A genetic algorithm for pancreatic cancer diagnosis

Charalampos Moschopoulos[1,2], Dusan Popovic[1,2], Alejandro Sifrim[1,2], Grigorios Beligiannis[3], Bart De Moor[1,2] and Yves Moreau[1,2]

[1] Department of Electrical Engineering-ESAT, SCD-SISTA, KU Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium.
[2] iMinds Future Health Department, KU Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium.
[3] Department of Business Administration of Food and Agricultural Enterprises, University of Western Greece, G. Seferi 2, GR-30100 Agrinio, Greece.
{Charalampos.Moschopoulos, Dusan.Popovic, Alejandro.Sifrim, Bart.DeMoor, Yves.Moreau}@esat.kuleuven.be, gbeligia@uwg.gr

Abstract. Pancreatic cancer is one of the leading causes of cancer-related death in the industrialized countries and it has the least favorable prognosis among various cancer types. In this study we aim to facilitate early detection of the pancreatic cancer by finding minimal set of genetic biomarkers that can be used for establishing diagnosis. We propose a genetic algorithm and we test it on gene expression data of 36 pancreatic ductal adenocarcinoma tumors and matching normal pancreatic tissue samples. Our results show that a minimum group of genes are able to constitute a high reliability pancreatic cancer predictor.

Keywords: genetic algorithm, support vector machines, pancreatic cancer, biomarkers, pancreatic ductal adenocarcinoma, microarrays.

1 Introduction

In present, pancreatic cancer is considered as one of the most lethal of common cancer types [1]. At this moment, pancreatic cancer holds the eighth most common cause of cancer related deaths with survival rate of less than 5%, five years after the diagnosis. Its lethality is largely due to the fact that is diagnosed at a later stage, which significantly decreases the chance of patient's survival. The most common type of pancreatic cancer, accounting for 95% of these tumors, is adenocarcinoma or PDAC. An additional problem is that PDAC has an extremely poor prognosis, as it seems that pancreas emits few clues to signal the carcinogenic process.

During the last years, several research teams have tried to detect molecular markers that facilitate early detection of the disease so that appropriate treatment could be applied timely [2-3]. There are several cases where such a set of biomarkers has been proposed, but none of them has been shown to be robust enough to constitute a diagnosis classifier [4]. Additionally, given the biomarker discovery problem, it is

rather difficult to extract knowledge from high throughput data as it suffers from curse of dimensionality and high level of noise [5].

In this contribution we apply a genetic algorithm on an publically available gene expression dataset (from GEO database [6] by [7]) and we try to obtain the minimum set of biomarkers that can be used for detection of pancreatic cancer. Due to the non-deterministic nature of the genetic algorithms, we performed extensive experiments and we show that the each time selection of biomarkers produces a classifier with a relatively robust performance. We also formed a list out of the most "popular" genes presented in the final results of each run and show the biological relevance of them for pancreatic cancer. Finally, we provide numerous performance metrics such as F-measure and accuracy. For these we obtain high values for almost all runs proving the efficiency of the proposed method.

The remaining of the paper is organized as follows: section 2 introduces the proposed genetic algorithm relative to chromosome encoding, initialization procedure, fitness function and genetic operators. Also, more information about the dataset used in our experiments is given. Section 3 presents our results and gives a overview of the biological significance of the biomarkers found. Finally, in the last section, we present our conclusions and directions for future work.

2 Methods

2.1 Microarray gene expression data for pancreatic cancer

The dataset used in our experiments contains pairs of normal and tumor tissue samples which were obtained at the time of surgery from resected pancreas of 36 pancreatic cancer patients [8]. All the patients were suffering from PDAC. Gene expressions were obtained using Affymetrix U133 plus 2.0 whole genome microarrays. Also, 6 control samples (3 normal and 3 tumor) were present in dataset, which were used to test the quality of the rest obtained samples. In total, this dataset includes 19898 genes and 78 genechip hybridizations have been performed. This dataset is freely available at GEO database where the reader can find more information (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471).

2.2 The proposed Genetic Algorithm (GA)

The efficiency of GAs has been proved by successful applications in many different scientific fields, including bioinformatics [9, 10], where they surpassed other algorithmic strategies for optimization and search. The GAs are stochastic algorithms that simulate the process of natural evolution. Based on this model, all GAs use three simple operators, which allows them to evolve and reach near optimal solutions [11, 12]. These are the Selection, Crossover and Mutation operators.

Initially, we randomly divided the expression dataset to 70% training (53 samples) and 30% test (25 samples) data. For the chromosome encoding we chose a one

dimensional binary array representation where each position corresponds to biological gene (Figure 1) to be included as a biomarker. We chose this representation because it gave us the opportunity to perform experiments with different mutation and crossover operators. In the end, as there were no significant changes in the performance of the GA, we chose to use the single point crossover and uniform mutation.

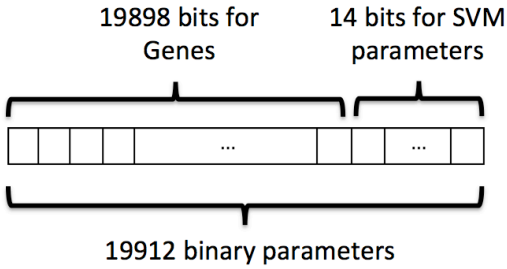

Fig. 1. Chromosome Encoding.

As a selection operator we chose to use a stochastic universal [13]. In the fitness function, which is the heart of each GA, we tried to achieve balance between the F-measure of a classifier and the size of the corresponding solution, penalizing solutions that contain many genes. Following is the formal expression of the fitness function used to evaluate each chromosome:

$$fitness\_function = 1 + F\_measure - \frac{size\_of\_chromosome}{mean\_size\_of\_initial\_population}$$

where the *size_of_chromosome* is the number of genes that are used as biomarkers in the chromosome and *mean_size_of_initial_population* is the mean number of genes that are used as biomarkers in the initial population of the GA, across all chromosomes. This particular form of fitness function pushes GA to select solutions that contain a small number of genes and are as accurate as possible. We experimented with different weights in order to boost one metric over the other but we obtained similar results, just with slower convergence. The F-measure metric has been preferred as it is geometric average of the precision and recall, both being of great interest when designing diagnostics tests. The efficiency of this strategy is clearly shown in the next section were our experimental results are presented in more detail. To classify samples given biomarkers selected in each chromosome of the population, we used a Support Vector Machine (SVM) with Gaussian Radial Basis Function (RBF) kernel function [14]. This also adds two parameters in the system that has to be tuned: the sigma and the penalty parameter c. These two parameters were represented, as the additional chromosome for each individual solution, constituted of 7 bits per each parameter, which were also tuned as the by-product of GA evolution.

The structure and operation of the proposed algorithm is presented in Figure 2. In the initialization procedure, a random population was created, where the 10% of the bits that represents the gene panel and 50% of the bits that represents the SVM parameters were set to 1. This way, an initial generated chromosome contains around 200 genes in average. We reassured that the individuals that constitute the first

generation were spread in the whole search space of possible solutions in order to avoid local optima solutions. In the next step, each chromosome is evaluated by the fitness function. Then, if the maximum number of generations is not reached, the three operators (selection, crossover and mutation) were applied consequently to create the new population. When the maximum number of generations is reached, the GA outputs the best solution generated throughout its execution. In all the performed experiments, the probability of was 0.7, while that of mutation was equal to 10^{-4} per bit, for all bits and all chromosomes. This means that in average two bits per chromosome were changing during the mutation. The size of the population has been set to 500 and the maximal number of generations to 500 (the number of generations was used as termination criterion).

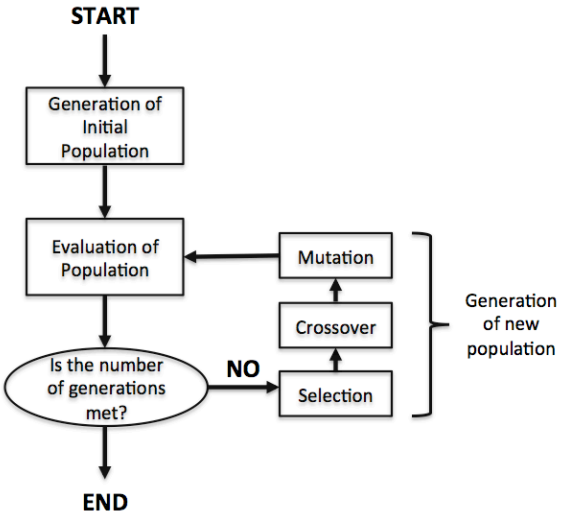

Fig. 2. Structure and operation of the proposed GA.

3 Results and Discussion

As the GAs are essentially stochastic, each run produced slightly different list of biomarkers. However, after performing 100 runs of the algorithm using the same operators and parameters, all the generated results had similar performance, achieving high values for performance metrics. Additionally, the number of genes that were selected in each run was approximately stable (around 16 with a deviation of 2). These values of the GA's parameters were selected by trial-and-error method after performing additional experiments.

The convergence of our algorithm and the reduction of the number of genes taking part in the best chromosome of each generation were similar in all our experiments, following the pattern presented in Figure 3. It is clear that the performance of the algorithm stabilizes approximately after 200 generations, and it remains literally

unchanged approximately after the 300th generation. After the training process, we use the best chromosome produced to classify the test data using a SVM classifier. Note that during the training process, the parameters of a SVM the also optimized.

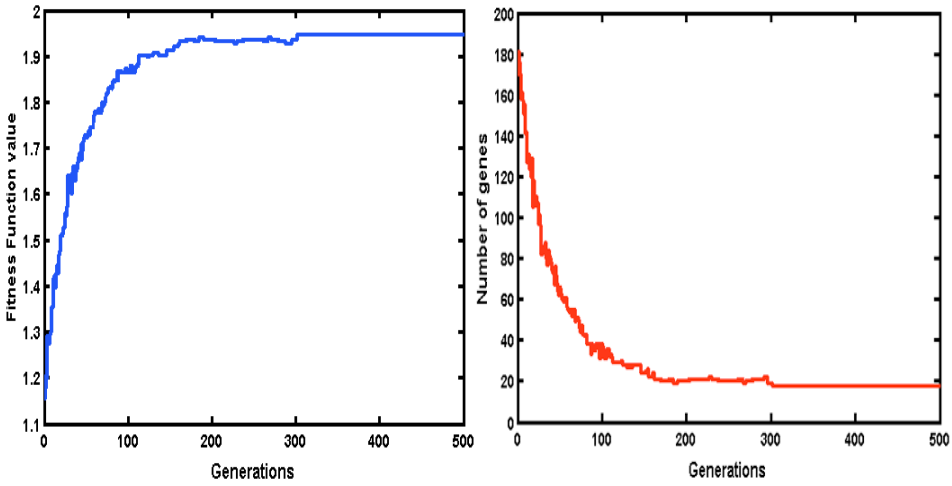

Fig. 3. Fitness function value and size of best chromosome during GA evolution. Most of our experiments followed the same pattern.

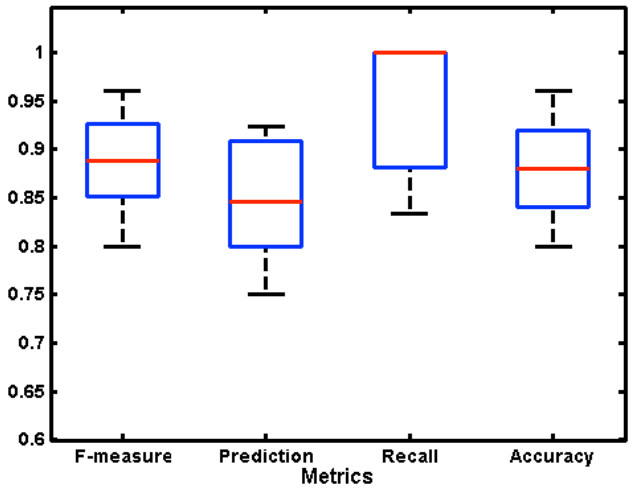

Fig. 4. Box plot presenting the variance of the evaluation metrics.

In order to evaluate the classification results, we used well-established metrics for classification such as F-measure, Prediction, Recall and Accuracy [15]. As it can be seen in Figure 4, the GA algorithm achieves on average 88% for F-measure and Accuracy metrics.

Even the list of biomarkers generated by the GA varies for every execution of the algorithm, one could expect that the genes that are somehow functionally related to pancreatic cancer should be present most of the times in results. To examine this, we counted genes that were present in solutions with corresponding fitness values higher than the average in the last generation of the GA. We repeated this procedure for every run of the algorithm. The 20 most "popular" genes are presented in Figure 5.

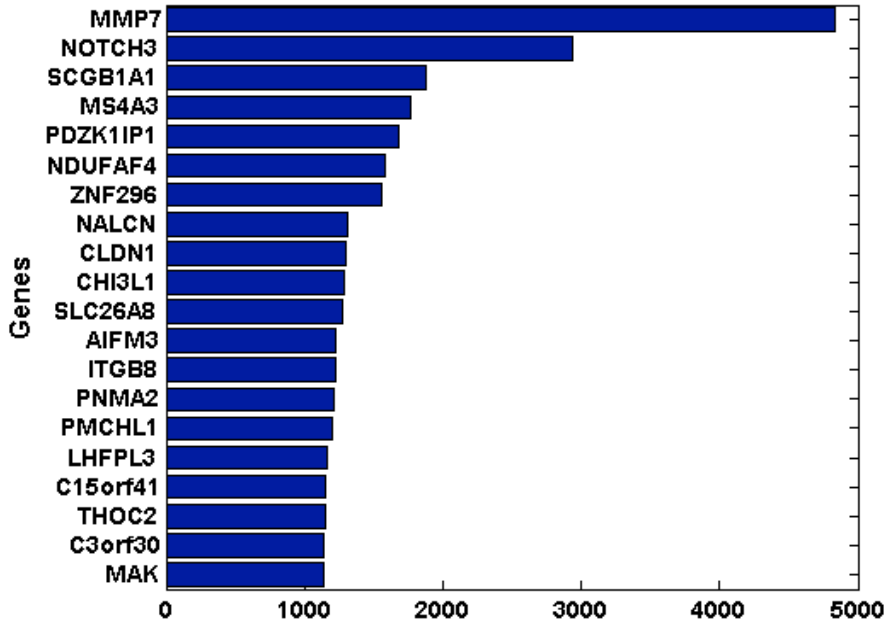

Fig. 5. The first 20 most popular genes in 100 different GAs runs.

We assessed the functional characteristics of the resulting set of genes using the Ingenuity Pathway Analysis suite. Out of the top 15 genes 7 genes were found to be associated with cancer (MMP7, NOTCH3, PDZK1IP1, NDUFAF4, CLDN1, NALCN, PNMA2). The top ranking gene, MMP7, has previously been reported to be overexpressed in pancreatic ductal adenocarcinomas and not in normal pancreatic tissue [16-17] and is believed to apply apoptotic pressure to epithelial cells. The second best ranking gene, NOTCH3, is part of the Notch signaling pathway, which is well studied in carcinogenesis of many different types of cancer [18-19]. NALCN, together with other genes involved in axon guidance, has recently been shown by Blankin et al. to show a significant higher number of aberrations in pancreatic cancer compared to control [20]. The remaining cancer-related genes (PDZK1IP1, NDUFAF4, CLDN1, PNMA2) were associated with non-pancreatic cancers or involved in apoptosis and/or cell proliferation pathways.

4 Conclusions and Future Work

In this manuscript a GA has been proposed, developed and applied on PDAC data to classify tissue samples. Our benchmark shows that the method results in robust classifiers for pancreatic cancer. In addition, the algorithm provides a list of biomarkers that play the most important role in this lethal disease as a by-product. However, due to stochastic nature of the GA, we decided to statistically measure the importance of the genes by measuring their appearances on "good" solutions generated by the GA in all our experiments, which is also suggested procedure if one should use the method for biomarker mining. Our results were verified by the biological significance of these genes in specific biological functions in human organism.

Our future plans include the optimization of the classification process of tissue samples by using the generated "popular" gene list as biomarkers and trying to optimize different kind of classifiers through a new genetic algorithm. Moreover, to further examine robustness of the method we are going to apply the resulting classifier on additional gene expression datasets on this type of cancer.

Acknowledgments. The authors would like to acknowledge support from:
Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc & fellow grants. Industrial Research fund (IOF): IOF/HB/10/039 Logic Insulin, IOF: HB/12/022 Endometriosis. Flemish Government: FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM, IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer, Hercules Stichting: Hercules III PacBio RS, iMinds: SBO 2013; Art&D Instance, IMEC: phd grant. Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate). COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network.

References

1. Jemal, A., Siegel, R., Xu, J., and Ward, E.: Cancer statistics, 2010. CA Cancer J Clin, 60(5): 277-300 (2010)
2. Brandt, R., Grutzmann, R., Bauer, A., Jesnowski, R., Ringel, J., Lohr, M., Pilarsky, C., and Hoheisel, J.D.: DNA microarray analysis of pancreatic malignancies. Pancreatology, 4(6): 587-597 (2004)
3. Bauer, A.S., Keller, A., Costello, E., Greenhalf, W., Bier, M., Borries, A., Beier, M., Neoptolemos, J., Buchler, M., Werner, J., Giese, N., and Hoheisel, J.D.: Diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis by measurement of microRNA abundance in blood and tissue. PLoS One, 7(4): e34151 (2012)
4. Bussom, S. and Saif, M.W.: Methods and rationale for the early detection of pancreatic cancer. Highlights from the "2010 ASCO Gastrointestinal Cancers Symposium". Orlando, FL, USA. January 22-24, 2010. JOP, 11(2): 128-130 (2010)
5. Phan, J.H., Moffitt, R.A., Stokes, T.H., Liu, J., Young, A.N., Nie, S., and Wang, M.D.: Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. Trends Biotechnol, 27(6): 350-358 (2009)

6. Edgar, R., Domrachev, M., and Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 30(1): 207-210 (2002)
7. Badea, L., Herlea, V., Dima, S.O., Dumitrascu, T., and Popescu, I.: Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepatogastroenterology, 55(88): 2016-2027 (2008)
8. Valentini, G., Tagliaferri, R., and Masulli, F.: Computational intelligence and machine learning in bioinformatics. Artif Intell Med, 45(2-3): 91-96 (2009)
9. Bandyopadhyay, S. and Pal, S.K.: Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence. Natural Computing Series: Springer. 311 (2007)
10. Bäck, T., Fogel, D.B., Michalewicz, Z., and Beck, T.: Evolutionary Computation 1: Basic Algorithms and OperatorsBristol: Institute of Physics Publishing (2000)
11. Bäck, T., Fogel, D.B., Michalewicz, Z., and Beck, T.: Evolutionary Computation 2: Advanced Algorithms and OperatorsBristol: Institute of Physics Publishing (2000)
12. Fishman, G.S.: Monte Carlo: Concepts, Algorithms, and Applications.New York: Springer (1995)
13. Baker, J.E.: Reducing Bias and Inefficiency in the Selection Algorithm. In 2nd International Conference on Genetic Algorithms and their Application, pp. 14–21, Cambridge, Massachusetts, USA (1987)
14. Parker, C., An Analysis of Performance Measures for Binary Classifiers, in IEEE 11th International Conference on Data Mining (ICDM)2011: Corvallis, OR, USA p. 517-526.
15. Xu, K., Cui, J., Olman, V., Yang, Q., Puett, D., and Xu, Y.: A comparative analysis of gene-expression data of multiple cancer types. PLoS One, 5(10): e13696 (2010)
16. Crawford, H.C., Scoggins, C.R., Washington, M.K., Matrisian, L.M., and Leach, S.D.: Matrix metalloproteinase-7 is expressed by pancreatic cancer precursors and regulates acinar-to-ductal metaplasia in exocrine pancreas. J Clin Invest, 109(11): 1437-1444 (2002)
17. Tan, X., Egami, H., Abe, M., Nozawa, F., Hirota, M., and Ogawa, M.: Involvement of MMP-7 in invasion of pancreatic cancer cells through activation of the EGFR mediated MEK-ERK signal transduction pathway. J Clin Pathol, 58(12): 1242-1248 (2005)
18. Doucas, H., Mann, C.D., Sutton, C.D., Garcea, G., Neal, C.P., Berry, D.P., and Manson, M.M.: Expression of nuclear Notch3 in pancreatic adenocarcinomas is associated with adverse clinical features, and correlates with the expression of STAT3 and phosphorylated Akt. J Surg Oncol, 97(1): 63-68 (2008)
19. Vo, K., Amarasinghe, B., Washington, K., Gonzalez, A., Berlin, J., and Dang, T.P.: Targeting notch pathway enhances rapamycin antitumor activity in pancreas cancers through PTEN phosphorylation. Mol Cancer, 10: 138 (2011)
20. Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J., Chang, D.K., Cowley, M.J., Gardiner, B.B., Song, S., Harliwong, I., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Gongora, M., Pajic, M., Scarlett, C.J., Gill, A.J., Pinho, A.V., Rooman, I., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, Q., Nones, K., Fink, J.L., Christ, A., Bruxner, T., Cloonan, N., Kolle, G., Newell, F., Pinese, M., Mead, R.S., Humphris, J.L., Kaplan, W., Jones, M.D., Colvin, E.K., Nagrial, A.M., Humphrey, E.S., Chou, A., Chin, V.T., Chantrill, L.A., Mawson, A., Samra, J.S., Kench, J.G., Lovell, J.A., Daly, R.J., Merrett, N.D., Toon, C., Epari, K., Nguyen, N.Q., Barbour, A., Zeps, N., Kakkar, N., Zhao, F., Wu, Y.Q., Wang, M., Muzny, D.M., Fisher, W.E., Brunicardi, F.C., Hodges, S.E., Reid, J.G., Drummond, J., Chang, K., Han, Y., Lewis, L.R., Dinh, H., Buhay, C.J., Beck, T., Timms, L., Sam, M., Begley, K., Brown, A., Pai, D., Panchal, A., Buchner, N., De Borja, R., Denroche, R.E., Yung, C.K., Serra, S., Onetto, N., Mukhopadhyay, D., Tsao, M.S., Shaw, P.A., Petersen, G.M., Gallinger, S., Hruban, R.H., Maitra, A., Iacobuzio-Donahue, C.A.,

Schulick, R.D., Wolfgang, C.L., Morgan, R.A., Lawlor, R.T., Capelli, P., Corbo, V., Scardoni, M., Tortora, G., Tempero, M.A., Mann, K.M., Jenkins, N.A., Perez-Mancera, P.A., Adams, D.J., Largaespada, D.A., Wessels, L.F., Rust, A.G., Stein, L.D., Tuveson, D.A., Copeland, N.G., Musgrove, E.A., Scarpa, A., Eshleman, J.R., Hudson, T.J., Sutherland, R.L., Wheeler, D.A., Pearson, J.V., McPherson, J.D., Gibbs, R.A. and Grimmond, S.M.: Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature, 491(7424): 399-405 (2012)