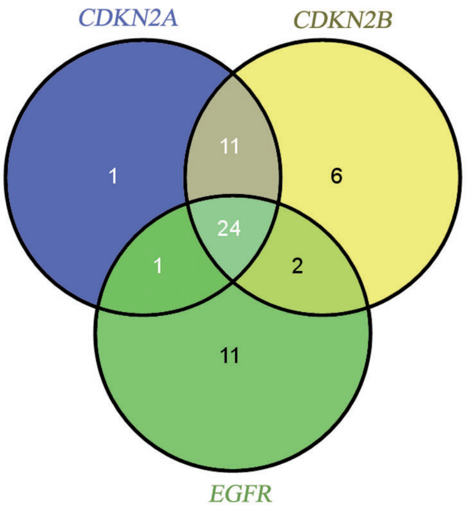

Figure 3. ***CDKN2A, CDKN2B*** **and** ***EGFR*** **form a metagene. This Venn diagram shows the overlapping relationship between genes in the metagene across analyzing the 84 TCGA GBM cases.**

The other value of integration of metagene is 0.

Step 5 Running integrated mutation matrix with the GA method and then driver pathway can be obtained.

We adopt random mutation data using the permutation test described by [16] to assess the significance of the identified gene set. We run a permutation test to assess the significance of the gene set. A permutation test gives a simple way to compute the sampling distribution for any test statistic. The statistic is the weight W(M) of the results and the null distribution was generated through the independent mutation data of arrangement, thus holding the frequency of mutation for each mutation metagene [16]. Rather than a fixed background mutation rate, we use the observed mutation frequency.

Step 6 Functionally annotate the final gene sets. In our study, we used the functional annotation tool of the Database for Annotation, Visualization and Integrated Discovery (DAVID) for Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment [24]. DAVID uses the so-called EASE score, which is a conservative adjustment to the Fisher's exact probability, to assess the significance of the mutated driver genes. A *P*-value, 0.01 and a false discovery rate, 5% were chosen as significant thresholds upon filtering the pathway data.

Results

Brief introduction to CoGA. The pipeline of the CoGA method is illustrated in Figure 2. We download somatic mutation and CNV data of a cancer on the website TCGA, and then construct a mutation matrix as indicated in Materials and Methods Section. After that, we construct co-occurrence mutation metagene using right-side Fisher's exact test between each pair of genes, and integration of these genes creates a new column into mutation matrix. We process integrated mutation matrix with the GA method. Driver pathway can be got. Finally, the biological significance of the final results and interpretation is obtained by utilizing DAVID and KEGG.

We analyzed a collection of 84 glioblastoma multiform samples from TCGA. Compared with the original GA model, our algorithm has more potential to identify driver pathway in cancer with biological significance.

Results on the glioblastoma multiform (GBM) data. We applied the proposed CoGA algorithm on84 GBM patients from TCGA. Somatic mutations in these patients were measured in 601 genes. Meanwhile, CNV data is obtained using GISTIC 2.0 [23] as described in Materials and Methods Section. After the preprocessing, the GBM dataset contained mutation and CNV data for 178 genes in 84 patients.

Firstly, we constructed the metagene, which contains three genes, i.e., *CDKN2A, CDKN2B, EGFR*. Figure 3 shows the

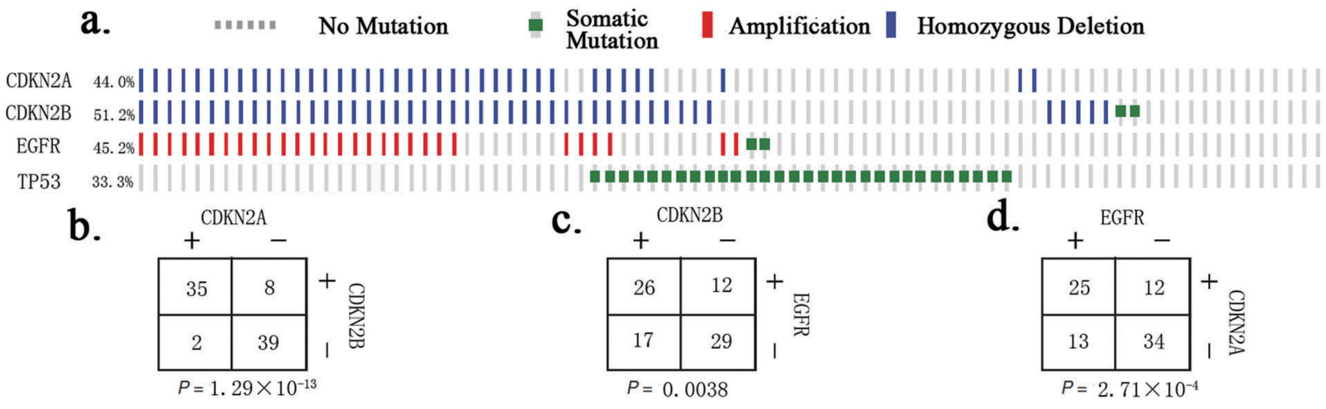

Figure 4. (a) The indentified pathway when k=2, which contains metagene (*CDKN2A, CDKN2B, EGFR*) and *TP53*. Mutually exclusive and co-occurring mutations are shown across analyzing the 84 TCGA GBM samples. (b) Incidence of *CDKN2A* mutations and co-occurring mutations in any *CDKN2B* gene. (c) Incidence of co-occurring of *CDKN2B* and *EGFR* mutations. (d) Incidence of co-occurring of *EGFR* and *CDKN2A* mutations

overlapping relationship between the genes in this metagene. *CDKN2A* and *CDKN2B* mutations are observed to co-occur with *EGFR* mutations.

Secondly, we explored the biological significance of combining the three genes into metagene. As is known, *CDKN2A* and *CDKN2B* are in the same chromosome arm (9p21.3). Because 9p21.3 homozygous deletion always happens in GBM, *CDKN2A* and *CDKN2B* can co-occurring mutate in the same sample. Lots of researchers have found that *EGFR* amplification and *CDKN2A* deletion are frequently simultaneous molecular alterations in GBM [25, 26]. A recent paper showed that the *CDKN2A* (p14/ARF protein) is frequently down regulated in cancer with *EGFR* mutation. Hence, this metagene has biological significance [27].

We used CoGA method to find gene sets of size 2≤k≤3, the results are shown in table 1. We also performed a permutation test, as described in Materials and Methods Section. The obtained *P*-values are less than 0.01. When k=2, the identified pathway is composed of the metagene and *TP53*. Figure 4 shows the mutually exclusive and co-occurring mutations pattern. These genes disrupt the *p53* and *RB* pathways. When k=3, we have several results. Some of them have biological significance. For instance, *CDK4*, *MDM2, RB1*can perturb one pathway for GBM [1].

We then used the tool DAVID to search for KEGG pathway enrichment (Figure 5).It is well known that the *p53* tumor suppressor pathway prevents the propagation of unstable genomes. Inactivation of the *p53* tumor suppressor pathway will cause cancer. The most common event is mutations and deletions of *TP53*, homozygous deletion of *CDKN2A*, and amplifications of *MDM2* [6, 28].Activation of *EGFR* frequently occurs in primary GBM. Among inactivation of the *RB* tumor suppressor pathway, meanwhile, mutations occurs in the form of homozygous deletion of the *CDKN2A/CDKN2B* locus on chromosome 9p21 and amplification of the *CDK4* locus.*RB1*is frequently altered in primary GBM, which obviates the genetic pressure for activation of upstream cyclin/cyclin-dependent kinases [6, 29, 30].

There are also other genes with biologically significant in GBM. The gene *TSPAN31* is considered to be taken part in growth-related cellular processes, because the encoded protein mediates signal transduction events resulting in regulation of cell development, activation and growth. Indeed, Zhang et al. have revealed that *TSPAN31* is associated with tumor genesis [31].*CENTG1* is substantially amplified in GBM cell line *TP366*, *LN-Z308*, and *CRL-2061* on chromosome 12. Amplification of chromosome 12q1 is frequently occurred in brain tumors. This chromosomal region contains the *MDM2*, *CDK2*, and *CENTG1* genes. It has been shown that *CENTG1* gene plays a vital role in GBM [32].

We used the functional annotation tool of DAVID for BIOCARTA pathway enrichment. We found that one of the significant BIOCARTA pathways, First Multivalent Nuclear Factor, is a novel pathway which hasn't been proved a connection with GBM previously. This pathway is a significant dysregulated pathway in another brain cancer, Astrocytomas [33]. Hence, our method can distinguish novel driver pathway (see Supplementary Materials for more details).

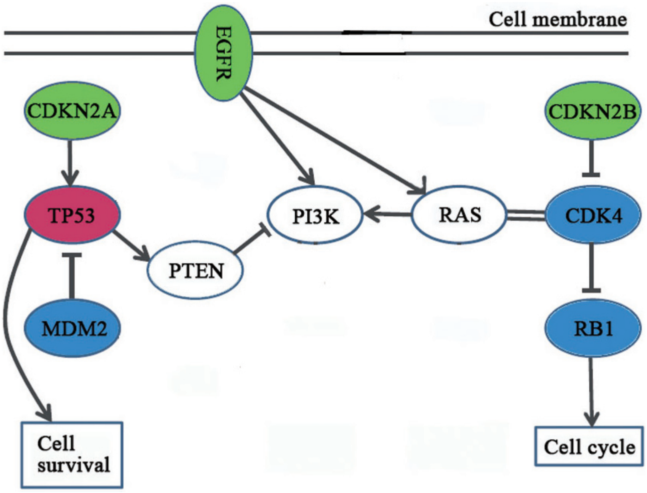

Figure 5. Illustration of *p53/Rb* signaling pathway. The double lines between *KRAS* and *CDK4* represent sometimes there is a synthetic lethal interaction between these two genes. The green nodes denote the co-occurring metagene. The red nodes denote identified mutation when k=2. The blue nodes denote identified mutation when k=3. Regulatory relations are extracted from the KEGG database and related literature.

In order to validate the efficacy of the CoGA methods, we analyzed all genomic data of gliomblastoma patients. Considering these data were generating from BI__IlluminaGA_DNASeq platform in TCGA, we applied our CoGA algorithm to analyze all genomic data of gliomblastoma patients from the same platform, which have 272 glioblastoma patients after filtering 4 samples with high mutation rate. We have applied our methods CoGA onto this GBM dataset. We can also find p53 signaling pathway (see Supplementary Materials for more details).

Discussion

It is significant to identify mutated driver pathway in cancer for clinical targeted therapeutics. In this paper, we propose a novel algorithm for detecting mutated driver patterns *de*

Table 1. Results on the GBM dataset when 2≤k≤3

k	driver pathway genes	maximum weight score of dataset GBMW(M)
k=2	*TP53*,metagene[a]	83
k=3	*CDK4,TP53,* metagene *MARCH9, TP53,* metagene *MDM2, TP53,* metagene *METTL1, TP53,* metagene *TSPAN31, TP53,* metagene *RB1, TSPAN31,* metagene *CENTG1, RB1,* metagene	85

[a]metagene is composed of *CDKN2A*,*CDKN2B*,and*EGFR*

novo using somatic mutation and CNV data from TCGA. This algorithm is based on two previous studies [16, 22]. We introduce a new strategy which constructs co-occurring mutated genes as a metagene in the maximum weight submatrix problem. We apply the proposed CoGA algorithm on 84 GBM patients from TCGA. The results show that our method has a great potential to identify driver pathway in cancer with biological significance.

Mutual exclusivity is a fairly strong assumption, and there are examples of co-occurring, and possibly cooperative mutations. Using the original GA method, the triplet (*CDKN2B, RB1, CYP27B1*) is the most significant one when k=3 [22]. The p53 signaling pathway genes, such as *CDKN2A* and *TP53*, cannot be detected. As shown in figure 3, *CDKN2A* and *CDKN2B* are co-occurring mutation. Because mutual exclusivity is a fairly strong assumption, *CDKN2A* and *CDKN2B* cannot be detected simultaneously with the original model (see Supplementary Materials for more details about difference between our method and the other two previous methods of MCMC and GA).

Recently, Zhang et al. proposed a new method CoMDP for identification of co-occurring mutated driver pathways in cancer [34]. However, it is different method between our method and Zhang's job. Zhang's paper makes the maximization of the weight *W* for each individual pathway. Meanwhile, it ensures the maximization of the inter-overlap between the pathway pair. Our new strategy constructs co-occurring mutated genes as a metagene using Fisher's exact test in the maximum weight submatrix problem.

These problems can be solved with CoGA method to some extent. We set strict conditions to construct 'metagene' for the following two reasons. First, some of the statistical significance of co-occurring mutations is very large. However, the numbers of these mutations are small in the total samples. They look like co-occurring mutation by chance. Hence, we can't construct these genes as "metagene". Second, even though mutual exclusivity is a strong assumption, the combinatorial mutational patterns are mutually exclusive in most cases. Only several observations violate this simple hypothesis. Based on these reasons, we construct 'metagene' with strict conditions. The choice of conditions about constructing co-occurrence mutation metagene is provided in Supplementary Materials.

To demonstrate the role of our algorithm in solving random somatic mutations, the simulation data set was constructed (see Supplementary Materials for more details). The results show that our algorithm can solve random somatic mutations which are commonly found in malignant cells.

However, some cancer data cannot be used to detect 'metagene'. For example, the lung adenocarcinoma dataset were obtained directly from a previous study [16]. We use a right-side Fisher's exact test to detect 'metagene' as described in Materials and Methods Section. The returned 'metagene' is an empty set. Hence if the metagene gene is null, we can use the initial GA method. This example also shows the flexibility of CoGA approach.

In future, we will incorporate other biological data, such as gene expression and DNA methylation data, with the aim of exploring other potential mutated driver pathways.

Supplementary information is available in the online version of the paper.

Acknowledgements: This work was supported by National Natural Science Foundation of China (61272339, 61271098 and 31301101), the Anhui Provincial Natural Science Foundation (1408085QF106), the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011), and the Technology Foundation for Selected Overseas Chinese Scholars from Department of Human Resources and Social Security of Anhui Province (No.[2014]-243).

References

[1] GREENMAN C, STEPHENS P, SMITH R, DALGLIESH GL, HUNTER C et al. Patterns of somatic mutation in human cancer genomes. Nature 2007; 446: 153–158. http://dx.doi.org/10.1038/nature05610

[2] INT CANC GENOME CONSORTIUM. International network of cancer genome projects. Nature 2010; 464: 993–998. http://dx.doi.org/10.1038/nature08987

[3] MARDIS ER, WILSON RK. Cancer genome sequencing: a review. Human Molecular Genetics 2009; 18: 163–168. http://dx.doi.org/10.1093/hmg/ddp396

[4] MEYERSON M, GABRIEL S, GETZ G. Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics 2010; 11: 685–696. http://dx.doi.org/10.1038/nrg2841

[5] BEROUKHIM R, GETZ G, NGHIEMPHU L, BARRETINA J, HSUEH T et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proceedings Of the National Academy Of Sciences Of the United States Of America 2007; 104: 20007–20012. http://dx.doi.org/10.1073/pnas.0710052104

[6] CHIN L, MEYERSON M, ALDAPE K, BIGNER D, MIKKELSEN T et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008; 455: 1061–1068. http://dx.doi.org/10.1038/nature07385

[7] GETZ G, HOFLING H, MESIROV JP, GOLUB TR, MEYERSON M et al. Comment on „The consensus coding sequences of human breast and colorectal cancers". Science 2007; 317: 1500. http://dx.doi.org/10.1126/science.1138764

[8] DING L, GETZ G, WHEELER DA, MARDIS ER, MCLELLAN MD et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008; 455: 1069–1075. http://dx.doi.org/10.1038/nature07423

[9] JONES S, ZHANG XS, PARSONS DW, LIN JCH, LEARY RJ et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008; 321: 1801–1806. http://dx.doi.org/10.1126/science.1164368

[10] BOCA SM, KINZLER KW, VELCULESCU VE, VOGELSTEIN B, PARMIGIANI G. Patient-oriented gene set analysis

for cancer mutation data. Genome Biology 2010; 11: R112. http://dx.doi.org/10.1186/gb-2010-11-11-r112

[11] EFRONI S, BEN-HAMO R, EDMONSON M, GREENBLUM S, SCHAEFER CF et al. Detecting Cancer Gene Networks Characterized by Recurrent Genomic Alterations in a Population. Plos One 2011; 6(1): e14437. http://dx.doi.org/10.1371/journal.pone.0014437

[12] CIRIELLO G, CERAMI E, SANDER C, SCHULTZ N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research 2012; 22: 398–406. http://dx.doi.org/10.1101/gr.125567.111

[13] DU JL, YUAN ZF, MA ZW, SONG JZ, XIE XL et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Molecular Biosystems 2014; 10: 2441–2447. http://dx.doi.org/10.1039/C4MB00287C

[14] JENSEN LJ, KUHN M, STARK M, CHAFFRON S, CREEVEY C et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research 2009; 37: 412–416. http://dx.doi.org/10.1093/nar/gkn760

[15] PRASAD TK, GOEL R, KANDASAMY K, KEERTHI-KUMAR S, KUMAR S ET AL. Human protein reference database—2009 update. Nucleic acids research 2009; 37: 767–772. http://dx.doi.org/10.1093/nar/gkn892

[16] VANDIN F, UPFAL E, RAPHAEL BJ. De novo discovery of mutated driver pathways in cancer. Genome Research 2012; 22: 375–385. http://dx.doi.org/10.1101/gr.120477.111

[17] LEISERSON MD, BLOKH D, SHARAN R, RAPHAEL BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol 2013; 9: e1003054. http://dx.doi.org/10.1371/journal.pcbi.1003054

[18] MASICA DL, KARCHIN R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 2011; 71: 4550–4561. http://dx.doi.org/10.1158/0008-5472.CAN-11-0180

[19] MILLER CA, SETTLE SH, SULMAN EP, ALDAPE KD, MILOSAVLJEVIC A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 2011; 4: 34. http://dx.doi.org/10.1186/1755-8794-4-34

[20] VOGELSTEIN B, KINZLER KW. Cancer genes and the pathways they control. Nature Medicine 2004; 10: 789–799. http://dx.doi.org/10.1038/nm1087

[21] YEANG CH, MCCORMICK F, LEVINE A. Combinatorial patterns of somatic gene mutations in cancer. Faseb Journal 2008; 22: 2605–2622. http://dx.doi.org/10.1096/fj.08-108985

[22] ZHAO J, ZHANG S, WU LY, ZHANG XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 2012; 28: 2940–2947. http://dx.doi.org/10.1093/bioinformatics/bts564

[23] MERMEL CH, SCHUMACHER SE, HILL B, MEYERSON ML, BEROUKHIM R et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology 2011; 12: R41. http://dx.doi.org/10.1186/gb-2011-12-4-r41

[24] HUANG DA W, SHERMAN BT, LEMPICKI RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009; 37: 1–13. http://dx.doi.org/10.1093/nar/gkn923

[25] BATCHELOR TT, BETENSKY RA, ESPOSITO JM, PHAM LDD, DORFMAN MV et al. Age-dependent prognostic effects of genetic alterations in glioblastoma. Clinical Cancer Research 2004; 10: 228–233. http://dx.doi.org/10.1158/1078-0432.CCR-0841-3

[26] VERHAAK RGW, HOADLEY KA, PURDOM E, WANG V, QI Y et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17: 98–110. http://dx.doi.org/10.1016/j.ccr.2009.12.020

[27] BLONS H, PALLIER K, LE CORRE D, DANEL C, TREMBLAY-GRAVEL M et al. Genome wide SNP comparative analysis between EGFR and KRAS mutated NSCLC and characterization of two models of oncogenic cooperation in non-small cell lung carcinoma. BMC Medical Genomics 2008; 1: 25. http://dx.doi.org/10.1186/1755-8794-1-25

[28] FURNARI FB, FENTON T, BACHOO RM, MUKASA A, STOMMEL JM et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. Genes & Development 2007; 21: 2683–2710. http://dx.doi.org/10.1101/gad.1596707

[29] HENSON JW, SCHNITKER BL, CORREA KM, VON DEIMLING A, FASSBENDER F et al. The retinoblastoma gene is involved in malignant progression of astrocytomas. Annals of neurology 1994; 36: 714–721. http://dx.doi.org/10.1002/ana.410360505

[30] REIFENBERGER G, REIFENBERGER J, ICHIMURA K, MELTZER PS, COLLINS VP. Amplification of multiple genes from chromosomal region 12q13–14 in human malignant gliomas: preliminary mapping of the amplicons shows preferential involvement of CDK4, SAS, and MDM2. Cancer research 1994; 54: 4299–4303.

[31] ZHANG JH, ZHANG SH, WANG Y, ZHANG XS. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. BMC Systems Biology 2013; 7(Suppl 2): S4. http://dx.doi.org/10.1186/1752-0509-7-S2-S4

[32] KNOBBE CB, TRAMPE-KIESLICH A, REIFENBERGER G. Genetic alteration and expression of the phosphoinositol-3-kinase/Akt pathway genes PIK3CA and PIKE in human glioblastomas. Neuropathology And Applied Neurobiology 2005; 31: 486–490. http://dx.doi.org/10.1111/j.1365-2990.2005.00660.x

[33] LIU ZY, YAO ZQ, LI C, LU YC, GAO CF. Gene Expression Profiling in Human High-Grade Astrocytomas. Comparative And Functional Genomics 2011; 2011: 245137 . http://dx.doi.org/10.1155/2011/245137

[34] ZHANG JH, WU LY, ZHANG XS, ZHANG SH. Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics 2014; 15: 217. http://dx.doi.org/10.1186/1471-2105-15-271

doi:10.4149/neo_2016_007

Supplementary Information

Identification of driver pathways in cancer based on combinatorial patterns of somatic gene mutations

H. T. LI[1,4], J. ZHANG[1], J. XIA[2,3,*], C. H. ZHENG[1,3,*]

[1]College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China; [2]Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China; [3]Center of Information Support & Assurance Technology, Anhui University, Hefei, Anhui 230601, China; [4]School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 273165, China

**Correspondence: jfxia@ahu.edu.cn, zhengch99@126.com*

1. Finding novel driver pathway mutations with CoGA

We used the functional annotation tool of the Database for Annotation, Visualization and Integrated Discovery (DAVID) for BIOCARTA pathway enrichment. We chose Benjamini test p<0.05, there are seven pathways in the result. The table is shown below.

Table 1. BIOCARTA pathway enrichment results

Pathway	Count	Benjamini
Cell Cycle: G1/S Check Point	5	8.61E-05
Tumor Suppressor Arf Inhibits Ribosomal Biogenesis	4	2.51E-04
p53 Signaling Pathway	4	3.95E-04
Cyclins and Cell Cycle Regulation	4	9.90E-04
RB Tumor Suppressor/Checkpoint Signaling in Response to DNA Damage	3	0.0082576
Telomeres, Telomerase, Cellular Aging, and Immortality	3	0.0157247
CTCF: First Multivalent Nuclear Factor	3	0.0166716

2. Apply CoGA to GBM data

We applied our CoGA algorithm to analyze all genomic data of gliomblastoma patients, which have 272 glioblastoma patients. We reported all the identified patterns with 2≤k≤3.

We constructed three metagenes, (CDKN2A, CDKN2B, EGFR), (CDKN2B, EGFR, PTEN), and (PTEN, PDGFRA), based on co-occurrence mutation patterns, one of which (CDKN2A, CDKN2B, EGFR), was also reported using the original 84 glioblastoma data.

The metagenes, (CDKN2B, EGFR, PTEN), and (PTEN, PDGFRA), have biological function connections with GBM. For example, it is well known that the occurrence of GBM is closely related to p16 (CDKN2) homozygous deletions[1], which was found frequently co-occurred with EGFR and PTEN alterations[2].

Then, we ran CoGA method for gene sets of size 2≤k≤3. The results are shown in Table 1.

Table 2. Results on the GBM dataset when 2≤k≤3

k	driver pathway genes
k=2	MICALCL, metagene1[a] MDM2, metagene1 HMP19, metagene1
k=3	HMP19, MDM2, metagene1

[a]metagene_1 is *CDKN2A*,*CDKN2B*,and*EGFR*

In our results, MDM2 and CDKN2A are the part of p53 signaling pathway. HMP19 (protein p19) has been demonstrated involved in dopamine receptor signaling. In addition, it was found highly expressed in neuroblastoma which is another brain cancer. Therefore, HMP19 may be crucial to GBM and brain development [3].

3. Conditions of construct co-occurrence mutation metagene

We set p<0.01 which is considered statistically significant, with the purpose of constructing ‘metagene’ with strict conditions (Coe B. P., et al. Nature genetics. 2014; 46: 1063-1071).

In some cases, though several mutations are detected with co-occurrence significance p<0.01, they are only a small proportion in the total samples, which indicates that these ‘co-occurring’ mutations just happen by chance. Consequently, we restricted the number of mutations to avoid this phenomenon. Below, we will show why 40% sample was used in this method.

The percentage threshold value that mutated genes account for the samples was set as 20%, 25%, 30%, 35%, 40% and 45%, respectively. Then the corresponding conditions of constructing ‘metagene’ were processed with our method. The results (when k=2 and k=3) are shown by Table 3 below.

Table 3. The results of different percentage threshold value with CoGA

Mutation genes number	Metagene	Results[a] (when k=2)	Results[a] (when k=3)
20%	SEC61G, EGFR, MTAP, CDKN2A, CDKN2B	TP53	CDC123, TP53

	MTAP, ELAVL2, CDKN2A, CDKN2B	TP53	MARCH9, TP53;
			TSFM, TP53;
			FAM119B, TP53

	TP53, MTAP, CDKN2A	TP53	CYP27B1, TP53
25% and 30%	SEC61G, MTAP, EGFR, CDKN2A, CDKN2B	TP53	CDC123, TP53
	TP53, MTAP, CDKN2A	TP53	CYP27B1, TP53
35%	MTAP, EGFR, CDKN2B	CYP27B1	CYP27B1, RB1
	TP53, MTAP, CDKN2A	TP53	CYP27B1, TP53
40%	CDKN2A, CDKN2B, EGFR	TP53	In our paper
45%	None	CDKN2B, CYP27B1	CDKN2B, RB1, CYP27B1

[a] Results are the genes and corresponding to the left metagene together.

In practice, no gold standard of known drivers is existence. However, well-studied cancer gene lists provide an approximate benchmark of known drivers. To help evaluate the quality of our results, we utilized a database of 547 known driver genes from the well-curated cancer gene list, CGC [4]. For each comparison, we used the precision measures.

$$\text{Precision} = \frac{(\#\textit{Mutated Genes found in CGC}) \cap (\#\textit{Genes found})}{(\#\textit{Genes found})}$$

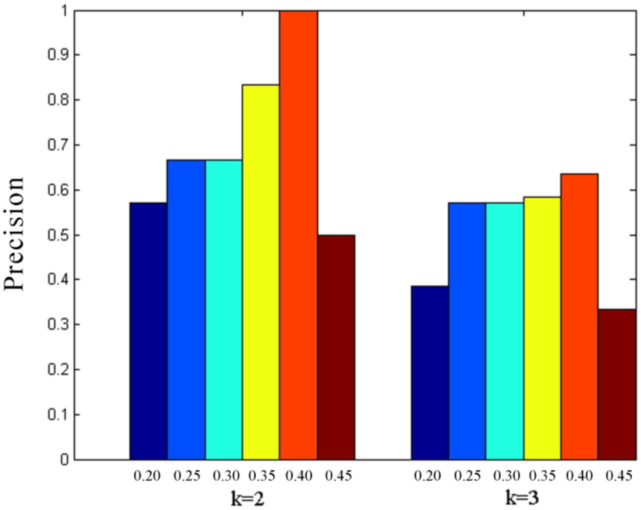

Figure 1. Comparison of the overlap between our results with different percentage threshold value and CGC

The Figure is shown the overlap between our results and CGC. We can see that the highest Precision was obtained when the percentage threshold value that mutated genes account for the samples was set as 40%. As a result, 40% sample was used in our method.

4. Simulation data study

To demonstrate the role of our algorithm in solving random somatic mutations, a simulation data set was constructed as below.

First, an empty m(samples)×n(genes) matrix was constructed(m=200,n=500 were used). Then we embedded p genes N_1, N_2,…,N_p (here p=3 was used) as the driver genes of common mutations. Through statistical analysis of GBM data, the mutation probability of the first gene was assumed to be 0.95. The probability that the following genes (N_2,…,N_p) have the same mutation number as the first gene was set to 0.8.The mutation probability of genes N_{p+1},…,N_q (here q=6 was used) was p_i(p_i=1-i*△, where △=0.05). The noisy probability of the rest genes (N_{q+1},…,N_n) was set from 0.02 to 0.12 in steps of 0.02. Because of the performances of GA is better than those of MCMC[5], the number of embedded genes detected in our algorithm was only compared with GA, as shown in the Figure below. (Figure 2)

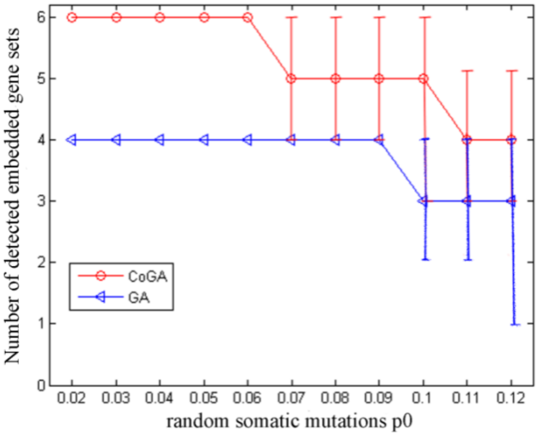

Figure 2. The number of embedded genes detected in CoGA and GA

We can see that as p_0 increases, the exclusivity among the genes in N_i decreases, so the find of the embedded gene sets N_i becomes increasingly difficult. In this figure, CoGA can precisely identify all six embedded gene sets when $p_0 \leq 0.06$. Meanwhile, in real data, the rate of noise mutation is mostly less than 0.07 [6]. So our algorithm can solve random somatic mutations in malignant cells.

5. Difference between our method and the other methods

Table 4 The methodological difference between our method and the other two previous methods of *Vandin et al.* (2012) [7] and *Zhao et al.* (2012) [5].

Table 4. Comparison of these three methods

Difference	Our method (CoGA)	MCMC	GA
co-occurring genes identified in driver pathway	Yes	No	No
Additional driver pathway (for example, p53 signaling pathway) identified without the requirement of removing prevalent genes	Yes	No	No
Data types used in the method	Somatic Mutation, CNV	Somatic Mutation, CNV	Somatic Mutation, CNV and Gene Expression Data
Cancer types used in the method	Any	Any	Any

References

[1] SIBIN M, BHAT DI, LAVANYA C, MANOJ MJ, AAKERSHITA S et al. CDKN2A exon-wise deletion status and novel somatic mutations in Indian glioma patients. Tumor Biology 2014; 35: 1467–1472. http://dx.doi.org/10.1007/s13277-013-1201-5

[2] ATTOLINI CSO, CHENG YK, BEROUKHIM R, GETZ G, ABDEL-WAHAB O et al. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. Proceedings Of the National Academy Of Sciences Of the United States Of America 2010; 107: 17604–17609. http://dx.doi.org/10.1073/pnas.1009117107

[3] FENG Y, HURST J, ALMEIDA-DE-MACEDO M, CHEN X, LI L et al. Massive human co-expression network and its medical applications. Chem Biodivers 2012; 9: 868–887. http://dx.doi.org/10.1002/cbdv.201100355

[4] FUTREAL PA, COIN L, MARSHALL M, DOWN T, HUBBARD T et al. A census of human cancer genes. Nat Rev Cancer 2004; 4: 177–183. http://dx.doi.org/10.1038/nrc1299

[5] ZHAO J, ZHANG S, WU LY, ZHANG XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 2012; 28: 2940–2947. http://dx.doi.org/10.1093/bioinformatics/bts564

[6] ZHANG JH, WU LY, ZHANG XS, ZHANG SH. Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics 2014; 15: 217. http://dx.doi.org/10.1186/1471-2105-15-271

[7] VANDIN F, UPFAL E, RAPHAEL BJ. De novo discovery of mutated driver pathways in cancer. Genome Research 2012; 22: 375–385. http://dx.doi.org/10.1101/gr.120477.111