International Journal of Hybrid Information Technology
Vol.9, No.4 (2016), pp. 291-304
http://dx.doi.org/10.14257/ijhit.2016.9.4.25

Multi-population Genetic Algorithm for Identifying Mutated Driver Pathways in Cancer

Shu-Lin Wang[1],* and Can-Jun Hu[2]

College of Computer Science and Electronics Engineering, Hunan University,
Changsha, Hunan, 410082, China
[1]smartforesting@gmail.com, [2]hcj8727@163.com

Abstract

Cancer is one of complex diseases that are a big threat to mankind's health and life so far, and it is well known that the somatic mutation is an important factor leading to cancer development. Finding these important somatic mutation or driver mutation is of great benefit to the gene therapy of cancer patients. However, it is difficult to distinguish driver mutations from a great number of passenger mutations because of mutational heterogeneity, which is the key factor to deal with the problem of cancer treatment. In this study, we present an efficient way Multi-population Genetic Algorithm (MPGA) integrated with chaos algorithm to find important mutated cancer genes, which can be transformed into the maximum weight submatrix problem. The experiments on the simulated and several real mutation datasets indicate that the presented methods performs more efficiently and can find more driver genes. Comparing with other relevant methods, MPGA method is proved the most robust one among these approaches. Analyzing the experimental results obtained indicates that these important pathways rediscovered play a key role in cancer development.

***Keywords**: Multi-population genetic algorithm; driver pathway; passenger mutation; Chaos algorithm; cancer*

1. Introduction

Cancer is a complex disease that is largely driven by many somatic mutations gradually accumulated during the lifetime of an individual. It can change person’s DNA structure or genome. As we know, infinite proliferation is a feature of cancer cells, and another dreadful feature is that it can spread to others through blood circulation and lymphatic system [1]. These mutations are divided into single nucleotide variants substitutions (SNVs), small indels, larger copy number aberrations (CNAs), structural aberrations (SVs) which is also called large-genome rearrangements, and so on. Numerous experimental works have identified lots of driver genes or pathways, which have dramatically expanded our knowledge about somatic mutations in cancer, such as The Cancer Genome Altas (TCGA). More importantly, some mutations have been successfully applied into medical treatment. For example, Imatinib has been used to target cells expressing the BCR-ABL fusion gene in chronic myeloid leukemia [2, 3], and Gefitinib is therapeutic to inhibit the epidermal growth factor receptor in lung cancer [4]. However, there are lots of works should be done because we have few knowledge about most cancer.

Generally, the somatic mutation can be divided into two types in another criterion: one is called as passenger mutation that has accumulated in somatic cells but has no influence to cell proliferation in cancer, and the other is driver mutation which is important for cancer development, and can promote the cancer cell to proliferate infinitely and diffuse

* Corresponding Author

ISSN: 1738-9968 IJHIT

slowly into other organs. A major challenge in finding mutated driver pathways in cancer is to distinguish driver mutations which are our target from sporadic passenger mutations. The methods for identifying mutated driver pathways, therefore, are needed urgently. Based on previous research, the study process mainly includes recurrent mutations and pathways study.

A variety of approaches have been developed for finding driver genes or pathways in cancer so far. One model to find driver mutations is to identify genes that are mutated at significant frequencies in a lot of cancer patients. To put it simply, lots of work about the driver gene problem focus on a single gene. The original method for finding driver gene is to identify recurrent mutations. There is a phenomenon that some cancer genes are mutated at high frequency (such as TP53), while some cancer genes are mutated at much lower frequency. The approach contains two difficulties: First, it is a challenge to get a reasonable estimate of the back ground mutation rate (BMR), because the rate of somatic mutation and selection or clonal amplification in the somatic evolution of a cancer should be both taken into account, but it is hard to obtain these data. Second, each of pathways contains more than one gene, and there are lots of combinations of driver mutations that play a key role in cancer. Mutational heterogeneity complexes the calculating work and it is difficult to distinguish passenger genes from driver genes in cancer [5]. In addition, testing the recurrence of individual mutations requires examining mutations which are part of cellular signaling and regulatory pathways. The method need prior knowledge, but the known pathways in database are incomplete. Because the superposition of all components in a pathway and the useful information for special cell types are incomplete so far, the method for identifying recurrent mutations are not efficient methods to measure the importance of a gene, particularly large cancer samples.

Compared with the approaches mentioned above, we should consider the combinations of mutations and find or develop new algorithms to discover driver pathways without relying on prior knowledge. Meanwhile, it is necessary to hold the point that we should consider the problem in pathway level rather than in gene level. Three approaches [2] have been used to solve the problem: Firstly, identify combinations of recurrent mutations in pre-defined gene sets which come from the databases of known pathways, *e.g.*, Recurrent Mutually Exclusivity (RME) [6] method focus on building sets of genes from cancer data. Secondly, identify the driver genes or pathways through the genome-scale interaction networks, *e.g.*, the approach Mutual Exclusivity Modules (MEMo) [7] proposed by Ciriello *et al.* find modules by considering mutual exclusivity between mutations for pair of genes that have recorded interactions in a protein interaction network. Thirdly, De novo algorithms are introduced to solve the so-called maximum weight submatrix problem. More recently, Vandin *et al.* developed De novo Driver Exclusivity (Dendrix) [8] to identify the driver genes and pathways with high coverage and mutual exclusivity feature. With the hypothesis that each driver pathway contain approximately one driver mutation per patient and an important driver pathway should be mutated in many patients promoted by Raphael *etc.*, the problem has changed to be focused on solving the maximum weight submatrix problem.

The submatrix problem was designed by Vandin *et al.* It was developed to de novo discover a single mutated driver pathway from mutation data. A weight function W was introduced by combining the coverage and exclusivity features they proposed: high coverage- most patients have at least one mutation the set; high exclusivity- nearly all patients have no more than one mutation in the set [8]. They define a measure on set of genes that quantifies the genes which satisfy the two features. Before the Dendrix algorithm, they have tried the greedy algorithm. When the algorithm is given a sufficiently large number of patients, the results shows an optional solution.

For Dendrix, they develop Markov chain Monte Carlo (MCMC) algorithm because of these statistical assumptions is too restrictive for some data, and the number of patients in currently available data sets are not incomplete. MCMC approach sample sets of genes according to a distribution that gives significantly higher probability to set of genes with high coverage and exclusivity features. It is a well-established technique to sample from combinatorial spaces. What's more, they think that this problem is computationally difficult to solve. The problem is too restrictive for analysis of real somatic mutation data.

In recent years, several studies have made contributions to solve the maximum weight submatrix problem [8, 9]: considering mutation data from cancer patients, they create a mutation matrix A with m rows and n genes, where each row is a patient and each column represents a gene. The corresponding entry A_{ij} in row i and column j is equal to 1 if gene j is mutated in patient i. In our study, we compare Dendrix, Simple Genetic Algorithm (SGA) and Multi-population Genetic Algorithm (MPGA) onto several data. It is well known that the Markov chain Monte Carlo (MCMC) using in Dendrix and the SGA algorithm are stochastic algorithm. They are easily trapped in local optimal solution or sometimes it cannot find some driver genes and pathways. Besides, SGA method always has the premature phenomenon and slow convergence deficiencies and the appropriate mutation rate is hard to define.

In order to overcome the problem inferred above, we combine MPGA algorithm and chaos algorithm advantages to improve the stability of these methods. In this way, the population diversity can be increased and the premature phenomenon might be avoided. The MPGA method divides populations into subpopulations. Among the subpopulations, the information of subpopulations can be exchanged through immigration. For example, some excellent individuals can be exchanged among the subpopulations to improve population diversity. Meanwhile, by importing the chaos operator, it has overcome the defect of precocity for SGA, for its particularly inherent randomness and ergodicity to skip the local optimization. The experimental results on several datasets indicate that our approach can find more relevant genes and can improve the stability of identifying driver genes and pathways in a certain degree.

2. Methods

2.1. The Problem

Driver mutations occur in minority genes, while passenger mutations occur randomly across all genes. As Vandin *et al*. introduced, we can identify the driver genes in $m \times n$ binary mutation matrix A with two features: high coverage and mutual exclusivity. They defined the coverage overlap as follows.

$$\omega(M) = \sum_{g \epsilon M} |\Gamma(g)| - |\Gamma(M)| \tag{1}$$

The maximum weight submatrix problem turns into measuring the trade-off between coverage and exclusivity, the weight scoring function is described as below.

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \epsilon M} |\Gamma(g)| \tag{2}$$

This problem is NP-hard problem[8]. $\Gamma(g) = \{i: A_{ig} = 1\}$ means the corresponding gene g is mutated in the matrix. Similarly, for a set M of genes, $\Gamma(M) = \cup_{g \epsilon M} \Gamma(g)$ denote the set of patients in which at least one of genes in M is mutated and $W(M)$ measures the coverage overlap of M [8]. As a result, the solution of the problem turns to be finding a submatix with higher weight value W.

2.2. Main Process of MPGA

MPGA simulates natural processes, such as selection, recombination, mutation, migration. Its individuals are selected according to their fitness function defined as the weight function W in Eq. (2). Selection means which individuals are chosen for mating or recombination and it is a preparation stage for mutation. In general, rank-based fitness assignment behaves in a more stable way than proportional fitness assignment, so we chose the rank-based fitness assignment in our approach. The migration model divides the populations into multiple subpopulations and it is a divide and conquer algorithm. For a certain number of generations, these subpopulations evolve independently from each other and then the migration will be distributed between the subpopulations. After recombination every offspring will be mutated in a low probability. We randomly alter one variable value with 1 to 0 and another variable value with 0 to 1 oppositely. When the chaos variables are put into practice to achieve chaos optimal search, there is no doubt that this algorithm will be much more superior to random search for its inherent ergodicity to skip the local optimization.

Additionally, because the chaos algorithm has the ergodic advantage, when we generate initial population, the chaos operator could be used in the process. We have used Logistic model to sample, and the parameter $\mu = 4$.

Figure 1 shows a brief structure of MPGA. MPGA works on populations of individuals instead of single solutions, which means that the process of MPGA can be designed in a parallel way. Individuals are selected according to their fitness. Parents are recombined to produce offspring. All offspring then will be mutated, and calculate the fitness of offspring. After producing a new generation, the immigration will be transformed among the subpopulations.

The cycle is performed until the max weight is reached. Genetic Algorithm is powerful and performs well on a broad class of problems, but the stability of the approach need to be increased. The MPGA simulates the evolution of a species in a way just like the evolution of the nature in our real life. With individuals are exchanged between the subpopulations, MPGA can get better performance.

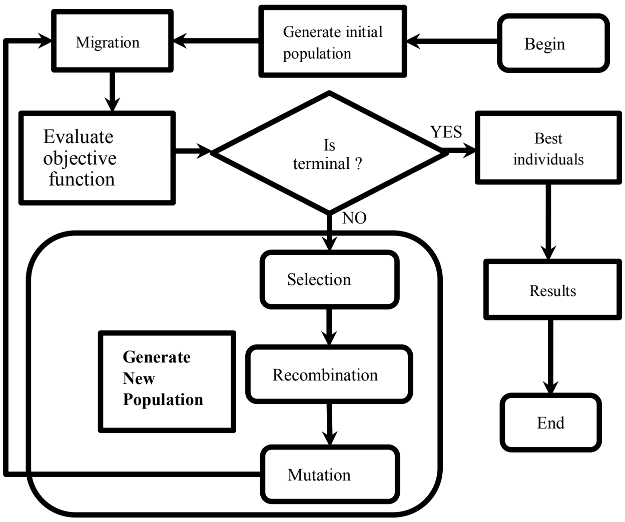


Figure 1. A Brief Structure of MPGA

The procedure of MPGA is described as follows.

1) Set the parameter of the program, *i.e.*, the number of population MP, population size P, mutation rate p_m, iteration of the algorithm $nger$, and submatrix size k.

2) Generate initial population. In each iteration, it will generate MP initial populations, here we will adopt chaos operator to sample, and each population is selected from the current population based on the selection probability, and generates offspring through recombination.

3) Immigrant is transformed among the populations.

4) Each offspring may optionally be mutated with a certain probability p_m.

5) All individuals are ranked according to their weight scoring value. The best result will be treated as the next generation which is used as the current population in the subsequent iteration.

6) Repeat 2-5 steps until the termination criterion is satisfied.

3. Experiments

3.1. Simulated Mutation Data

We first compared MPGA to SGA and Dendrix on simulated data. In general, because of the stochastic of approaches, these may not get the best solution in a run. So we could measure the stability of these approaches in two features: average score and well-run. Average sore (*Z*) is the result of dividing the sum of max weight by total number of runs (*N*), m represents the sum of max weight pathways in a run. Well-run means the run of an approach that can find the max weight pathway.

$$Z = \frac{\sum_{j=1}^{N}(m \times W_j)}{N} \tag{3}$$

where W_j represents the weight value of the j-th run. There are two ways to evaluate the stability performance of these methods. One way is the comparison of average score, and the other way is counting the well-runs of one known pathway with plenty of runs. In our experiments, the results can be divided into two situations according to the number of pathways with max weight value: single pathway case and multi pathways case. Average score feature adapts to both cases mentioned above.

In order to guarantee the fair comparison, the total number of individuals should be equal in each subpopulation. In experiments, the parameters are set as: MP=20, P=30**floor*(n/100), p_m=0.008, and $nger$=2000. The experimental results are described in Tables 1-6.

Table 1. Average Score of each Method on Simulated Data. *N* Represents the Number of Runs of Program. *S* Represents Samples. *S*=48 Corresponds To Multi Pathway Case, *S*=86 Corresponds to Single Pathway Case

N (Runs)	Average score *Z* (*S*=48)			Average score *Z*(*S*=86)		
	Dendrix	*SGA*	*MPGA*	*Dendrix*	*SGA*	*MPGA*
50	6.120	7.480	9.520	9.540	11.660	15.900
100	7.820	8.840	19.720	12.720	14.310	24.380
1000	6.494	12.036	30.702	10.123	17.596	15.794
1500	8.817	9.089	20.898	13.709	14.628	16.536
2000	8.551	9.452	20.094	13.356	14.866	18.576

Table 2. Average Score of each Method on Simulated Data. N Represents the Number of Runs of Program. S Represents Samples. S=50 Corresponds to Multi Pathway Case, S=90 Corresponds to Single Pathway Case

N	Average score *Z* (*S*=50)			Average score (*S*=90)		
(Runs)	*Dendrix*	*SGA*	*MPGA*	*Dendrix*	*SGA*	*MPGA*
50	7.600	10.640	18.240	9.280	10.440	26.680
100	9.500	12.160	19.380	11.020	13.340	28.420
1000	8.094	11.590	19.114	9.048	13.978	23.432
1500	7.549	11.298	18.037	8.784	12.524	24.940
2000	7.619	11.438	18.981	8.758	11.629	23.142

Table 1 and Table 2 show the results comparison of Dendrix, SGA and MPGA on simulated mutation data in different samples (*S*). The average score of these methods are compared varying with the number of runs. Here we considered both cases mentioned above. Overall, it is easy to know the performance of our method is higher than others, which means our method can identify the pathway efficiently. Secondly, MPGA identify the max weight pathway with great probability, while other methods even could not get the solution, which is the obvious difference among these methods. For example, MPGA can identify a pathway whose max weight value *W* is 35 when *S*=48 and find a pathway whose max weight value equals to 58 when *S*=90. However, Dendrix could not get it.

As discussed before, there is a situation that multi-pathways correspond to the same weight score. For MPGA, it can identify lots of max weight pathways, while other methods run with higher randomness and could not get the driver pathway in some cases. For example, when *S*=86, MPGA can identify 4 pathways with the same max weight value, SGA always can find one or two pathways, Dendrix performance is worse than SGA. Such phenomenon lies in the instability of MCMC and SGA. As shown in Table 2, MPGA's average score is the highest among these methods. As a whole, MPGA achieves our goal and is a much more stable method for identifying driver pathway in cancer.

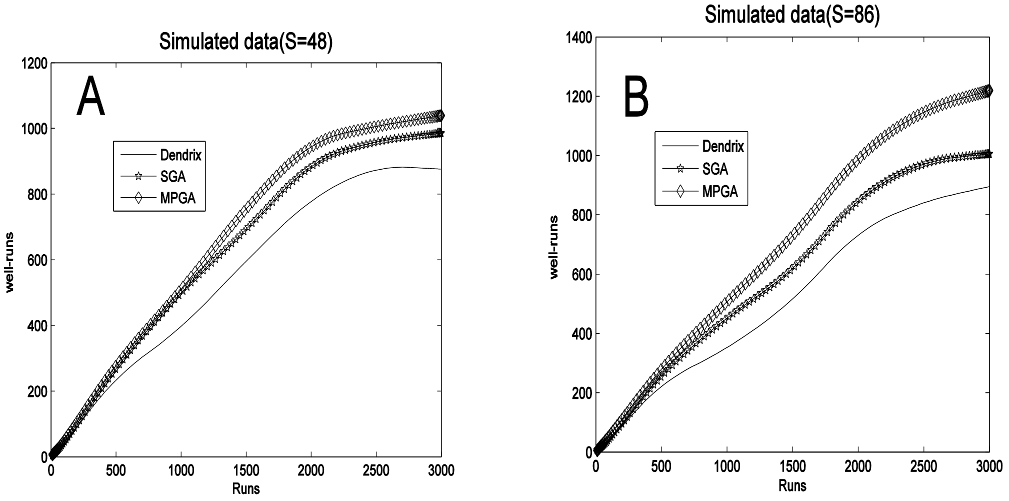


Figure 2. Well-runs of Dendrix, SGA, and MPGA vary with the Number of Runs from 10 to 3000 with Different Samples(*S*). *N*: the Number of Runs of Program (A) *S*=48, Corresponding to the Multi Pathways Case (B) *S*=86, Single Pathway Case

Figure 2 shows the change of well-runs in different samples. We calculate the available runs in a test. For example, if we run the program 50 times, and we found 36 of them can identify the max pathway, we took the 36 as well-runs. It is easy to know MPGA runs with high performance from Figure 2. In summary, when the samples are large, the corresponding well-runs will be higher. Just as shown in Figure 2, the results of subplot B is larger than A. In addition, when samples are small, MPGA also can identify max weight pathways with high performance. In summary, the approach MPGA achieves our goal and is a much more stable method for identifying driver pathway in cancer.

3.2. Lung Cancer Data

Then we applied MPGA onto lung cancer, glioblastoma (GBM), Head and neck squamous cell carcinoma (HNSCC) and ovarian data. The lung data is consists of 163 rows and 346 columns [8]. The average score and well-run features as compared for each method. From Table 3, the results are similar and show that these methods run with good performance. When the parameter k = 3, the max weight value in our experiments is 32. The performance in these methods is close. Because on the one hand, the data is small, on the other hand, the max weight pathway is single. From Table 2, the results are similar to results on the data before. Here when we set S =120, and find MPGA can identify 7 or 8 max weight pathways. SGA can identify 6 pathways on average. Specially, we find that MPGA even get the whole well-runs in some case. From this experiment, we began to discover that Dendrix and SGA depend more on data type or structure than MPGA. It is a disparity among them in other cases. In some case, these results are close. Above all, from another aspect we deduce that MPGA are more applicable to different cancer data.

Figure 3 shows part of pathway in lung cancer. Gene sets (EGFR, KRAS, and STK11) and (ATM, TP53) can be discovered. Because EGFR, ATM, TP53, KRAS, and STK11 are famous driver gene in lung cancer. When we remove (EGFR, STK11, KRAS, ATM, TP53) and set k=5 (k is number of genes in M), we identify a pathway (CDKN2A, GNAS, LRP1B, NF1 and NTRK3) whose weight value is 44. CDKN2A and NTRK3 are driver genes in lung cancer to control cell cycle [10]. GNAS play an important role in calcium signaling pathway. NF1 can lead to neurofibromatosis.

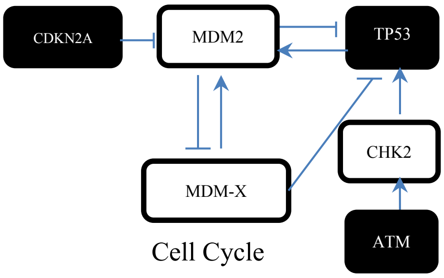


Figure 3. Part of Cell Cycle Pathway in Lung Cancer

Table 3. Average Score of each Method on Lung Cancer Vary with Runs when *S*=120

N (Runs)	Average score *Z* (*S*=120)		
	Dendrix	*SGA*	*MPGA*
10	23.100	23.100	26.400
50	29.700	33.000	36.300
100	26.400	31.680	34.320
1000	26.532	29.568	32.175
2000	28.083	31.251	31.416

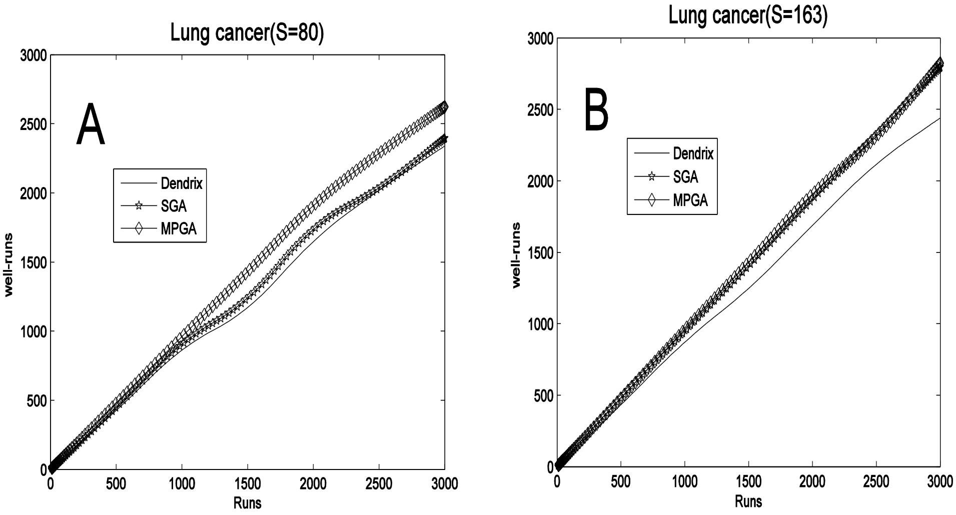


Figure 4. Well-runs of Dendrix, SGA, and MPGA Vary with Runs from 10 to 3000 on Lung Cancer Data (A) S =80 (B) S=163

As shown in Figure 4, the performance of SGA and MPGA is close when S is the max value 163. The trend of subplot A and B is similar. When the samples are large, the corresponding well-runs will be higher. The value of subplot B is larger than A. In subplot A, the well-runs value is clearly greater than other methods. As a whole, the performance of MPGA is superior to SGA. There is a phenomenon that the famous driver gene or pathway of lung cancer is easy to identify.

3.3. Glioblastoma Cancer Data

The glioblastoma dataset is downloaded from bioinformatics [11]. After processing the dataset, we got a mutation matrix with 90 samples and 1126 genes through an excel file, Then we can run our program with the file. Here we set the iteration $nger$=1000, number of child population MP=20, the number of individual in one population $popsize$=25, and k=3. When S=84, we ran the MPGA, and can get the max weight $W(M)$ = 62. When S=44, the result $W(M)$ = 34, corresponding 3 max weight ways. Many experiments are carried out in our research. The results are described below. The results are similar to the results of simulated data (Table 4).

As shown in Table 4, MPGA can get the highest score, while the Dendrix run with the lowest results and it depends on the samples highly. The average weight values of MPGA, SGA, Dendrix are about 56, 47, 33 in Table 4, respectively. So the performance of these approaches are ranked like MPGA > SGA > Dendrix. It is a good idea that we identify more feasible solutions as soon as possible, and solve it with additional information, such as gene expression data [11]. In real mutation data, there are multiple optimal solutions. In addition, because of the noise in the data or other factors, the max weight pathway may not be the best one in biological mutation data. So we should identify more feasible solutions as soon as possible.

Similarly, we divide it into two conditions: S=84 and S=44, and then compared the stability of finding the optimal pathway (CDKN2B, RB1, CDK4) [8]. Both we checked the well-runs of each method varying with the number of samples. Firstly, we run the methods based on two loops, the times of outer loop is 5, and the inner loop is {10, 50, 100, 1000, 1500, 2000}, then calculate their average results for each case. Generally, the results are similar to the experiments on simulated data. In addition, when the samples S=44, the MPGA's performance is much better than other methods. In other words, this phenomenon shows MPGA is also efficient even in small sample data.

Table 4. Average Score of each Method on GBM data. *N* Represents the Number of Runs of Program in a Test. *S*=84 Corresponds to Multi Pathway Case, *S*=44 Corresponds to Single Pathway Case

N	Average score *Z* (*S*=84)			Average score (*S*=44)		
(Runs)	*Dendrix*	*SGA*	*MPGA*	*Dendrix*	*SGA*	*MPGA*
10	37.200	49.600	55.800	10.200	13.600	23.800
50	34.720	48.360	58.280	14.280	17.340	23.120
100	31.620	45.880	57.660	12.92	13.230	21.080
1000	30.318	48.112	56.792	11.390	13.532	21.318
2000	31.186	46.934	56.854	10.404	12.512	20.094

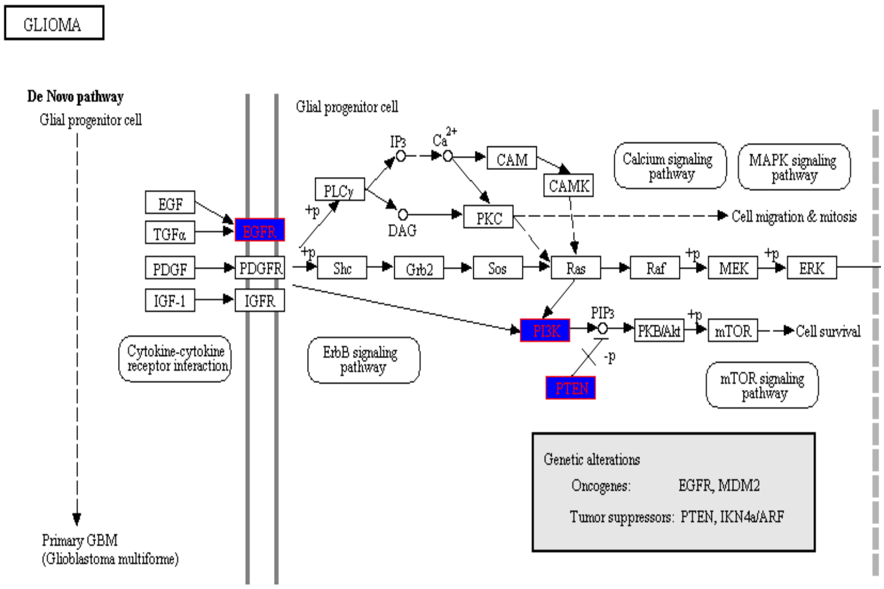


Figure 5. Part of RTK/RAS/PIK Signal Pathway in GLIOMA (KEGG)

As shown in Table 4, the smaller the samples are, the higher difference among these methods you get. The value are ranked as MPGA>SGA>Dendrix. The result clearly shows that MPGA successfully avoids trapping in local optimal solution and is much more applicable to different samples.

Besides, when the famous driver genes (CDK4, CDKN2B, TSPAN31, RB1, and TP53) and the metagene CYP27B1 are moved, on the remaining genes, we set k = 5, then run the MPGA and find three pathways whose weight are 50, including 6 genes totally. Other than ERBB2, the rest genes (EGFR, GRIA2, PIK3CA, PIK3R1, and PTEN) play an important roles in glioblastoma cancer [11-13]. The identified known pathways are shown in Figure 5.

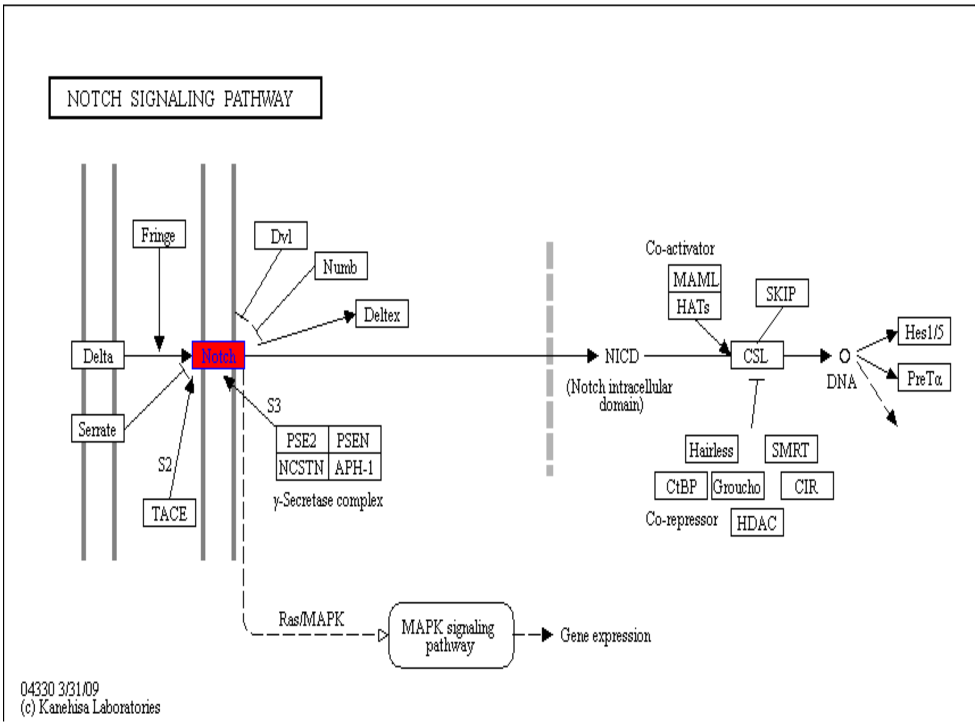
3.4. Head and Neck Squamous Cell Carcinoma Data

HNSCC is the sixth most common deadly cancer in the world. The survival rates for many HNSCC patients have made little increase over the past 40 years [14]. This mutation data matrix includes 74 rows and 4920 columns, which is sparse.

Some significantly mutated genes had previously been detected in HNSCC data, such as TP53, TTN, and CDKN2A. TP53 and TTN are mutated in the majority (46/74), (23/74) of samples respectively. Therefore we remove the genes TP53 and TTN because of the prevalence of mutation [11]. When k=3, we get gene set (CDKN2A, PCLO, SYNE1). SYNE1 was observed in 8% of HNSCC samples, it have been implicated in the regulation of nuclear polarity. PCLO mutation was seen in 12% of cases, and it is important for terminal squamous differentiation [14]. When *k*=7, we got the max weight *W*=46. MPGA can identify 7 optimal pathways. NOTCH1 mutations have been reported that it occurs in 10% to 15% of head and neck squamous cell carcinomas. The result of the paper [15] shows a bimodal pattern of NOTCH pathway transformations, which will be more suitable for HNSCC treatment (Figure 6).

Table 5. Average Score of each Method on HNSCC Data. *N* Represents the Number of Runs of Program in a Test. *S*=44 Corresponds to Multi Pathway Case, *S*=70 Corresponds to Single Pathway Case

N	Average score *Z* (*S*=44)			Average score (*S*=70)		
(Runs)	*Dendrix*	*SGA*	*MPGA*	*Dendrix*	*SGA*	*MPGA*
10	57.200	66.600	75.900	33.200	43.600	53.860
50	54.720	68.360	78.280	34.480	47.340	53.150
100	51.620	65.880	77.760	32.920	43.236	51.080
1000	50.318	68.112	76.782	34.390	43.582	51.328
2000	51.186	66.934	76.854	36.400	45.512	50.694



04330 3/31/09
(c) Kanehisa Laboratories

Figure 6. Part of NOTCH Signaling Pathway in HNSCC

As we did above, we also carried out the same experiments on HNSCC data. Unlike the results on lung cancer data, the performance of these methods are similar to the results on simulated data. This phenomenon demonstrates that the performance of these methods is similar.

3.5. Ovarian Carcinoma Data

The data comes from bioinformatics [11]. They have analyzed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number aberrations [16]. The data matrix covers 313 samples and 5385 genes.

We also applied the MPGA algorithm onto ovarian carcinoma patients from The Cancer Genome Atlas (TCGA, 2011). The experiments are similar to the results of HNSCC data. TP53 is mutated in 251 patients and the analysis of gene TTN shows that the mutations of it are more likely artifacts [11, 16]. We ran the MPGA algorithm with the range of *k* (2≤*k*≤10). For *k*=2, we find the gene set (CCNE1, MYC), they are key genes in ovarian cancer; for *k*=3, the optimal gene set is CCNE1, MYC, and NINJ2. They play an important role in the ovarian carcinoma. After removing the above genes, when *k*=4, we identify a set of four mutation groups (KRAS, PPP2R2A, PRPF6 and RYR2) that is altered in 102/313 of the patients. When *k=5*, the gene set (KRAS, MAPK8IP2, NF1, MUC16, STMN3) is identified.

Table 6. Average Score of each Method on Ovarian Data. *N* Represents the Number of Runs of Program. *S*=313 Corresponds to Multi Pathway Case, *S*=90 Corresponds to Single Pathway Case.

N	**Average score *Z* (*S*=313)**			**Average score (*S*=160)**		
(Runs)	***Dendrix***	***SGA***	***MPGA***	***Dendrix***	***SGA***	***MPGA***
10	47.200	59.600	65.600	30.200	43.604	53.806
50	44.720	58.360	68.260	34.286	47.342	53.122
100	41.240	55.880	67.660	32.922	43.232	51.280
1000	40.328	58.112	66.792	31.380	43.522	51.016
2000	41.166	56.934	66.840	30.404	42.502	50.082

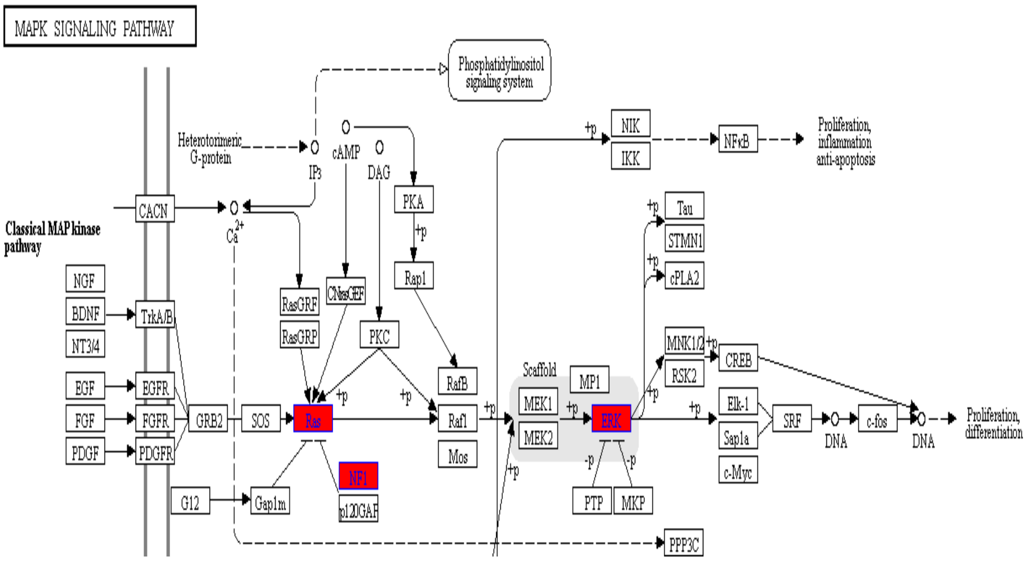


Figure 7. Part of MAPK Signaling Pathway in Ovarian Cancer

It is well known that KRAS, NF1 and MAPK8IP2 are part of MAPK signaling pathway (Figure 7). The mutation of STMN3 is related to malignant progression of multiple cancer types [17]. The abnormal of STMN3 can affect multiple cancer and play a key role in EMT. The mutation of MUC16 is related to Wnt signaling pathway, and play an important role in ovarian carcinoma [18]. Thus, MPGA can identify extra pathways and shows stable performance.

4. Conclusions

This study mainly focus on some of the challenges in finding driver mutations and driver genes in cancer. We discuss several computational approaches that are used to detect somatic mutations and pathways in our research. With the development of large-scale cancer sequencing projects, the rapidly computational identification method of driver mutations is needed urgently. It could be an important step in determining patient prognosis and treatment.

Dendrix is a Markov Chain Monte Carlo (MCMC) algorithm that samples sets of k genes according to their submatrix weight value W. While the MCMC algorithm could not identify optimal gene sets in some cases. sMPGA can also produce sets M with strictly larger weight. What's more, there are a variety of cases might occur. Firstly, there may be multiple gene sets with maximum weight on a run, while SGA only finds part of them. Secondly, what the SGA identified in a run may not be the optimal solution when considered in isolation or the algorithm is stochastic. While MPGA can basically find all of these feasible solutions on real somatic mutation data. As a result, MPGA is much more stable method for identifying driver pathway in cancer.

The maximum weight submatix problem model has its limitation that the model could not integrate biological information effectively. Therefore our future work may combine the interaction network model with the max weight submatirx model to design a new model to discover the fact hiding in cancer data.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61472467, 60973153 and 61471169) and the Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province.

References

[1] I. J. Fidler, "Timeline - The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited," Nature Reviews Cancer, vol. 3, **(2003)** June, pp. 453-458.

[2] B. J. Raphael, J. R. Dobson, L. Oesper and F. Vandin, "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine", Genome Medicine, vol. 6, **(2014)** January 30.

[3] J. M. Goldman and J. V. Melo, "Mechanisms of disease - Chronic myeloid leukemia - Advances in biology and new approaches to treatment", New England Journal of Medicine, vol. 349, **(2003)** October 9, pp. 1451-1464.

[4] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich and S. Gabriel, "EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy", Science, vol. 304, pp. 1497-1500, **(2004)** June 4.

[5] F. Vandin, E. Upfal and B. J. Raphael, "Finding driver pathways in cancer: models and algorithms", Algorithms for Molecular Biology, vol. 7, **(2012)** September 6.

[6] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape and A. Milosavljevic, "Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors", Bmc Medical Genomics, vol. 4, **(2011)** April 14.

[7] G. Ciriello, E. Cerami, C. Sander and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," Genome Research, vol. 22, **(2012)** February, pp. 398-406.

[8] F. Vandin, E. Upfal and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer", Genome Research, vol. 22, **(2012)** February, pp. 375-385.

[9] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control", Nature Medicine, vol. 10, **(2004)** August, pp. 789-799.

[10] M. K. Otnaess, S. Djurovic, L. M. Rimol, B. Kulle, A. K. Kahler and E. G. Jonsson, "Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia", Neurobiology of Disease, vol. 34, pp. 518-524, **(2009)** June.

[11] J. Zhao, S. Zhang, L. Y. Wu and X. S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer", Bioinformatics, vol. 28, **(2012)** November 15, pp. 2940-7.

[12] F. Beretta, S. Bassani, E. Binda, C. Verpelli, L. Bello and R. Galli, "The GluR2 subunit inhibits proliferation by inactivating Src-MAPK signalling and induces apoptosis by means of caspase 3/6-

dependent activation in glioma cells", European Journal of Neuroscience, vol. 30, pp. 25-34, **(2009)** July.

[13] S. Maas, S. Patt, M. Schrey, and A. Rich, "Underediting of glutamate receptor GluR-B mRNA in malignant gliomas," Proceedings of the National Academy of Sciences of the United States of America, vol. 98, **(2001)** December 4, pp. 14687-14692.

[14] N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis and A. Sivachenko, "The Mutational Landscape of Head and Neck Squamous Cell Carcinoma", Science, vol. 333, **(2011)** August 26, pp. 1157-1160.

[15] W. Y. Sun, D. A. Gaykalova, M. F. Ochs, E. Mambo, D. Arnaoutakis and Y. Liu, "Activation of the NOTCH Pathway in Head and Neck Cancer", Cancer Research, vol. 74, **(2014)** February 15, pp. 1091-1104.

[16] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. W. Cramer and F. Dao, "Integrated genomic analyses of ovarian carcinoma", Nature, vol. 474, **(2011)** June 30, pp. 609-615.

[17] F. Fang, A. J. Flegler, P. Du, S. Lin and C. V. Clevenger, "Expression of Cyclophilin B is Associated with Malignant Progression and Regulation of Genes Implicated in the Pathogenesis of Breast Cancer," American Journal of Pathology, vol. 174, **(2009)** January, pp. 297-308.

[18] M. Comamala, M. Pinard, C. Theriault, I. Matte, A. Albert and M. Boivin, "Downregulation of cell surface CA125/MUC16 induces epithelial-to-mesenchymal transition and restores EGFR signalling in NIH:OVCAR3 ovarian carcinoma cells", British Journal of Cancer, vol. 104, **(2011)** March 15, pp. 989-999.

[19] R. D. Prasasya, D. Tian and P. K. Kreeger, "Analysis of cancer signaling networks by systems biology to develop therapies", Seminars in Cancer Biology, vol. 21, **(2011)** June, pp. 200-206.

[20] W. Ren and Q. Zhao, "A note on 'Algorithms for connected set cover problem and fault-tolerant connected set cover problem", Theoretical Computer Science, vol. 412, **(2011)** October 21, pp. 6451-6454.

[21] V. Bansal, A. L. Halpern, N. Axelrod and V. Bafna, "An MCMC algorithm for haplotype assembly from whole-genome sequence data", Genome Research, vol. 18, **(2008)** August, pp. 1336-1346.

[22] K. Burkitt and M. Ljungman, "Phenylbutyrate interferes with the Fanconi anemia and BRCA pathway and sensitizes head and neck cancer cells to cisplatin", Molecular Cancer, vol. 7, **(2008)** March 6.

[23] L. Chin, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways (vol. 455, pg 1061, 2008)", Nature, vol. 494, **(2013)** February 28, pp. 506-506.

[24] S. Efroni, R. Ben-Hamo, M. Edmonson, S. Greenblum, C. F. Schaefer and K. H. Buetow, "Detecting Cancer Gene Networks Characterized by Recurrent Genomic Alterations in a Population", Plos One, vol. 6, **(2011)** January 4.

[25] A. Gonzalez-Perez and N. Lopez-Bigas, "Functional impact bias reveals cancer drivers", Nucleic Acids Research, vol. 40, **(2012)** November.

[26] A. M. Hudson, T. Yates, Y. Y. Li, E. W. Trotter, S. Fawdar and P. Chapman, "Discrepancies in Cancer Genomic Sequencing Highlight Opportunities for Driver Mutation Discovery", Cancer Research, vol. 74, pp. 6390-6396, **(2014)** November.

[27] B. E. Johnson, M. G. Kris, L. D. Berry, D. J. Kwiatkowski, A. J. Iafrate and M. Varella-Garcia, "A multicenter effort to identify driver mutations and employ targeted therapy in patients with lung adenocarcinomas: The Lung Cancer Mutation Consortium (LCMC)", Journal of Clinical Oncology, vol. 31, **(2013)** May 20.

[28] L. J. Lancashire, D. G. Powe, J. S. Reis, E. Rakha, C. Lemetre and B. Weigelt, "A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks", Breast Cancer Research and Treatment, vol. 120, **(2010)** February, pp. 83-93.

[29] F. S. Lichtenegger, M. Krempasky, K. Spiekermann, J. Braess, W. Hiddemann and M. Subklewe, "Costimulatory Expression Profiling of AML Blasts Identifies Surface Markers with High Correlation to Isolated NPM1 Mutation", Blood, vol. 118, **(2011)** November, pp. 1075-1075.

[30] L. E. MacConaill, C. D. Campbell, S. M. Kehoe, A. J. Bass, C. Hatton and L. L. Niu, "Profiling Critical Cancer Gene Mutations in Clinical Tumor Samples", Plos One, vol. 4, **(2009)** November 18.

[31] H. Shi, T. Tao, D. Huang, Z. Ou, J. Chen and J. Peng, "A naturally occurring 4-bp deletion in the intron 4 of p53 creates a spectrum of novel p53 isoforms with anti-apoptosis function", Nucleic Acids Res, vol. 43, **(2015)** January, pp. 1035-43.

[32] T. P. Shuai and X. D. Hu, "Connected set cover problem and its applications", Algorithmic Aspects in Information and Management, Proceedings, vol. 4041, **(2006)**, pp. 243-254.

[33] K. Suda, K. Tomizawa and T. Mitsudomi, "Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation", Cancer and Metastasis Reviews, vol. 29, **(2010)** March, pp. 49-60.

[34] Y. L. Yin, C. E. Soteros, and M. G. Bickis, "A clarifying comparison of methods for controlling the false discovery rate", Journal of Statistical Planning and Inference, vol. 139, **(2009)** July 1, pp. 2126-2137.

[35] W. Zhang, W. L. Wu, W. Lee and D. Z. Du, "Complexity and approximation of the connected set-cover problem", Journal of Global Optimization, vol. 53, **(2012)** July, pp. 563-572.

[36] Z. Zhang, X. F. Gao, and W. L. Wu, "Algorithms for connected set cover problem and fault-tolerant connected set cover problem", Theoretical Computer Science, vol. 410, **(2009)** March 1, pp. 812-817.

[37] S. Zou, Y. M. Xu, F. Zou, X. Q. Jiang, H. W. Deng and Y. Yu, "A Distributed Approximate Algorithm for Minimal Connected Cover Set Problem in Sensor Networks", 2009 Wri International Conference on Communications and Mobile Computing: Cmc 2009, vol. I, **(2009)**, pp. 556-562.