# Artificial Neural Networks and Ranking Approach for Probe Selection and Classification of Microarray Data

Alisson Marques Silva*†, Alexandre Wagner C. Faria*, Thiago de Souza Rodrigues†,
Marcelo Azevedo Costa*, Antônio de Pádua Braga*,
*Graduate Program in Electrical Engineering, Federal University of Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
†Federal Center of Technological Education of Minas Gerais, CEFET-MG
Av. Amazonas 5253, 30421-169, Belo Horizonte - MG - Brazil
Email: alissonmarques@cpdee.ufmg.br, axlwagner@gmail.com, tsouza@decom.cefetmg.br,
azevedo@est.ufmg.br, apbraga@cpdee.ufmg.br

***Abstract*—Acute leukemia classification into its Myeloid and Lymphoblastic subtypes is usually accomplished according to the morphological appearance of the tumor. Nevertheless, cells from the two subtypes can have similar histopathological appearance, which makes screening procedures very difficult. Correct classification of patients in the initial phases of the disease would allow doctors to properly prescribe cancer treatment. Therefore, the development of alternative methods, to the usual morphological classification, is needed in order to improve classification rates and treatment. This paper is based on the principle that DNA microarray data extracted from tumors contain sufficient information to differentiate leukemia subtypes. The classification task is described as a general pattern recognition problem, requiring initial representation by causal quantitative features, followed by the construction of a classifier. In order to show the validity of our methods, a publicly available dataset of acute leukemia comprising $72$ samples with $7,129$ features was used. The dataset was split into two subsets: the training dataset with $38$ samples and the test dataset with $34$ samples. Feature selection methods were applied to the training dataset. The $50$ most predictive genes, according to each method, were selected. Artificial Neural Network (ANN) classifiers were developed to compare the feature selection methods. Among the $50$ genes selected using the best classifier, $21$ are consistent with previous work and $4$ additional ones are clearly related to tumor molecular processes. The remaining $25$ selected genes were able to classify the test dataset, correctly, using the ANN.**

***Keywords*—*artificial neural networks, classification, microarray analysis.***

## I. INTRODUCTION

The main challenge of cancer treatment is to find specific therapies to treat distinct tumor types in order to maximize efficacy and minimize toxicity. Therefore, cancer classification is essential for treatment success.

In this work two different types of Leukemia were chosen as case studies. Leukemia is a type of cancer that starts in blood-forming tissues, such as bone marrow, and causes large numbers of blood cells to be produced. The different types of leukemia are grouped by the speed at which the disease develops (chronic versus acute). Types are also categorized according to which blood cells are affected (lymphoid versus myeloid) [1].

The acute subtype is a disease of the leukocytes that is characterized by the appearance of immature abnormal cells in the bone marrow, peripheral blood and, frequently, in the liver, spleen, lymph nodes, and other parenchymatous organs [1]. In acute leukemia, the blood cells cannot carry out their normal functions and the number

[1] http://www.cancer.gov/cancertopics/types/leukemia

978-1-4799-3194-1/13 $31.00 © 2013 IEEE
DOI 10.1109/BRICS-CCI-CBIC.2013.105

of abnormal cells increases rapidly. There are two common types of acute leukemia: acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML). Acute lymphocytic leukemia is a type of cancer in which the bone marrow makes too many lymphocytes, a type of white blood cells. Although acute lymphocytic leukemia is the most common type of leukemia in young children, it can also affect adults. Acute myeloid leukemia is a cancer of the blood and bone marrow. AML usually progresses quickly if it is not treated, and it occurs in both adults and children [2].

Both types of acute leukemia cells appear identical under the microscope, which, for many years, has led them to be considered as a single disease. However, they are genetically very distinct and can follow significantly different clinical courses and show different responses to therapy. Therefore, *ALL* must be distinguished from *AML* as soon as possible. Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis [2], [3].

DNA Microarray, a high throughput genomic measurement method, has made possible the identification of gene expression signatures associated with distinct clinical subtypes of leukemia [4]. Because microarray assays can analyze the expression of multiple genes in parallel, they have been proposed as a robust test for diagnosis. The high number of genes and small sample size problem is a challenge for modeling microarray data. This leads to the problem of relevant gene set selection for the development of low cost and fast diagnostic tests.

In this paper a predictor based on Artificial Neural Networks for classification of *Acute Lymphoblastic Leukemia* and *Acute Myeloid Leukemia* patients is proposed. DNA microarray expression levels are used as input data, in which feature selection methods are applied to reduce the input size. The $50$ most relevant genes previously selected were used to classify the patients in the test set, using an Artificial Neural Network. Among the $50$ selected genes, $21$ are consistent with previous work and $4$ additional ones are clearly related to tumor molecular processes.

## II. Material and methods

The acute leukemia dataset [5] [3], containing $7,129$ expression levels of $72$ patients, is used to determine the more relevant genes to *ALL* and *AML* classification. The dataset was split into a *training set*, of $38$ patients ($27$ *ALL* and $11$ *AML*), and a *test set*, of $34$ patients ($20$ *ALL* and $14$ *AML*) [5].

Six standard feature selection methods were applied to the *training dataset*: *Gini* [6], *t-test* [7], *rank features* [8], *Kruskall-Wallis* [9], *stepwise regression* [10] and *Fisher scoring* [11]. For five of the methods, the genes were ranked according to the corresponding metric, and the best $50$ genes were selected; for the *Stepwise Regression* method $31$ gene expression levels were selected.

Feature selection methods are usually assessed using the rank metric contained within each method. Nevertheless, rank ranges may differ among methods and a direct comparison of rank values is not usually possible. Therefore, one general approach is to assess the performance of each classifier in the prediction of an independent data set.

Artificial Neural Network (ANN) predictors, one for each set of selected genes were trained in order to compare the performance of the feature selection methods. All ANNs have the same topology: multi-layer perceptron with $5$ neurons in the hidden layer (Log-sigmoid activation function) and $1$ neuron in the output layer (Hyperbolic tangent sigmoid activation function). The *Levenberg-Marquardt* [12] learning algorithm was used with $300$ iterations. Two-fold cross-validation [13] was also applied and the ANN with the best performance was chosen. The *test* data set was then used to assess the performance of each trained ANN.

## III. Results and Discussion

Regarding the *test* data set. The ANN trained with the probes selected using the *Gini* method was able to classify all samples correctly. The error rate for all ANNs are presented in Figure 1, suggesting that the gene set selected is consistent with the *ALL/AML* class discrimination problem.

Assuming that the induced classification function approximates the universal classifier for the

[2] http://www.nlm.nih.gov/medlineplus

[3] http://www.broadinstitute.org

underlying problem, the corresponding subset of genes points to genes that are able to solve the general problem of differentiating *AML* and *ALL*. Probes selected using the methods previously presented by Golub et al. [5] were not able to induce a classifier with the same performance. The best classifier obtained for Golub's probes resulted in 2.94% error rate (see Figure 1). This suggests that the probes selected in this paper using *Gini* represent a more general subset for solving the general classification problem.

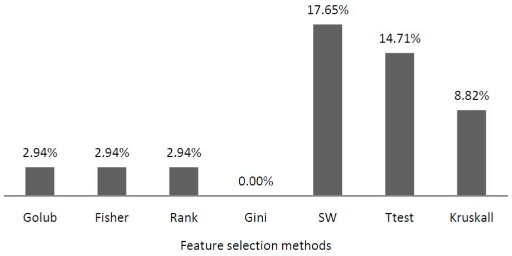

Figure 1. Error rate in the *independent data set*.

Figure 2 shows the boxplots of the 10 most relevant genes, using *Gini* for *ALL* and *AML* patients. These data refer to the *training set*. Results show no overlapping between boxplots, indicating good data separability. Additionally, 7 of 10 most relevant genes are under-expressed in the *ALL* class, which is confirmed in Figure 3.

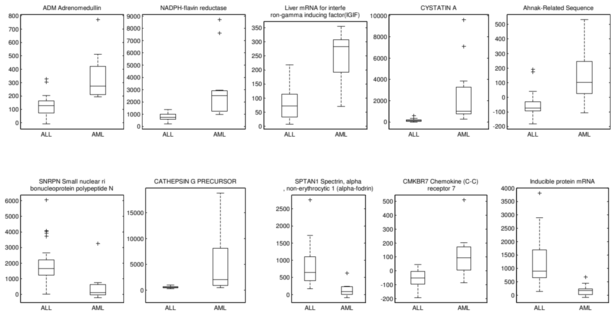

Figure 2. Boxplots for the 10 more relevant genes in the training dataset.

The expression level of 50 genes differentially expressed between *ALL* and *AML* patients can be seen in Figure III. The genes where normalized across the samples. The scale ranges from green (under-expressed) to red (over-expressed). The expression levels for each gene appear according to *AML* and *ALL* class distribution, in which the first fifteen genes are under-expressed in *AML* patients and the last thirty five genes are under-expressed in the *ALL* patients. However, there is no gene with the same expression level across the classes, indicating the multi-gene prediction requirement.

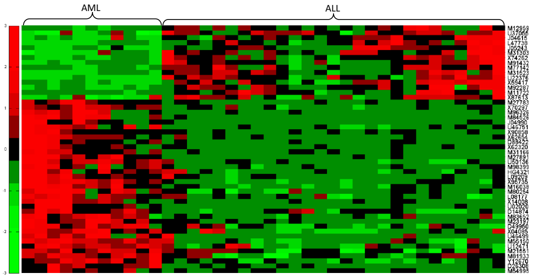

Figure 3. Expression levels for the 50 genes selected by *Gini* in the training dataset.

Some of the best predictive genes selected are known to be related to leukemia and to molecular biological processes associated with cancer. The *ADM Adrenomedullin* gene (D14874) is correlated with differentiation in human leukemia cell lines and peripheral blood monocytes [14]. The *CYSTATIN A* (D88422) gene was reported by [15] as under-regulated in cases of T-LGL leukemia. The *Ahnak-related Sequence* (HG4321) gene is a known tumor antigen [16]. The *SNRPN* (J04615) gene is associated with differential DNA methylation which is a fundamental regulator of gene transcription [17]. The *c-Myb* (U22376) gene is described as a proto-oncogene, frequently observed as strongly expressed in a variety of tumor types [18]. The *E2A* (M31523) gene is a known oncogenic gene [19]. Both *CD33* and (M23197) genes are related to antibodies that are useful to distinguish lymphoid from myeloid lineage cells [5]. The *leptin receptor* (Y12670) gene has anti-apoptotic function in hematopoietic cells [20]. The *Cystatin C* (M27891) gene is important to predict nephrotoxicity in children with acute leukemia [21]. The *Zyxin* (X95735) gene is a critical regulator of HIPK2 which activates the apoptotic arm of the DNA damage response [22]. Among all these genes, the last two have been reported as the strongest predictors [23], [24], [25] and [26].

Table I. Relevant probes selected by *Gini* method. The genes in bold were selected by Golub et al. [5]

| Genes | Description |
|---|---|
| D14874 | ADM Adrenomedullin |
| D26308 | NADPH-flavin reductase |
| D49950 | Liver mRNA for interferon-gamma inducing factor(IGIF) |
| D88422 | CYSTATIN A |
| HG4321 | Ahnak-Related Sequence |
| J04615 | SNRPN Small nuclear ribonucleoprotein polypeptide N |
| J04990 | CATHEPSIN G PRECURSOR |
| J05243 | SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin) |
| L08177 | CMKBR7 Chemokine (C-C) receptor 7 |
| **L47738** | Inducible protein mRNA |
| M11722 | Terminal transferase mRNA |
| **M16038** | LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog |
| M21551 | Neuromedin B mRNA |
| **M23197** | CD33 antigen (differentiation antigen) |
| **M27891** | CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) |
| M31166 | PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta |
| **M31303** | Oncoprotein 18 (Op18) gene |
| M54995 | PPBP Connective tissue activation peptide III |
| **M55150** | FAH Fumarylacetoacetate |
| M77142 | NUCLEOLYSIN TIA-1 |
| **M80254** | PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR |
| M81933 | CDC25A Cell division cycle 25A |
| **M84526** | DF D component of complement (adipsin) |
| **M91432** | ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain |
| **M92287** | CCND3 Cyclin D3 |
| **M96326** | Azurocidin gene |
| U02020 | Pre-B cell enhancing factor (PBEF) mRNA |
| U12471 | Thrombospondin-p50 gene extracted from Human thrombospondin-1 gene, partial cds |
| U46499 | GLUTATHIONE S-TRANSFERASE, MICROSOMAL |
| **U46751** | Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA |
| **U50136** | Leukotriene C4 synthase (LTC4S) gene |
| **X04085** | Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11 |
| **X59417** | PROTEASOME IOTA CHAIN |
| X62320 | GRN Granulin |
| X62654 | ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen |
| X70297 | CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7 |
| **X74262** | RETINOBLASTOMA BINDING PROTEIN P48 |
| X87613 | Skeletal muscle abundant protein |
| X90858 | Uridine phosphorylase |
| **X95735** | Zyxin |
| **Y12670** | LEPR Leptin receptor |
| **U22376** | C-myb gene extracted from Human (c-myb) gene, complete primary cds |
| L09209 | APLP2 Amyloid beta (A4) precursor-like protein 2 |
| U37055 | Hepatocyte growth factor-like protein gene |
| M12959 | TCRA T cell receptor alpha-chain |
| M27783 | ELA2 Elastatse 2, neutrophil |
| **M83652** | PFC Properdin P factor, complement |
| M98399 | CD36 antigen (collagen type I receptor, thrombospondin receptor) |
| X14008 | Lysozyme gene (EC 3.2.1.17) |
| **M31523** | TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) |

Table 1 shows the 50 genes selected using *Gini* along with their functional descriptions. Among them, 21 are consistent with those obtained by Golub et al. [5] (genes in bold in the first column). In addition, the selected genes *ADM Adrenomedullin*, *CYSTATIN A*, *Ahnak-related Sequence* and *SNRPN*, which are not in the intersection with Golub's subset, are also associated with tumor molecular process. Previous evidences from Golub is reinforced by the overlap with the subset presented in this paper. Furthermore, and the additional 4 genes above suggest that the resulting subset of 25 genes is relevant to leukemia classification.

## IV. Conclusion

The large number of potential predictors and the small sample size of microarray data represent additional challenges for the development of machine learning and statistical models. If statistical evidence of association between gene expression levels, and the discrimination of different types of cancers is found then patient-oriented diagnosis are possible using predictive models.

In this paper we present a classification method based on feature ranking and Artificial Neural Networks. Results show that the proposal was able to select a subset of 50 genes, which are known predictors to differentiate *AML* and *ALL* acute leukemia types. Among the selected 50 genes, 25 are functionally consistent with leukemia, cancer and associated molecular processes. Among those, 21 genes are consistent with those previously reported by Golub et al. [5]. The remaining 25 genes require further investigation about their relationship with leukemia. These genes might have been selected due to their high sample correlation to the outcome. In conclusion, the methodology presented in this paper, which considers evidences from ANN classifiers to select input features, reinforces previous results obtained by Golub et al [5], and suggests 4 additional as potential predictors in leukemia classification procedures.

## Acknowledgements

The authors thank FAPEMIG, CNPq and CAPES for financial support.

## References

[1] T. Mughal, J. Goldman, and S. Mughal, *Understanding leukemias, lymphomas, and myelomas*. Taylor & Francis, 2006.

[2] A. T. Look, "Oncogenic transcription factors in the human acute leukemias," *Science*, vol. 278, no. 5340, pp. 1059–1064, 1997.

[3] J. Rowley *et al.*, "Molecular genetics in acute leukemia," *Leukemia*, vol. 14, no. 3, pp. 513–517, 2000.

[4] B. J. Wouters, B. Löwenberg, and R. Delwel, "A decade of genome-wide gene expression profiling in acute myeloid leukemia: flashback and prospects," *Blood*, vol. 113, no. 2, pp. 291–298, 2009.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[6] Wikipedia, "Wikipedia - gini coefficient," Disponível em en.wikipedia.org/wiki/Gini_coefficient. Acesso: junho, 2010.

[7] Wikiversity, "Wikiversity - t-test," Disponível em http://en.wikiversity.org/wiki/T-test. Acesso: junho, 2010.

[8] MathWorks, *Bioinformatics Toolbox*. Natic, MA, USA: The MathWorks, 2009.

[9] W. H. K. W. A. Wallis, "Use of ranks in one-criterion variance analysis," *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, vol. 47, no. 260, pp. 583–621, December 1952.

[10] H. Demuth, M. Beale, and M. Hagan, *Neural Network Toolbox 6*. Natic, MA, USA: The MathWorks, 2008.

[11] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, New York, 2001.

[12] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2.

[13] S. Haykin, *Redes Neurais*, 2nd ed. Porto Alegre, RS, Brasil: Bookman, 2001.

[14] A. Kubo, N. Minamino, Y. Isumi, K. Kangawa, K. Dohi, and H. Matsuo, "Adrenomedullin production is correlated with differentiation in human leukemia cell lines and peripheral blood monocytes," *FEBS letters*, vol. 426, no. 2, pp. 233–237, 1998.

[15] D. O'Malley, "T-cell large granular leukemia and related proliferations," *American Society for Clinical Pathology*, vol. 127, no. 6, pp. 850–859, 2007.

[16] D. H. Chang, *Tumor Markers Research Focus*, 1st ed., N. Biomedial, Ed., 2008.

[17] L. Benetatos, E. Hatzimichael, A. Dasoula, G. Dranitsaris, S. Tsiara, M. Syrrou, I. Georgiou, and K. L. Bourantas, "Cpg methylation analysis of the meg3 and snrpn imprinted genes in acute myeloid leukemia and myelodysplastic syndromes," *Leukemia Research*, vol. 2, no. 34, 2010.

[18] R. RG and G. TJ, “Myb function in normal and cancer cells,” *Nat Rev Cancer*, vol. 8, no. 523, 2008.

[19] J. de Boer, J. Yeung, J. Ellu, R. Ramanujachar, B. Bornhauser, O. Solarska, M. Hubank, O. Williams, and H. J. M. Brady, “The e2a-hlf oncogenic fusion protein acts through lmo2 and bcl-2 to immortalize hematopoietic progenitors,” *Leukemia*, vol. 25, pp. 321–330, 2010.

[20] M. Konopleva, A. Mikhail, Z. Estrov, S. Zhao, D. Harris, G. Sanchez-Williams, S. M. Kornblau, J. Dong, K.-O. Kliche, S. Jiang, H. R. Snodgrass, E. H. Estey, and M. Andreeff, “Expression and function of leptin receptor isoforms in myeloid leukemia and myelodysplastic syndromes: Proliferative and anti-apoptotic activities,” vol. 93, no. 5, pp. 1668–1676, 1999.

[21] E. ÜNAL, Ümran ÇALISKAN, and Y. KÖKSAL, “The importance of cystatin-c for predicting nephrotoxicity in children with acute leukemia and non-hodgkin lymphoma,” *International Journal of Hematology and Oncology*, vol. 19, no. 2, pp. 69–74, 2009.

[22] J. Crone, C. Glas, K. Schultheiss, J. Moehlenbrink, E. Krieghoff-Henning, and T. G. Hofmann, “Zyxin is a critical regulator of the apoptotic hipk2-p53 signaling axis,” vol. 71, no. 6, pp. 2350–2359, 2011.

[23] C. A.C., P. G., C. E.C., C. T.G., and H. D.G., “Between-group analysis for microarray data.” *Bioinformatics*, vol. 18.

[24] C. S.B. and W. H-H., “Cancer classication using ensemble of neural networks with multiple signi¯cant gene subsets.” *Applied Intelligence*, vol. 26.

[25] L. W and Y. Y, “How many genes are needed for a discriminant microarray data analysis?” *Techniques for Microarray Data Analysis*, vol. 137.

[26] M. Y., Z. X., P. D., S. Y., and W. S.T.C., “Cancer classi¯cation by using fuzzy support vector machine and binary decision tree with gene selection,” *Biomedicine and Biotechnology*, vol. 160.