International Journal of Applied Science & Environmental Engineering – IJASET 2018

Volume 1 : Issue 1 Publication Date : 28 December, 2018

A NSGA-II application with different gene expression technologies integration

Daniel Castillo[1], Juan Manuel Galvez[1], Olga Valenzuela[1] and Ignacio Rojas[1]

***Abstract* — Thanks to genetic expression, biomarkers related to much kind of diseases could be obtained. However, depending on the previous experimental constraints, these biomarkers can become too many, especially if they are intended to be used in a classification process. Multi-Objective Genetic Algorithm (MOGA) has been thought of as a suitable solution to get a trade-off among candidate biomarkers and classification accuracy. In this sense, genetic expression and Non Sorting Genetic Algorithm (NSGA) are combined in order to obtain the minimum number of possible genes that achieve the maximum classification accuracy. Only with a few genes, the classifier is expected to be more computationally efficient and faster than with all the genes, besides not endangering the final accuracy.**

***Keywords* — breast cancer, genetic algorithm, NSGA-II, classification, genes expression, microarray, RNA-Seq, SVM, genes selection**

I. Introduction

The human genome discovered has exponentially augmented the possibilities of analysis on human genetics through the bioinformatics support. This is so that the biologic knowledge about the human being is increasing due largely to advances in sequencing technologies as microarray or RNA-seq.

The analysis of gene expression profiles from these technologies have become fundamental because they allow to identify key genes that are though as potential biomarkers of the analysed disease.

Although sequencing technologies offer the opportunity to see the whole biological system through the quantification of entire human genome, it is intrinsically an intractable problem due to the high dimensionality of the data.

The gene expression quantification in both microarrays and RNA-seq is arranged as massive parallel information, where the set of features (genes) is usually much larger than the number of samples available.

This very high difference among genes and samples is named as the curse of dimensionality [2,9] or more widely well-known in the literature as (NP)-hard problem. The 'large-p small-n' paradigm has been addressed from different approaches in order to reduce its manifestation: on the one hand, through the integration of samples from multiple platforms and technologies; on the other hand, through the selection of relevant genes depending on the pursued objective.

1: Information and Communications Technology Centre (CITIC-UGR). Univeristy of Granada, Spain.

The first option is being developed extensively in the field of bioinformatics, which seeks to expand the total repertoire of samples and achieve a greater statistical significance of the available population sample. Integration can be done at different levels: platforms (Affymetrix, Illumina, etc.), technologies (microarrays, RNA-seq, etc.), omics (genomic, transcriptomic, methylation, etc.).

The second option implies the considerable reduction of the repertoire of genes (usually in the order of thousands) from a selection carried out using machine learning techniques. This selection not only allows to reduce the complexity of the problem, but also maintains the recognition efficiency in addition to eliminating irrelevant or noisy genes. Therefore, genes selection seems to be very valid to reduce the dimensionality, thereby discarding the ambiguous genes for achieving a high classification accuracy [6].

Genetic algorithms (GAs) [12,13] evaluate and evolve the population using machine learning techniques. Thanks to this, the algorithm can produce robust solutions that are important in fields like bioinformatics because real disease and people are involved.

Data reductions techniques are very important when there are gene expression data due to the high dimensionality of the features comparing to the samples available. Therefore, machine learning, GA, feature selection techniques have been very used in the last years to reduce the dimension of feature genes and to avoid the curse of dimension.

Most real-world optimization problem are affected by more than one conflicting objectives. For this cases, there are a kind of GA called Multiple Objectives Genetic Algorithm (MOGA) [11] that use different objectives for search the optimal or sets of optimal solutions (Pareto front). In this sense, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [4] will be used in this study. This algorithm allows to minimize the number of genes but maximizing the accuracy achieved by this genes in classification.

Before the apply the GA is necessary to extract the data from their respective technologies. Two sequencing technologies has been used to compute the genes expression, which are explained below.

A. *Microarray technology*

Microarray is a method that allows the measurement of the value expressions of a large number of genes simultaneously from a collection of microscopic DNA spots attached to a solid surface. This technology is based on the DNA

hybridization process, so that DNA is hybridized for each of the spots that represent one gene value expression. Once the step is finished with a laser, the expression values are read and written in a file with the extension .CEL.

Once microarray data are available, all of them are processed and filtered from a quality analysis to be later normalized. Once this was done, the last step is the integration of all microarrays. VirtualArray tool [7] has been used for the integration process.

B. *RNA-seq technology*

This technology appeared as a revolutionary tool for transcriptome and as a natural evolutionary step in the study of the genome after the massive use of microarray technology. In this sense, one of the most advantageous aspects is that although RNA-seq can be used only for transcriptome profiling, it also can be combined with other functional genomics methods to enhance the analysis of gene expression. Expression is quantified by counting the number of reads mapped to each locus in the transcriptome assembly step. This expression level can be calculated for exons or genes using contigs or reference transcript annotations. These observed RNA-seq read counts have been robustly validated against previous technologies such as microarrays or quantitative polymerase chain reaction (qPCR) [10].

II. Material and Methods

The (NP)-hard problem was minimized and approached from a combination of both options: sample integration and gene selection.

All analyzed RNA samples were obtained from NCBI GEO web platform [1]. 108 samples from microarray series and 6 samples from RNA-seq samples were finally integrated.

Table 1 shows a summary about the series used and their origin. As it can be seen, there are series from different countries, and thus there are samples from different ethnic groups. Furthermore, there are different sequencing technologies in the experiment including samples from Affymetrix [5] and Illumina [8]. Moreover, there are data from different generation sequencing. In summary, samples have been integrated from different generation sequencing, technologies, platforms and countries, bringing all of them heterogeneity to the study.

Both microarray and RNA-seq data have passed a strict pipeline. Microarray samples require restrictive quality analysis to discard non-representative samples which took place due to incorrect acquisition, as well as normalization during pre-processing in order to adapt the range of quantification variability of the samples considered.

In our experiment, 98 genes comply the statistical restrictions of logarithmic fold change ($|\ logFC \geq 2\ |$) and p-value ≤ 0.001 to form the final ranking of relevant genes considered as potential biomarkers of the disease. logFC represents the difference between breast cancer and control expressed values, whilst p-value represents the probability of obtaining a result equal or higher than what it was observed when the null hypothesis is true.

Series	Technology	Quality Samples	Excl. Outliers	Samples Origin
GSE52712	Microarray	19	1	Manchester UK
GSE40987	Microarray	10	0	Boston USA
GSE52262	Microarray	16	0	Houston USA
GSE12790	Microarray	20	1	San Francisco USA
GSE46834	Microarray	8	0	New York USA
GSE68651	Microarray	35	1	Southampton UK
GSE78011	RNA-seq	3	0	Louisville USA
GSE81593	RNA-seq	3	0	New York USA
TOTAL		114	3	

TABLE I. INPUT SERIES, TECHNOLOGY, NUMBER OF SAMPLES/OUTLIERS AND SAMPLES ORIGIN

Once the 98 expressed genes have been calculated, a GA was used. All the process followed for achieved this is explained in the subsections below.

A. *Non-dominated Sorting Genetic Algorithm II (NSGA-II)*

GAs are algorithms for optimization inspired by the biological mechanisms of reproduction and evolution. Normally, GAs try to maximize or minimize an objective using a function, but this is only when the problem have one objective.

International Journal of Applied Science & Environmental Engineering – IJASET 2018
Copyright © Institute of Research Engineers and Doctors, SEEK Digital Library
Volume 1 : Issue 1 Publication Date : 28 December, 2018

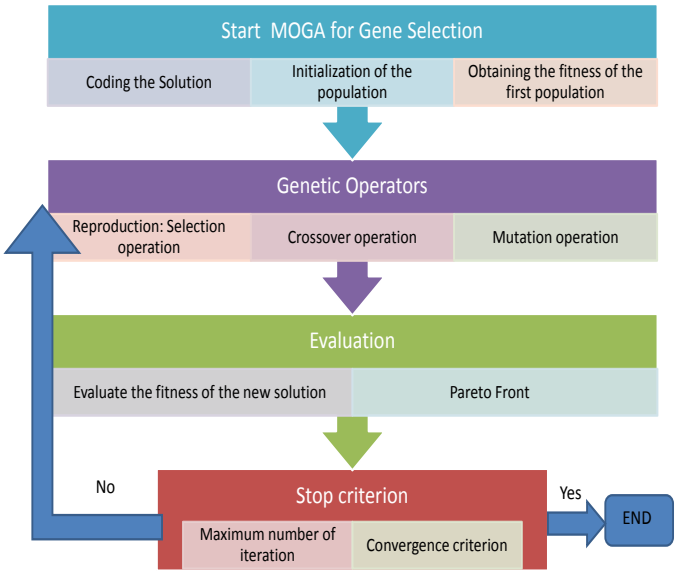

Figure 1. Pipeline followed for feature selection using NSGA-II

In the case of this study talks about a multi-objective problem due to is necessary minimize the number of genes but maximizing the final accuracy achieved when a classifier is used with these genes.

Hence, NSGA-II have been used. This GA allows reach the optimal solutions of a optimization situation when a multi-objective problem appears and calculated a Pareto front or optimal non-dominated solutions front. So, if a point is better than other points in all of objectives, this is non-dominated by these worst points. The non-dominated points of the last generations make the optimal Pareto front. All the intermediate process is explained then (the block diagram is presented in Fig.1).

First, the initial population is generate with the fixed-length binary strings for N individuals (the binary codification is presented in Fig.2) .

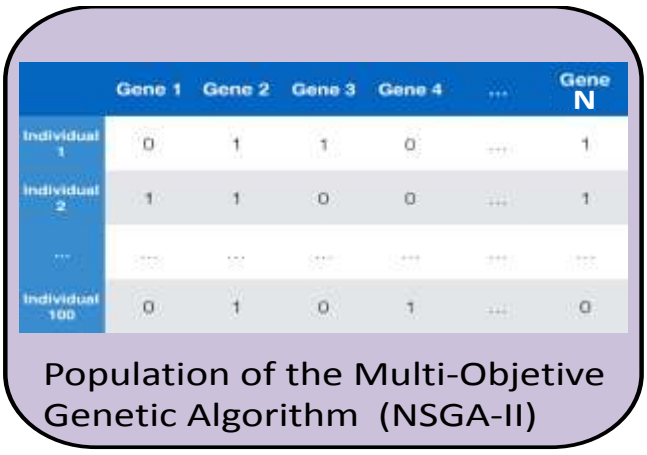

Figure 2. Binary codification of the NSGA-II for selecting the most relevant genes

Each string shows a feature subset and each position in the string are coded as one in this case if the gene have been selected or zero if not. The next step is calculate the fitness for the survival of each feature subset. Such as it said before, two objective will be used so there are two fitness functions, one for minimize the genes and other for maximize the accuracy. The best subsets will be selected for the crossover or mutation for the next generation. The mutation changes some of the values in a subset randomly. In the other hand, crossover join different features from a two subsets (parents) into a new subset (child). This process will be repeat in an iterative process until the maximum number of generation will be reached or until a stop criterion will be achieved.

B. *Support Vector Machine (SVM)*

For calculated the second objective function, a SVM has been used in order to obtain the accuracy with the selected genes of the first objective functions. This algorithm is based on the idea of separating the different categories in a problem through a hyperplane. The algorithm calculates the maximum-margin hyperplane that maximizes the distance between different classes. In one dimension, this hyperplane would be a single point. With two dimensions, a line, and with three dimensions a plane would be needed in order to separate the classes. This model could be extrapolated mathematically to higher dimensions [14].

The Fig.3 shows the pipeline followed for this study.

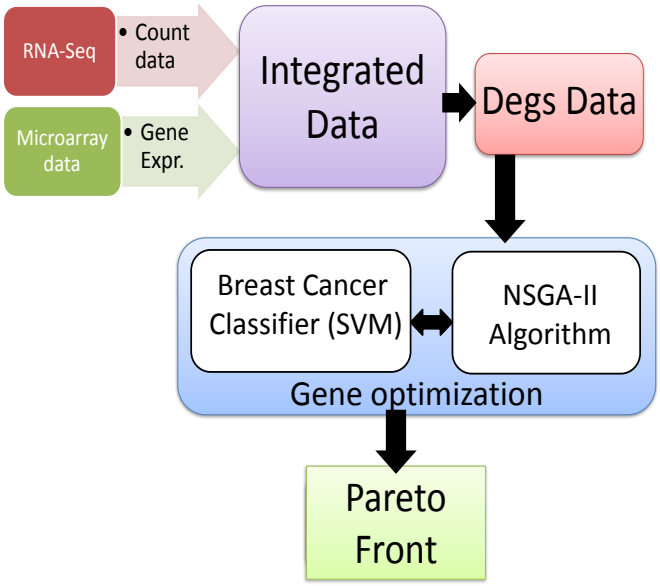

Figure 3. Pipeline followed for simultaneously using microarray data and RNA-Seq data, in conjuntion with NSGA-II, for selecting the most representative genes in breast cancer

III. Results and Discussion

The final set of 98 expressed genes have been analysed. An exhaustive list is shown in Table 2. At this point, it should be remembered that the 98 genes shown in this table correspond to those genes that appeared as differentially expressed from the analysis of the complete dataset.

This table is formed by five statistics values computed by the limma package from Bioconductor [3]. The log-fold change (logFC) represents the difference between breast cancer and control expressed values. If $|\ logFC \geq 2\ |$ it means that there exists significance differences between cancer and control values. The second value in Table 2 is the the moderated t-statistic, which has the same interpretation as the normal t-statistic but the standard errors have been reduced between the genes, effectively obtaining information from the set of genes to help with inference about each individual gene. The next value is the P-Value (P.Val) which represents the probability of obtaining a result equal or higher than what it was observed when the null hypothesis is true. The B-statistic (B) is the log-odds that a given gene is deferentially expressed.

Genes names	$\|logFC\|\geq 2$	t	p-val	B
KRT19	7.993	11.072	8.124E-21	36.607
KRT6A	-7.800	-13.558	3.347E-27	51.214
NNMT	-7.584	-11.544	4.951E-22	39.384
VIM	-7.261	-15.117	3.917E-31	60.213
AKR1B1	-6.943	-11.437	9.357E-22	38.753
SFRP1	-6.866	-18.820	4.925E-40	80.570
TGFBI	-6.701	-14.299	4.424E-29	55.515
MT1E	-6.650	-15.281	1.537E-31	61.142
C3	-6.569	-15.928	3.857E-33	64.805
BMP7	6.406	13.058	6.330E-26	48.292
KRT5	-6.229	-9.125	7.460E-16	25.273
CXCL1	-6.145	-13.526	4.030E-27	51.030
S100A2	-6.016	-9.582	5.249E-17	27.902
KRT7	-5.991	-11.975	3.850E-23	41.922
TNS4	-5.866	-25.125	1.651E-53	111.284
EEF1A2	5.764	8.956	1.979E-15	24.307
CLMP	-5.631	-11.238	3.037E-21	37.583
IFI16	-5.543	-9.230	4.073E-16	25.872
LAMC2	-5.426	-12.346	4.247E-24	44.112
IGFBP4	5.412	13.779	9.173E-28	52.501
FAM83A	-5.328	-14.042	1.974E-28	54.028
SYTL2	5.283	11.883	6.617E-23	41.384
SNAI2	-5.169	-9.731	2.204E-17	28.762
DNER	-5.152	-11.859	7.620E-23	41.244
PRKCDBP	-5.105	-10.241	1.105E-18	31.730
ALOX15B	-5.088	-16.524	1.353E-34	68.133
IGFBP5	5.085	8.165	1.755E-13	19.871
BNC1	-5.072	-16.335	3.889E-34	67.085
GFRA1	5.021	6.872	1.958E-10	12.955
DSC3	-4.999	-17.145	4.296E-36	71.561
PTGES	-4.990	-17.489	6.479E-37	73.440
TFF1	4.925	4.857	3.168E-06	3.497
RAB25	4.864	8.521	2.368E-14	21.851
KRT14	-4.863	-6.445	1.768E-09	10.794
EFEMP1	-4.855	-10.020	4.059E-18	30.440
SLPI	-4.793	-10.194	1.455E-18	31.457
SDPR	-4.728	-12.002	3.264E-23	42.086
FBP1	4.707	6.789	3.017E-10	12.530
EPCAM	4.662	8.150	1.906E-13	19.790
GNA15	-4.570	-15.676	1.614E-32	63.382
HTRA1	-4.527	-10.906	2.178E-20	35.627
RAC2	-4.524	-11.727	1.669E-22	40.465
CLCA2	-4.411	-9.272	3.189E-16	26.115
GPX1	-4.384	-6.773	3.281E-10	12.448
EMP3	-4.383	-9.299	2.728E-16	26.269
SERPINB5	-4.371	-8.314	7.600E-14	20.698
TSPYL5	4.317	6.297	3.735E-09	10.062
GSTP1	-4.242	-5.846	3.433E-08	7.892
SLC2A10	4.216	11.411	1.088E-21	38.602
LDHB	-4.182	-5.892	2.745E-08	8.111
VSTM2L	-4.146	-11.277	2.409E-21	37.813
BIRC3	-4.079	-13.064	6.110E-26	48.327
ABLIM3	-4.000	-12.337	4.481E-24	44.059
TFCP2L1	-3.874	-11.847	8.202E-23	41.171
DSG3	-3.820	-8.387	5.035E-14	21.105
SLC26A2	-3.798	-13.491	4.947E-27	50.826
C3orf14	3.763	7.772	1.558E-12	17.715
IL20RB	-3.667	-8.868	3.262E-15	23.812
FXYD5	-3.623	-5.585	1.191E-07	6.679
GSTM3	3.590	9.622	4.161E-17	28.133
ADRB2	-3.572	-9.968	5.512E-18	30.136
EMP1	-3.535	-7.622	3.543E-12	16.905
IGFBP7	-3.530	-4.676	6.866E-06	2.751
GJB5	-3.517	-12.456	2.225E-24	44.755
HENMT1	3.514	7.953	5.732E-13	18.702
ZBED2	-3.507	-6.452	1.705E-09	10.830
MSLN	-3.504	-8.558	1.917E-14	22.061
IL18	-3.415	-9.270	3.223E-16	26.104
TRIM29	-3.395	-9.588	5.081E-17	27.934
OSR2	3.346	8.380	5.238E-14	21.066
LAMB1	-3.346	-6.972	1.162E-10	13.468
UCP2	3.332	5.788	4.539E-08	7.620
CPVL	-3.331	-7.870	9.043E-13	18.253
KRT81	-3.320	-5.133	9.424E-07	4.670
S100A8	-3.292	-5.698	6.982E-08	7.200
TP53I3	-3.242	-11.149	5.160E-21	37.057
FOXA1	3.226	5.576	1.241E-07	6.640
SLC24A3	3.211	6.190	6.356E-09	9.541
PNLIPRP3	-3.200	-7.998	4.470E-13	18.948
INHBB	3.180	7.756	1.698E-12	17.630
RAB38	-3.129	-9.539	6.781E-17	27.649
ZBTB16	-3.112	-8.869	3.251E-15	23.816
PLD5	-3.070	-11.039	9.925E-21	36.408
DFNA5	-3.047	-7.565	4.835E-12	16.599
FKBP5	-2.988	-10.435	3.528E-19	32.863
CD109	-2.986	-7.196	3.541E-11	14.637
CASP1	-2.955	-6.388	2.367E-09	10.509
SULT1E1	-2.903	-7.749	1.763E-12	17.594
FAM174B	2.779	5.557	1.353E-07	6.555
PDZK1IP1	-2.752	-7.028	8.611E-11	13.743
TNNI2	-2.750	-7.896	7.842E-13	18.393
CAV1	-2.727	-5.028	1.503E-06	4.217
IRX4	-2.714	-7.628	3.433E-12	16.936
KRT80	2.706	5.268	5.131E-07	5.259
FOXO1	-2.649	-8.921	2.408E-15	24.113
SNCA	-2.635	-8.533	2.211E-14	21.919
TBL1X	2.565	9.676	3.043E-17	28.442

TABLE II. List of 98 expressed genes obtained with limma as the intersection of microarray, RNA-Seq and integrated dataset

Such as said before, two objective functions have been used in order to achieve the optimal solution for the problem. The first function is the number of active chromosome genes, in this case the number of genes used for the classification. The second function is the accuracy reached with the number of genes selected by the first function. So, the GA will

minimize the number of selected genes but maximizing the accuracy achieved.

In this sense, the chromosome have 98 genes because the study has 98 expressed genes and an initial population of 200 individuals. Furthermore, the GA has been done crossover between the population during 40 generations with a crossover rate equal to 0.7 and a mutation probability of 0.1.

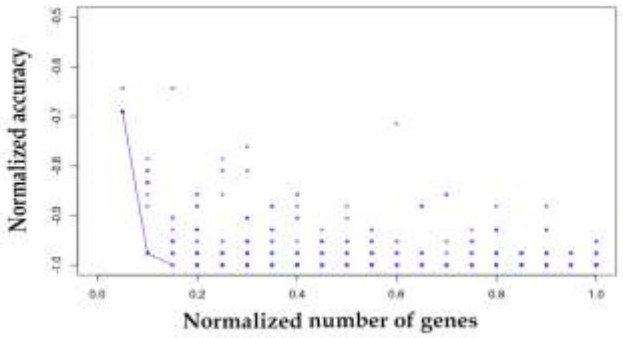

Figure 4. Optimal Pareto front calculated by the Genetic Algorithm NSGA-II

The result can be seen at the Fig. 2. Three non-dominant points appear in the plot, these points show the best solutions for the problem of this study called Pareto front. This representation has one point with one gene, one point with two genes and one point with three genes. This last point have two overlapped points because there are two combinations of three genes that achieved the maximum accuracy. These genes and their results are shown by the Table 3. Also, the expression levels of these genes for cancer and control samples have been represented at the Fig. 4. The values show that all genes except one of them (IFI16) have the expression levels with many differentiation between cancer and control samples. This means that they are good biomarkers for breast cancer classification.

Gene Names	Accuracy
PTGES	69.047 %
DSC3, SLC26A2	97.619 %
FKBP5, GSTM3, IFI16	100 %
INHBB, PLD5, PTGES	100 %

TABLE III. Optimum genes calculated by the GA for achieving the maximum accuracy

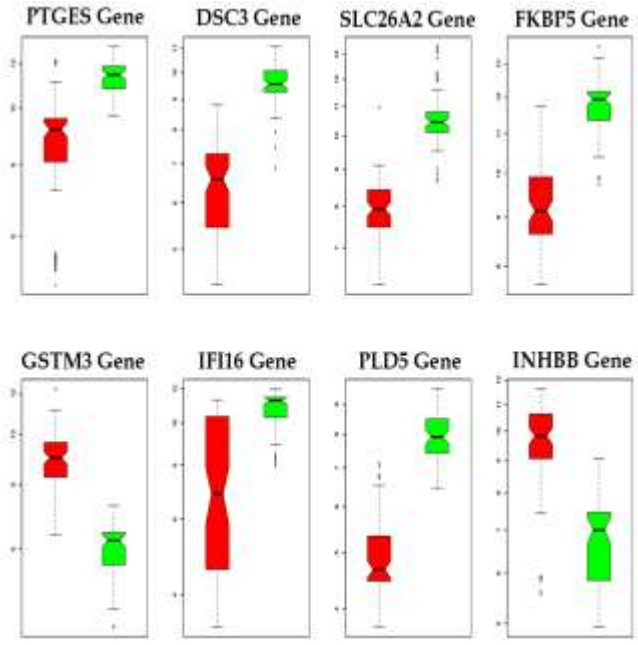

Figure 5. Expression levels of the genes that appear in Pareto front

IV. Conclusions

An heterogeneous data integration from different technologies (microarray and RNA-seq) that quantify the quantity of RNA in human biological samples is carried out in this work.

Once this integration was done 98 expressed genes that allow difference between breast cancer and control samples were achieved.

A multi-objective genetic algorithm (NSGA-II) has been used in order to minimize the number of genes needed in the classification process but without endanger the accuracy.

Finally, a Pareto front with 4 non dominated point has been calculated. This points ensures the maximum accuracy but minimizing the number of genes. Only with one genes the accuracy is 70 %, but the most important question is that only using two or three genes the final accuracy is between 97\% and 100 %, so the main purpose of this study have been achieved and the fact of use a GA is a great way of reduce the dimensionality of the feature in biological data like genes. The most important novelty is apply the GA to an integrated dataset from different technologies like microarray and RNA-seq.

Acknowledgment

This work has been partially supported by the project from the Ministry of Spain with the reference TIN2015-71873 and from J. Andalucia with reference P12-TIC-2082.

References

[1] Tanya Barre., Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. 2007. NCBI GEO: mining tens of millions of expression profiles database and tools update. Nucleic acids research 35, suppl 1 (2007), D760–D765.

[2] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. 2003. Bayesian factor regression models in the €large p, small nfi paradigm. Bayesian statistics 7 (2003), 733–742.

[3] BioConductor. 2004. Limma moderated t-statistics and B-statistics. (2004). h.ps://stat.ethz.ch/pipermail/bioconductor/2004-September/006132.html

[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation 6, 2 (2002), 182–197.

[5] Hinrich Gohlmann and Willem Talloen. 2009. Gene expression studies using Affymetrix microarrays. CRC Press.

[6] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classi€cation using support vector machines. Machine learning 46, 1 (2002), 389–422.

[7] Andreas Heider and R¨udiger Alt. 2013. virtualArray: a R/bioconductor package to merge raw data from di.erent microarray platforms. BMC bioinformatics 14, 1 (2013), 75.

[8] Illumina. 2009. Illumina Genes Expression arrays. (2009). h.p://www.illumina.com/techniques/microarrays/gene-expression-arrays.html

[9] Patrenahalli M. Narendra and Keinosuke Fukunaga. 1977. A branch and bound algorithm for feature subset selection. IEEE Trans. Comput. 26, 9 (1977), 917–922.

[10] Stuart N Peirson and Jason N Butler. 2007. .antitative polymerase chain reaction. Circadian Rhythms: Methods and Protocols (2007), 349–362.

[11] Silvia Poles. 2003. MOGA-II an improved multi-objective genetic algorithm. ESTECO-Technical Report (2003).

[12] Feng Tan, Xuezheng Fu, Yanqing Zhang, and Anu G Bourgeois. 2006. Improving feature subset selection using a genetic algorithm for microarray gene expression data. In Evolutionary Computation, 2006. CEC 2006. IEEE Congress on. IEEE, 2529–2534.

[13] Wang Wei and Liu Hong. 2010. Genetic algorithm and support vector machinebased gene microarray analysis*,... Journal of Clinical Rehabilitative Tissue Engineering Research 14, 17 (2010), 3099–3103.

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.