A self-tuning genetic algorithm with applications in biomarker discovery

Dusan Popovic, Charalampos Moschopoulos, Ryo Sakai, Alejandro Sifrim, Jan Aerts,Yves Moreau, Bart De Moor
Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics / iMinds Medical IT
KU Leuven
Leuven, Belgium

***Abstract* — Recent developments in the field of -omics technologies brought great potential for conducting biomedical research in very efficient manner, but also raised a plethora of new computational challenges to be addressed. Extremely high dimensionality accompanied with poor signal-to-noise ratio and small sample size of data resulting from high-throughput experiments pose previously unprecedented problem, creating an increasing demand for innovative analytical strategies. In this work we propose an island model-based genetic algorithm for multivariate feature selection in the context of -omics data, which accommodates to a particular classification scenario via dynamic tuning of its parameters. We demonstrate it on two publicly available data sets containing gene expression profiles corresponding to the two distinct biomedical questions. We show that the algorithm consistently outperforms two additional feature selection schemes across data sets, regardless to which method is used in the subsequent classification step.**

Keywords—genetic algorithm, self-tuning, island model, feature selection, biomarker discovery

I. INTRODUCTION

The rapid development of high-throughput technologies during the past few decades significantly accelerated biomedical research by facilitating studies of biological phenomena at a molecular level. Consequently, translational efforts also benefited, resulting in diagnostic tools based on the gene transcription measuring [1], therapeutic agents that target disease-causing mutations with the associated tests [2] and more. However, the full potential of these technologies is yet to be reached, as the main difficulty associated with their utilization now arises on the analytical side of the biomarker discovery process.

Typical high-throughput experiment results in several thousands measures per sample. Also, in many cases only a small number of samples are available in the context of a single study, which severely complicates standard statistical analysis. Furthermore, when using certain technologies (ex. microarrays), these measures could be extremely noisy and might vary significantly from a batch to a batch. In the same time, it is highly impractical and costly to perform confirmatory functional experiments on long lists of potentially interesting biological targets. The discussed factors eventually created a growing interest for machine learning methods which could aid biomarkers discovery [3].

The computational biomarker mining can take one of the two major forms, depending on the overall goal of a study. In the first case, the aim can be identification of biomarkers which are causally related to a disease of interest, thus of these that can also serve as potential therapeutic targets. This approach heavily relies on an extensive in vivo validation, as the causality has to be proven via controlled experiments. The role of computational methods in this type of studies is to support the process by narrowing down the search space. Alternatively, selecting variables that are relevant in the context of diagnostics or predictive models can be aim on its own. Although informative for the classification or regression, features resulting from an application of methods that are appropriate in this case do not have to be necessary directly related to a biological phenomena under the study [4]. However, they can still potentially reveal valuable functional information.

Having the later perspective in mind, we treat the biomarker mining as an instance of the feature selection problem; setting our focus on the two main goals - improved classification performance and parsimony of the resulting biomarker sets. In general, finding the optimal subset of variables given the objective function while taking into account all possible interactions is a NP complete combinatorial optimization problem, thus especially suitable for application of various meta-heuristics. The genetic algorithms are a class of methods belonging to this broad category which have been successfully applied in the feature selection context before.

The genetic algorithms (GA [5],[6]) are iterative stochastic optimization meta-heuristics which mimic natural evolution process. Each candidate solution of an optimization problem is represented as an individual in a pool that constitutes a GA population. A solution (individual) is characterized by number of genes, which are sometimes organized in chromosomes[1]. Depending on its fitness value, a single individual is selected for proliferation and subjected to the genetic operators – typically selection, crossover and mutation. The application of these operators results in a next generation of solutions whose

[1] Note that the terms individual and chromosome are sometimes used interchangeably in the literature. In this text we assume clear distinction between the two – when using the term individual we refer to a solution as a whole, while certain distinct aspects of it (various parameters, selected features) might be represented by smaller groups of genes, i.e. chromosomes.

average fitness is ideally better than that in the previous generation. The fitness value of an individual reflects a level to which optimization goal(s) are realized given the corresponding solution, and it is supplied by the fitness function. The fitness function encodes optimization goals and evaluates a degree to which these are fulfilled during the execution; so it is the crucial part of the GA-based solver design.

The genetic algorithms have been successfully applied in the several subfields of bioinformatics, including RNA structure prediction [7], multiple sequence alignment [8], microarray data classification [9]-[11] and biomarker discovery [12]-[14]. However, in each of these applications a new GA has been developed and its parameters have been specifically tuned for the problem at hand. Thus, having a method that could be readily used by unspecialized user for a wider class of bioinformatics tasks would be of a great utility. Several formulations of genetic algorithms with self-adjusting parameters have been developed before [15],[16]; but to our best knowledge the given type of GA has never been applied in this field. In this work we propose one such an algorithm, while keeping focus on the particularities of the biomarker discovery problem.

II. METHODS

A. *The outline of the algorithm*

The general working mechanism of the method is depicted on Fig. 1. Initially, a class-balanced bootstrap replicate [17] is extracted from a full dataset. This sample is then subjected to a feature selection as defined by individuals in a population, after which reduced data sets are passed to one nearest-neighbor (1-NN) classifiers [18]. The out-of-bag performance measures for all of the solutions are then used for fitness values calculations, consequently guiding selection of individuals to be included in a mating pool. The selection is preformed on sigma-scaled [19] fitness values using the stochastic universal sampling [20]. Finally, the crossover, mutation and immigration operators are applied, resulting in a next generation of individuals. This process is repeated until the maximum number of generations is reached.

Each solution is represented as an individual that is composed of four chromosomes. The first chromosome is a binary string whose length equals the total number of variables in the full data set, and it encodes a selection of features to be used for subsequent classification step. The values of genes on this chromosome are initially randomly assigned such that in average every variable is selected once in a population of a single island. The next three chromosomes encode probability multipliers of the genetic operators (see subsection B), and they are randomly initialized with a chance of 0.1 for each gene to be activated (i.e. taking value 1 instead of 0).

As displayed on Fig. 1, the whole population is divided to several subpopulations (islands), which interact via genetic operator called immigration. The main role of this division is to preserve genetic diversity while increasing efficiency of the search by expanding the population. In addition, it eases parallelization of the algorithm. It has been shown that application of this strategy results in better performance compared to single-population GAs, and that wider class of problems can be solved using it [21],[22].

B. *Genetic operators, parameters and self-tuning mechanism*

The proposed GA utilizes three genetic operators – crossover, mutation and immigration. The probability that each operator will be triggered is defined as the product of its basic probability and a value of the multiplier. The probability multiplier is an integer ranging from 2^0 to $(2^{10}-1)$ that is encoded by 10-bits chromosome. So, in addition to chromo-

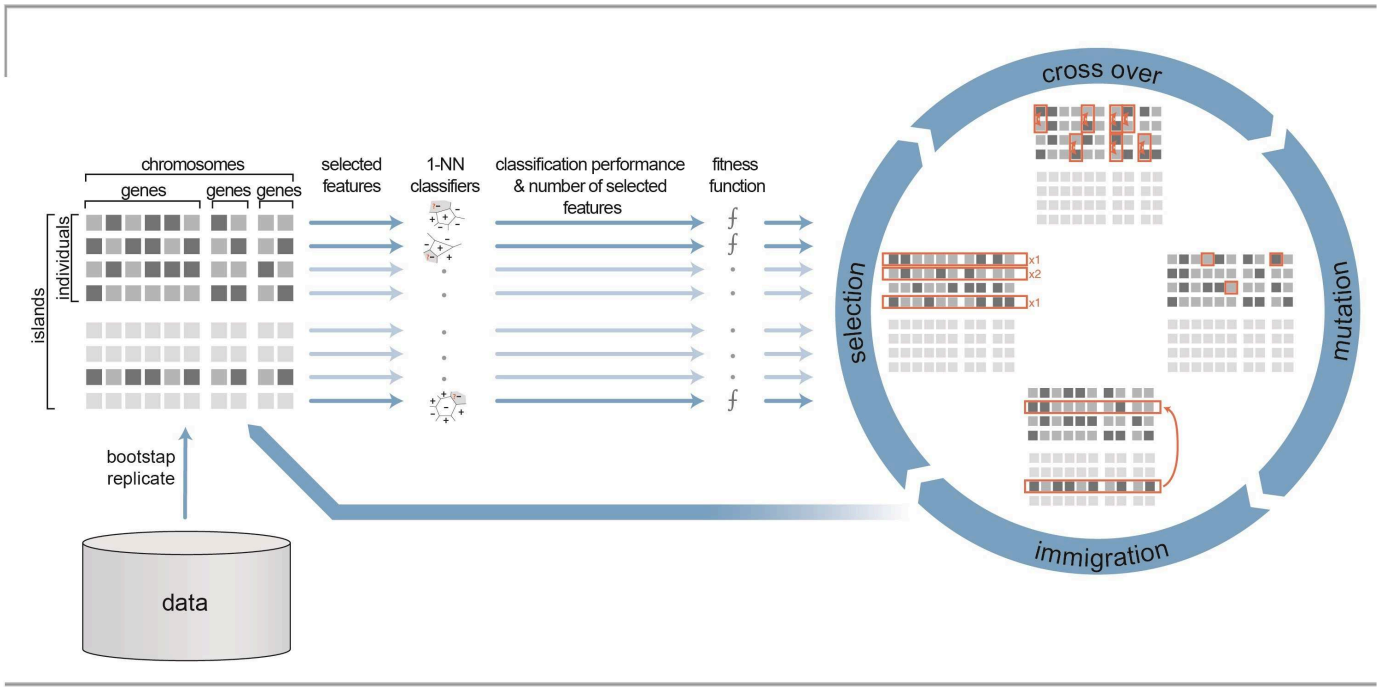

Fig. 1. The outline of the method

some that guides feature selection, each individual contains three additional chromosomes corresponding to these three genetic operators. That is, every individual is characterized by its own unique affinity for mutation, crossover and immigration; and this affinity is also subjected to the evolutionary pressure.

We hypothesize that auto regulation would occur naturally within this setup, provided that the values of basic probabilities are reasonably initialized. For example, if an individual has high affinity for mutation it would ultimately decrease its own fitness trough destruction of high-quality solutions. Conversely, if an individual decreases its mutation potential too much, it will be eventually suppressed by these that adapt faster. This reasoning easily generalizes to other operators as well.

The basic probabilities are introduced to control the effect of a large span size on the values of operator parameters. That is, if the allowed value domain is set too wide, a small change could have profound impact on the behavior of the corresponding operator. Thus, reasonable limits should be enforced beforehand. Formally, if s stands for the binary indicator function that is equal to 1 if feature f (f=1..f_{max}) is selected by the individual i (i=1..i_{max}) that resides on the island p (p=1..p_{max}) during the generation g (g=1..g_{max}), and to zero otherwise; the basic probabilities (P) of the crossover, mutation and immigration (denoted by C,M and I, respectively) are given by the following expressions :

$$P\left(C|i,j,p,g\right) = 1/1013. \tag{1}$$

$$P\left(M|s(f,i,p,g)=0,f,i,p,g\right) = 1/\left(i_{\max}\cdot g_{\max}\right). \tag{2}$$

$$P\left(M|s(f,i,p,g)=1,f,i,p,g\right) = \left(\frac{1}{i_{\max}\cdot g_{\max}}\right)\cdot\left(f_{\max}\Big/\sum_{f=1}^{f_{\max}} s(f,i,p,g)-1\right). \tag{3}$$

$$P\left(I|i,p,g\right) = \left((p_{\max}-1)\cdot g\right)/\left(1023\cdot i_{\max}\cdot g_{\max}\right). \tag{4}$$

In plain words, as the probability of crossover is always between 0 and 1 for an individual, the value of its basic probability reduces to the scaling constant (1). The total probability that two selected individuals will engage in the (uniform) crossover is given by the geometric mean of their total crossover probabilities. The geometric mean is chosen over the simple multiplication to moderate influence of extremely low affinities. Uniform crossover is utilized instead of its n-point counterpart for its recombination potential and exploratory power.

The basic probability of the mutation (expressed per gene) of an individual depends on the activity status of that gene (1 or 0). In a case when a gene is not active, it is set in such a way that every gene turns on once (in average) during the whole evolution on a single island (2). The basic probability of the opposite case (replacing 1 with 0) is given as the product of the first basic probability and the ratio of active versus inactive genes in an individual (3). These two-track mutation rates are set as such to avoid inflation of selected features. That is, the number of potentially interesting variables in biological problems is usually for several orders of magnitude smaller than their total number, which superimposes the likelihood of turning inactive gene active over likelihood of the opposite case. Note that this effect is not linear, so it is hard to control by parsimony pressure only.

The basic probability of the migration (4) is adjusted dynamically; depending on how far is the current generation from the start of an optimization process. This mechanism is introduced to prevent early homogenization, thus to promote divergent evolution on distinct islands during the initial phase of execution. In average, if an immigration multiplier is set to the maximum and if the last generation is reached, every island should send one immigrant to every other island. A single individual is selected for migration according to its total immigration probability, after which it is replicated on a randomly selected island by replacing a random native individual.

C. *The fitness function*

The fitness function formulation reflects the trade-off between the two optimization goals defined before, namely the high classification performance and the small size of the resulting solutions. The values of the performance measure are obtained via application of the 1-NN classifier on the data matching selected features. In particular, the balanced accuracy of the classifier on out-of-bag examples serves as the indicator of the classification performance. The size of a solution is expressed in terms of a relative increase (or decrease) in number of activated genes with the respect to the average number of these in the initial population. The parsimony of solutions is achieved trough constant penalization (reward) of relative size gain (loss).

The 1-NN has been chosen as the base classifier for several practical reasons. Firstly, it is capable of capturing nonlinearity in data and it displays surprisingly good performance on a wide range of heterogeneous classification problems [23]. Asymptotically, the error rate of 1-NN classifier never exceeds twice the Bayes rate [24]. The second, training time of 1-NN is effectively zero and the prediction time can be relatively short, provided that the training set is reasonably sized. Finally, the 1-NN classifier has no parameters to be tuned; meaning that there is no need for embedding an internal parameter optimization loop. This extra loop typically increases complexity of the algorithm for an order of magnitude and reduces total amount of data available for training.

The previous discussion can be summarized by a formal expression of the fitness function that takes the following form (using nomenclature introduced before) :

$$F\left(i,p,g\right) = 1 + A\left(i,p,g\right) - \frac{f_{\max}\cdot i_{\max}\cdot p_{\max}\cdot\sum_{f=1}^{f_{\max}} s(f,i,p,g)}{\sum_{f=1}^{f_{\max}}\sum_{i=1}^{i_{\max}}\sum_{p=1}^{p_{\max}} s(f,i,p,0)}. \tag{5}$$

Here F stands for fitness function evaluated for the individual i residing on the island p and belonging to the generation g. Similarly, A stands for the classification performance of the same individual, expressed in terms of the balanced accuracy achieved on the out-of-bag data. The third term in the above expression (5) is actually a number of genes that are active in an individual divided by the average number

of active genes per individual in the initial population. This term introduces a small but constant shrinking pressure on the solutions which already reached the classification performance optimum. Finally, the constant term is added to assure that every possible value of *F* remains positive.

III. EXPERIMENTS

A. The benchmark setup

In order to estimate performance of the proposed algorithm we set the following benchmark. We test all combinations of three feature selection methods and seven classifiers on the two publicly available data sets containing gene expression profiles. Each of the datasets is initially randomly partitioned into the three segments - one for the feature selection, one for the classifier training and one for the testing. To assure stability of the result we repeat this procedure one hundred times for both data sets, each time making a different random data split. For facilitating a complete insight into a classification performance, ROC (receiver operating characteristic) curves are harvested during the execution of benchmark, and finally aggregated over iterations using the threshold averaging [25].

In addition to our algorithm, we utilize two complementary feature selection methods – the Random forest feature importance measures (RF-FI, [26]) and the genetic algorithm specially tailored for the microarray data feature selection [13]. The estimation of variable importance with Random forest relies on measuring the difference in prediction error on out-of-bag indices between two cases - when the values of a single feature are shuffled and that on undisturbed data. This method is multivariate in nature and characterized by the ability to cope with a huge number of features efficiently, which is the reason behind its growing popularity within bioinformatics community [27]. As it produces rankings rather than closed subset of features, it still has to be chosen how many variables will be used in the classification. This number can be either arbitrary assigned (ex. the first hundred features), obtained via statistical modeling or determined by posing certain thresholds (ex. on p-values). To keep the comparison fair, we chose to use the *n* top-ranked features as supplied by the method, where *n* stands for the cardinality of the feature subset resulting from an application of the self-tuning GA.

To account for possible interactions between certain feature selection methods and classifiers we apply seven different classification algorithms on each of the biomarker lists obtained from the previous phase. Namely, Linear discriminant analysis (LDA, [28]), Quadratic discriminant analysis (QDA, [29]), Logistic regression [30], Decision trees [31], Naïve Bayes [32], Random forests [26] and Feed-forward artificial neural networks [33],[34] are used. These differ in complexity, learning biases and dependence on explicit distributional assumptions; so the consistency of performance across these classifiers can rule out a possibility that the obtained result comes from a specific feature selection/classifier combination.

B. The data sets

The method has been tested on the two publicly available microarray data set from the GEO (Gene Expression Omnibus [35]) database. The first data set contains 243 healthy liver and 268 tumourus tissue samples obtained from hepatocellular carcinoma (HCC) cancer patients (GSE25097, [36]-[38]). The six samples from healthy donors and 40 samples of cirotic liver tissue have been removed. The second set includes paired (tumourus/healthy) gastric glands tissue samples from 134 patients (GSE29272, [39]). The first data set has been created by using Rosetta/Merck Human RSTA Affymetrix 1.0 microarrays, while the second one comes from Affymetrix HG U133 2.0 Plus platform. Both data sets have been preprocessed by the RMA [40]; and in both of the cases, the tissue status (tumourus/healthy) was considered as the outcome of interest.

IV. RESULTS AND DISCUSSION

Fig. 2 displays ROC curves obtained by application of the three feature selection methods in conjunction with the seven different classifiers on the two biomedical problems. Table 1 enlists corresponding numerical values of the area under the ROC curve (AUC). The best performing method on the liver cancer data set was the combination of self-tuning GA and LDA classifier, which achieved the AUC value of 0.99. The best performing method on the gastric cancer data set was the same feature selection algorithm (and QDA), with average AUC value of 0.96. Overall, the self-tuning GA won in 15 out of 16 possible classification scenarios.

It is immediately apparent from the table that the self-tuning GA outperforms Random forest feature importance measures by notable margin on both data sets, irrespective to a classifier used. We hypothesize that the main reason for this is ill behavior of the later when a correlation between variables is present. That is, RF-FI tends to underestimate importance of features that are correlated among themselves [41], even in the

TABLE I.

	Classifier	FS method		
		Adaptive GA	*Simple GA*	*RF-FI*
Liver cancer	LDA	**<u>0.9900</u>**	0.9882	0.8685
	QDA	**0.9862**	0.9855	0.8786
	LR	**0.9872**	0.9866	0.8843
	DT	**0.9672**	0.9624	0.8321
	NB	**0.9862**	0.9854	0.8768
	RF	**0.9889**	0.9885	0.8897
	FFNN	**0.9839**	0.9838	0.8937
Gastric cancer	LDA	0.9605	**0.9623**	0.9247
	QDA	**<u>0.9629</u>**	0.9615	0.9279
	LR	**0.9619**	0.9616	0.9264
	DT	**0.9110**	0.9055	0.8588
	NB	**0.9577**	0.9575	0.9196
	RF	**0.9486**	0.9483	0.9091
	FFNN	**0.9506**	0.9497	0.9083

The average values of AUC obtained by testining the three FS methods in conjuction with the seven classifiers on the two data sets. The abbreviations LDA, QDA, LR, DT, NB, RF and FFNN stand for Linear discriminant analysis, Quadratic discriminant analysis, Logistic regression, Decision tree, Naïve Bayes, Random forest and feed-forward neural networks; respectively. The bold typing indicates the best feature selection method given a classifier, while the bold underlined typing marks the overall best value of AUC obained on one data set.

situations when those are also individually strongly correlated with the outcome. In the same time, many biomedical data sets contain several groups of inter-correlated variables (ex. genes sharing the same genetic pathway); and very often some of these are good discriminators. Conversely, within the same setup, and if a parsimony pressure is imposed, a genetic algorithm tends to preserve at least one feature from the correlated group while selecting-out redundant ones.

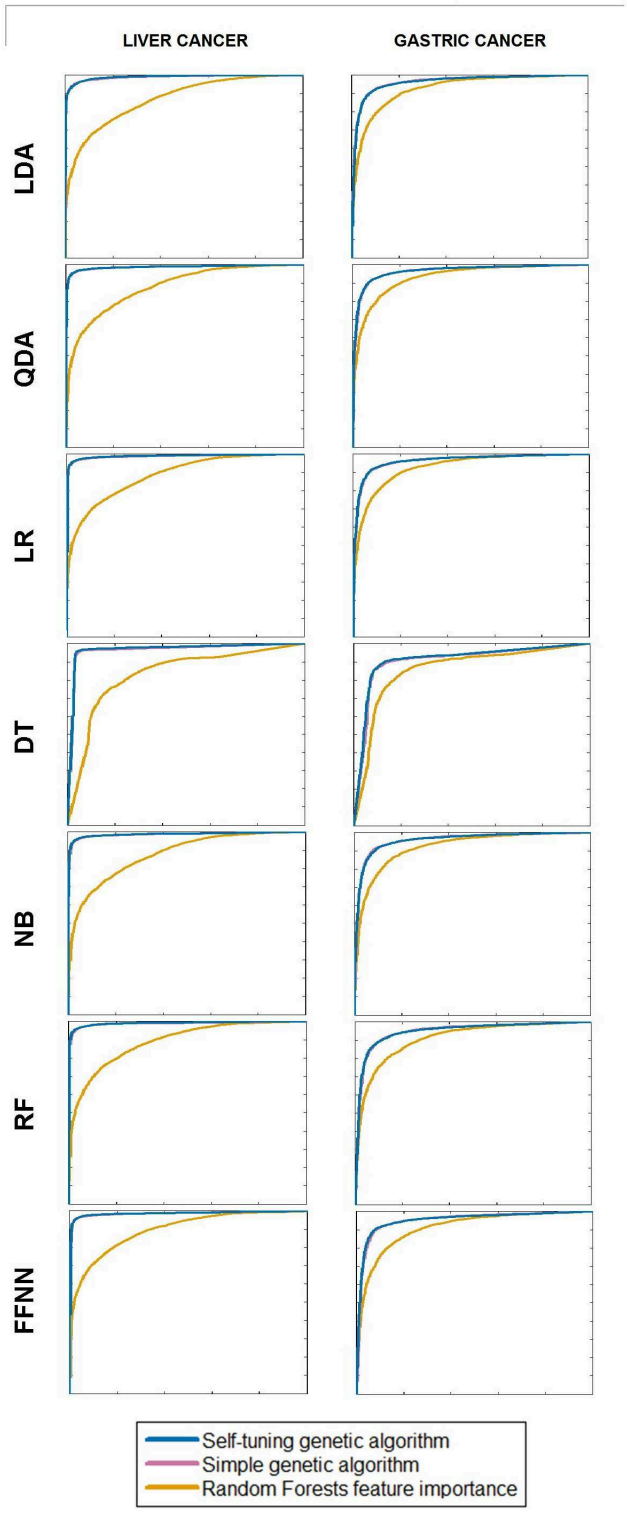

Fig. 2. The average ROC curves obtained by testining the three FS methods in conjuction with the seven classifiers on the two data sets. The abbreviations for classifiers are the same as in Table I. Note that the ROC curves of the self-tuning genetic algorithm almost completely oclude these of the simple genetic algorithm

The difference in performance between the self-tuning and the customized GA is less obvious on the utilized data. This is not surprising as the customized GA has been developed and tuned specially for the microarray feature selection. However, self-tuning GA still managed to adapt to a problem at hand; and even to marginally suppress the former in terms of the obtained AUCs on both data sets. This indicates that the method indeed can be used in various settings with no additional adjustments, while achieving the performance similar to that of problem-tailored genetic algorithms.

V. Conclusion

In this work we proposed a method for multivariate feature selection that is based on the adaptive genetic algorithm and which operates in the context of high-throughput data classification. We demonstrated that the method consistently outperforms the Random forest feature importance measures, and behaves at least equally well as the genetic algorithm that is designed especially for this application. However, in contrast to the later, it requires no tuning or other customizations; so it can be readily used by an inexperienced user. In the future we plan to extend the application scope of self-tuning genetic algorithm beyond the classification problem discussed here, as we strongly believe that its utility extends to a wider class of computational problems within the field of bioinformatics.

Acknowledgment

Bart De Moor and Yves Moreau are full professors at the Katholieke Universiteit Leuven, Belgium.
Research supported by:

- Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016, CoE PFV/10/016, SymBioSys, PhD/Postdoc grants
- Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin
- Flemish Government:
  - FWO: projects: G.0871.12N (Neural circuits); PhD/Postdoc grants
  - IWT: TBM-Logic Insulin(100793), TBM Rectal Cancer(100783), TBM IETA(130256); O&O ExaScience Life Pharma; PhD/Postdoc grants
  - Hercules Stichting: Hercules 3: PacBio RS, Hercules 1: The C1 single-cell auto prep system, BioMark HD System and IFC controllers (Fluidigm)
  - iMinds Medical Information Technologies SBO 2014; ICON b-SLIM
  - VLK Stichting E. van der Schueren: rectal cancer
  - IOF: IOF_KP
- Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate)
- COST: Action: BM1104 (Mass Spectrometry Imaging), BM1006 (NGS Data analysis network)

The scientific responsibility is assumed by its authors.

REFERENCES

[1] Pusztai L. Current status of prognostic profiling in breast cancer. Oncologist 2008;13:350–60.

[2] Margaret A. Hamburg and Francis S. Collins. (2010) The Path to Personalized Medicine. N Engl J Med; 363:301-304

[3] He Z, Yu W (2010) Stable feature selection for biomarker discovery. Comput Biol Chem 34: 215–225.

[4] Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

[5] Fraser A. (1957) Simulation of genetic systems by automatic digital computers. I. Introduction. Aust. J. Biol. Sci. 10, 484–491

[6] Holland J.H. (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press

[7] Van Batenburg FH, Gultyaev AP, Pleij CW (1995). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. Journal of Theoretical Biology 174 (3): 269–280. doi:10.1006/jtbi.1995.0098. PMID 7545258.

[8] Gondro C, Kinghorn BP (2007). A simple genetic algorithm for multiple sequence alignment. Genetics and Molecular Research 6 (4): 964–982. PMID 18058716.

[9] Laurence Rodrigues do Amaral , Geraldo Sadoyama , Foued Salmen Espindola , Gina Maira Barbosa de Oliveira, Oncogenes classification measured by microarray using Genetic Algorithms, Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, February 11-13, 2008, Innsbruck, Austria

[10] Edmundo Bonilla Huerta , Béatrice Duval , Jin-Kao Hao, A hybrid LDA and genetic algorithm for gene selection and classification of microarray data, Neurocomputing, v.73 n.13-15, p.2375-2383, August, 2010

[11] Guoqiang Yu , Yuanjian Feng , David J. Miller , Jianhua Xuan , Eric P. Hoffman , Robert Clarke , Ben Davidson , Ie-Ming Shih , Yue Wang, Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases, The Journal of Machine Learning Research, 11, p.2141-2167, 3/1/2010

[12] Lecocke M, Hess K. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. Cancer Informatics. 2006;2(3):313–327.

[13] Popovic D, Sifrim A, Pavlopoulos GA, Moreau Y, De Moor B(2012). A simple genetic algorithm for biomarker mining. In Proceedings of the 7th IAPR international conference on Pattern Recognition in Bioinformatics (PRIB'12),Springer-Verlag, Berlin, Heidelberg, 222-232.

[14] Liu J, Cutler G, Li W, et al. (2205) Multiclass cancer classification and biomarker discorvery using GA-based algorithms, Bioinformatics, 21(11), 2691–2697.

[15] T. Bäck. (1998) An overview of parameter control methods by self-adaptation in evolutionary algorithms. Fundamenta Informaticae, 35(1-4):51–66

[16] James E. Smith.(2008) Self-Adaptation in Evolutionary Algorithms for Combinatorial Optimisation. Adaptive and Multilevel Metaheuristics Studies in Computational Intelligence Volume 136, pp 31-57

[17] Efron B, Tibshirani R (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton FL

[18] Fix, E., Hodges, J.L. (1951) Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas

[19] Mitchell M.: An Introduction to Genetic Algorithms. MIT Press (1996)

[20] Baker J.E.: Reducing Bias and Inefficiency in the Selection Algorithm. Proceedings of the Second International Conference on Genetic Algorithms and their Application (Hillsdale, New Jersey: L. Erlbaum Associates), 14–21 (1987)

[21] D. Whitley, S. Rana, and R. B. Heckendorn. The island model genetic algorithm: On separability, population size and convergence. Journal of Computing and Information Technology, 7(1):33–47, 1999.

[22] Z. Skolicki and K. De Jong. The influence of migration sizes and intervals on island models. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2005). ACM Press, 2005.

[23] Stanfill C. and Waltz D. (1986), Toward memory-based reasoning. Commun. ACM, 29(12), 1213-1228

[24] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, IEEE Transactions on Information Theory IT-11: 21–27.

[25] Tom Fawcett (2006), An introduction to ROC analysis. Pattern Recognition Letters 27 pp. 861–874

[26] Breiman, Leo (2001). Random Forests. Machine Learning 45 (1): 5–32.

[27] Boulesteix Anne‐Laure, Janitza Silke, Kruppa Jochen, König Inke R.. (2012), Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining Knowl Discov, 2: 493-507.

[28] Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics 7 (2): 179–188.

[29] McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1.

[30] Hilbe, J. M. (2009). Logistic Regression Models. Chapman & Hall/CRC Press. ISBN 978-1-4200-7575-5.

[31] Breiman L, Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[32] Thomas Bayes, (1763), "An essay towards solving a Problem in the Doctrine of Chances." , Philosophical Transactions of the Royal Society of London, Vol. 53, p. 370

[33] Rosenblatt, F. (1958). "The Perceptron: A Probalistic Model For Information Storage And Organization In The Brain". Psychological Review 65 (6): 386–408.

[34] Werbos, P.J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University

[35] Edgar R., Domrachev M., Lash A.E. : Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 1;30(1), 207-10 (2002)

[36] Tung EK, Mak CK, Fatima S, Lo RC et al. Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. Liver Int 2011 Nov;31(10):1494-504.

[37] Lamb JR, Zhang C, Xie T, Wang K et al. Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One 2011;6(7)

[38] Sung WK, Zheng H, Li S, Chen R et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 2012 May 27;44(7):765-9.

[39] Wang G, Hu N, Yang HH, Wang L et al. Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. PLoS One 2013;8(5)

[40] 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)

[41] Strobl C, Boulesteix AL, Zeileis A, Hothorn T, (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8:25