A Novel Approach for Detecting Driver Mutated Pathways in Glioblastoma Multiform

Yassine EL Kati[1], Shu-Lin Wang[1], and Fouad Kharroubi[2]
[1]School of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, Hunan, China
[2]Ecole Nationale des Sciences Appliqu ées, ENSAJ, LTI Lab, Chouaib Doukkali University, El Jadida, 24002, Morocco
Email: elkati.yassine@outlook.com, {smartforesting, fouad.kharroubi}@gmail.com

***Abstract*—Quite recently, considerable attention has been paid to finding the distinction between driver mutations that lead to tumorigenesis and passenger mutations that are neutral and do not play any role in the cancer proliferation. The main objective of this work is to come up with a new method to solve "The Maximum Weight Submatrix Problem". To that end, we introduce a new constraint named "approximate exclusivity" that helps to determine precisely the number of mutations that each patient has in the pathway. Depending on this constraint, we present a novel algorithm that detects driver mutated pathways based on an exact approach. We describe the details about our algorithm, then we compare the results with a Genetic Algorithm and a Binary Linear Programming model in both simulated and genetic data. Our exact algorithm has shown a good performance in terms of maximizing the weight and detecting all the possible driver pathways.**

***Index Terms*—driver pathways, maximum weight submatrix problem, exact algorithm, genetic algorithm, binary linear programming, glioblastoma multiform**

I. Introduction

Cancer stands for a group of more than 100 diseases caused by genetic changes, such as somatic mutations in DNA. In such wise, when cells make a copy of themselves during the cell division, some mutations occur. Analyzing these mutations in order to tell apart neutral mutations from the mutations that lead to cancer propagation has become a challenging task.

Previous studies indicate that driver mutation lead to tumorigenesis, while passenger mutations are neutral and do not play any role in the cancer proliferation. To that end, it is very effective to test the biological function of the mutation in order to decide whether it is a driver or a passenger mutation. The literature on testing the biological function shows a variety of techniques and methods that have been developed. We can list frequency-based methods [1], [2], methods that require a prior knowledge about pathways [3], [4] and methods that find mutated genes and pathways without any prior knowledge of pathways or other interaction between genes [5], [6].

Quite recently, there has been a growing interest in algorithms that do not require any prior knowledge, since we cannot get all the information about pathways and interactions between genes. The most interesting approach to this issue has been proposed by Vandin *et al.* [5] and was named "The Maximum weight Submatrix Problem". It is about maximizing a scoring function that combines two properties: high coverage and mutual exclusivity [6].

To solve the "Maximum weight Submatrix Problem", many researchers have proposed various methods in this area. We count, Dendrix [7], some random search based algorithms [8]-[11] and a method based on gene networks construction [12].

In this paper, we introduce a new constraint that helps to determine precisely the number of mutations that each patient has in the pathway. We also introduce an algorithm that we have designed based on an exact approach in order to detect all the possible driver pathways. In order to emphasize the performance of our algorithm, we compared the results with both the BLP and the GA both cited in [6].

II. ProblemDescription

Vandin *et al.* [5] indicated that "The Maximum Weight Submatrix Problem" is a very effective approach for detecting driver pathways without any prior knowledge. It plays a vital role in detecting new pathways that have not been detected by other methods and that can be a real case of study for future works [13].

The "Maximum Weight Submatrix Problem" takes into consideration two constraints. "High coverage" which is about identifying the number of patients with at least one mutation in the group of genes, and "high exclusivity" that implies finding a group of genes where each patient has at most one mutation in the pathway.

By constructing a binary mutation matrix A(m,n), based on somatic mutation data with m rows (patients) and n columns (genes), the Maximum Weight Submatrix Problem is about finding a submatrix M(m,k) with m

Manuscript received August 13, 2018; revised December 2, 2018.

doi: 10.18178/ijpmbs.8.1.17-21

rows and k columns from the mutation matrix A(m,n) by maximizing the fitness function:

$$W(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$$

where $|\Gamma(g)| = \{i = Aij = 1\}$ denotes the set of patients with a mutation in gene g.

$|\Gamma(M)|$ Indicates all thesets of patients that have mutations in the set M of genes.

III. Our Proposed Method

A. Description

Even though the efficiency of detecting new driver pathways using the scoring functions has improved in recent years, most improvements have been achieved by maximizing the "Maximum Weight Submatrix Problem". Nevertheless, it is possible to further improve the efficiency by coming up with a new method based on an exact approach. With this goal, this work seeks to define a new constraint that we have named "approximate exclusivity " and by developing an exact approach that can produce all the possible driver pathways with the maximum weight.

Fig. 1 depicts a mutation matrix that we have simulated at random with 10 patients (rows) and 11 genes (columns). When k stands as 2, the group of genes {i,j} illustrated by the submatrix B respects the constraint of high coverage and it is mutually exclusive. The weight of the submatrix B is W(B) = 8.

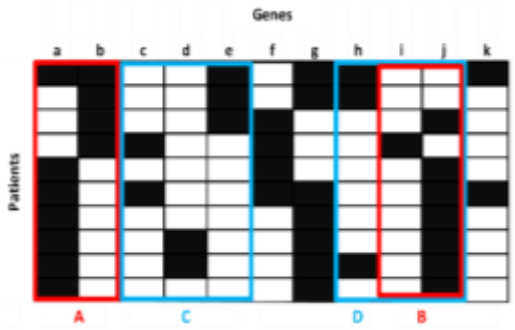

Figure 1. Illustration of the binary mutation matrix A with 10 patients and 11 genes.

When k is 3, the set of genes {c, d, e} which is represented by the submatrix C is also high coverage and mutually exclusive with a weight W(C) = 7. It is clearly noticeable that for both k = 2 and k = 3, the submatrices B and C has the greater weights W(B) and W(C).

However, for k = 2, the submatrix A is approximately exclusive with a weight W(A) = 9. Also when k represents 3, the submatrix D shows an approximate exclusivity gene set with a weight of W(D) = 9.

We conclude that, regardless the value of k, the gene sets with approximate exclusivity have always a weight greater than those with mutual exclusivity. In this direction, we can notice that it is needful to define a new constraint that we call "approximate exclusivity degree" or "co-occurrence degree" α which assists in finding exactly mutation number that each patient has in the pathway. Note that if $\alpha = 1$, the gene set is mutually exclusive.

B. ExactAlgorithm

The main objective of this work is to find groups of pathways with maximum weight and appropriate co-occurrence degree. For this purpose, we implemented an algorithm that filters all submatrices given the value of k and the co-occurrence degree α.This co-occurrence degree is subject to the following constraint: $1 \leq \alpha \leq k$. If $\alpha = 1$the submatrix M is mutually exclusive, if$\alpha \neq 1$ the submatrix is approximatively exclusive. Table I illustrates the steps of our exact algorithm.

TABLE I. Illustration of the Eight Steps of the Exact Algorithm.

Exact Algorithm Detecting submatrices with the maximum weight
Input: The mutation matrix A(m,n). The size of the gene set k. The co-occurrence degreeα. **Output:** All the submatrices with maximum weight and respect the co-occurrence degree α. 1. Remove the genes with a frequency of mutation less than 5%; 2. Make all the combination of genes to get the submatrixM(m,k). 3. Calculate the number of ones per row N1 for every submatrixM(m,k). 4. Delete the submatrices with N1 less than α in M(m,k). 5. Count how many rows are not all zeros $\Gamma(m)$; 6. Count the number of ones in all the submatrix$\sum_{g \in M} |\Gamma(m)|$; 7. Calculate the weight W of every submatrix M2 using the formula $2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$; 8. Return all the submatrices with the highest weight W;

In the first step, we removed the genes with the frequency of mutation is less than 5% because genes altered in only one or few cancer patients may not be driver mutations and possibly could be passenger ones [12]. In the second step, we made all the possible combinations of genes according to the k value. After that, we proceeded by deleting the matrices that do not fit with the co-occurrence degree. Then for the next steps, it is about calculating the weight and producing set of genes with the highest weight.

IV. Experimental Results and Comparison

We ran our experiments on a 2.4GHz i7-5500U CPU PC, and we applied it on both simulated data and biological data, then we compared the results with the BLP model and the GA, both [6].

A. Simulated Data

We generated mutation data for m=100 patients and n=(100,200,300,400,500) genes. Our experiments have shown that the major drawback of the GA is that it shows an instability in maximizing the weight, this clearly noticeable from Fig. 1. However, our Exact Algorithm and the GA always produce the same weight.

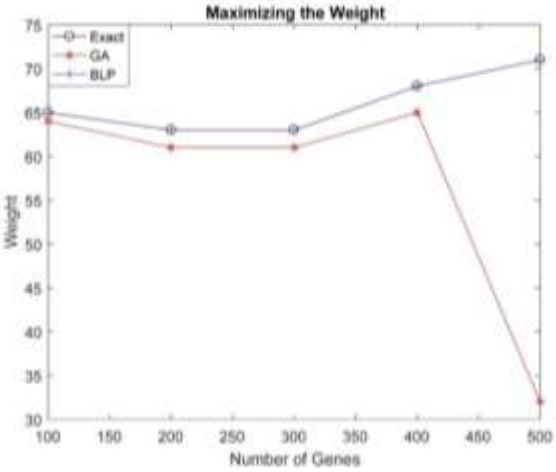

Figure 2. Comparison of the weights obtained by the three methods.

In this scenario, the x-axis represents the number of genes and the y-axis is the weight. The red line denotes the weight of GA, the black line with circles represents our Exact Algorithm, and the blue line with '+' denotes BLP.

B. Biological Data

In this study, in order to assess the performance of our algorithm. We applied the BLP model, the GA and our Exact Algorithm tothe Glioblastoma multiform data 1, 2 and 3obtained from Zhao *et al.* [6]. Glioblastoma data 1 and 3 are two mutation matrices of 84 patients and 178 genes while Glioblastoma data 2 is a mutation matrix of 90 patients and 1126 genes.

1) Glioblastoma Multiform Data 1

When k is 2, the three algorithms produce (CDKN2B, CYP27B1) as optimal gene set. The weight of this gene set is 54 and it is mutated in 57 samples. CDKN2B is known to be the core member of the cell cycle and the p53 signaling pathway [14]. Furthermore, Wu H *et al.* [12] have shown that CYP27B1 is also a member of the Glioblastoma copy number up.

When k stands as 3, (CDKN2B, CYP27B1, RB1) was the unique gene set produced by the algorithms. This gene set is important, since it is altered in 79% of the patients.

For k=4, only 5 gene sets were produced by the GA while our exact algorithm has detected 6. As emphasized in Figure 3 (CDK4, CDKN2B, ERBB2, RB1) is the significant gene set missed by the BLP and the GA. This set of genes covers 83% of the diagnosed patients. In addition, CDK4, CDKN2B and RB1 are the core members of the cell cycle part of the p53 signaling pathway [15]-[17], while ERBB2 is the member of the Glioblastoma copy number up [12].

The major drawback of the GA and the BLP model is that, when the size of the gene set increases, their ability of sampling new maximum weighted gene sets decreases significantly. This is clearly noticeable when k is 5. Our Exact Algorithm has detected 31 optimal gene set while the BLP has detected one and the GA has detected only 10 optimal solutions. The gene sets were sampled with a co-occurrence degree of 3 and a weight of 60. Furthermore, ERBB2 is a part of the p53 signaling pathway and the genes CDKN2B and RB1 are core members of the cell cycle and ERBB2. These three genes were universally detected in every optimal gene set sampled by our Exact Algorithm.

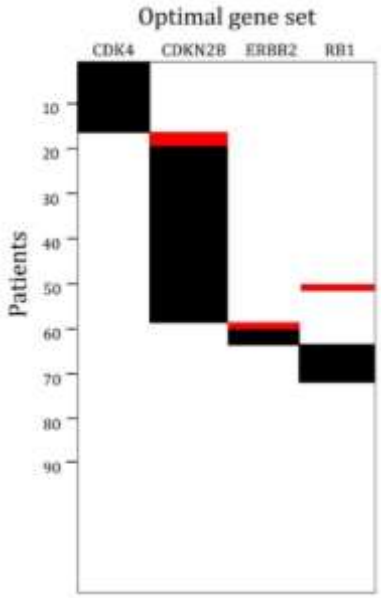

Figure 3. Submatrix of the optimal gene produced by our exact algorithm.

We show the mutation characteristics between patients and genes: (black) exclusive mutation; (red) co-occurring mutation; (white) no mutation.

2) Glioblastoma Multiform Data 2

When k represents 2, the three methods have produced three identical gene sets with a weight of 58. We have noticed that the genes CDK4, CDKN2B, CDKN2A, CDKN2B and TP53 are dominant in most of the gene sets sampled. These genes are the core members of the cell cycle. However, Zhang *et al.* [6] have proven that until now, no relationship has been proven between TSPAN31 and CDKN2B.

When k is 3, 2 optimal gene sets have been sampled. These sets of genes cover 73 % of the diagnosed patients. The pair of genes (CDK4, CDKN2B) that we detected whenk was 2 contains RB1. These sets were detected as a potential pathways and were a subset of the set (CDK4, CDKN2B, ERBB2, RB1) in Glioblastoma Data 1. CDK4, CDKN2B and RB1 are all the core members of the cell cycle [13].

Table II reveal that when k is set as 4 and $\alpha = 3$, our Exact Algorithm samples 5 maximum weighted gene sets with 59 as the maximum weight. All the genes that form the quadruplet (CDK4, CDKN2A, RB1, TP53) are the core members of the Glioma which is a tumor that starts in the brain and spinal cord. In addition, three gene sets have the triplet (CDK4, CDKN2B, RB1) as a core subset with the genes CPT1B, NF1 and PIK3R1 each one separately. As for CPT1B, it is a part of the AMPK signaling pathway. PIK3R1 is also a core member of the glioma while NF1 is a part of the MAPK signaling pathway. The last quadruplet (CDKN2A, MDM2,

PIK3R1, TP53) encloses the gene MDM2 which is also a core member of the glioma and the cell cycle. On top of that, it is also it is a part of different signaling pathways such as the p53 signaling pathway [13].

TABLE II. GLIOBLASTOMA DATA 2 RESULTS WHEN K IS 2, 3 AND 4

Optimal gene sets	Set size	Weight	Coverage
CDK4, CDKN2B	2	58	67%
CDKN2A, TP53	2	58	71%
CDKN2B, TSPAN31	2	58	67%
CDK4, CDKN2B, RB1	3	63	73%
CDKN2B, RB1, TSPAN31	3	63	73%
CDK4, CDKN2B, MAN1A1, RB1	4	60	74%
CDK4, CDKN2B, MET, RB1	4	60	74%
CDK4, CDKN2B, NMBR, RB1	4	60	74%
CDKN2B, MAN1A1, RB1, TSPAN31	4	60	74%
CDKN2B, MET, RB1, TSPAN31	4	60	74%
CDKN2B, NMBR, RB1, TSPAN31	4	60	74%

It is important to say that, Fork=5 the gene set (CDKN2A, MDM2, MDM4, PIK3R1, TP53) was sampled by our Exact Algorithm. This gene covers 80% of the patients, and it is composed by the quadruplet (CDKN2A, MDM2, PIK3R1, TP53), that was previously sampled when k was 4. MDM4 as well as MDM2, are both parts of the p53 signaling pathway.

3) Glioblastoma Multiform Data 3

For k = 2, the three algorithms have produced (EGFR, NF1) as optimal gene set with aweight of 50. This set of genes covers 61% of the diagnosed patients. The detection of EGFR is important because this gene serves as a stimulus for cancer growth and it plays an important role in the regulation of cellular homeostasis. It is also a core member of different signaling pathways such as ErbB signaling pathway, FoxO signaling pathway, and MAPK signaling Pathway while NF1 is a core member of the Ras signaling pathway and the MAPK signaling pathway [13].

As Table III reveals that4 triplets of genes have been produced with a weight of 54 and a co-occurrence degree $\alpha = 2$. Two triplets cover 80% of the diagnosed patients while the remaining sets cover 70%. These 4 gene sets are formed with the genes MTAP, NF1, TSFM, TSPAN31 and PTEN.

TABLE III. GLIOBLASTOMA DATA 3WHEN K IS 2, 3 AND 4

Optimal gene sets	Set size	Weight	Coverage
EGFR, NF1	2	50	61%
MTAP, NF1, TSFM	3	54	70%
MTAP, NF1, TSPAN31	3	54	70%
MTAP, PTEN, TSFM	3	54	80%
MTAP, PTEN, TSPAN31	3	54	80%
DOCK1, GLI1, MTAP, PTEN	4	55	77%

When k is 4, the three algorithms have detected only one optimal gene set (DOCK1, GLI1, MTAP, PTEN). This quadruplet covers 77 % of the samples. DOCK1 is a significant gene and plays an important role in cell proliferation and gene expression. GLI1, it is a core member of the Hedgehog signaling pathway, it helps to control cell proliferation and stem cell maintenance and development. MTAP is known to play an important role in Cysteine and methionine metabolism. Several studies has shown that the mutation of PTEN has a relationship with several types of cancers including Breast cancer, Endometrial cancer, Lung cancer and Prostate cancer. Concerning Glioblastoma multiform, PTEN also plays an important role in the glioma cell proliferation.

When k stands as 5, our Exact Algorithm detected solution (DOCK1, GLI1, KDR, MTAP, PTEN) as optimal solution with a co-occurrence degree $\alpha = 3$. This gene set was sampled with a weight of 54 and covers 67 patients. This gene set is formed by KDR and the set of genes (DOCK1, GLI1, MTAP, PTEN). This quadrupletwas previously detected when k = 4. KDR is a part of the Ras signaling pathway, the PI3K-Akt signaling pathway, the Rap1 signaling pathway, and even others. In addition, the same as EGFR and PTEN, KDR contributes in cellular homeostasis [13].

V. CONCLUSION

The main concern of the paper was to propose a new approach in identifying driver genes and driver pathways that can be readily used in practice to contribute in determining the potential causes of the glioblastoma multiform cancer particularly and the human cancer generally, and then, naturally assist in designing cancer treatments.

By conducting some experiments, we have proven that whatever the k value is, the weight of the mutually exclusive gene sets is always less than those with approximate exclusivity. Based on that, we have shown that if we want to identify multiple driver pathways and find gene sets whose weights are maximized, it is necessary to define a new constraint; we called it "approximate exclusivity". It assists in finding groups of genes respecting the co-occurrence degree (approximate exclusivity degree) α that determines the number of mutations, which each patient has in the pathway. Then depending on this constraint, we designed an algorithm based on an exact approach in order to solve the maximum weight submatrix problem.

To assess the effectiveness of our method, we compared the results of our algorithm with the GA and the BLP model [6]. In both simulated data and glioblastoma multiform data, our method has shown a great ability in sampling all the possible driver pathways and finding group of genes whose weights are maximized. A key limitation of the BLP and the GA is that, when the size of the gene set raises, they cannot guarantee the optimal results in terms of maximizing the weight and sampling all the possible driver pathways. This was clearly noticeable in Glioblastoma Data 1. When the size of the gene set was 5, our exact algorithm produced 31

optimal results while the BLP and GA have sampled only 10. To that effect, more tests and experiments will be needed in order to elucidate if there is any relationship between some of the genes sampled and the human cancer.

ACKNOWLEDGMENTS

The authors would like to acknowledge the valuable grants of the National Science Foundation of China (Grant Nos. 61472467 and 61672011) and the Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province.

REFERENCES

[1] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, *et al.*, "The consensus coding sequences of human breast and colorectal cancers," *Science*, vol. 314, pp. 268-274, 2006.
[2] F. Vandin, E. Upfal, and B. J. Raphael, "Finding driver pathways in cancer: Models and algorithms," *Algorithm MolBiol.*, vol. 7, 2012.
[3] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *J. ComputBiol.*, vol. 18, pp. 507-522, 2011.
[4] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Res.*, vol. 22, pp. 398-406, 2012.
[5] F. Vandin, E. Upfal, B. J. Raphael, "Algorithms and genome sequencing: Identifying driver pathways in cancer," *Computer*, vol. 45, pp. 39-46, 2012.
[6] J. F. Zhao, S. H. Zhang, L. Y. Wu, and X. S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, pp. 2940-2947, 2012.
[7] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Res.*, vol. 22, pp. 375-385, 2012.
[8] C. Yan, H. T. Li, A. X. Guo, W. Sha, and C. H. Zheng, "Simulated annealing based algorithm for mutated driver pathways detecting," Lect. Notes ArtifInt., vol. 8589, pp. 658-663, 2014.
[9] F. Kharroubi, J. He, J. Tang, M. Chen, and L. Chen, "Evaluation performance of genetic algorithm and tabu search algorithm for solving the Max-RWA problem in all-optical networks," *J. Comb. Optim.*, vol. 30, pp. 1042-1061, 2015.
[10] F. Kharroubi, J. He, and L. Chen, "Performance analysis of GA, ROA, and TSA for solving the max-RWA problem in optical networks," in *Optical Fiber Communication Conference*, San Francisco, 2014, paper W2A.48.
[11] S. L. Wang and Y. Y. Tan, "Dynamically heuristic method for identifying mutated driver pathways in cancer," *Lect. Notes ComputSc.*, vol. 9771, pp. 366-376, 2016.
[12] H. Wu, L. Gao, F. Li, F. Song, X. F. Yang, and N. Kasabov, "Identifying overlapping mutated driver pathways by constructing gene networks in cancer," *Bmc. Bioinformatics*, vol. 16, 2015.
[13] Y. El Kati, S. L. Wang, F. Kharroubi, and Y. Tan, "An efficient exact method for identifying mutated driver pathways in cancer," *J. Appl. Bioinforma. Comput. Biol.*, vol. 6, 2017.
[14] M. Wrensch, R. B. Jenkins, J. S. Chang, R. F. Yeh, Y. Xiao, P. A. Decker, *et al.*, "Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility," *Nat. Genet.*, vol. 41, pp. 905-908, 2009.
[15] C. Ling; B. L. Carlson, M. A. Schroeder, J. L. Ostrem, G. J. Kitange, A. C. Mladek, *et al.*, "p16-Cdk4-Rb axis controls sensitivity to a cyclin-dependent kinase inhibitor PD0332991 in glioblastomaxeno graft cells," *Neuro-Oncology*, vol. 14, pp. 870-881, 2012.
[16] J. Dean, C. Thangavel, A. McClendon, C. Reed, and E. Knudsen, "Therapeutic CDK4/6 inhibition in breast cancer: Key mechanisms of response and failure," *Oncogene*, vol. 29, pp. 4018-4032, 2010.
[17] W. R. Wiedemeyer, I. F. Dunn, S. N. Quayle, J. H. Zhang, M. G. Chheda, G. P. Dunn, *et al.*, "Pattern of retinoblastoma pathway inactivation dictates response to CDK4/6 inhibition in GBM," in *Proc. Natl. Acad. Sci. USA*, 2010, pp. 11501-11506.

EL KATI YASSINE Born in Morocco on December the 22nd 1991, El Kati Yassine has studied Mathematics, Physics and Engineering Sciences in the Preparatory Classes, Mohammedia, Morocco in 2012. After that, he continued his engineering studies at the Higher School of Textile and Clothing Industries (ESITH) where he received his State Engineer diploma majored in Industrial Engineering, Casablanca, Morocco in 2015.
On August 2018, he moved to China to continue his studies at Hunan University. He received his master degree majored in Computer Science and Technology, after completing three years of research in the field of Bioinformatics, precisely in detecting mutated driver pathways that are responsible of cancer proliferation, Changsha, People's Republic of China in 2018.

Shu-Lin Wang was born in Sichuan province, China. Currently, he is working in Hunan University as a professor. He received his PhD degree in Computer Science and Technology from the National University of Defense Technology, China, in 2008 (Advisor: Prof. Huowang Chen and Prof. Ji Wang). He also received his MSc degree in Computer Application from the National University of Defense Technology, China, in 1997, and obtained his BSc degree in Computer Application from China University of Geosciences in 1989. From 2008-1010 he was a postdoctoral student at the Institute of Hefei Intelligent Machines, Chinese Academy of Sciences, and from 2012-2013 he also was a postdoctoral student at Applied Bioinformatics Laboratory in Kansas University, USA. His current research interest includes Bioinformatics, Artificial intelligence, and Complex System.

Foaud Kharroubi was born in Morocco, on February 3rd 1985. He received a B.S. degree in Computer Science and Mathematics Sciences from Ibno Zohr University, Agadir, Morocco in 2006. He obtained a M.Sc. degree with honors in Computer Science from the Graduate School of Computer Science and Systems Analysis (ENSIAS) Mohammed V University of Souissi, Rabat, Morocco, in 2008. In 2014, he received his PhD degree with honors in Computer Science and Communication from Hunan University, Changsha, China.
He is working now as an Assistant Professor in the Department of Telecommunication, Networks and Computer Ccience (TRI) at Ecole National des Sciences Appliquees (ENSAJ) of Chouaib Doukkali University at El Jadida, Morocco. Prior to this, he was an Assistant Professor at the universities of CSUFT and Yichun at the People's Republic of China. He has authored and co-authored several conference and journal papers. His main research interests include optimization of Things in the fields of: Bioinformatics, Military, Optical Fiber and Marine. Fouad Kharroubi is a member of the Optical Society of America (OSA).