A Simple Genetic Algorithm for Biomarker Mining

Dusan Popovic, Alejandro Sifrim, Georgios A. Pavlopoulos,
Yves Moreau, and Bart De Moor

ESAT-SCD / IBBT-KU Leuven Future Health Department, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium
{Dusan.Popovic,Alejandro.Sifrim,Georgios.Pavlopoulos,
Yves.Moreau,Bart.DeMoor}@esat.kuleuven.be

Abstract. We present a method for prognostics biomarker mining based on a genetic algorithm with a novel fitness function and a bagging-like model averaging scheme. We demonstrate it on publicly available data sets of gene expressions in colon cancer tissue specimens and assess the relevance of the discovered biomarkers by means of a qualitative analysis. Furthermore, we test performance of the method on the cancer recurrence prediction task using two independent external validation sets. The obtained results correspond to the top published performances of gene signatures developed specially for the colon cancer case.

Keywords: genetic algorithm, feature selection, biomarker discovery, gene expressions, colon, cancer, gene signature, k-nearest neighbours, bagging.

1 Background

The recent advances in high-throughput technologies have opened a wide space of opportunities for studying complex diseases, such as cancer, at the molecular level. These led to the successful development of clinically approved diagnostic tests based on gene expression, such as the MammaPrint [1,2] for breast carcinoma. However, the complexity of resulting data from next generation sequencing or microarray experiments still poses a great analytical challenge. High dimensionality that characterizes high-throughput data, together with usually low number of available samples, renders classical statistical methodology nearly helpless when faced with data analysis tasks in this domain. This creates the increasing demand for data-driven modelling approaches capable of facilitating search for prognostics biomarkers. In this study we propose a methodology for mining cancer biomarkers from high-throughput data and demonstrate it on microarray samples in colon cancer.

Colorectal cancer is the third most common cancer type worldwide [3]. The disease starts as a benign polyp that develops to advanced adenoma and finally to invasive carcinoma. Although fairly curable if discovered on time (prior to stage III), a long term survival of initially successfully treated colorectal cancer patients critically depends on the stage of the disease at the time of diagnosis. As the current staging system does not always accurately reflect patient's individual risks [4], there is a

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 222–232, 2012.
© Springer-Verlag Berlin Heidelberg 2012

growing need for patient-tailored diagnostics and prognostics tests. This resulted in increased efforts in the development of the gene signatures for this type of cancer [5-7].

The main objective of biomarker mining is to aid in the discovery of genes, proteins or other biological indicators that could be potentially associated with a particular clinical condition. By performing a part of this process in an automated fashion the costs of wet-lab analysis and the clinical trials could be sustainably reduced, which motivated a myriad of recent research initiatives in this direction. In general, one can distinguish between the two main types of tasks and the corresponding methods that fall within a category of biomarker mining. The first includes approaches for the identification of causative factors of disease development and progression, thus of potential therapeutic targets. The second consists of methods for searching biomarkers of which alternations are indicative with, but not necessarily directly involved, in disease onset. These are mostly used for diagnostics or prognostics purposes, which renders this task closely related to feature selection as known in the field of machine learning.

In this work we present a genetic algorithm-based method that facilitates the later approach to biomarker mining. It essentially searches through the space of possible gene combinations to optimize prediction accuracy, taking into account multivariate relations between genes. Also, in contrast to similar existing methods, it explicitly enforces short gene signatures through the fitness function with a constant shrinkage pressure. Furthermore, we employ an iterative randomized procedure similar to bootstrapping to enhance robustness of resulting gene signatures.

Genetic algorithms have been frequently used for feature selection as they scale well with increasing data dimensionality and do not rely on a particular decision surface form. This renders them suitable for solving multidimensional, non-differentiable, non-continuous and other types of problems of arbitrary complexity; such as in genetic biomarker discovery. Jourdan et al. [8] use GA for feature selection, taking into account spatial correlation between neighbouring genes on the chromosome. In [9] Jirapech-Umpai and Aitken proposed an evolutionary approach without cross-over for the same task and demonstrated it on two microarray data sets on cancer. They also compared it against a simple wrapper method based on genetic algorithm. However, both described approaches assume a fixed number of features. Ooi and Tan [10] partially address this problem in an implicit way - by introducing the gene that controls the size of a solution, but still within a predefined range.

This paper is organised as follows. The second section describes the method (2.1) and the datasets (2.2) used. Discussion on the method starts with an introduction to genetic algorithms, followed by a top-level view on the system, a detailed description of the fitness function, other particularities of our implementation and the experimental framework for the external evaluation. The sub-section on data sets (2.2) contains a description of the data together with the details on preprocessing. The third section discusses results in terms of qualitative biological analysis, followed by quantitative external evaluation. Finally, in the fourth section we present our conclusions.

2 Materials and Methods

2.1 Introduction to Genetic Algorithms

Genetic algorithms (GA) [11,12] are a class of search and optimization methods inspired by the "survival of the fittest" concept as known in evolutionary biology. They mimic the process of natural selection by repeatedly generating sets of solutions, called *populations*, from which the fittest *individuals* (sometimes also called *chromosomes*) are selected for producing the next generation. Here each and every individual represents one candidate solution of the optimization problem, usually by an array of binary values called *genes*. It is an iterative process that terminates when the given objective is achieved or when some stopping criteria is met.

The particular implementation of a genetic algorithm is completely characterized by its fitness function and the types of genetic operators used. The fitness function reflects the quality of a single individual (i.e. of a single solution) and thus affects the probability that it later would be kept in the next generation or selected for combining with other well adapted individuals. This function is essential for guiding the search process and therefore its form represents an important algorithm design choice.

The genetic operators play a crucial role in the diversification of the solution pool through chromosomal structure alterations. The two most important are the *crossover* and the *mutation*, while additional custom operators, such as a *random immigrant*, are also used sometimes. Crossover is a mechanism of exchanging genes between two individuals (*parents*) in a random manner to produce child solutions (Fig. 1). It could take various forms given the particular implementation of genetic algorithm, such as single-point, two points "cut and splice", half-uniform, uniform or other. The mutation operator affects one or more genes of a single chromosome in a way that is analogous to natural mutations. Usually, the value of a single bit of individual solution is flipped according to the predefined probability (Fig. 1).

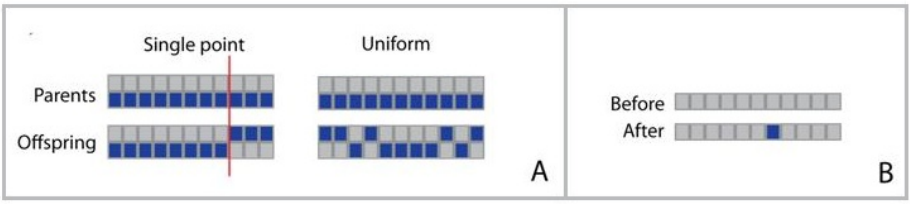

Fig. 1. Genetic operators: crossover (A) and mutation (B)

2.2 The Method

Our strategy for biomarker mining could be summarized by the following workflow (see Fig. 2). The core of our method is a genetic algorithm that optimizes a feature subset given the data and preferred classification performance metrics. This GA utilizes a customized fitness function based on supervised classification and the minimization of genetic signature length. The described optimization process is repeated iteratively, following a procedure similar to bagging [13] to facilitate robustness of the final result.

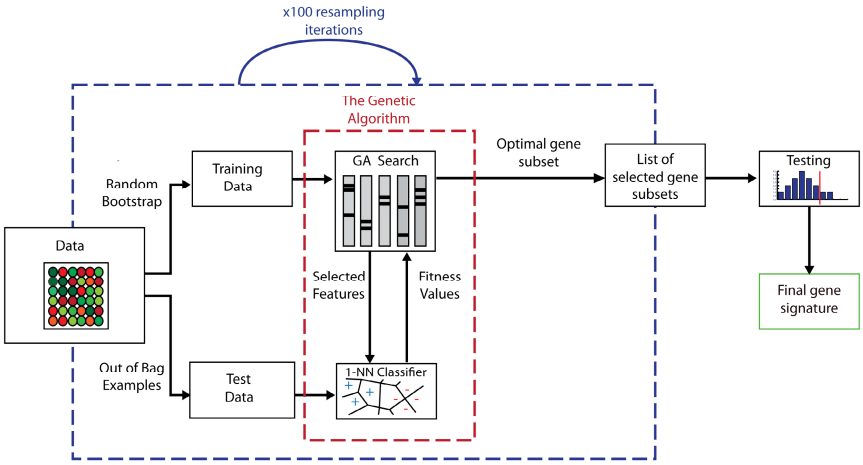

Fig. 2. The workflow of biomarker mining

We represent presence/absence of a biological gene in the signature by a value of a corresponding binary variable (*gene*) in a chromosome of the genetic algorithm. Thus a chromosome (candidate solution) works as a feature mask vector, having ones on the places of features (biological genes) to be selected and zeros elsewhere. A particular instance of potential predictive gene subset is then evaluated by the fitness function and discarded or retained for proliferation with chances proportional to its fitness. This is repeated for several chromosomes during many generations until GA reaches the execution limit, after which the most optimal genetic signature found is returned. We repeat described procedure one hundred times, saving these individual signatures from every iteration.

Each of GA optimization runs that we perform uses a different random sample from the whole training data for internal training of the classifier embedded in the fitness estimation procedure. For this we use Monte Carlo resampling with replacement, where the size of a resample is equal to that of the whole data set (bootstrap [14]). This leaves approximately 36.8% of total examples out, so that they can be utilized for the internal testing (*out-of-bag* examples). As we keep the counts on selected genes over all hundred runs, and use these for the final estimation of a particular gene importance, our procedure for model averaging emulates the bootstrap aggregation principle (bagging).

Counts across candidate genes approximately follow a negative binomial distribution which can be used for determining the threshold for selection. In general, the negative binomial distribution has relaxed assumptions compared to the Poisson distribution, which renders it appropriate for modelling a wider class of count data. Here we decide to include in the final signature genes that were selected more times than the 99% quantile of the estimated negative binomial distribution, which in this case corresponds to 17 or more (Fig 3). However, these counts could be also used as non-parametric ranks if one does not need to pose hard threshold for his/hers particular application, as is often the case in gene prioritization tasks.

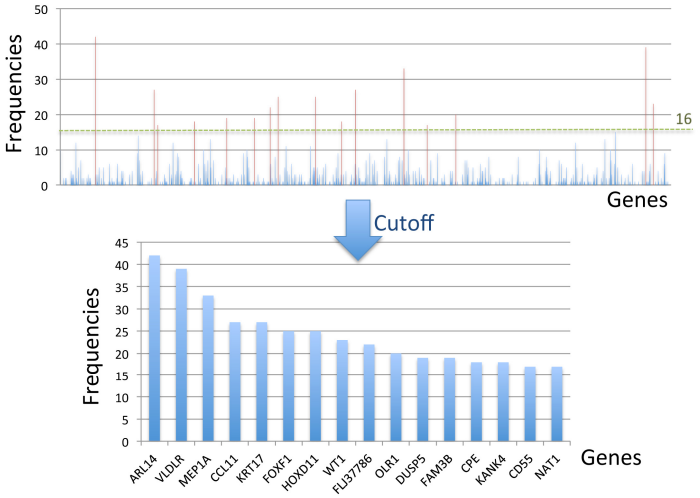

Fig. 3. Extracting the gene signature. The top figure shows how many times each gene has been selected in a signature (out of 100 independent GA runs). The figure on bottom shows these that suppressed the threshold, together with their names and frequencies (the final signature).

The Fitness Function. We use a fitness function that is based on the size of the individual solution and its performance on the independent test set. Firstly, we select genes based on a candidate solution and train one nearest neighbour (1-NN) classifier [15] on a bootstrap sample from the original data set. Then we measure performance of a trained classifier on the out-of-bag examples in terms of balanced accuracy :

$$B_{(j_c,g_c)} = \left(\frac{tp_{(j_c,g_c)}}{tp_{(j_c,g_c)} + fn_{(j_c,g_c)}} + \frac{tn_{(j_c,g_c)}}{tn_{(j_c,g_c)} + fp_{(j_c,g_c)}} \right) / 2 \tag{1}$$

where B stands for the balanced accuracy corresponding to a classifier based on a chromosome j_c from a generation g_c , and tp,tn,fp,fn for the obtained numbers of true positives, true negatives, false positives and false negatives, respectively.

We choose balanced instead of standard accuracy due to its robustness in presence of highly skewed class distributions, which is often a problem with the biomedical data sets in general. Furthermore, we choose 1-NN over more complex classification algorithms as it is very fast to evaluate and still able to capture non-linear relationships in data. When a new data point is presented to the trained algorithm, it simply assigns the outcome value of the closest (usually in terms of the Euclidean distance) example from training set to it. Thus, it also does not require any parameter tuning and, consequently, nested loops in algorithm. In addition, kNN asymptotically achieves Bayes error within a constant factor [16] and there is a body of empirical evidence suggesting that it could not be consistently outperformed by several more complex classification algorithms [17].

Furthermore, we penalize longer, and reward shorter solutions in terms of relative size gain or loss (*S*) when compared to average size of individuals from the initial generation:

$$S_{(j_c,g_c)} = \frac{Nc\sum_{i=1}^{Ng} f_{(i,j_c,g_c)}}{\sum_{j=1}^{Nc}\sum_{i=1}^{Ng} f_{(i,j,0)}} \tag{2}$$

where *Nc* and *Ng* stand for number of chromosomes in a single generation and number of gene positions per chromosome, respectively; *f* is a binary variable that equals to one if a gene has been selected given the position (*i*), chromosome (*j*) and generation (*g*); *g* stands for a generation number (here zero and current generation – g_c). In this way, in addition to maximizing the performance measure we force algorithm to converge toward smaller solutions, hoping that this would lead to more robust and general feature subsets. Finally, given (1) and (2), the fitness function (*F*) takes the following form (3):

$$F_{(j_c,g_c)} = B_{(j_c,g_c)} - S_{(j_c,g_c)} + 1 \tag{3}$$

where the same weight is given to size and accuracy, while the constant 1 is added to assure that every possible fitness function value remains positive.

Implementation Details of the Proposed GA. We build up our code basing it on the SpeedyGA.m 1.2 Matlab script [18] that implements a simple genetic algorithm as described in [19]. Our initial population counts 200 randomly generated individuals with chance of 0.2 for each feature to be present in one. The probability of mutation per bit of individual chromosome has been set to 0.5 divided by the maximal length of solution (1000). We use uniform crossover, with the probability of reproduction without it set to zero. Selection is preformed proportionally to the sigma-scaled value [19] of the fitness function using the stochastic universal sampling [20]. We restrict the maximal number of generations to 500 and keep track on the best solution over all generations.

External Evaluation. To estimate the generalisation ability of the method we fit a simple linear regression to the selected biomarkers using all samples from the training data set and apply it to two independent test sets. In addition, we compare our algorithm against another frequently used multivariate feature selection method that utilizes bagging and supervised classification performance - namely Random Forest (RF) feature selection [21] on the same data sets. It estimates the importance of the single variable by comparing accuracy of each and every tree in the trained ensemble on corresponding out-of-bag examples against accuracy that is obtained when the values of former are randomly shuffled. To avoid influence of a solution size to the unbiased assessment, we set the number of genes to be selected by the RF to that obtained with our method and number of trees to be generated to hundred.

2.3 Data Sets

We utilize three independent publicly available microarray data sets containing colon cancer samples from the Gene Expression Omnibus (GEO) [22]. The set under GEO accession number GSE17536 [23] has been used for deriving the gene signature where sets GSE17537 [23] and GSE5206 [24] have been considered for the external evaluation of our method. All three data sets were generated on the Affymetrix HG U133 Plus 2.0 microarray platform. Prior to a public release, the first two data sets were preprocessed using the MAS5.0 [25] and the third one by the RMA [26]. In addition, we discard probes that correspond to multiple genes from all three data sets and average values over multiple probes associated with a single gene.

We use samples from the Moffitt Cancer Center (GSE17536) as the training set. This data set contains 177 samples from which 145 with known relapse status (36 out of 145 patients relapsed). Prior to application of the method, we pre-filtered it with the Wilcoxon Rank Sum test by keeping thousand of the most significant genes. The p-value of this particular non-parametric test corresponds to the area under ROC curve, so we use it here due to its robustness. The data from the Vanderbilt Medical Center (GSE17537) are used as one of our external validation sets. Here, the relapse status is determined for all 55 patients with 19 of them having developed recurrent cancer within a five years period. The second validation set (GSE5206) contains samples from 105 patients. We exclude non-diseased subjects and cases where the location of major diagnosis was not the colon, resulting in 74 retained examples in total, from which 16 with recorded recurrence.

3 Results and Discussion

Our final signature consists of 16 genes, namely (in the order of relative importance) : ARL14, VLDLR, MEP1A, CCL11, KRT17, FOXF1, HOXD11, WT1, FLI37786, OLR1, DUSP5, FAM3B, CPE, KANK4, CD55, NAT1. Firstly, we performed functional analysis to estimate the biological relevance of this result. We used Ingenuity Pathway Analysis (IPA) to determine if the signature was significantly enriched for particular pathways or functions of interest. We also performed a transcription factor association analysis in IPA. Out of the 16 signature genes, 10 genes (CCL11,CD55,DUSP5,FOXF1,HOXD11, KRT17, MEP1A, NAT1, OLR1, WT1) were functionally associated with cancer (p=4.84E-04), of which 3 were associated with colorectal cancer in particular (DUSP5,FOXF1, MEP1A, p-value=4.75E-02).

Interestingly the NF-κB complex regulates 5 (DUSP5,FOXF1, CCL11, OLR1,KRT17) of the 16 genes, of which 2 are associated with colorectal cancer: DUSP5 and FOXF1. NF-κB plays a well-studied role in the immune response, cell proliferation and cell survival by inhibition of apoptosis. DUSP5 [27,28] is a kinase phosphatase which negatively regulates members of the mitogen-activated protein (MAP) kinase family, which are associated with cellular proliferation and differentiation. Forkhead box F1 (FOXF1) is a gene associated with multiple cancer types and plays a role as a putative tumor suppressor gene [29,30]; also its inactivation causes megacolon, colorectal muscle hypoplasia and agangliosis [31]. FOXF1 has also been involved in paracrine signalling in association with the WNT signalling pathway,

known to be involved in colorectal cancer development [32]. We found FOXF1 to be downregulated in our dataset coinciding with the hypothesis of it being a tumor-suppressor gene. Although no strong evidence supports associating other signature genes with colon cancer, their performance in the signature is likely related to their coexpression with functionally relevant markers, as we can see with the NF-κB regulated genes.

To assess reliability of our approach we test the gene signature that we obtained and another one generated by RF feature selection on the two independent test sets in a way previously described in "external evaluation" sub-section with following results (Fig 4): the GA based feature selection produces an AUC of 0.7705 on GSE5206 data and an AUC of 0.7266 on GSE17537, while the corresponding values for the RF feature selection are 0.7188 and 0.6564. Here we use the area under the ROC curve (AUC) as our preferred metrics for comparing classifiers due to its independence from a biased choice for a decision threshold. On these figures one can notice that, comparing to the Random Forest feature selection, our method yields better results on both testing data sets. In addition, it produces a stable set of biomarkers on repeated runs which the RF does not do.

Furthermore, our results are comparable or better to those already reported in literature [6,7], [33,34]. In [6] the 30-genes signature gives prediction accuracy of 80 and 76,3%, depending on a cross-validation scheme used. Wang et al. [7] suggest a gene signature that includes 23 genes and has corresponding AUC of 0.741. Jiang et al. [33] proposes further refinement of this signature (7 genes) and achieves an AUC of 0.66 on an independent validation set. In a study by Lin et al. [34], the authors test different combinations of classifiers and gene signatures augmented with clinical data on two data sets, resulting in AUCs of 0.73 and 0.80. They do not report AUCs obtained on gene expression data only.

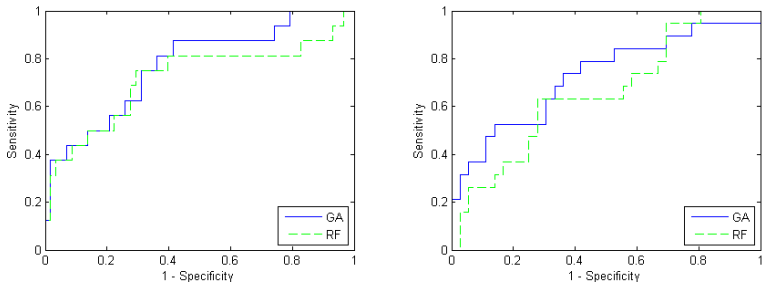

Fig. 4. ROC curves for linear regression based classification using the two feature selection methods on two test data sets (left - GSE5206, right - GSE17537)

However, most of these results are obtained using a single data set and some sort of internal validation. The predictive performance estimation in [6] and [7] relies on a training/validation split scheme (with addition of Monte Carlo crossvalidation in [6]), while [34] employs leave-one-out crossvalidation. We strongly believe that in order to prove robustness of a predictor and to avoid overestimation of its performance, one should test against external data set that originates from different cohort of patients.

Some of these studies [33,34] utilize additional or prior information, while some optimize choice on classifier to be used with biomarkers [34]. Finally, our gene signature is shorter than those reported in [6,7].

4 Conclusions

We present a simple genetic algorithm that is potentially applicable for a variety of biomarker discovery tasks and demonstrate it on the colon cancer recurrence prediction problem. The resulting gene signature displayed similar or better prediction performance than several of these proposed in the literature. Furthermore, in contrary to most of studies on the given problem, we utilize independent test sets for assessment of our method, which gave us indication of strong generalization properties of the resulting predictors. We also demonstrate biological relevance of particular biomarkers by means of a qualitative functional analysis.

In our future work we plan to improve the algorithm via finer tuning of its components and to introduce a dynamic version of the proposed fitness function to facilitate faster convergence. Furthermore, we will test it in conjunction with several popular classifiers to obtain fully optimized and complete classification system. In addition, we look forward to test the method on a wider class of biomarker mining problems and on data originating from various high-throughput platforms.

Acknowledgements. The authors would like to acknowledge support from:

- Research Council KUL: ProMeta, GOA MaNet, KUL PFV/10/016 SymBioSys , START 1, OT 09/052 Biomarker, several PhD/postdoc & fellow grants.
- Flemish Government:
  - IOF: IOF/HB/10/039 Logic Insulin
  - FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits) research community MLDM; G.0733.09 (3UTR); G.0824.09 (EGFR)
  - IWT: PhD Grants; TBM-IOTA3, TBM-Logic Insulin
  - FOD: Cancer plans
  - Hercules Stichting: Hercules III PacBio RS
- EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED
- COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network

The scientific responsibility is assumed by its authors.

References

1. Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)

2. Glas, A.M., Floore, A., Delahaye, L.J., Witteveen, A.T., Pover, R.C., Bakx, N., Lahti-Domenici, J.S., Bruinsma, T.J., Warmoes, M.O., Bernards, R., Wessels, L.F., Van't Veer, L.J.: Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7, 278 (2006)
3. Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Thun, M.J.: Cancer statistics, 2009. CA Cancer J. Clin. 59, 225–249 (2009)
4. O'Connell, J.B., Maggard, M.A., Ko, C.Y.: Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. J. Natl. Cancer. Inst. 96, 1420–1425 (2004)
5. Kerr D., Gray R., Quirke P., Watson D., Yothers G., Lavery I.C., Lee M., O'Connell M.J., Shak S., Wolmark N.: A quantitative multigene RT-PCR assay for prediction of recurrence in stage II colon cancer: Selection of the genes in four large studies and results of the independent, prospectively designed QUASAR validation study. J. Clin. Oncol. 27(suppl.), 169s, abstr 4000 (2009)
6. Barrier, A., Boelle, P.Y., Roser, F., Gregg, J., Tse, C., Brault, D., Lacaine, F., Houry, S., Huguier, M., Franc, B., Flahault, A., Lemoine, A., Dudoit, S.: Stage II colon cancer prognosis prediction by tumor gene expression profiling. J. Clin. Oncol. 24, 4685–4691 (2006)
7. Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M.G., Talantov, D., Jiang, J., McLeod, H.L., Atkins, D.: Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J. Clin. Oncol. 22, 1564–1571 (2004)
8. Jourdan, L., Dhaenens, C., Talbi, E.-G.: A genetic algorithm for feature selection in datamining for genetics. In: Proceedings of the 4th Metaheuristics International Conference Porto (MIC 2001), Porto, Portugal, pp. 29–34 (2001)
9. Jirapech-Umpai, T., Aitken, S.: Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics 6, 148 (2005)
10. Ooi, C.H., Tan, P.: Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19(1), 37–44 (2003)
11. Fraser, A.: Simulation of genetic systems by automatic digital computers. I. Introduction. Aust. J. Biol. Sci. 10, 484–491 (1957)
12. Holland, J.H.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press (1975)
13. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
14. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton (1993)
15. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
16. Stone, C.J.: Consistent nonparametric regression. The Annals of Statistics 5(4), 595–620 (1977)
17. Stanfill, C., Waltz, D.: Toward memory-based reasoning. Commun. ACM 29(12), 1213–1228 (1986)
18. Keki, M.B.: Generative Fixation: A Unified Explanation for the Adaptive Capacity of Simple Recombinative Genetic Algorithms. Ph.D. Thesis, Brandeis University (2009)
19. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press (1996)
20. Baker, J.E.: Reducing Bias and Inefficiency in the Selection Algorithm. In: Proceedings of the Second International Conference on Genetic Algorithms and their Application, pp. 14–21. L. Erlbaum Associates, Hillsdale (1987)
21. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)

22. Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 1:30(1), 207–210 (2002)
23. Smith J.J., Deane N.G., Wu F., Merchant N.B., Zhang B., Jiang A., Lu P., Johnson J.C., Schmidt C., Bailey C.E., Eschrich S., Kis C., Levy S., Washington M.K., Heslin M.J., Coffey R.J., Yeatman T.J., Shyr Y., Beauchamp R.D.: Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138(3), 958–968, PMID: 19914252 (2010)
24. Kaiser, S., Park, Y.K., Franklin, J.L., Halberg, R.B., Yu, M., Jessen, W.J., Freudenberg, J., Chen, X., Haigis, K., Jegga, A.G., Kong, S., Sakthivel, B., Xu, H., Reichling, T., Azhar, M., Boivin, G.P., Roberts, R.B., Bissahoyo, A.C., Gonzales, F., Bloom, G.C., Eschrich, S., Carter, S.L., Aronow, J.E., Kleimeyer, J., Kleimeyer, M., Ramaswamy, V., Settle, S.H., Boone, B., Levy, S., Graff, J.M., Doetschman, T., Groden, J., Dove, W.F., Threadgill, D.W., Yeatman, T.J., Coffey Jr., R.J., Aronow, B.J.: Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. Genome Biol. 8(7), R131, PMID: 17615082 (2007)
25. Hubbell, E., Liu, W.M., Mei, R.: Robust estimators for expression analysis. Bioinformatics 18(12), 1585–1592 (2002)
26. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2), 249–264 (2003)
27. Mandl, M., Slack, D.N., Keyse, S.M.: Specific inactivation and nuclear anchoring of extracellular signal-regulated kinase 2 by the inducible dual-specificity protein phosphatase DUSP5. Mol. Cell. Biol. 25(5), 1830–1845 (2005)
28. Ueda, K., Arakawa, H., Nakamura, Y.: Dual-specificity phosphatase 5 (DUSP5) as a direct transcriptional target of tumor sup-pressor p53. Oncogene 22(36), 5586–5591 (2003)
29. Watson, J.E., Doggett, N.A., Albertson, D.G., Andaya, A., Chinnaiyan, A., van Dekken, H., Ginzinger, D., Haqq, C., James, K., Kamkar, S., Kowbel, D., Pinkel, D., Schmitt, L., Simko, J.P., Volik, S., Weinberg, V.K., Paris, P.L., Collins, C.: Integration of high-resolution array com-parative genomic hybridization analysis of chromosome 16q with expression array data refines common regions of loss at 16q23-qter and identifies underlying candidate tumor suppressor genes in prostate cancer. Oncogene 23, 3487–3494 (2004)
30. Lo, P.K., Lee, J.S., Liang, X., Han, L., Mori, T., Fackler, M.J., Sadik, H., Argani, P., Pandita, T.K., Su-kumar, S.: Epigenetic inactivation of the potential tumor suppressor gene FOXF1 in breast cancer. Cancer Res. 70, 6047–6058 (2010)
31. Ormestad, M., Astorga, J., Landgren, H., Wang, T., Johansson, B.R., Miura, N., Carlsson, P.: Foxf1 and Foxf2 control murine gut development by limiting mesenchymal Wnt signaling and promoting extracellular matrix production. Development 133, 833–843 (2006)
32. Madison, B.B., McKenna, L.B., Dolson, D., Epstein, D.J., Kaestner, K.H.: FoxF1 and FoxL1 link hedgehog signaling and the control of epithelial proliferation in the developing stomach and intestine. J. Biol. Chem. 284, 5936–5944 (2009)
33. Jiang, Y., Casey, G., Lavery, I.C., Zhang, Y., Talantov, D., Martin-McGreevy, M., Skacel, M., Manilich, E., Mazumder, A., Atkins, D., Delaney, C.P., Wang, Y.: Development of a clinically feasible molecular assay to predict recurrence of stage II colon cancer. J. Mol. Diagn. 10, 346–354 (2008)
34. Lin, Y.H., Friederichs, J., Black, M.A., Mages, J., Rosenberg, R., Guilford, P.J., Phillips, V., Thompson-Fawcett, M., Kasabov, N., Toro, T., Merrie, A.E., van Rij, A., Yoon, H.S., McCall, J.L., Siewert, J.R., Holzmann, B., Reeve, A.E.: Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. Clin. Cancer. Res. 13, 498–507 (2007)