

**Biomedical robots.**  
**Application to translational medicine.**



Universidad de Oviedo

**Enrique J. deAndrés-Galiana**

Supervisors: Prof. Juan Luis Fernández-Martínez

&

Prof. Oscar Luaces

This dissertation is submitted under the PhD program of  
*Mathematics and Statistics*

May 2016





## RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: Diseño de robots biomédicos. Aplicaciones en medicina traslacional.	Inglés: Biomedical robots. Application to translational medicine.
2.- Autor	
Nombre: Enrique Juan de Andrés Galiana	DNI/Pasaporte/NIE:
Programa de Doctorado: Matemáticas y Estadística.	
Órgano responsable: Departamento de Matemáticas.	

### RESUMEN (en español)

Esta tesis trata sobre el análisis y diseño de robots biomédicos y su aplicación a la medicina traslacional. Se define un robot biomédico como el conjunto de técnicas provenientes de la matemática aplicada, estadística y ciencias de la computación capaces de analizar datos biomédicos de alta dimensionalidad, aprender dinámicamente de dichos datos, extraer nuevo conocimiento e hipótesis de trabajo, y finalmente realizar predicciones con su incertidumbre asociada, cara a la toma de decisiones biomédicas. Se diseñan y analizan diferentes algoritmos de aprendizaje, de reducción de la dimensión y selección de atributos, así como técnicas de optimización global, técnicas de agrupamiento no supervisado, clasificación y análisis de incertidumbre. Dichas metodologías se aplican a datos a pie de hospital y de expresión génica en predicción de fenotipos para optimización del diagnóstico, pronóstico, tratamiento y análisis de toxicidades.

Se muestra que es posible establecer de modo sencillo el poder discriminatorio de las variables pronóstico, y que dichos problemas de clasificación se aproximan a un comportamiento linealmente separable cuando se reduce la dimensión al conjunto de variables principales que definen el alfabeto del problema biomédico y están por tanto relacionadas con su génesis. Se analiza la robustez de dichos métodos con respecto a dos fuentes principales de ruido (en los datos y en la asignación de clases), así como errores en la modelización dado que se desconoce a priori el clasificador perfecto (si existiese). Además se demuestra el impacto en la identificación de genes altamente predictivos y de las rutas metabólicas asociadas, de las principales técnicas de preprocesado de microarreglos de expresión en la predicción de fenotipos. Finalmente se muestra que la metodología de robots biomédicos que se basa en técnicas de predicción por consenso, que explotan el espacio de incertidumbre de los problemas de predicción asociados, es la manera adecuada de abordar este tipo de problemas y por tanto de descubrir nuevo conocimiento.



## RESUMEN (en Inglés)

In this PhD we present the analysis and design of "Biomedical Robots" and its application to translational medicine. A Biomedical Robot is defined as the ensemble of methodologies and bioinformatic algorithms, coming from applied mathematics, statistical methods and computer science, able to treat different types of very high dimensional data (biomedical big data), to learn dynamically, discover new knowledge and working hypothesis, and make predictions with their corresponding uncertainty to improve biomedical decision making processes. Different learning algorithms, dimension reduction and feature selection techniques were studied and analyzed, as well as global optimization, clustering, classification and uncertainty assessment algorithms. Those methodologies were applied to clinical data gathered in hospitals and genetic expression data to phenotype prediction in order to optimize diagnosis, prognosis, treatment and toxicity analysis.

We demonstrated that is possible to establish the discriminatory power of prognostic variables in a simply way, and the corresponding classification problems approximate a linear separable behavior when the dimension is reduced to the principal variables that define the alphabet of the biomedical problem, and therefore are related to its genesis. We also analyzed the robustness of the methodology with respect to two main sources of noise (noise in the data and in the class assignment), as well as the modeling errors since the perfect classifier, if there exists, is a priori unknown. Moreover, we demonstrated the impact in the identification of high predictive genes and, consequently their associated pathways, of the main microarrays preprocessing techniques in phenotype prediction. Finally, we showed that the methodology that is based on consensus prediction techniques that explores the uncertainty space of the associated prediction problems, is the right way of addressing these types of problems and, therefore, discovering knowledge and improving medical decision-making.

**SR. DIRECTOR DE DEPARTAMENTO DE MATEMÁTICAS**  
**SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN MATEMÁTICAS Y ESTADÍSTICA**

to my beloved family...



## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 90,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 40 figures.

Enrique J. deAndrés-Galiana

May 2016





# General Setup

This thesis is included in the PhD Programme in Mathematics and Statistics (RD 99/2011) of the University of Zaragoza, the University of La Laguna, the University of Oviedo, the University of the Basque Country and the Public University of Navarra. In the elaboration of the manuscript of this PhD thesis we followed the requirements fixed by Article 26 of the Regulations of PhD Studies, agreement of June 17 of 2013 (BOPA 146 / 25-VI-2013) about the nature of the PhD thesis, that literally states: "*1. La tesis doctoral consistirá en un trabajo original de investigación elaborado por el doctorando en cualquier campo del conocimiento. La tesis debe capacitar al estudiante de doctorado para el trabajo autónomo en el ámbito de la I+D+i. 2. En su elaboración, habrán de ser tenidas en cuenta las siguientes normas mínimas: a) La memoria que recoge la labor realizada en la tesis doctoral se redactará en español. No obstante, la Comisión de Doctorado podrá autorizar su redacción en otro idioma oficial de la Unión Europea, previo informe de la Comisión Académica del Programa de Doctorado, y siempre que se garantice que los miembros del Tribunal están en condiciones de juzgarla. En este caso, la memoria deberá contener el resumen y las conclusiones en español. En las mismas condiciones, y de acuerdo con el artículo 6.2 de los Estatutos de la Universidad de Oviedo, la redacción podrá hacerse en lengua asturiana. b) En la cubierta de la memoria figurará Universidad de Oviedo, junto con el escudo institucional, el nombre del Programa de Doctorado, el título de la tesis y el nombre del autor. c) Los datos anteriores aparecerán también en la portada, y en las páginas siguientes figurará la autorización de la Comisión Académica del Programa de Doctorado, del tutor y del director del trabajo para la presentación de la tesis.*"

The main objective of this thesis is to design and build a dynamic tool called "Biomedical Robots" capable of analyze huge amount of data, generate work hypothesis and discover new knowledge in the field of biomedicine. This work belongs to the line of research dedicated to Biomedical Applications, of the research group for Inverse Problems, Optimization and Learning, directed by Professor Juan Luis Fernández-Martínez from the Mathematics Department of Oviedo University.

Since the beginning of my Bachelor's degree in Computer Science, I knew that Mathematics, Computer Science, and Medicine would have a convergence point. Thereby, my final project in the Artificial Intelligence subject of my Bachelor was to build a Lung cancer predictor. Lately, during my MSc in Soft Computing and Intelligence Data Analysis, I designed as a final research project, a genome simulator able of predict the probabilities of developing a given genetic disease depending on the mutations that are present in the genome. In the mid of 2012 I started my PhD in Mathematics with the invaluable guidance of Professors Juan Luis Fernández-Martínez and Oscar Luaces. During almost four years we walked a hard and long path, full of closed doors and hitting walls, but with simplicity and tenacity we were able to open all the doors and knock down all the walls. The final work is presented herein as a result of people with a great enthusiasm and a desire of solving problems that help to have a better world.

As the present work has a multidisciplinary scope, the main results of this research were published (or in revision) in international journals of different categories:

- **CANCER INFORMATICS**. Published. Categories (no cataloguing on Journal Citation Reports but with a H-index of 18):
  - COMPUTATIONAL BIOLOGY
  - ONCOLOGY
- **CLINICAL AND TRANSLATIONAL ONCOLOGY**. Published. Categories: ONCOLOGY (Q3).
- **BIOLOGICAL RESEARCH FOR NURSING**. Published. Categories: NURSING (Q1)
- **JOURNAL OF COMPUTATIONAL BIOLOGY**. Accepted for publication. Categories:
  - STATISTICS & PROBABILITY (Q1)
  - MATHEMATICAL & COMPUTATIONAL BIOLOGY (Q2)
  - COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (Q2)
  - BIOTECHNOLOGY & APPLIED MICROBIOLOGY (Q3)
  - BIOCHEMICAL RESEARCH METHODS (Q3).
- **JOURNAL OF BIOMEDICAL INFORMATICS**. Published. Categories:
  - COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS (Q1)

– MEDICAL INFORMATICS (Q2)

- **JOURNAL OF GENE MEDICINE.** Under review. Categories: ONCOLOGY (Q1).

The selection of these journals represents the multidisciplinary character of this research, that address problems from applied mathematics and computer science (Mathematical and Computational Biology and Medical Informatics), to genomics and translational medicine (Nursing and Oncology).

Additionally, we also presented an oral communication to the International Conference on Man-Machine Interactions celebrated in Poland on October 6-9 2015. The communication paper is a chapter of a book edited by Springer.

The core of this thesis are seven manuscripts that are either published or in revision in international journals. Consequently, the organization of the thesis is based on those papers, presented them in original form (see appendix) and preceded by an explanation of the methodology that is used, and the main original results that were achieved, focussing in each case on some specific topics and concepts used in the research, that needed a more detailed description. Accordingly the structure of this manuscript is as follows:

- **Chapter I. Introduction.** In this section we described the problem background, the main target of the thesis, the steps followed for developing it, and finally a formal description of both the problem and the methodology that was designed.
- **Chapter II. Application to clinical data.** This section is devoted to the application of the methodology to clinical data. We described two research articles related to:
  - "Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems."
  - "On the prediction of Hodgkin lymphoma treatment response."
- **Chapter III. Application to genetic data.** In this section we applied the methodology to the analysis of gene expression data. This part is developed in two manuscripts:
  - "Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy-Related Fatigue in Patients with Prostate Cancer."
  - "Genomic Data Integration in Chronic Lymphocytic Leukemia."
- **Chapter IV. Sensitivity analysis.** This research is exposed in two different papers:
  - "Sensitivity analysis of gene ranking methods in phenotype prediction."

- "Impact of microarray preprocessing techniques in unraveling biological pathways."
- Chapters **V. Design and application of biomedical robots to phenotype prediction problems**. This is the cornerstone of the dissertation and exposes the methodology that was designed to address the modeling of biomedical big data in phenotype prediction.
- Chapter **VI. Conclusions and future research**.
- Appendix **A**. We included all the publications in the original format in the appendix.
- Appendix **B**. Concentrations for the Spike-In experiment used in chapter IV.

## **Acknowledgements**

And I would like to thank my supervisors Juan Luis Fernández-Martínez and Oscar Luaces who guide me through the "not knowing" of this novel world of modeling biomedical data. I was very lucky of having the advice and support of these two great minds. Now I completely understand the quotation "standing on the shoulders of giants". I would also like to thank to all our collaborators, specially to Dr. Stephen T. Sonis from Biomodels (Massachusetts, USA), and Dr. Leorey N. Saligan from the National Institute of Nursing Research, National Institutes of Health (Maryland, USA) for supporting us, giving constructive criticisms, and providing clinical perspectives. I also thank the close collaboration of Dr. Ana Pilar González Rodríguez from Hospital Universitario Central de Asturias (Oviedo, Spain) and Dr. Segundo González from the University of Oviedo, Instituto Universitario de Oncología del Principado de Asturias, for providing us the clinical data treated in this PhD and their support and interpretation of the clinical results. Finally, I would like to thank the committee members for having accepted to take part of my PhD dissertation.



# Resumen

Esta tesis trata sobre el análisis y diseño de robots biomédicos y su aplicación a la medicina traslacional. Se define un robot biomédico como el conjunto de técnicas provenientes de la matemática aplicada, estadística y ciencias de la computación capaces de analizar datos biomédicos de alta dimensionalidad, aprender dinámicamente de dichos datos, extraer nuevo conocimiento e hipótesis de trabajo, y finalmente realizar predicciones con su incertidumbre asociada, cara a la toma de decisiones biomédicas. Se diseñan y analizan diferentes algoritmos de aprendizaje, de reducción de la dimensión y selección de atributos, así como técnicas de optimización global, técnicas de agrupamiento no supervisado (clustering), algoritmos de predicción y clasificación, y análisis de incertidumbre. Dichas metodologías se aplican a datos a pie de hospital y de expresión génica en predicción de fenotipos para optimización del diagnóstico, pronóstico, tratamiento y análisis de toxicidades.

Se muestra que es posible establecer de modo sencillo el poder discriminatorio de las variables pronóstico, y que dichos problemas de clasificación se aproximan a un comportamiento linealmente separable cuando se reduce la dimensión al conjunto de variables principales que definen el alfabeto del problema biomédico y están por tanto relacionadas con su génesis. Se analiza la robustez de dichos métodos con respecto a dos fuentes principales de ruido (en los datos y en la asignación de clases), así como errores en la modelización dado que se desconoce a priori el clasificador perfecto (si existiese). Además se demuestra el impacto en la identificación de genes altamente predictivos y de los pathways asociados, de las principales técnicas de preprocesado de microarreglos de expresión en la predicción de fenotipos. Finalmente se muestra que la metodología de robots biomédicos que se basa en técnicas de predicción por consenso, que explotan el espacio de incertidumbre de los problemas de predicción asociados, es la manera adecuada de abordar este tipo de problemas y por tanto de descubrir nuevo conocimiento.





## **Abstract**

In this PhD we present the analysis and design of "Biomedical Robots" and its application to translational medicine. A Biomedical Robot is defined as the ensemble of methodologies and bioinformatic algorithms, coming from applied mathematics, statistical methods and computer science, able to treat different types of very high dimensional data (biomedical big data), to learn dynamically, discover new knowledge and working hypothesis, and make predictions with their corresponding uncertainty to improve biomedical decision making processes. Different learning algorithms, dimension reduction and feature selection techniques were studied and analyzed, as well as global optimization, clustering, classification and uncertainty assessment algorithms. Those methodologies were applied to clinical data gathered in hospitals and genetic expression data to phenotype prediction in order to optimize diagnosis, prognosis, treatment and toxicity analysis.

We demonstrated that is possible to establish the discriminatory power of prognostic variables in a simply way, and the corresponding classification problems approximate a linear separable behavior when the dimension is reduced to the principal variables that define the alphabet of the biomedical problem, and therefore are related to its genesis. We also analyzed the robustness of the methodology with respect to two main sources of noise (noise in the data and in the class assignment), as well as the modeling errors since the perfect classifier, if there exists, is a priori unknown. Moreover, we demonstrated the impact in the identification of high predictive genes and, consequently their associated pathways, of the main microarrays preprocessing techniques in phenotype prediction. Finally, we showed that the methodology that is based on consensus prediction techniques that explores the uncertainty space of the associated prediction problems, is the right way of addressing these types of problems and, therefore, discovering knowledge and improving medical decision-making.



# Table of contents

<b>General Setup</b>	<b>vii</b>
<b>Resumen</b>	<b>xiii</b>
<b>List of figures</b>	<b>xxi</b>
<b>List of tables</b>	<b>xxv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and state of the art . . . . .	1
1.2 Objectives . . . . .	4
1.3 Methodology . . . . .	6
1.3.1 Problem definition . . . . .	6
1.3.2 The effect of modeling errors . . . . .	8
1.3.3 The effect of noise in data . . . . .	9
1.3.4 The ill-posed character of the classification problem . . . . .	11
1.3.5 Biomedical robots . . . . .	12
1.3.6 Feature selection . . . . .	16
<b>2 Application to clinical data</b>	<b>21</b>
2.1 Introduction . . . . .	21
2.2 Methodology applied to both practical cases using clinical data . . . . .	22
2.2.1 Data pre-processing . . . . .	23
2.2.2 Risk assessment . . . . .	24
2.3 Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems . . . . .	26
2.3.1 Introduction to Chronic Lymphocytic Leukemia related problems . . . . .	26
2.3.2 CLL clinical data . . . . .	27
2.3.3 CLL results for Chemotherapy treatment . . . . .	28

2.3.4	CLL results for Autoimmune Disease development . . . . .	31
2.3.5	Conclusions for CLL related problems . . . . .	34
2.3.6	Additional results for survival analysis . . . . .	35
2.4	On the prediction of Hodgkin Lymphoma treatment response . . . . .	40
2.4.1	Introduction to Hodgkin Lymphoma treatment response . . . . .	41
2.4.2	HL clinical data . . . . .	41
2.4.3	ROC-based PSO optimization of the classifier . . . . .	42
2.4.4	HL results . . . . .	45
2.4.5	Conclusions for HL treatment response prediction . . . . .	47
<b>3</b>	<b>Application to genetic data</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.2	Methodology applied to both practical cases using genetic data . . . . .	50
3.2.1	Gene discriminatory power . . . . .	51
3.2.2	Gene selection . . . . .	52
3.2.3	Correlation networks . . . . .	53
3.3	Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy - Related Fatigue in Patients with Prostate Cancer . . . . .	54
3.3.1	Introduction to the cancer treatment-related fatigue prediction problem	55
3.3.2	CRTF gene expression data . . . . .	56
3.3.3	CRTF results . . . . .	56
3.3.4	CRTF conclusions . . . . .	65
3.4	Genomic data integration in Chronic Lymphocytic Leukemia . . . . .	66
3.4.1	Introduction to Genomic data Integration in Chronic Lymphocytic Leukemia . . . . .	67
3.4.2	CLL gene expression data . . . . .	67
3.4.3	CLL results . . . . .	68
3.4.4	CLL conclusions . . . . .	78
3.4.5	Additional results for NOP16 mutational status . . . . .	78
<b>4</b>	<b>Sensitivity analysis</b>	<b>83</b>
4.1	Introduction . . . . .	83
4.2	Sensitivity analysis of gene ranking methods in phenotype prediction . . . . .	84
4.2.1	The effect of noise in phenotype prediction . . . . .	85
4.2.2	Gene selection ranking methods and noise . . . . .	86
4.2.3	The synthetic and diseases datasets . . . . .	87

4.2.4	Results using synthetic dataset . . . . .	89
4.2.5	Results using disease datasets . . . . .	94
4.2.6	Conclusions for noise analysis . . . . .	96
4.3	Impact of microarray preprocessing techniques in unraveling biological pathways . . . . .	97
4.3.1	Microarrays preprocessing techniques . . . . .	98
4.3.2	Results with the Spike-in experiment . . . . .	100
4.3.3	Results for the cancer related fatigue dataset . . . . .	102
4.4	Conclusions of the microarrays preprocessing techniques impact analysis . . . . .	108
<b>5</b>	<b>Design and application of biomedical robots to phenotype prediction problems</b>	<b>109</b>
5.1	Introduction . . . . .	109
5.2	Noise Sensitivity Analysis of Biomedical Robots . . . . .	110
5.3	Predicting IgHV mutation with biomedical robots . . . . .	111
5.4	Predicting Inclusion Body Myositis and Polymyositis with Biomedical Robots	115
5.5	Predicting Amyotrophic Lateral Sclerosis with Biomedical Robots . . . . .	118
<b>6</b>	<b>Conclusions and future research</b>	<b>123</b>
	<b>References</b>	<b>129</b>
	<b>Appendix A Publications</b>	<b>139</b>
A.1	Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems . . . . .	139
A.2	On the prediction of Hodgkin Lymphoma treatment response . . . . .	150
A.3	Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy - Related Fatigue in Patients with Prostate Cancer . . . . .	159
A.4	Genomic data integration in Chronic Lymphocytic Leukemia . . . . .	172
A.5	Sensitivity analysis of gene ranking methods in phenotype prediction . . . . .	215
A.6	Impact of microarray preprocessing techniques in unraveling biological pathways . . . . .	247
A.7	Design and application of biomedical robots to phenotype prediction problems	266
	<b>Appendix B Concentrations in the Spike-In experiment</b>	<b>289</b>



# List of figures

1.1	Conceptual scheme for the design of biomedical robots. . . . .	13
1.2	Uncertainty space in a 2D ill-conditioned linear regression problem. . . . .	15
2.1	Flow diagram for the prediction model. The methodology is composed of three steps: 1) Data pre-processing, 2) Feature selection and 3) Risk assessment. ROC-based PSO classifier optimization step is only applied in the case of Hodgkin Lymphoma. The different sub steps are also detailed. . . . .	23
2.2	A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Chemotherapy Treatment. The optimum result ( $TPR = 63.4$ and $PPV = 64.3$ ) is obtained for $p_{th} = 0.47$ . . . . .	30
2.3	A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Autoimmune Disease occurrence. The optimum result ( $TPR = 62.5$ and $PPV = 90.1$ ) is obtained for $p_{th} = 0.5$ . . . . .	32
2.4	A) ROC curve. B) Sensitivity (or True Positive Rate - TPR) and Specificity (or True Negative Rate - SPC) for 3-year survival. The optimum result ( $TPR = 99.1$ and $SPC = 53$ ) is obtained for $p_{th} = 0.48$ . Nevertheless, other probability thresholds could be adopted depending on the TPR/SPC balance. . . . .	39
3.1	Flow diagram for the prediction model. The methodology is composed of 3 steps: 1) Obtain the gene discriminatory power. 2) Select the genes according to the discriminatory power. 3) Create the correlation networks between the selected genes. . . . .	51
3.2	Data visualization in decibels ( $\log_2$ of the expression). HF is composed of 18 samples, LF 9 samples and Validation 17 samples. . . . .	57
3.3	Gene expression histograms in $\log_2$ scale for the Low Fatigue and High Fatigue subjects. . . . .	58
3.4	Fisher's ratio curve for the Low Fatigue-High Fatigue phenotype discrimination. . . . .	59

3.5	Fold change-Fisher's ratio plot of genes in the learning dataset with absolute fold change greater than 0.52 that corresponds to the 0.005 and 99.5% tails of the fold change distribution. In this case the Fisher's ratio plays a similar role than $-\log(P \text{ value})$ for the volcano plot analysis (Cui and Churchill, 2003).	59
3.6	Leave-One-Out-Cross-Validation (LOOCV) learning predictive accuracy of the first 360 gene sets with the highest discriminatory power. The shortest list with the highest accuracy (92.6%) contains only the first 14 genes. . . .	60
3.7	Histograms (in $\log_2$ scale) for the Low Fatigue (LF) and High Fatigue (HF) patients, of the first 360 most discriminatory genes. . . . .	61
3.8	(A) PCA plot for the learning set in the reduced base of the 14 most discriminatory genes. (B) PCA plot for the learning set in the reduced base of the 35 most discriminatory genes. A linear separability with a similar structure can be observed in both cases. Low Fatigue samples lie between P1A and xrt18A. Xrt25A might be a biological or behavioral outlier. High Fatigue (HF) samples lie between 13A and xrtp2A. Xrt20A marks the HF limit towards the west of the plot. Additional data are needed to perfectly delineate this PCA plot. . . . .	62
3.9	High Fatigue (HF)/Low Fatigue (LF) median expression signatures and misclassified samples at validation. It can be observed that sample xrt33 is closer to LF median signature, while xrt14, xrt36 and xrt39 are closer to the HF median signature (values for the expressions are given in tables 2 and 3).	65
3.10	Correlation network of the most discriminatory genes for the IgVH mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information. . . . .	70
3.11	Correlation network of the most discriminatory genes for the NOTCH1 mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information. . . . .	74
3.12	Correlation network of the most discriminatory genes for the SF3B1 mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information. . . . .	76
3.13	Intersection among the most discriminatory genes of the IgVH, NOTCH1 and SF3B1 mutations. The three main mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with three edges (surrounded by a dot rectangle) are common to these three main mutations. . . . .	77



3.14	Intersection among the most discriminatory genes of the IgVH, NOTCH1, SF3B1 and NOP16 mutations. The four mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with four edges (surrounded by a dot rectangle) are common to these four mutations. . . . .	81
4.1	Flow diagram of the noise analysis methodology . . . . .	88
4.2	Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for Gaussian noise. . . . .	91
4.3	Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for log-Gaussian noise. . . . .	92
4.4	Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for class assignment noise. . . . .	92
4.5	Flow chart of the methodology . . . . .	98
4.6	Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes ranked by the FC/FR methods for each comparison and different types of data. . . . .	102
4.7	Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using raw data. . . . .	106
4.8	Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using preprocessed data with RMA. . . . .	107
4.9	Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using preprocessed data with RMA. . . . .	108
5.1	IgVH classification in CLL: A) Considering all the genes of the microarray, the classification problem is nonlinear. B) Using the most discriminatory genes (13 probes) the classification problem approximates a linear separable behavior. . . . .	113
5.2	Correlation network for IgVH mutational status in Chronic Lymphocytic Leukemia. . . . .	114
5.3	Correlation network for Inclusion Body Myositis/Polymyositis. . . . .	116
5.4	Classification of IBM, PM and Control: A) PCA graphic for IBM+PM versus control samples. B) PCA graphic for IBM versus PM. . . . .	119

5.5 Correlation network for Amyotrophic Lateral Sclerosis. Probe names are used when gene names are unknown. . . . . 121

5.6 PCA graphic for ALS versus control samples . . . . . 122

# List of tables

2.1	Clinical variables description by group and their corresponding symbols and sampling frequency (Samp. Freq.). Discrete variables are shown in bold faces.	29
2.2	Chemotherapy Treatment.	31
2.3	Autoimmune disease development.	33
2.4	Summary of the results.	34
2.5	Variable Selection for one-year survival. Figures shows mean values.	37
2.6	Variable Selection for three-year survival. Figures shows Median / Interquartile range (IQR).	38
2.7	Variable selection for five-year survival. Two groups of variables are shown. First, the main reduced base with the highest accuracy (85.6%) and below, other relevant variables obtained with Entropy method. Figures shows Median / Interquartile range (IQR).	40
2.8	Main characteristics of the patients (number of patients / percentage), including Hasenclever International Prognostic Score (IPS)	42
2.9	Clinical variables description by group and their corresponding symbols and sampling frequency (Samp. Freq.). Discrete variables are shown in bold faces.	43
2.11	Mean values of the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), and weights ( $\omega_b$ ) for the optimum NN-classifier without weights optimization.	46
2.10	Best results for all the comparisons obtained without weights optimization.	46
2.12	Best results for the comparisons obtained after weights optimization.	47
2.13	Mean values of the true positives, true negatives, false positives and false negatives and optimized weights $\omega_a$ of the optimum NN classifier after weight optimization.	47
3.1	Mean values for the 14 most discriminatory genes.	63
3.2	Misclassified samples.	64
3.3	IgVH mutational status prediction	69

3.4	NOTCH1 mutational status prediction. . . . .	71
3.4	NOTCH1 mutational status prediction. . . . .	72
3.4	NOTCH1 mutational status prediction. . . . .	73
3.5	SF3B1 mutational status prediction. . . . .	75
3.6	NOP16 mutational status prediction. . . . .	79
4.1	Synthetic modeling precision. Precision for each of the noise types at different noise levels. . . . .	90
4.2	Synthetic modeling accuracy. Mean LOOCV predictive accuracy for each of the noise types at different noise levels. . . . .	91
4.3	Synthetic modeling precision with combined noise . . . . .	94
4.4	Synthetic modeling accuracy with combined noise . . . . .	94
4.5	Mean LOOCV accuracy / Number of selected probes for CLL, IBM, and ALS datasets. . . . .	95
4.6	Precision on the selection of the differential expressed genes using raw data or preprocessed data with RMA and MAS5. The data is the Affymetrix Latin Square Data for Expression Algorithm Assessment. The selection is performed between the first group and the rest to include all the differences between the spike-in concentrations. . . . .	101
4.7	Probe/Gene name and Accuracy (Acc %) of the selected probes for raw data and preprocessed data with RMA and MAS5 . . . . .	105
5.1	Noise results . . . . .	110
5.2	CLL, IBM & PM and ALS results . . . . .	118

# Chapter 1

## Introduction

### 1.1 Background and state of the art

The advance of high-throughput technologies in the last 20 years, have provided a huge increase of information that needs to be properly managed. Such advance has impacted in every single field of science, especially in Medicine. New technologies have allowed to improve data collection, from research centers to hospitals. Nowadays, medical doctors can retrieve data from the patient faster. Clinical data, such as electronic health records, clinical trials or disease registries, are publicly available and can be retrieved in a safer and more efficient manner. Information technologies allow to make these clinical data available through biobanks and electronic medical records. For instance, in the Hospital Central de Asturias the implementation of program Millenium, designed by Cerner corporation for managing the electronic health record, had an original cost around 17 millions euros. Nevertheless, this program does not allow to mine this information in order to solve different kind of problems, such as the estimation of surgical risk (see for instance <http://riskcalculator.facs.org/>) based on individual health records and/or analysis of prognostic variables for particular diseases based on customized data bases that are specifically created by medical experts. These are some examples about the complexity of extracting information from hospital data and bringing back the results with a translational approach. In this thesis we provide two examples of the application of clinical data to the analysis of Hodgkin Lymphoma and Chronic Lymphocytic Leukemia related problems.

Genomic data has also become a key element in the medical research field. Since the discovery of the DNA structure in 1953 there have been important developments in the field of genomics, being a milestone the whole human genome sequencing in 2000. Data related to DNA sequence, RNA sequence and expression, protein sequence, structure, modification, and small molecule metabolite structure are available in valuable resources such as Genbank

(<http://www.ncbi.nlm.nih.gov/genbank/>), Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>), Protein Data Bank (<http://www.wwpdb.org/>), and many others. Such progress has enabled to set down the genetic basis of a wide range of common diseases, leading to identification of the genes and biomarkers that might be responsible for the development of complex diseases. Particularly, it is important to have at disposal methods that allow to break the bottlenecks for the application of genetics into clinics. In this PhD thesis we present several examples of the use of genetic information translationally: 1) Radiotherapy-related fatigue prediction in patients with prostate cancer. 2) Genomic data integration of the main mutations that impact survival in patients with Chronic Lymphocytic Leukemia. 3) Genetic analysis of rare and neurodegenerative diseases in the search for orphan drugs and new therapeutical targets.

Medical doctors and researchers have at disposal high dimensional and heterogeneous biomedical data that needs to be mined and converted into knowledge to support their decision making processes. The management of biomedical data currently used in most research settings are labor intensive and rely upon technologies that have not been designed to handle such multi-dimensional data. Furthermore, novel molecular-based tools are emerging and rapidly entering the clinic and creating new paradigm in healthcare. Circulation tumor cells, nucleic acids and exosomes in blood of cancer patients have received increasing attention as new diagnostic tools enabling the so called "liquid biopsies", avoiding thus other invasive methods like tissues biopsies, and obtaining even more information by a simple blood test (Alix-Panabieres and Pantel, 2013). One of the main challenges is the creation and delivery of information management platforms capable of adapting different data sources, supporting workflows, and generating new hypothesis to support decision making processes, connecting, therefore, the molecular/cellular world with the clinical research providing them a translational approach.

It is considered that one of the most important revolutions of the 21st century will be related to the field of translational medicine, defined as the basic research with an impact over the global healthcare system. Genomic and clinical data resources are now allowing to consider individual variations, and not simply population averages, leading to improved diagnosis, prognosis, and treatment. The translational approach of the knowledge mined within the biomedical data will allow the creation of new medical devices, molecular diagnostics based in small-scale genetic signatures, small molecule therapeutics, biological therapeutics, vaccines, and others. Particularly, the analysis of the pharmaco-genomics (mechanism of actions) and pharmaco-kinetics aspects (minimization of toxicities) of new drugs, is crucial important in translational medicine and constitute the last step towards what has been known as personalized medicine and more recently precision medicine.

Since the birth of modern computer science, biology and computer science have gone hand in hand, and both areas have influenced each other. On one hand, most of the recent discovery in biology and especially in genetics, have been made possible thanks to computer science techniques and algorithms. For instance, the sequencing of the human genome would not have been possible without high-performance computational facilities (Venter et al., 2001). On the other hand, biology has influenced the computer science, with the developments of tools such as Artificial Neural Networks, Swarm Intelligence and/or Genetic Algorithms. Nowadays computer science based methods and technologies can allow researchers to access and extract domain knowledge and applying these results to generate and test hypotheses. During the last 10 years, Artificial Intelligence as a part of Applied Mathematics and Computer Science, has had an important roll in both medical research and translational medicine fields. They provide through the optimization of diagnosis, treatment, planning, and prediction of prognosis, a natural way of representing the uncertainties involved in the classical medical procedures. Clinical and genetic data has become increasingly fundamental, and we must tackle it from all the possible approaches, in which Applied Mathematics and Computer Science, have an important role. As the medicine advance towards a more personalized medicine where data and information have a key role in that progress, we must introduce and adapt the classical procedures of treating data to the new personalized medicine.

The majority of the research works in the field of translational medicine where it was applied Artificial Intelligence are related to data mining processes. And most of them, deal with the goal of analyzing gene expression data coming from gene expression analysis through hybridization microarrays or RNA sequencing, consisting of thousands of genes for each patient, with the aim to diagnose (sub)types of diseases and to obtain a prognosis which may lead to individualized therapeutic decisions.

The published papers are mainly related to oncology, where there is a strong need for defining individualized therapeutic strategies (Bellazzi et al., 2011). One of the most important work in this area was that of Golub et al. (1999) where they were able to build a classification model based on a weighted-voting approach relying on a list of about 50 genes related to acute myeloid leukemia and acute lymphoblastic leukemia. Another important work was carried out by Futschik et al. (2003). They used both clinical and microarray data to build two models for the prediction of diffuse large B-cell lymphoma.

According to PubMed statistics, more than 65000 publications are related to Artificial Intelligence and Medicine. We can find a wide range of publications, beginning from the definition of the well-known Perceptron (Rosenblatt, 1958) to a recent publication in where an artificial intelligence methodology is applied for detecting and characterizing epistasis in genetic association studies (Moore and Hill, 2015).

As noted by Eli and Edythe of the Broad Institute for Biomedical Research of Harvard and MIT: "We have an historic opportunity and responsibility to transform medicine by using systematic approaches to dramatically accelerate the understanding and treatment of disease". In this PhD we introduce the concept of "Biomedical Robots" as a methodological framework for solving any medical problem, independently of the type of data and problem. Moreover, the Biomedical Robot has not any character of a black-box but the capability of inferring solutions from data with the medical doctor's understanding, with the main target of taking those solutions to the side of hospitals (translational medicine) and research centers, from the "bench" to the "bedside".

## 1.2 Objectives

The main purpose of this thesis is to describe, develop and apply the novel concept of Biomedical Robot. A Biomedical Robot is defined as the ensemble of methodologies and bioinformatic algorithms, coming from applied mathematics, statistical methods and computer science, capable to treat different types of very high dimensional data (biomedical big data), to learn dynamically and make predictions with their corresponding uncertainty. The techniques involved by the biomedical framework are:

1. Machine Learning, classification problems (supervised and unsupervised), and ensemble learning.
2. Feature selection and model reduction.
3. Global optimization algorithms.
4. Receiver Operator Characteristic (ROC) curves and uncertainty analysis.

Within this framework it is possible to analyze dynamically (as a function of time) any type of data independently of their dimensionality, discovering knowledge and generating new medical working hypothesis, and finally supporting medical research and decision making approaches with its corresponding uncertainty assessment (risk analysis). Generating new working hypothesis could include for instance the analysis of biomarkers and mechanisms of action involved in a specific problem or discovering pathways and druggable targets in phenotype prediction problems. Also, a benefit of this approach could be the design of intelligent systems to support medical doctors/researchers in the decision making process of new incoming uncatalogued samples to decide crucial questions related to their diagnosis, prognosis and treatment optimization before any decision was taken. These techniques can



help for instance in segmenting patients with respect to drug response based on genetic signatures, to predict the development of induced toxicities, to predict the surgical risk, etc. . . , among many different applications that we can imagine. Particularly in the case of hospital data it is important to be able to filter and interpolate missing data and also to design classifiers using the most significant medical prognostic variables, and penalizing a given criterion, for instance the probability of having false positive or negatives.

In the case of genetic data, two different aims are complementary:

1. Finding robust small-scale genetic signatures for personal diagnosis, prognosis and treatment optimization.
2. Understanding the biological pathways involved in the mechanisms of action of phenotype prediction problems corresponding to disease development, treatment response and development of toxicities.

Given the high underdetermined character of any kind of phenotype prediction problem, it is not correct to provide a unique gene that is responsible for the disease development. As these kind of problems are ill-posed (Hadamard, 1902), the correct answer would be to address the corresponding classification or prediction problem with its uncertainty assessment. That way, the gene networks, that is, the set of genes that are interrelated, have a high discriminatory power and work synergistically for the phenotype prediction problem, are the right solution for assessing the uncertainty. Based on these networks it would be possible to find the biological pathways that are affected. However, a Biomedical Robot must have a dynamic character and it must be updated as the level of knowledge of the problem we want to solve increases. These types of optimization and learning problems are subordinated to the Non-free-lunch theorem (Wolpert and Macready, 1997), therefore, although the techniques are common, their applications should be custom designed.

The steps followed to develop this methodology were the following:

1. Developing learning algorithms, dimensionality reduction, global optimization, and clustering/classification techniques.
2. Application of these methods to different types of biomedical data:
  - Clinical data collected in hospitals (immunohistochemical, biochemical, demographic, . . .). Which lead us to the following research works:
    - Treatment response prediction in patients with Hodgkin Lymphoma. Treatment optimization.

- Need of chemotherapy prediction and autoimmune disease occurrence prediction in patients with Chronic Lymphocytic Leukemia. Diagnosis and treatment optimization.
  - Genetic data (gene expression). The following researches were developed:
    - Toxicity analysis of radiotherapy treatments in patients with prostate cancer. Treatment optimization.
    - Prediction of the main mutations that impact survival in patients with Chronic Lymphocytic Leukemia. Diagnosis and prognosis optimization.
3. Sensitivity analysis of the methodology, using both synthetic and real data:
- Impact of different kind of noise in phenotype prediction problems.
  - Impact of main microarrays preprocessing techniques in the discovering of biological pathways.
4. Design, development and analysis of biomedical robots and their application to phenotype prediction problems.

## 1.3 Methodology

### 1.3.1 Problem definition

The first class of problems found in the biomedical field are regression-type problems, that are typically solved using nonlinear multivariate regression techniques with statistical packages as SPSS. The problem consists in giving a set of variables  $\mathbf{x} \in \mathbb{R}^N$  parameterizing the samples, finding the estimator  $f^*(\mathbf{x})$  such as  $\|\mathbf{y}^{pre} - \mathbf{y}^{obs}\|_p$  is minimum, where

$$\mathbf{y}^{obs} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \mathbf{y}^{pre} = \begin{bmatrix} f^*(\mathbf{x}_1) \\ f^*(\mathbf{x}_2) \\ \vdots \\ f^*(\mathbf{x}_m) \end{bmatrix}$$

for a given type of regression models,  $f^*$ . This is, for instance, the case of the regression models that are used to predict survival (Kaplan and Meier, 1958).

In this PhD we adopted the decision of approaching most of the biomedical problems as supervised classification problems, since we found it is a more versatile way of modeling.

Besides, the uncertainty related to a classification problem might be lower than the corresponding regression problem, since predicting the unknown class of a sample is generally a better-posed problem than predicting the value of given decision variable. Nevertheless, to properly understand the supervised classification problems, we have cast them as general inverse or parameter identification problem.

A medical problem, posed as a classification problem, consists in a set of patients that have a given peculiarity, such as a disease condition or treatment response, which is described by its corresponding class value established by medical experts, who usually want to know what is causing those peculiarities by comparing them with other types of patients, known as healthy controls. The classification problem does not need necessarily to be binary, that is, it could be multi-class.

The first source of uncertainty in a classification problem comes from the fact that the perfect classifier is usually a priori unknown, that is, no physical relationship is at disposal to predict the class of the observed data. Accordingly, the classification problem is nonlinear, since the classifier and the features that serve to achieve an optimum prediction are unknown. That way, a classification problem can be typically catalogued as a non-linear inverse problem (Aster et al., 2012). Therefore, as a first step, a given type of classifier (nearest-neighbor, neural networks, SVM, ...) should be built ad-hoc. This can be considered an additionally source of uncertainty.

Let us imagine that we have at disposal a set of  $n$  features (clinical data, genetic expressions, ...) for a set of  $m$  samples whose classes were provided by medical expert annotations. This information is typically organized in the matrix  $E \in M_{m \times n}(\mathbb{R})$ , usually with  $m \ll n$ , and in the class vector  $\mathbf{c}^{obs} \in \mathbb{R}^m$ . The classifier,  $L^*(\mathbf{f})$  can be formally defined as an application between the set of features  $\mathbf{f} \in M \subset \mathbb{R}^s$  and the set of classes  $C = \{c_1, c_2, \dots, c_n\}$ :

$$L^*(\mathbf{f}) : \mathbf{f} \in \mathbb{R}^s \rightarrow C = \{c_1, c_2, \dots, c_n\}. \quad (1.1)$$

However, not all the features are involved in the inverse problem. Furthermore, when all the variables parametrizing the samples are considered, the corresponding classification problem becomes nonlinear separable, that is, it is not possible to define in the feature space a set of hyperplanes that optimally separates the samples.

Importantly, not all the features provide useful information for the class prediction. These extraneous features are noisy and can be analytically disruptive. Fortunately, it is possible to discard irrelevant features, that is, those that do not provide any useful information for the discrimination, since they introduce ambiguity in the classification. The relevant features would be defined as the ones that minimize a given target function  $O(\mathbf{f})$  related to the class prediction vector:

$$\mathbf{f}: O(\bar{\mathbf{f}}) = \min_{\mathbf{f} \in \mathbb{R}^s} O(\mathbf{f}), \quad (1.2)$$

$$O(\mathbf{f}) = \|\mathbf{L}^*(\mathbf{f}) - \mathbf{c}^{obs}\|_p \quad (1.3)$$

$$\mathbf{L}^*(\mathbf{f}) = (L^*(\mathbf{f}_1), \dots, L^*(\mathbf{f}_i), \dots, L^*(\mathbf{f}_m)), \quad (1.4)$$

where  $\mathbf{c}^{obs}$  is the set of observed classes,  $p$  is the norm applied in the distance criterion,  $\mathbf{L}^*(\mathbf{f})$  is the set of predicted classes,  $\mathbf{f}_i \in \mathbb{R}^s$  is the set of features of size  $s$  corresponding to sample  $i$ , and  $L^*(\mathbf{f}_k)$  is the classifier prediction for sample  $k$ . Otherwise said, the relevant features would be the ones that allow us to predict the class of new incoming samples. This process in machine learning is called generalization.

Three different aspects are particularly relevant in the design of the classifier:

1. The effect of modeling errors.
2. The effect of noise in data.
3. The ill-posed character of the classification problem.

### 1.3.2 The effect of modeling errors

In most biomedical problems the forward problem is unknown, that is, no physical relationship is available relating input and output variables. This translates in classification problems in the fact that the correct classifier  $L_{true}^*$  is a priori unknown.

Let us imagine that the relationship between features and classes is linear:

$$L_{true}^* \mathbf{f}^{true} = \mathbf{c}^{obs}.$$

If we consider a classifier that has a modeling error  $\delta L^*$  and it is related to the true classifier as follows:  $L_p^* = L_{true}^* + \delta L^*$ . Then we have:

$$\begin{aligned} (L_p^* - \delta L^*) \mathbf{f}^{true} &= \mathbf{c}^{obs} \\ L_p^* \mathbf{f}^{true} &= \mathbf{c}^{obs} - \delta L^* \mathbf{f}^{true} \end{aligned}$$

that is, the classifier  $L_p^*$  used in practice to achieve  $\mathbf{f}^{true}$  will need to correct the observed class  $\mathbf{c}^{obs}$  by the term  $\delta L^* \mathbf{f}^{true}$ , which is a priori unknown. Otherwise said, if we solve  $L_p^* \mathbf{f} = \mathbf{c}^{obs}$  then  $\mathbf{f} \neq \mathbf{f}^{true}$ . Only if  $\mathbf{f}^{true} \in \ker(\delta L^*)$  the classifier  $L_p^*$  will achieve  $\mathbf{f}^{true}$  from  $\mathbf{c}^{obs}$ . Obviously this simple analysis is theoretical but explains the importance of choosing the "correct" classifier.

In this PhD dissertation we decided to use the principle of parsimony, that is, between the set of all possible classifiers that could be employed we will try to choose the simplest one. Particularly, we try to avoid the use of wrapper and embedded classifiers, whose design imposes an additional uncertainty analysis, due to the optimization processes that are involved.

### 1.3.3 The effect of noise in data

Biomedical data is notorious for containing noise which has historically contributed to issues around reproducibility, especially as related to clinical/gene phenotype relationships. Noise also impedes accurate mechanistic conclusions, for example in the case of genetic data, by partially falsifying biological pathways. This topic is formally developed in the next section.

There are two main sources of noise:

- First, **noise in the feature data** that is introduced by the process of data treatment (preprocessing techniques) and measurement. The observed feature data of a sample,  $\mathbf{f}^{obs}$ , can be expressed as the sum of the noiseless data  $\mathbf{f}^{true}$  and the measurement noise  $\delta\mathbf{f}$ :  $\mathbf{f}^{obs} = \mathbf{f}^{true} + \delta\mathbf{f}$ . Therefore, using a simple Taylor expansion we get:

$$\begin{aligned} L^*(\mathbf{f}^{obs}) &= L^*(\mathbf{f}^{true}) + \delta L^*(\mathbf{f}^{true}) = \\ &= L^*(\mathbf{f}^{true}) + \sum_{k=1}^s \frac{\partial L^*}{\partial f_k}(\mathbf{f}^{true}) \delta f_k + o(\delta\mathbf{f}), \end{aligned}$$

where  $o(\delta\mathbf{f})$  vanishes when the noise term  $\delta\mathbf{f} \rightarrow 0$ . Therefore, given a classifier  $L^*(\mathbf{f})$ , the noise in the feature data involves a modeling error whose first order approximation is:

$$\begin{aligned} \delta L^*(\delta\mathbf{f}) &= \sum_{k=1}^s \frac{\partial L^*}{\partial f_k}(\mathbf{f}^{true}) \delta f_k \\ &= \nabla L^*(\mathbf{f}^{true}) \cdot \delta\mathbf{f}. \end{aligned}$$

Obviously  $\delta L^*(\delta\mathbf{f}) \rightarrow 0$  when  $\delta\mathbf{f} \rightarrow 0$ . This analysis is theoretical because  $\mathbf{f}^{true}$  and  $\delta\mathbf{f}$  are unknown.

- Secondly, **noise in the class assignment**  $\delta\mathbf{c}$ , typically due to an incorrect labeling of the samples by the medical experts. Therefore the observed class vector can be expressed as the sum of the true class vector  $\mathbf{c}^{true}$  and the class assignment noise  $\delta\mathbf{c}$ :  $\mathbf{c}^{obs} = \mathbf{c}^{true} + \delta\mathbf{c}$ . For instance, sometimes the classification problem is parameterized as binary when in fact there are more than two classes. Therefore, assigning two

different classes to the samples will input noise in the classification. In this case, finding a predictive accuracy lower than 100% would be the expected result, otherwise the algorithm will find a wrong set of features in order to fit (or explaining) the wrong class assignment. Obviously this situation is always difficult to detect, since the strategy that one might expect consists in achieving a perfect classification, and overfitting the noise. This is not the point of view presented herein.

It is straight forward to show that both kinds of noise ( $\delta\mathbf{f}$  and  $\delta\mathbf{c}$ ) induce a modeling error in the classifier. In the case of class assignment noise the cost function writes:

$$\begin{aligned} O^p(\mathbf{f}) &= \|L^*(\mathbf{f}) - \mathbf{c}^{obs}\|_p = \\ &= \|L^*(\mathbf{f}) - \mathbf{c}^{true} - \delta\mathbf{c}\|_p = \\ &= \|L^*(\mathbf{f})\|_p + \delta L^*(\mathbf{f}) = O^f(\mathbf{f}) + \delta L^*(\mathbf{f}), \end{aligned}$$

where  $O^p(\mathbf{f})$ ,  $O^f(\mathbf{f})$  stand respectively for the perturbed and noise-free cost functions, and  $\delta L^*(\mathbf{f})$  for the modeling error term induced by the noise in the class assignment. For instance, if the squared Euclidean norm is used to define the cost function, we have:

$$\begin{aligned} O^p(\mathbf{f}) &= \|L^*(\mathbf{f}) - \mathbf{c}^{obs}\|_2^2 = \|L^*(\mathbf{f}) - \mathbf{c}^{true} - \delta\mathbf{c}\|_2^2 = \\ &= (L^*(\mathbf{f}) - \mathbf{c}^{true} - \delta\mathbf{c})^\top (L^*(\mathbf{f}) - \mathbf{c}^{true} - \delta\mathbf{c}) = \\ &= \|L^*(\mathbf{f}) - \mathbf{c}^{obs}\|_2^2 - 2(L^*(\mathbf{f}) - \mathbf{c}^{true})^\top \delta\mathbf{c} + \delta\mathbf{c}^\top \delta\mathbf{c}. \end{aligned}$$

Therefore the modeling error is in this case:

$$\delta L^*(\mathbf{f}) = \delta\mathbf{c}^\top \delta\mathbf{c} - 2(L^*(\mathbf{f}) - \mathbf{c}^{true})^\top \delta\mathbf{c}$$

and  $\delta L^*(\mathbf{f}) \rightarrow 0$  when  $\delta\mathbf{c} \rightarrow 0$ .

In presence of these types of noise the set of features with the highest predictive accuracy (and therefore the lowest misfit error) will never perfectly coincide with the set(s) of features that explains the disease (noise-free classification problem). For that reason it is desirable to look also for other sets of features with lower predictive accuracy than the optimum. Besides, the classifier  $\mathbf{L}^*$  is built ad-hoc and it is just a mathematical abstraction used to discover the features that are involved in the discrimination problem, but it is not the reality itself. As it has been shown in section 1.3.2 devoted to the analysis of the modeling errors.

### 1.3.4 The ill-posed character of the classification problem

Typically the number of samples is finite due to economic constrains (hundreds of samples) and the number of prognostic variables (genes or genetic probes) is much higher (hundred of thousands). The ill-posed character of the classification is due to the high underdetermined character of the inverse problem involved. This is not always the case when working with hospital data when the number of variables is usually lower than the number of samples. However, it is not necessarily the case that these variables would carry enough information about the decision problem that is going to be solved.

Addressing this analysis as a linear system:

$$L^* \mathbf{f}^{true} = \mathbf{c}^{obs}, \quad L^* \in M_{m \times n}(\mathbb{R}),$$

the main typologies of these problems would correspond to:

1. **Case of genetic data:**  $m \ll n$  and  $\exists \ker(L^*)$  whose dimension is  $n - \text{rank}(L^*)$ . The  $\ker(L^*)$  forms the uncertainty space of the classifier. If the  $\text{rank}(L^*) = m$ , that is the samples are independent, the problem is purely underdetermined. Then, the minimum norm solution applies:

$$\mathbf{f}_{MN} = L^{*\top} (L^* L^{*\top})^{-1} \mathbf{c}^{obs}.$$

In this case the principle of parsimony applies since  $\mathbf{f}_{MN}$  has not component on the  $\ker(L^*)$ . If  $\text{rank}(L^*) = r < m$ , the samples are redundant, and the kernel of the classifier increases its dimension to  $n - r$ . In any case the problem is highly underdetermined.

2. **Case of the hospital data:**  $m > n$  and  $\text{rank}(L^*) = r \leq n$ . In the case where  $\text{rank}(L^*) = n$ , the problem is purely overdetermined, all the prognostic variables will be independent predictors, and the least square solution applies:

$$\mathbf{f}_{LS} = (L^{*\top} L^*)^{-1} L^{*\top} \mathbf{c}^{obs}.$$

The classifier has a null kernel in this case ( $L^*$  is injective), but the uncertainty space still exists. If  $\text{rank}(L^*) < n$ , then the prognostic variables are dependent and the problem becomes rank deficient, similar to the rank deficient underdetermined case.

Fernández-Martínez et al. (2012, 2013) analyzed the uncertainty space of linear and nonlinear inverse and classification problems showing that the topography of the cost function  $O(\mathbf{f})$  in the region of lower misfits (or higher predictive accuracies) correspond to one or several flat elongated valleys with null gradients, where the high predictive sets of features reside. This valley is unique and rectilinear if the classification/inverse problem is linear,

and bends and might be composed of several disconnected basins if the inverse problem is nonlinear and the classification problem becomes nonlinear separable. Also, if we are somehow able to define the discriminatory power of the different features, a classification problem could be interpreted as the Fourier expansion of a signal, that is, there will be features that provide high accuracy for the classification problem alone (head features), while others will assist in expanding the high frequency details (helper features) in order to improve the predictive accuracy. Nevertheless, there is a time when adding more details to the classifier do not increase its predictive accuracy. The smallest scale signature is the one that has the least number of highest discriminatory features. This knowledge could be important for diagnosis and treatment optimization since it allows a fast and cheap data gathering.

As a result of the foregoing we will need a tool able to manage the underlying uncertainty of the problem, due to its ill-posed character. Moreover, the tool will must be robust against the different sources of data noise. Consequently and in response to such challenge, we developed a methodological framework called Biomedical Robots.

### 1.3.5 Biomedical robots

We defined a biomedical robot as the ensemble of methodologies and bioinformatics algorithms, derived from applied mathematics, statistics and computer science that are capable of dynamically analyzing high dimensional data, discovering knowledge, generating new biomedical working hypothesis, and supporting medical decision making with its corresponding uncertainty assessment. It is important to remark that we are not interested in building a black-box methodology, but being able of inferring the mechanisms of action that are involved in the specific biomedical problem.

Figure 1.1 shows a conceptual scheme of how biomedical robots can be generated and applied for instance to a phenotype prediction problem. From a training data set we built  $N_r$  robots. The robots is in this case are a set of classifiers characterized by their small-scale set of features  $\mathbf{f}$ , and their corresponding set of parameters needed to perform the classification of the incoming samples. These robots will be deduced from the dataset by applying different supervised feature selection methods and dimensional reduction algorithms. Each robot will be also characterized by its predictive accuracy according to the classification cost function  $O(\mathbf{f})$  built in a testing dataset. The design of the cost function is important because the sets of features found might depend on that design. The average error will depend on the type of experiment (Cross-Validation, Hold-Out with repetitions, . . . ) that we used to define the cost function.



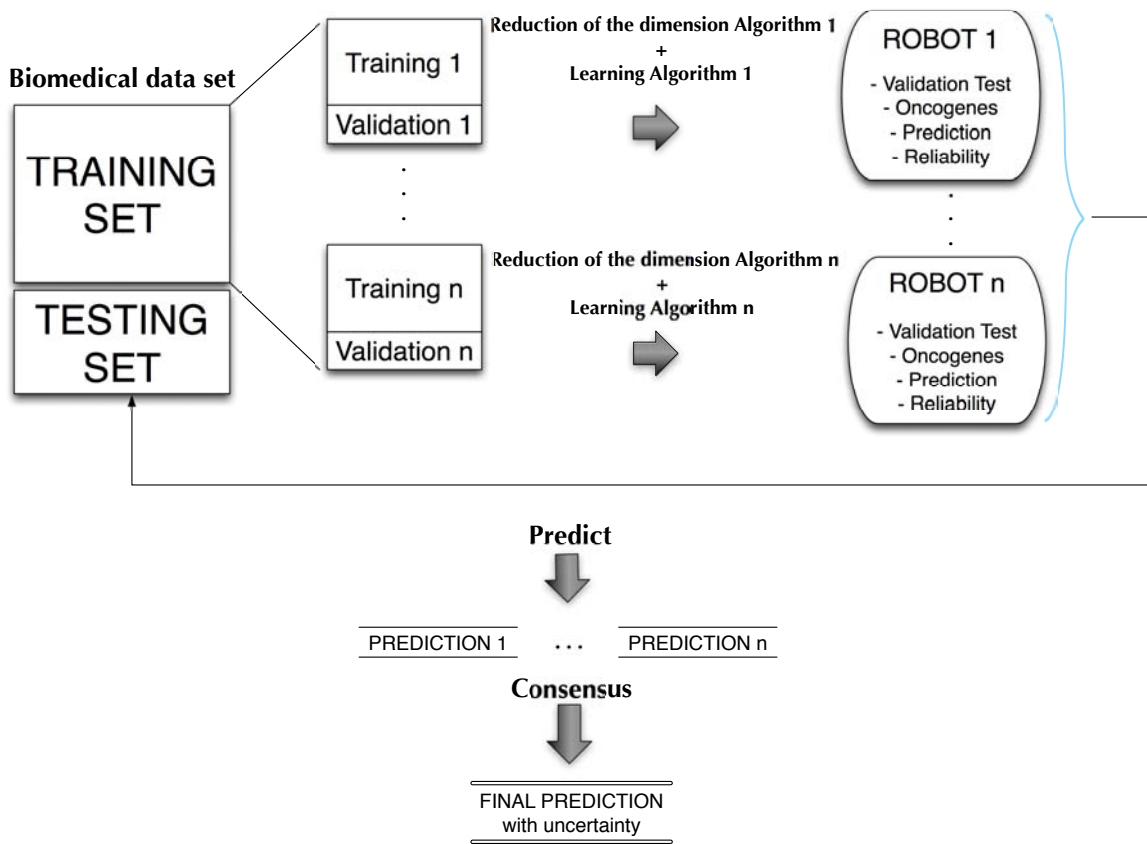


Fig. 1.1 Conceptual scheme for the design of biomedical robots.

This design enables the sampling of the uncertainty space corresponding to the classification problem. This can be shown through a simple linear regression problem. Let us suppose that we have at disposal a set of points  $\{(x_1, y_1) \dots (x_m, y_m)\}$  and we define the linear estimator  $y^*(x; m, b) = mx + b$ . The model parameters  $(m, b)$  are found by least squares, solving the normal equations.

$$F^T F \begin{pmatrix} b^{ls} \\ m^{ls} \end{pmatrix} = F^T \mathbf{y},$$

with  $F = [\mathbf{1}, \mathbf{x}]$ ,  $F \in M_{m \times 2}(\mathbb{R})$ . Matrix  $F^T F$  is symmetric, therefore, it admits orthogonal diagonalization:

$$F^T F = V \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} V^T, \quad \lambda_1 > \lambda_2 > 0.$$

Besides  $F$  can be written using SVD:  $F = U \Sigma V^T$ . The least square solution is:

$$\begin{aligned} \begin{pmatrix} b^{ls} \\ m^{ls} \end{pmatrix} &= (F^T F)^{-1} F^T \mathbf{y} = \\ &= V \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix} V^T V \Sigma^T U^T \mathbf{y} = \\ &= V \begin{pmatrix} y_{1u}/\sqrt{\lambda_1} \\ y_{2u}/\sqrt{\lambda_2} \end{pmatrix} = \frac{y_{1u}}{\sqrt{\lambda_1}} \mathbf{v}_1 + \frac{y_{2u}}{\sqrt{\lambda_2}} \mathbf{v}_2, \end{aligned} \quad (1.5)$$

where  $y_{1u}, y_{2u}$  are the two first coordinates of vector  $\mathbf{y}$  referred to the orthogonal basis set  $U$  of  $\mathbb{R}^m$ , and  $\mathbf{v}_1, \mathbf{v}_2$  are the eigen vectors of  $F^T F$ .

Two considerations are relevant:

- The conditioning of the normal equations depends on the ratio  $\lambda_1/\lambda_2$  and the region of linear equivalence of value  $tol$  (Fernández-Martínez et al., 2012) referred to the  $V$  base is:

$$\frac{(b - b^{ls})_V^2}{\left(\frac{tol}{\sqrt{\lambda_1}}\right)^2} + \frac{(m - m^{ls})_V^2}{\left(\frac{tol}{\sqrt{\lambda_1}}\right)^2} = 1.$$

Therefore, the axis of the maximum uncertainty corresponds to the direction  $\mathbf{v}_2$  associated to the smallest eigenvalue of  $F^T F$ ,  $\lambda_2$ , and the center of the ellipse is the least squares solution of the linear system. The noise is amplified mainly in the direction

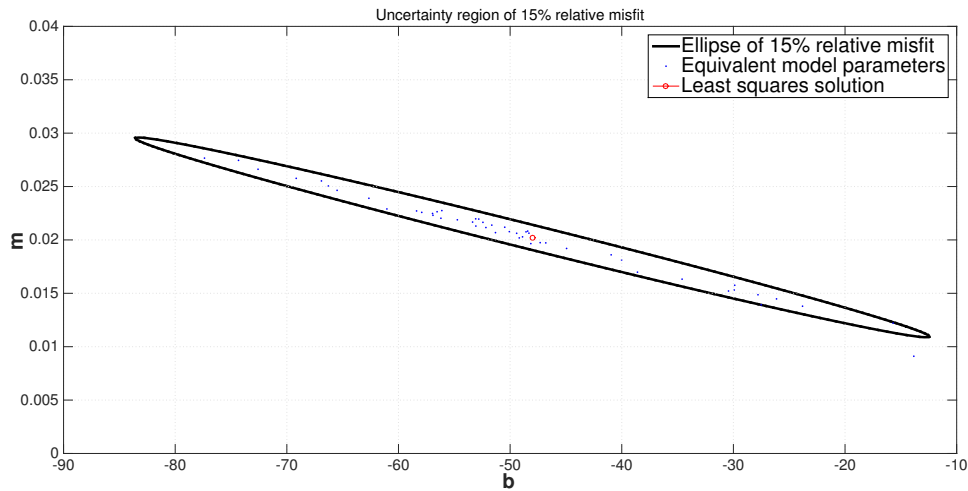


Fig. 1.2 Uncertainty space in a 2D ill-conditioned linear regression problem.

of  $\mathbf{v}_2$ , perturbing the location of the least squares solution (Fernández-Martínez et al., 2014a,b).

- Taking into account relationship (1.5) we see that the least squares solution will change if we consider different bags of the training dataset. These solutions belong to region of uncertainty of the linear regression problem. The same can be concluded for the biomedical robots in a specific classification problems. This idea is numerically illustrated in figure 1.2 for a 2D ill-conditioned linear regression problem where all the equivalent model parameters are sampled along the maximum uncertainty axis direction, using different training data bags.

The final decision approach is as follows: given a new incoming sample, each of the equivalent robots will perform a prediction. A final prediction with its uncertainty assessment will be given using all these predictions via a consensus strategy such as majority voting. Ensemble classification and majority vote decisions are based on Condorcet's jury theorem, which is a political science theorem about the probability of a given group of individuals arriving at a correct decision (Ladha, 1992). In the context of biomedical robots and ensemble learning, it implies that the probability of being correct for a majority of independent voters is higher than the probability of any of the individual voters, and tends to 1 when the number of voters (or weak classifiers) tends to infinite. In this case the weak classifiers are any of the biomedical robots of the ensemble that have a high predictive accuracy. These classifiers are guaranteed to be independent since they use different high discriminatory set of features, measured by their corresponding discriminatory power.

More in detail, the algorithm for building a biomedical robot consists in three main steps:

1. Applying several filter methods to find different lists of high discriminatory features.
2. Establishing the predictive accuracy of these lists of features using a validation cost function (cross-validation accuracy for instance) via any machine learning classifier (like a k-Nearest-Neighbor k-NN). This sampling procedure of the prediction uncertainty space aims at obtaining from these lists different biomedical robots with their corresponding predictive accuracy. For that purpose we can use feature elimination techniques and/or random sampling methodologies.
3. Selecting robots above a certain predictive accuracy (or below a given error tolerance) and performing the consensus prediction through a voting system (like majority voting).

According to the definitions stated in (1.1), (1.2), (1.3), and (1.4) we can formally define a biomedical robot as the set of classifiers:

$$L_{tol} = \{L^*(\mathbf{f}_k) : k = 1, \dots, m\}, \quad (1.6)$$

whose predictive error (the number of misclassified samples) is lower than a given bound  $tol$ . The prediction problem with uncertainty estimation consists in, giving an incoming sample  $\mathbf{s}_{new}$ , applying the set of Biomedical robots  $L_{tol}$  (with predictive accuracy higher than  $(100 - tol)\%$ ) and performing the consensus classification. Supposing that the uncertainty analysis was correctly performed, this procedure also provides the uncertainty in the class prediction. For instance if the class vector is composed by  $n$  classes, the probability of  $\mathbf{s}_{new}$  to belong to class  $c_i$  is calculated as the number of robots that predicted the sample to belong to class  $c_i$  divided by the total number of selected robots in the set  $L_{tol}$ .

### 1.3.6 Feature selection

Following the two first steps of the algorithm for building a biomedical robot, we present in this section the methods we applied for filtering and selecting features and, consequently form essential part of the biomedical robot.

There are different kind of feature selection methods. In the case of filter methods, the feature selection and the classifier for the prediction are independent (uncoupled). However, wrapper and embedded techniques are most sophisticated approaches where the selection is the solution of an optimization problem; therefore selection and classification are coupled. Wrapper and embedded methods usually involve the use of neural network, support vector machines, decision trees and global optimization algorithms. Filter methods rank different features according to different measures of their discriminatory power in phenotype prediction

problems. Besides, filtering/ranking methods provide clear interpretation, low computational cost, and the possibility of being applied to both, discrete and continuous variables. However, other types of filtering/ranking algorithms could be used. A survey about feature selection methods can be consulted in (Saeys et al., 2007). These algorithms can be easily generalized for multiclass classification problems. A future work will be devoted to this important subject. In the present case we did not need to tackle this problem since all the cases were modeled as binary.

Firstly, features are first ranked according to different filter/ranking methods for binary classification problems:

- Maximum Fisher's ratio (Fisher, 1936; Yang and Mao, 2011): The Fisher's ratio (FR) of a feature  $j$ , in a two-class problem,  $c_1, c_2$ , is defined as follows:

$$FR_j(c_1, c_2) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \quad (1.7)$$

where,  $\mu_{j1}$ , and  $\mu_{j2}$  are measures of the center of the distribution (means) of feature  $j$  in classes 1 and 2, and  $\sigma_{j1}^2$ , and  $\sigma_{j2}^2$  are measures of the dispersion (variance) within these classes. This method looks for prognostic features that separate the classes further apart and are very homogeneous within classes (low intra class variance).

- Fold Change (Schena et al., 1996): The Fold Change (FC) of a feature  $j$  is defined as follows:

$$fc_j(c_1, c_2) = \log_2 \frac{\mu_{j1}}{\mu_{j2}} \quad (1.8)$$

where,  $\mu_{j1}$ , and  $\mu_{j2}$  are measures of the center of the distribution (means) of feature  $j$  in classes 1 and 2. This method selects features according to their absolute FC value  $|fc_j(c_1, c_2)|$ .

- Minimum class Entropy (Quinlan, 1993; Shannon, 1948): Entropy (EN) is a measure of the number of specific ways in which a system may be rearranged, and it is often considered a measure of disorder, or progression towards thermodynamic equilibrium. In the case of a binary classification problem, the entropy of each feature is defined as follows:

$$E_j(c_1, c_2) = - \sum_{k=1}^2 \sum_{j=1}^{N_c} p_{kj} \log_2 p_{kj}, \quad (1.9)$$

where  $N_c$  are the number of bins used to describe the probability distribution of feature  $j$  in class  $k$ , and  $p_{kj}$  is the probability that this feature takes the center class value

$x_{kj}$ . The algorithm to compute the entropy is based in ordering the features according to their value and calculating the mismatch to the class vector. A perfect ordering occurs when the values correspond perfectly to the class vector. Features with higher ordering (or lower entropy) are therefore the most discriminatory. The algorithm used for Entropy ranking in this PhD is a simpler modification of this method, and it is based on the optimum order of prognostic feature with respect to the class vector.

- **Maximum Percentile Distance (MPD):** This a novel method proposed here in and it is based on selecting the features with higher distances between the corresponding cumulative probability functions (percentile array) within each class, defined for feature  $j$  as follows:

$$d_j(c_1, c_2) = \frac{\|\mathbf{p}_{j1} - \mathbf{p}_{j2}\|_2}{\max(\|\mathbf{p}_{j1}\|_2, \|\mathbf{p}_{j2}\|_2)}, \quad (1.10)$$

where  $\mathbf{p}_{ji}$  stands for the percentile vector  $j$  in class  $i$ , and  $\|\mathbf{p}_{ji}\|_2$  its Euclidean norm. Percentiles vary from 5 to 95 to avoid the possible effect of outliers. This method can be considered as a generalization of a Mann-Whitney selection test, which is only based in the median (percentile 50).

- **Significance Microarray Analysis (SAM Tusher et al. (2001)):** SAM uses as score the absolute difference between the means in both classes divided by the sum of the total standard deviation ( $\sigma_j^T$ ) and a tunable exchangeability factor ( $\sigma_{j0}$ ) used to damp the effect of outliers, that is, genes with very small  $\sigma_j^T$  that will bring an anomalous score:

$$SAM_j(c_1, c_2) = \frac{|\mu_{j1} - \mu_{j2}|}{\sigma_j^T + \sigma_{j0}}. \quad (1.11)$$

Once the most discriminatory features are determined and ranked in decreasing order by their discriminatory power, the aim is to determine the shortest (having the smallest number of features) list of prognostic features with the highest predictive accuracy. The algorithm to find the minimum-size list of features we chose is the Backwards Feature Elimination (BFE), which is similar to the Recursive Feature Elimination algorithm proposed by Guyon et al. (2002). Feature elimination tries to unravel the existence of redundant or irrelevant features to yield the smallest set of prognostic features that provide the greatest possible classification accuracy.

The BFE algorithm works as follows:

1. Beginning by the tail of the ranked list of prognostic features, the algorithm iteratively generates increasingly shorter lists by eliminating one prognostic feature at a time, calculating their classification accuracy.
2. Finally, the list with the optimum accuracy and minimum size is therefore selected.

This way of proceeding is based on the following idea: prognostic features with higher discriminatory ratios span low frequency features of the classification, while features with lowest discriminatory ratios account for the details in the discrimination (high frequency features). This method determines the minimum amount of high frequency details that are needed to optimally discriminate between classes.

For the predictive accuracy estimation, we applied a Leave One Out Cross-Validation experiment (LOOCV), using the average distance of the reduced set of features to each training class set. The goal of cross-validation is to estimate how accurately a predictive model (classifier) will perform in practice. LOOCV involves using a single sample from the original dataset as the validation data (sample test), and the remaining samples as training data for each fold until all the samples were predicted. The class assignment is based in a nearest-neighbor classifier in the reduced base, that is, the class with the minimum distance in the reduced base to the sample test is assigned to the sample test. As the clinical data has a heterogeneous character, the Euclidean distance is not always an appropriate metric, since it works well with continuous attributes. Therefore, the average LOOCV predictive accuracy is calculated by iterating over all the samples using the Heterogeneous Value Difference Metric (HVDM) (Wilson and Martinez, 1997). This metric in the case of continuous variables coincides with the Euclidean distance between the corresponding normalized variables. For that purpose the weights used to normalize the variables are the inverse of two times the prior variability (standard deviation) of the prognostic features. These weights serve to scale the different kinds of measurements into approximately the same range in order to give to each variable a similar influence on the overall distance measurement. The distance between a new sample  $\mathbf{s}_{new}$  and the average signature  $\mathbf{m}_j$  in class  $j$  is:

$$d(\mathbf{s}_{new}, \mathbf{m}_j) = \|\omega(\mathbf{s}_{new} - \mathbf{m}_j)\|_2, \quad (1.12)$$

where  $\omega$  is a diagonal matrix with  $\omega(k, k) = \frac{1}{2\sigma_k}$ , where  $\sigma_k$  is the standard deviation of the  $k$ -th discriminatory prognostic variable. Although, other more sophisticated classifiers could be used like SVM (Vapnik, 1995), ELM (Huang et al., 2006) or Proximal algorithms (Parikh and Boyd, 2013), we decided to use the above explained classifier due to its simplicity and clear interpretation.

In this procedure the feature selection method is executed only once using all training samples before estimating the accuracy by means of a leave-one-out procedure. For each new sample the classifier computes the average distance to the training samples of each class, being  $d_1$  the average distance to class 1, and  $d_2$  the average distance to class 2. Based on these distances the probability of a new sample  $\mathbf{s}_{new}$  to be in class 1 can be written as:

$$P(\mathbf{s}_{new} \in c_1) = \frac{d_2}{d_1 + d_2}. \quad (1.13)$$

The procedure to decide the class assignment is as follows:

$$\mathbf{s}_{new} \in c_1 \iff P(\mathbf{s}_{new} \in c_1) > p_{th} = 0.5. \quad (1.14)$$

Otherwise,  $\mathbf{s}_{new} \in c_2$ . The threshold probability  $p_{th}$  can be considered as a continuous variable to establish the Receiver Operator Characteristic (ROC) curve for this classifier (Swets, 1996). Finally, the reduced base might be tested over different randomly chosen training and testing dataset, and averaging the results over a set of independent simulations.

Although this simple classifier seems to be similar to a nearest neighbor algorithm (k-NN), it is not obviously the same, since neither the centroid definition of the distributions, nor the way of adopting the decisions coincide. Notice that in this process, the feature selection method is executed only once using all training samples, before estimating the accuracy by means of a leave-one-out procedure. Our goal is to study the effectiveness of feature selection methods in finding the groups of prognosis variables with higher predictive accuracy. Also, if the feature selection process was performed each time the classifier was executed (i.e. in each of the folds of the leave-one-out), different sets of features would be obtained, thus, it would more difficult to assess the goodness of any concrete group of prognosis features. The only way will be performing frequency analysis of the selected prognostic variables and applying BFE to this set of variables ranked by decreasing order of their posterior frequency. Besides, since the accuracy is established by LOOCV the selected features within each fold of the LOOCV will not be so different from selecting them using the whole dataset, considering that the training set of each of fold in a LOOCV is composed by all the samples but one.

Finally, following the steps described in section 1.2 we presented above the results of applying the developed methods to different types of biomedical data: Clinical and genetic data. Then we present the sensitivity analysis of the methodology against the main sources of noise and how is affected by the main preprocessing techniques. Finally, we describe the results of applying Biomedical Robots to phenotype prediction problems.



# Chapter 2

## Application to clinical data

### 2.1 Introduction

Clinical data consists of health records, clinical trials or disease registries. They are usually retrieved in hospitals or clinics by the specialist. They have an heterogeneity character and they frequently present different sampling frequency. Namely, they express different values in different measures with different bounds, and they are not usually available in all the samples/patients, that is, there are some clinical data that have not been retrieved for some patients. These heterogeneity makes data preprocessing techniques the clue for solving the problem. It will be of paramount importance finding the appropriate normalizing and imputing methods in order to correctly address problems related to clinical data.

Following the steps described in section 1.3.5, "we developed and applied different learning, dimensionality reduction, global optimization and classification algorithms to clinical data gathered in different hospitals". Firstly, we tackled the prediction of two decision making problems that are very common on patients with Chronic Lymphocytic Leukemia: The need of chemotherapy treatment, and the Autoimmune disease occurrence. This work in collaboration with Cabueñas Hospital (Gijón, Asturias, Spain), was reflected in a paper called: "Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems". Secondly, in collaboration with eight hospitals in Asturias we addressed the prediction of the treatment response in patients with Hodgkin Lymphoma. As a result we published a manuscript titled "On the prediction of Hodgkin Lymphoma treatment response". In both cases we developed and applied the different classification and feature selection algorithms described in section 1.3.6. In the case of Hodgkin Lymphoma we also applied optimization techniques that allowed us to improve the classifier, taking into account the confusion matrix and the ROC curve. The simplicity of these methods in both cases, gave us the possibility of implementing them in platforms like spreadsheets, as well as, allowing

an easy understanding for medical doctors. This is the cornerstone of the whole methodology and the key for taking the results of the research work to the hospital and laboratory side (translational medicine).

This chapter is structured in three parts. Firstly we present the common methodology applied in both practical cases: Chronic Lymphocytic Leukemia and Hodgkin Lymphoma. Secondly we introduce the problems addressed for the Chronic Lymphocytic Leukemia and present the results and conclusions. Finally, we proceed in the same way with the Hodgkin Lymphoma case.

## **2.2 Methodology applied to both practical cases using clinical data**

The common methodology applied in both practical cases is composed of three main steps: (1) Data pre-processing, (2) Feature selection and (3) Risk assessment. Figure 2.1 shows the flowchart of the methodology. Moreover, an additional step only applied in the case of Hodgkin Lymphoma, is shown in dashed box. The feature selection step is explained in detail in section 1.3.6.

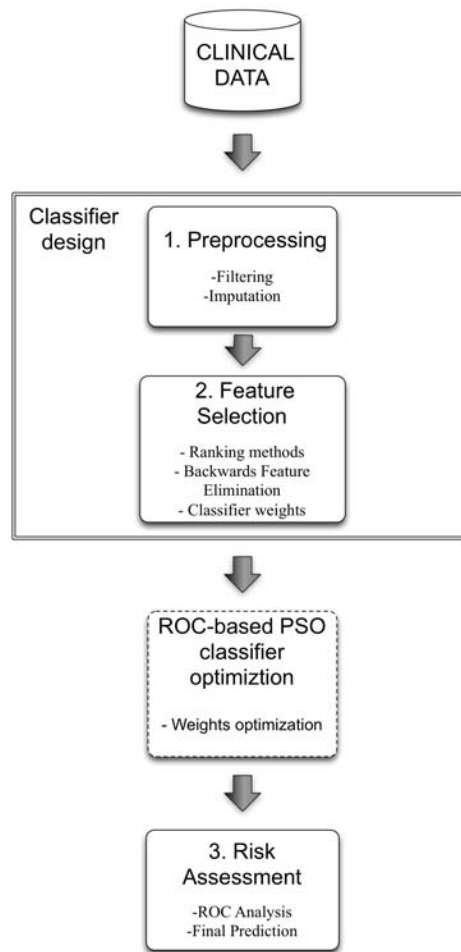


Fig. 2.1 Flow diagram for the prediction model. The methodology is composed of three steps: 1) Data pre-processing, 2) Feature selection and 3) Risk assessment. ROC-based PSO classifier optimization step is only applied in the case of Hodgkin Lymphoma. The different sub steps are also detailed.

### 2.2.1 Data pre-processing

Data preprocessing is applied to improve the quality of data used for performing feature selection, prediction and optimization. It includes two main sub steps that can be applied or not depending on their impact on the prediction:

- **Filtering:** All the features that had certain number of missing values (sampling frequency) are removed. The filtering cut offs used were 30, 40 and 50%.
- **Imputation:** This technique consists in interpolating all the missing values using a Nearest-Neighbor algorithm (Troyanskaya et al., 2001). Given a partially-informed sample (with missing values) the algorithm finds the closest sample within the set of

fully-informed samples and gives the values of the missing variables in this closest sample to the imputed sample. The similarity between samples is measured using the standard Euclidean dot product in N-dimensional vector spaces, where N is the number of fully-informed variables. This way of interpolation has the advantage of not introducing additional outliers that are not originally present in the dataset before imputation. Although the success of the different imputed algorithms might be data-driven, imputing the data improved the accuracy in the predictions and did not alter the prognostic variables that were involved providing shorter lists with higher discriminatory power.

The imputation algorithm is as follows:

1. Finding the subset  $S_{fi}$  of samples (patients) that are fully-informed for all the control variables.
2. For each patient  $k$  that is not fully-informed, finding the set of variables  $\mathbf{m}_k(\text{var}_1 : \text{var}_q)$  that are missed. These variables are interpolated using the values of the same variables corresponding to the nearest fully-informed patient  $f_k$  in  $S_{fi}$ :

$$\mathbf{m}_k^*(\text{var}_1 : \text{var}_q) = \mathbf{m}_{f_k}(\text{var}_1 : \text{var}_q).$$

3. To measure the similarity between patients we use the cosine criterion induced by the Euclidean scalar product defined over the set of fully-informed variables in the current sample (patient):

$$\cos(\mathbf{m}_k, \mathbf{m}_j) = \frac{\mathbf{m}_k \cdot \mathbf{m}_j}{\|\mathbf{m}_k\|_2 \|\mathbf{m}_j\|_2},$$

where  $\mathbf{m}_k$  and  $\mathbf{m}_j$  stand for the vectors of fully-informed variables in patients  $k$  and  $j$ .

### 2.2.2 Risk assessment

In the feature selection step (see section 1.3.6), maximizing the predictive accuracy according to the LOOCV criterion allowed to determine the best reduced base of prognostic variables. However, it is also important to analyze the confusion matrix, obtained from the set of predictions of the training set using the LOOCV method. The confusion matrix is composed by: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). From the confusion matrix we can calculate different rates that are very useful to understand the risk in the prediction:

- True Positive Rate or Sensitivity (TPR): measures the proportion of actual positives that are correctly predicted as such.
- True Negative Rate or Specificity (SPC): measures the proportion of negatives that are correctly predicted as such.
- Positive Predicted Value (PPV): is the proportion of positives values that are true positives.
- False Positive Rate (FPR): fraction of false positives out of the total actual negatives.
- False Negative Rate (FNR): fraction of false negatives out of the total actual positives.
- False Discovery Rate (FDR): fraction of false positives out of the total actual positives.

Based in these rates it is possible to construct a Receiver Operating Characteristic curve (or ROC curve), which is a graphical plot that illustrates the performance of a binary classifier as a function of the cut-off probability. This idea allowed us to create a ROC methodology for this simple distance-based classifier. The curve is created by plotting the TPR against the FPR or fall-out. A perfect classifier has as ROC curve the step function at the origin. ROC analysis is related to cost/benefit analysis of diagnosis/prognosis/treatment decision making. TPR and SPC values are important due to the impact on the patients of the decision taken by physicians.

The selected attributes are used to provide simple biomedical discriminatory rules for diagnosis and prognosis since for each classification problem we provide the bounds for the four groups of the confusion matrix. This knowledge can be used by the physicians in their decision-making process. Additionally to the LOOCV results, we also performed the mean accuracy obtained for 100 random holdouts 75/25 (75% for training and 25% for testing). In any case, and independently of how the predictive accuracy is established, it is crucially important to understand that there exist different combinations of prognostic variables with similar predictive accuracy whose knowledge might be useful to understand the genesis of the problem from a medical point of view. The existence of these different lists is related to the uncertainty analysis of the solutions in any decision-making problem (Fernández-Martínez et al., 2012, 2013).

It is possible to optimize the TPR and/or TNR (improving at the same time the overall predictive accuracy) by optimizing the parameters of the classifier. The idea is to balance/improve the confusion matrix by optimizing the prior weights assigned by the HVDM metric to the best reduced-base that has been found applying the LOOCV approach. This optimization was performed in the Hodgkin Lymphoma problem via a powerful family of

Particle Swarm Optimizers (PSO, Fernández-Martínez and García Gonzalo (2008); Kennedy and Eberhart (1995)).

Finally, it is remarkable the simplicity of the methodology of selecting the shortest list of prognostic variables that could be easily interpreted by medical doctors to perform prognostic predictions with their corresponding risk assessment. The success of the methodology is not based on the sophistication of the classifier but on selecting the most discriminatory variables in each case and building the classifier based on these variables. By selecting the most important prognostic variables, it has been shown that the classification problem approximates a linear separable behavior. This is also a novel result since the methodology currently used (for instance SVM) acts on the opposite direction by transforming the data into an infinite dimensional feature space where the problem becomes linearly separable. An illustration of this idea is that a polynomial function of the type  $f(x, y) = a_0 + a_1x + a_2y + a_3xy + a_4x^2 + a_5y^2$  becomes an hyperplane in  $\mathbb{R}^6$  if the terms in  $xy$ ,  $x^2$  and  $y^2$  are considered independent variables.

## 2.3 Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems

In this research work we show how using the methodology explained in section 2.2 and clinical data obtained from a large population of well-studied Chronic Lymphocytic Leukemia patients (Gonzalez-Rodriguez et al., 2010) can be efficiently applied to address diagnosis problems in medical practice by capturing the hidden implicit relationships between the clinical variables and the corresponding class of the different patients that have been established by medical experts. The problems to be addressed are the need of chemotherapy treatment and the autoimmune disease development, thereby optimizing the treatment and diagnosis. This work has been published in the "Journal of Biomedical Informatics" (see Appendix A.1).

### 2.3.1 Introduction to Chronic Lymphocytic Leukemia related problems

Chronic Lymphocytic Leukemia (CLL) is the most common adult Leukemia in western countries, and it is characterized by the accumulation of malignant B-cells in blood and lymphoid organs. The clinical course of CLL is highly heterogeneous since the survival of some patients is only slightly affected by the disease, whereas other patients have a progressive disease associated with infectious and autoimmune complications. These progressive patients

have poor prognosis, but they could benefit from an earlier or more intense chemotherapeutic treatment. It has been reported that many poor prognostic factors, due to their high cost and complexity, are not used in most hospitals on regular basis. To overcome this problem in the clinical practice staging systems using few, simple, cheap and accessible clinical variables have been popularized. The Rai staging system (Rai et al., 1975) and the Binet classification (Binet et al., 1981) are useful to predict the prognosis of CLL patients, to stratify them, and to achieve comparisons for interpreting specific treatment results. Staging systems stratify subsets of patients who have significant differences in the overall survival but they fail to identify patients who have a high risk of progression in early stages of the disease. Additionally, no current prognostic factors exist to predict the development of some severe complications such as the development of Autoimmune Diseases (AD), or the need for Chemotherapy Treatment (CT). Consequently, the identification of currently available clinical variables to assess the medical decisions in these CLL-related diagnosis problems is a key goal in the management of this disease.

The development of AD or the need of CT is not known at diagnosis. So far, only with the evolution of the patient during the 5 years follow up, medical doctors can answer these questions. Therefore, the interest of the methodology previously presented consists of being able to predict both CLL related problems at diagnosis. Particularly, AD problem was very hard to predict, and up to our knowledge no previous research was successful to explain this phenomenon using biochemical variables.

### **2.3.2 CLL clinical data**

The CLL clinical data we managed were a cohort of two hundred sixty-five Caucasians who were diagnosed in the Cabueñes Hospital (Gijón, Spain) with CLL between 1997 and 2009. The population distribution by gender and age was the following: 154 males and 111 females, with ages ranging from 42 to 92, and 47 to 94 years old respectively. Clinical characteristics of patients including time for diagnosis to first treatment, need of chemotherapy treatment and appearance of autoimmune complications were also taken into account in this study. Additionally, thirty-six different clinical and biological variables were measured at diagnosis of the disease. Table 2.1 shows the variables description used in this study. Some variables reflect the malignant characteristic of leukemia cells; others measure the immunological characteristics of CLL patients, and some may be associated with the presence or development of autoimmune complications (autoimmune haemolytic anemia and immune-thrombocytopenia). Finally, some of the variables are demographic and biochemical. Most of them have a sampling frequency higher than 80%, however, the reticulocyte count (RET) and ZAP-70 are the ones that show the lowest sampling frequency.

Particularly, ZAP-70 is only sampled in 21.9% of the patients (58 out of 265), showing that this popular CLL prognostic factor is not always available in medical practice. Although some of these variables were not at disposal at diagnosis (LD for instance), they have been used for analytical purposes.

### 2.3.3 CLL results for Chemotherapy treatment

In CLL, there can be some patients that have an indolent disease and they do not require CT. Other patients who present a progressive disease may require an intense CT. The identification of those patients at early stages of the disease with a high risk of rapid disease progression may help to significantly improve their prognosis. Thus, we try to establish the prognostic variables and criteria to assess the need for CT, assuming that the clinical decisions on the 71 (out of 259, therefore there are 6 missing values since the total cohort is 265) patients that have received CT were correct.

Using the methodology explained in section 2.2 we found that Fisher's ratio ranking method provided the minimum-size set of prognostic variables with the highest accuracy of 80.3%: B2M, WBC, ALC and MBC. The True Positives (TP) are formed by the group of patients that need CT (+) and are correctly predicted, and the True Negatives (TN) are formed by the groups of patients that do not need CT (-) and are correctly predicted. Thus, False Positives (FP) are the patients that do not need CT (-) and are not correctly predicted and False Negative (FN) are the patients that need CT (+) and are not correctly predicted.

Figure 2.2 shows the ROC curve and the Recall (or True Positive Rate -TPR) against Precision (or Positive Predicted Value - PPV) curves for several probability thresholds in the CT classification problem. The optimum result ( $p_{th} = 0.47$ ) shows that 63.4% (TPR) of the patients that need CT and 86.7% (True Negative Rate or Specificity - SPC) of the patients that do not need CT were correctly predicted. Besides, with that probability threshold we got a Precision (or Positive Predicted Value - PPV) of 64.3%. Nevertheless, other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery Rate (FDR) was 36.62%. The confusion matrix is shown below:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 45 & 25 \\ 26 & 163 \end{pmatrix} \quad (2.1)$$

CT is recommended in patients with advanced and progressive disease. Thus, the amount of malignant leukemia cells that it is measured by the different counts of leucocytes; particularly WBC (White Blood Cells count), ALC (Absolute Lymphocyte Count) and MBC



Table 2.1 Clinical variables description by group and their corresponding symbols and sampling frequency (Samp. Freq.). Discrete variables are shown in bold faces.

Group	Variable Name	Samp. Freq.	
Biochemical	ALB - Albumin (g/L)	98.49%	
	ALC - Absolute Lymphocyte Count (cells/microL)	100.00%	
	ALP - Alkaline phosphatase (U/L)	95.47%	
	B2M - Beta 2 Microglobulin (mg/L)	93.58%	
	BU - Bilirubin (mg/dL)	96.23%	
	CR - Creatinine (mg/dL)	99.62%	
	GOT - Glutamic-Oxaloacetic Transaminase (U/L)	98.11%	
	GPT - Glutamic-Pyruvic Transaminase (U/L)	99.25%	
	HGB - Hemoglobin (g/dL)	100.00%	
	IgA - Immunoglobulin A (g/L)	96.60%	
	IgG - Immunoglobulin G (g/L)	96.60%	
	IgM - Immunoglobulin M (g/L)	96.60%	
	K - Potasium (mEq/L)	90.94%	
	LDH - Lactate Dehydrogenase (U/L)	96.98%	
	MBC - Monoclonal B cell Count (cells/microL)	90.94%	
	MCV - Mean Corpuscular Volume (fl)	100.00%	
	NA (mEq/L)- Sodium	90.57%	
	NCC - Natural killer Cell Count (cells/microL)	90.94%	
	PLT - Platelets (cells/microL)	100.00%	
	RET - Reticulocyte count (cells/microL)	75.47%	
	SNC - Segmented Neutrophils Count (cells/microL)	100.00%	
	T8C - CD8 T cell Count (cells/microL)	86.42%	
	TLC - Total Lymphocyte Count, CD8 + CD4 (cells/microL)	96.60%	
	UA - Uric acid (mg/dL)	97.36%	
	UR - Urea (mg/dL)	99.25%	
	WBC - White Blood cells Count (cells/microL)	100.00%	
	CLL Specific	<b>CD38</b> - CD38 positive	81.51%
		<b>COOMBS</b> - Coombs test	94.34%
<b>LD</b> - Time for duplication of the number of lymphocytes		96.98%	
<b>MOR</b> - Morphology		98.49%	
<b>MP</b> - Monoclonal Peak		98.87%	
<b>NLymph</b> - Number of affected lymph nodes		99.62%	
<b>SMG</b> - Splenomegaly		99.62%	
ZAP70 - Zeta-chain-associated protein kinase 70 (%)		21.89%	
Personal	AGE - Age	100.00%	
	<b>SEX</b> - Sex	100.00%	

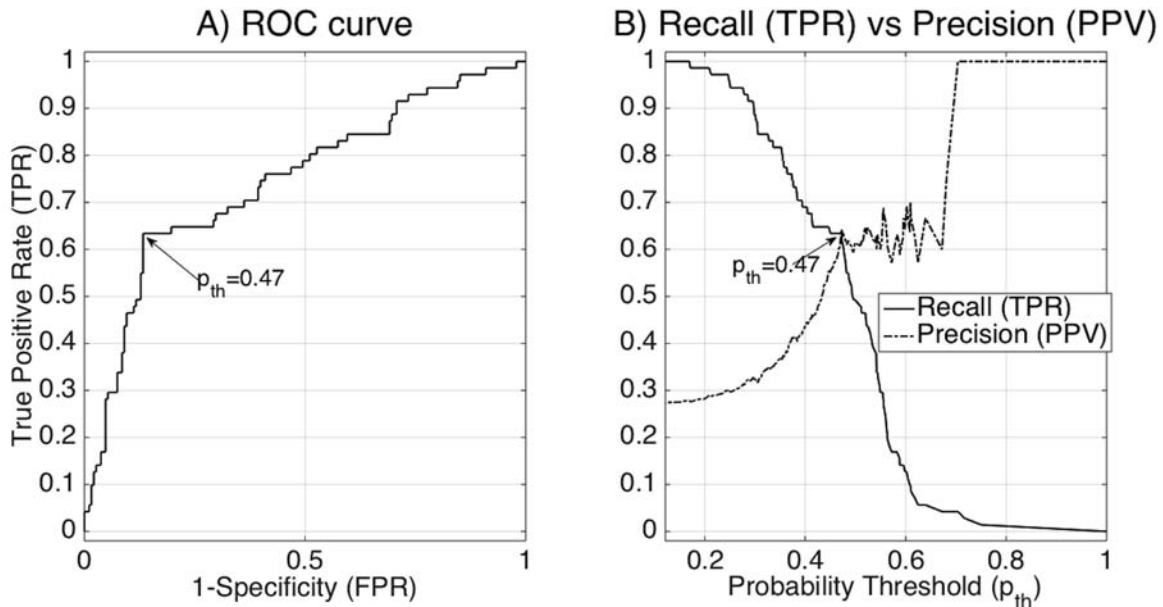


Fig. 2.2 A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Chemotherapy Treatment. The optimum result ( $TPR = 63.4$  and  $PPV = 64.3$ ) is obtained for  $p_{th} = 0.47$ .

(Monoclonal B Cell Count) are key clinical parameters. Nevertheless, these variables are not currently used to select patients who may benefit from CT. On the other hand, AGE, B2M and ZAP70 are traditional clinical parameters that have demonstrated their prognostic importance independently of the clinical stage. Our results also indicated the great prognostic significance of variables that are mainly related with the characteristics of the immune system and are not currently used as prognostic markers in this disease.

Table 2.2 shows the median/mean signatures for the 4 groups of the confusion matrix for the main decision variables found by the methodology. We can observe that there exists a significant distance between the mean signatures of the TP and TN groups, being the median/mean signatures in all the decision variables much higher in the TP group. Moreover, the distance between the median and the mean values of the decision variable distributions is much higher in the TP and in the FP groups, meaning a higher variability in these groups. The mean signatures of the FN group (patients that need CT and are incorrectly predicted) are very close to the mean signatures of the TN group. These patients will never be correctly predicted according to this classifier.

To understand the ambiguity in the CT prediction, it should be taken into account that the criteria used to establish the need of CT (Hallek et al., 2008) sometimes have not correlation with the biological data. The reason is that some patients are diagnosed in early stages

Table 2.2 Chemotherapy Treatment.

Variables	TP	TN	FP	FN
B2M	<b>3.9 / 4.24</b>	2.06 / 2.15	4.37 / 4.58	2.0 / 2.18
WBC (K)	<b>34.1 / 61.8</b>	14.3 / 16.8	18.3 / 28.3	14.2 / 15.5
ALC (K)	<b>24.7 / 47.6</b>	9.0 / 11.2	12.4 / 21.8	8.5 / 10.4
MBC (K)	<b>21.7 / 40.3</b>	6.1 / 8.4	10.1 / 18.4	6.9 / 7.8

This table shows the list of most discriminatory variables with a predictive accuracy of 80.3%. Median and mean values (median/mean) of the prognostic variables for the different groups of the confusion matrix are also given. Variables with (K) are expressed in kilo units. Bold faces indicate the highest value for each prognostic variable in the TP and TN groups. Bounds for the decision correspond to the TP and TN groups.

of the disease when a low burden tumor mass has been detected but they have a very fast progression which implies the need of CT.

### 2.3.4 CLL results for Autoimmune Disease development

In CLL, an autoimmune response against red blood cells (known as autoimmune haemolytic anemia), and an autoimmune response against platelets (known as immune thrombocytopenia) are severe complication of this disease. To the best of our knowledge no prognostic factors capable to predict the presence or development of an autoimmune disease in CLL patients have been currently disclosed. In our cohort only 16 patients (out of 263, therefore there are 2 missing values since the total cohort is 265) have shown autoimmune disorders. This classification problem is highly unbalanced, corresponding to the genesis of the disorder. The classifier has to be able to learn this fact. Some strategies exist to artificially balance the training data set (Chawla et al., 2002; Estabrooks et al., 2004; He et al., 2008; Liu et al., 2006; Ting, 2002), but in this case the results did not improve.

The shortest list of prognostic variables with the highest accuracy (97.3%) was found by the Fisher's ratio method and includes 13 clinical variables: PLT, RET, ALB, HGB, BU, UR, MCV, NCC, K, WBC, LDH, ALC and MBC. The True Positives (TP) group is formed in this case by the patients that present AD (+) and are correctly predicted and True Negatives (TN) correspond to the patients that do not have AD (-) and are correctly predicted. Similarly, the False Positives (FP) are the patients that do not have AD (-) and are not correctly predicted and the False Negatives (FN) correspond to the patients that present AD (+) and are not correctly predicted.

Figure 2.3 shows the ROC and the Recall (or True Positive Rate -TPR) against Precision (or Positive Predicted Value - PPV) curves throughout all possible probability thresholds for

the AD classification problem. The optimum result ( $p_{th} = 0.5$ ) shows that 62.5% (TPR) of the patients that have AD and 99.6% (True Negative Rate or Specificity - SPC) of the patients that do not have AD are correctly predicted. Moreover, over that probability threshold we get a Precision (or Positive Predicted Value - PPV) of 90.1%. However, other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery Rate (FDR) in this case is 9.1%. The confusion matrix is the following one:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 6 & 246 \end{pmatrix} \quad (2.2)$$

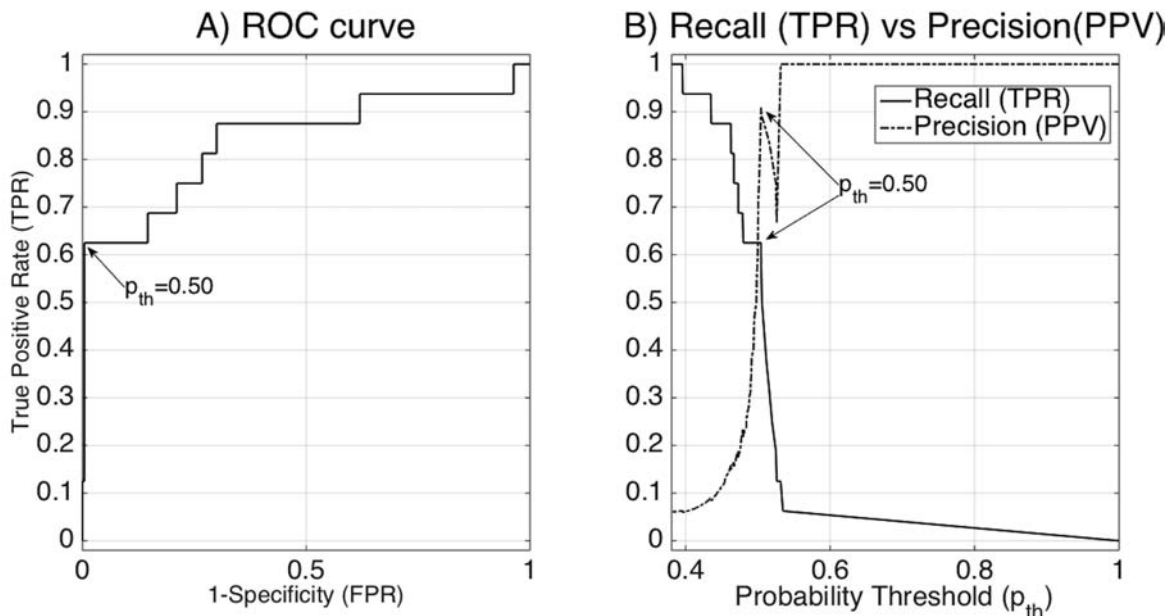


Fig. 2.3 A) ROC curve. B) Sensitivity (or True Positive Rate -TPR) and Precision (or Positive Predicted Value - PPV) for Autoimmune Disease occurrence. The optimum result ( $TPR = 62.5$  and  $PPV = 90.1$ ) is obtained for  $p_{th} = 0.5$ .

Table 2.3 shows the medians and means for the 13 prognostic variables for the 4 groups of the confusion matrix. The differences between the means in TP and TN groups decrease with the Fisher's ratio. Prognostic variables with lower Fisher's ratios (secondary variables) also contribute to improve the discrimination. Except for the main variable, PLT, and the secondary variables HGB and K, the mean and median values are higher in the group with autoimmune disease (TP). The analysis of the two main prognostic variables shows that patients that develop AD and are correctly predicted (TP) have much lower medians and means PLT values (97.7/95.0 Kcells/microL). The normal platelet count lays in the range 150-

Table 2.3 Autoimmune disease development.

Variables	TP	TN	FP	FN
PLT (K)	97.7 / 95.0	<b>191 / 202.2</b>	138	163 /147.2
RET (K)	<b>128.0 / 135.7</b>	67.2 / 69.8	101.3	54.4 /71.8
ALB	<b>42.0 / 40.4</b>	38.0 / 37.4	41.1	39 /39.7
HGB	14.0 / 11.5	14.0 / <b>13.6</b>	13.6	14 / 13
BU	1.0 / <b>1.1</b>	1.0 / 0.6	0.6	1.0 /0.76
UR	<b>52.0 / 64.1</b>	43 / 46.7	49	42 /43.7
MCV	<b>93.0 / 98.1</b>	90 / 89.6	88.9	87 /86.7
NCC	<b>966 / 2251</b>	576 / 741	1657	338 /393.4
K	4.0 / 4.09	4.0 / <b>4.33</b>	4.0	4.0 /4.33
WBC (K)	<b>23.1 / 56.0</b>	15.4 / 24.7	23.6	13.5 /13.9
LDH	<b>360 / 398.1</b>	325 / 343.4	288	333 / 333
ALC (K)	<b>16.1 / 42.2</b>	10.1/ 17.8	18.4	8.5 / 6.7
MBC (K)	<b>10.2 / 36.3</b>	7.3 / 14.2	14.7	5.2 / 4.6

This table shows the list of most discriminatory variables with a predictive accuracy of 97.3%. Median and mean (median/mean) values of the prognostic variables for the different groups of the confusion matrix are also given. FP is composed only by 1 sample (median and mean coincides). Variables with (K) are expressed in kilo units. Bold faces indicate the highest value for each prognostic variable in the TP and TN groups. Bounds for the decision correspond to the TP and TN groups.

450 Kcells/microL, being the average 237 Kcells/microL in men, and 266 in women. On the other hand, the reticulocyte count (RET) in the TP group almost doubles (136 Kcells/microL) the average RET count in patients with no AD (70 Kcells/microL). Median values also show similar tendencies.

The False Positives (FP group) is composed in this case only by 1 sample, whose signature is closer for all the 13 variables to the TP group, except for PLT, RET that are somewhere in between the median/mean values for TP and TN. This fact points out the difficulty of classifying this sample, and it can be proposed as a "biological" outlier. On the other hand, the FN group is composed by 6 samples. The mean PLT count (147 Kcells/microL) of the FN group lies between the mean value for the TP (95 Kcells/microL) and TN (202.2 Kcells/microL) groups. The RET count is however closer to the TN group showing a tendency to very low median values (54.4 Kcells/microL).

These results show the importance of variables associated with the characteristics of platelets and red cells, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia, such as PLT, HGB, MCV and RET. Other variables depend on the presence of autoantibodies or products or symptoms derived from the lysis of blood cells

(BU and LDH). Moreover, some variables associated with the immunological characteristics of patients, such as NCC, constitute a relevant subset of variables that may predict an autoimmune disease occurrence. The association of these variables with an autoimmune disease is not unexpected based on the biology of CLL, but we would like to highlight that no prognostic factors or system may currently predict the development of an autoimmune disease in the clinical practice. To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an autoimmune disease.

### 2.3.5 Conclusions for CLL related problems

Table 2.4 summarizes the main results found for both classification problems (CT and AD): the optimum reduced set of features, the LOOCV accuracy, the Hold Out (HO) mean accuracy over 100 different random simulations using 75% and 25% of samples for training and testing the Sensitivity or True Positive Rate (TPR), and the Specificity or True Negative Rate (SPC) statistics.

Table 2.4 Summary of the results.

Problem	Variables	TPR / SPC	LOO Acc.	HO-100 Acc.
CT(+) Vs. No CT (-)	<b>B2M WBC</b> <b>ALC MBC</b>	63.4% / 86.7%	80.30%	76.10%
AD (+) Vs. No AD (-)	<b>PLT RET ALB HGB</b> <b>BU UR MCV NCC</b> <b>K WBC LDH ALC MBC</b>	62.5% / 99.6%	97.30%	92.80%

Sensitivity or True Positive Rate (TPR) and Specificity or True Negative Rate (SPC) together with the mean accuracy (Acc.) for both experiments leave one out (LOO) and 100 repetitions of a hold-out 75/25 (HO, 75% for training and 25% for testing); and the positive and negative case description of each problem. Bold face indicates the prognostic variables that have been discussed in the text.

The results show that the accuracies are rather high and the difference between both experiments LOOCV and 100 repetitions of a Hold Out (75/25) is quite low, which highlights the robustness of the methodology. In addition, risk assessment ROC curves are provided for each problem and show a good balance between False Positives and False Negatives.

From a medical point of view, the methodology allow the identification of clinical variables obtained at diagnosis of CLL patients, which may predict the development of AD and the need of CT. These variables were obtained at diagnosis of CLL patients on a regular

basis, and consequently, their use does not increase the cost or complexity of the diagnosis in CLL patients.

The need of CT seems to be related to the amount of malignant leukemia cells that are measured by the different leucocytes counts. Although the results concerning these prognostic variables (B2M, WBC, ALC and MBC) are well known in other plasma disorders, this analysis served to conclude that these variables only carry partial information to adopt this important decision, that most of the times, is taken based on criteria that have not correlation with the biological data.

To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an AD, which is a severe and currently unpredictable complication. These results show the importance of variables associated with the characteristics of platelets, reticulocytes and natural killers (PLT, RET and NCC), which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia.

Additionally, the methodology focuses on the relevance of some variables, such as the immunological ones, which may have an important impact on the prognosis of CLL patients, but they are not currently used by hematologists. This analysis has also shown that the low sampling frequency of RET and ZAP-70 could be troubling given their predictive significance in all the problems that have been treated: RET is a key factor for predicting AD, whilst ZAP-70 seems to be important for predicting the need of CT.

In conclusion, the methodology allow an easy accurate prediction of risk in CLL related problems. Moreover, it may establish the relevance of clinical variables that are not widely used as prognostic factor in this disease. The prognostic significance of these variables may probably reflect the relevance of some clinical aspects of this disease that are more important for prognosis than it is currently thought.

This methodology can be adapted to different pathologies as it is shown for the case of Hodgkin Lymphoma.

### 2.3.6 Additional results for survival analysis

Survival analysis is a branch of the applied mathematics that attempts to answer what is the proportion of a population that will survive past a certain time and which are the particular features or characteristics that influence the probability of survival. Particularly the population can include different sub-cohorts with different survival times. The object of primary interest is the survival function  $S(t) = P(T > t)$  which is the probability that the death time  $T$  exceeds a given time threshold  $t$ . Moreover, in survival studies it is also

important the force of mortality or hazard function that provides the instantaneous rate of occurrence of the death.

The Kaplan-Meier estimator (Kaplan and Meier, 1958) can be used to estimate the survival function from lifetime data. The Kaplan-Meier estimator with a large enough sample size approaches the true survival function of a population. For a sample of size  $N$  of a population with observed times until death  $t_1 \leq t_2 \leq \dots \leq t_N$ , the Kaplan-Meier estimator of the survival function is:

$$S^*(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i},$$

where  $n_i$  is the number of survivors prior to time  $t_i$  (when there is no censoring) and  $d_i$  the number of deaths at time  $t_i$ .

Censoring occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. The Kaplan-Meier estimator can be easily adapted to this case taking  $n_i$  as the number of survivors minus the number of censored cases. Kaplan-Meier curves are often used in medical research to measure the fraction of patients living for a certain amount of time after treatment, or to perform the segmentation of a population into subpopulations with different survival times. In our case the aim consists in finding the prognostic variables that better explain the different survival of the CLL population at different time thresholds: 1, 3 and 5 years.

Logistic regression is usually applied to predict survival times. It was developed by D.Cox (1958) to estimate the probability of a binary response based on a set of features. The logistic regression is just a linear regression of the logit of the probability:

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_n x_{in},$$

where  $(x_{i1}, x_{i2}, \dots, x_{in})$  are the attribute values of the sample  $i$  and  $p_i$  its (survival) probability, and  $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_n)$ . The logistic regression implies the solution of the linear system of the kind:

$$\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ & & \vdots & & \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_n) \end{pmatrix}$$

As any regression problem, logistic regression is ill-posed, that is, there exist different set of features or attributes providing a similar predictive accuracy. In addition, all the equivalent features are located for a given error tolerance within the linear hyper-quadratic



(Fernández-Martínez et al., 2012, 2013). Moreover, the ill-condition character of the logistic regression inverse problem comes from the fact that not all the attributes from the samples are relevant for the prediction. In diagnosis, it is important not only finding the optimum prediction of  $p_i$ , but also learning which are the most predictive discriminatory variables.

We approach this problem as a binary classification problem, that is, given a survival time threshold (1, 3 or 5 years in this case) we divide the population into two classes: the ones that survived more than this time threshold, and the ones that did perish before. Censoring is automatically performed since the individuals that are censored are not taken into consideration for the analysis.

### One-year survival

This is a highly unbalanced problem since only 18 patients died (out of 265) during the first year. However, the identification of the subset of patients with risk of such severe disease progression has obviously important clinical consequences. The best prediction was achieved using the following ranking methods and variables: 1. Entropy (94.3%): LD, CD38, SEX; 2. Fisher's ratio (94%): NLymph, MP, MOR and LD; 3. Maximum Percentile Distance (94%): NLymph, MP, LD and SMG. Particularly, the number of affected lymph nodes (NLymph) in patients who died during the first year was higher compared with the ones that survived. Table 2.5 shows a brief description of the selected variables.

Variables	Survive	Not Survive	Description
LD	1.8	1.77	1 - Positive / 2 - Negative
CD38	1.7	1.55	1 - Positive / 2 - Negative
SEX	1.42	1.33	1 - Male / 2 - Female
NLymph	0.64	1.62	0 - 3 affected lymph nodes
MP	1.11	1.38	1 - Positive / 2 - Negative
MOR	1.14	1.27	1 - Typical / 2 - Atypical
SMG	1.84	1.66	1 - Splenomegaly / 2 - No splenomegaly

Table 2.5 Variable Selection for one-year survival. Figures shows mean values.

### Three-year survival

The aim, in this case, was to find the most important features that allow a certain patient to overtake 3-years survival. This is a highly unbalanced problem as well, since the number of deaths (34) is far from the number of survivors (231). We also show the comparison with best prognostic variables for 1 and 5-year survival. The shortest subset of features with the

highest accuracy (93.2%) was found by the Fisher's ratio method, and it was composed of 8 prognostic variables: B2M, AGE, HGB, ALP, UR, LDH. Table 2.6 shows the median and the IQR values for the main variables previously commented.

Variables	Survive	Not Survive
B2M	2.2 / 1.2	3.8 / 4.7
AGE	71 / 12	80 / 16
HGB	14 / 2	12 / 3.5
ALP	64 / 28	80 / 50
UR	43 / 15	53.5 / 18
LDH	321 / 72.7	385 / 129

Table 2.6 Variable Selection for three-year survival. Figures shows Median / Interquartile range (IQR).

As in the CT problem the most discriminatory prognostic variable is beta-2 microglobulin (B2M). Higher median values correspond to the patients that according to the classifier will not survive more than three years. As it was already mentioned, elevated values (>4 mg/L) of B2M is an indicator of poor survival prognosis for multiple myeloma and lymphoma (Hallek et al., 1996). In this group we can also observe levels of hemoglobin lower than the normal HGB range (11-15 g/dL). Also, the median and mean LDH values are abnormally high with respect to its normal range (105-333 U/L). LDH is a protein linked to tumor initiation and metabolism, therefore, patients who have abnormally high levels of LDH could develop more rapidly the disease and die during the first three years.

Figure 2.4 shows the ROC curve and Sensitivity (or True Positive Rate -TPR) against Specificity (or True Negative Rate -SPC) throughout all possible probability thresholds for 3-year survival classification problem. The optimum result, obtained for a probability threshold of 0.48 shows that 99.1% of the patients that survive and 53% of the patients that do not survive during the first 3 years are correctly predicted. Nevertheless, other probability thresholds could be adopted depending on the TPR/SPC balance. The FDR in this case (False Discovery Rate) is 6.5%.

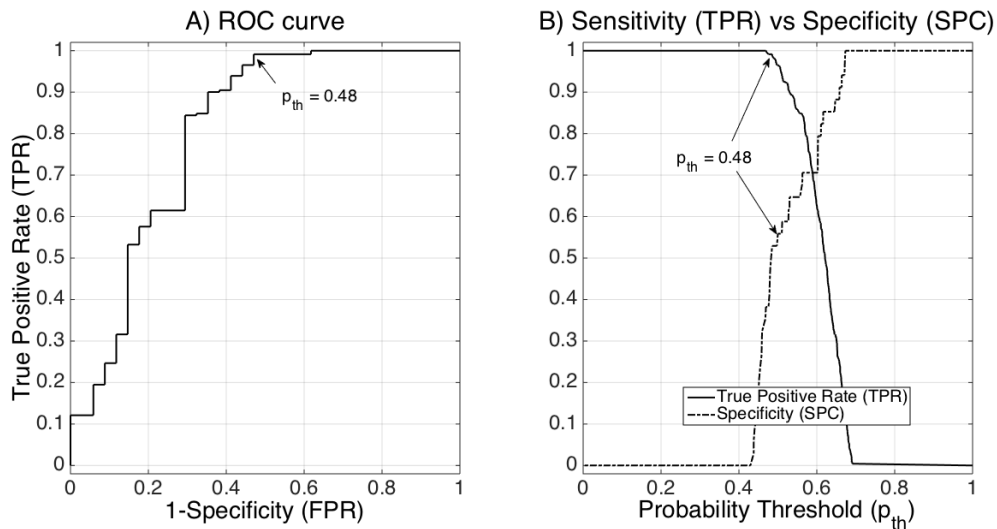


Fig. 2.4 A) ROC curve. B) Sensitivity (or True Positive Rate - TPR) and Specificity (or True Negative Rate - SPC) for 3-year survival. The optimum result (TPR = 99.1 and SPC = 53) is obtained for  $p_{th} = 0.48$ . Nevertheless, other probability thresholds could be adopted depending on the TPR/SPC balance.

### Five-year survival

For the five-year survival problem the difference between the number of dead and survivors was lower but still unbalanced (58 dead and 207 survivors). Fisher's ratio method obtained the best subset of variables in terms of accuracy (85.6%): AGE, B2M, HGB, ALB, ALP, UR and LDH. The entropy method also found a group of variables composed of 6 prognostic variables with similar accuracy (82.3%): B2M, AGE, LDH, GOT, GPT and ALP. Table 2.7 shows the difference between the median and IQR values of those biomarkers. Notice that variables with very similar median and IQR figures (GOT and GPT) are obtained from methods that do not take into account neither median nor IQR values (Entropy).

Variables	Survive	Not Survive
AGE	71 / 13.5	80 / 11
B2M	2.18 / 2.21	3 / 3.04
HGB	14 / 1.8	12.4 / 2.5
ALB	39 / 5.5	35 / 5
ALP	65 / 27.7	73 / 51
LDH	320 / 72.7	361 / 102
UR	42 / 15	51 / 17
GOT	22 / 6.7	22 / 11
GPT	20 / 8	19 / 13

Table 2.7 Variable selection for five-year survival. Two groups of variables are shown. First, the main reduced base with the highest accuracy (85.6%) and below, other relevant variables obtained with Entropy method. Figures shows Median / Interquartile range (IQR).

Overall, the results obtained for prediction of three- and five-years survival coincide. Of note, some new variables related with the renal or hepatic function, such as UR, GOT, GPT and ALP, which are not frequently altered in CLL patients also affect their survival. These variables are affected by co-morbidities of these organs or by the invasion of the kidney or liver by leukemia cells; and these results suggest that the adequate identification and treatment of these complications may play a more important role in the survival of CLL patients than expected.

## 2.4 On the prediction of Hodgkin Lymphoma treatment response

In this case the methodology was applied to figure out prognostic variables for Hodgkin Lymphoma treatment response using the clinical data of a retrospective study of a cohort of 263 caucasians. Besides, in this case the methodology incorporates the weight optimization of the classifier according to the ROC curve to improve risk assessment in the decision-making process, that is, to provide a very high predictive accuracy with an optimum balance between the different rates of the confusion matrix (the true-positive and false-positive rates defining the corresponding ROC curve). The aim is to find the shortest list of clinical variables providing the highest predictive accuracy for Hodgkin lymphoma first-line treatment response (at diagnosis). Therefore, we could use the results to the treatment of Hodgkin Lymphoma

patients. This work has been published in the journal "Clinical & Translational Oncology" (see Appendix A.2).

### 2.4.1 Introduction to Hodgkin Lymphoma treatment response

Hodgkin lymphoma (HL) is characterized by the presence of the so-called malignant Reed-Sternberg cells, surrounded by an inflammatory infiltrate consisting of lymphocytes, neutrophils, eosinophils, plasma cells, macrophages and fibroblasts, constituting a model of interaction of tumor cells with their microenvironment. This kind of cancer is most commonly diagnosed in young adults between the ages of 15 and 35 years and in older adults over 50 years. The cure rate in HL patients is high, but the response along the treatment is still unpredictable and varies from patient to patient. Besides, a small minority is resistant or relapses before treatment. Detecting those patients with a poor prognosis at early stages (diagnosis) could bring improvements in their treatment and prognosis.

There was an international effort to identify the prognostic factors to accurately predict the development and treatment of HL, mainly in patients with advanced stage. The identified adverse prognostic factors were: male older than 45 years, stage IV disease, hemoglobin lower than 10.5 g/dl, lymphocyte count lower than 600/ $\mu$ l (or less than 8%), albumin lower than 4.0 g/dl and white blood count greater than 15,000/ $\mu$ l (Hasenclever et al., 1998; Schreck et al., 2009).

Several research works highlighted the importance of the identification of prognostic variables to predict patients who will suffer relapse and the adaptation of treatments to individual risks (Josting, 2010; Provencio et al., 2004; Smolewski et al., 2000; Zander et al., 2002). Particularly, the result of treatment optimization provoked some criteria modification, with the disappearance of some factors that were considered to be of poor prognosis and with the proposal of new ones that allowed establishing groups with differing risks of relapse and different treatments.

### 2.4.2 HL clinical data

The HL clinical data we dealt with, belongs to a cohort of 263 Caucasians who were diagnosed with classical Hodgkin lymphoma in Asturias (Spain) and enrolled in this study between 2002 and 2012. The treatment response was divided into three categories according to international standards (Cheson, 2008): 237 of the patients were in Complete Remission (CR), 17 in Partial Remission (PR) and only in 9 cases the disease progressed without any relevant change. This last category was named as Progressive Disease (PD). Table 2.8 describes the main characteristics of the patients: age, sex, stage at diagnosis, percentage of

Table 2.8 Main characteristics of the patients (number of patients / percentage), including Hasenclever International Prognostic Score (IPS)

Age	Median: 37 Males range: 9-82 Females range: 10-83
Sex	Males: 171 / 65% Females: 92 / 35%
Stage at diagnosis	Stage I: 42 / 16% Stage II: 92 / 35% Stage III: 82 / 31% Stage IV: 47 / 18%
Early disease: 113 / 43%	Favourable: 57 / 22% Desfavourable: 56 / 21% 150 / 57%
Advanced disease: 150 / 57%	$IPS \leq 2$ : 81 / 31% $IPS > 2$ : 69 / 26%

early favor- able and early unfavorable and percentage of advanced disease depending on Hasenclever Prognostic Score.

Progression-free survival (PFS) was calculated from the date of diagnosis to the date of progression, relapse or death by of any cause. Overall survival (OS) was calculated from the date of diagnosis to the date of death from any cause or last follow-up. OS and PFS distribution curves were estimated using the product-limit method of Kaplan-Meier. The median PFS and OS for the entire group were, respectively, 150 and 160 months. The probabilities of PFS and OS at 7 years were 57 and 76%, correspondingly.

Thirty-five clinical and biological variables were measured at diagnosis and before treatment. These variables were classified into five groups: biochemical, immunohistochemical, Hodgkin lymphoma specific, treatment specific and host information. Table 2.9 shows the description of all these variables, boldfacing those that take discrete predefined values. Most of the variables had a sampling frequency higher than 90%. However, others were scarcely sampled, such as CRP(14%), immunoglobulins and Ki67(20%).

### 2.4.3 ROC-based PSO optimization of the classifier

As it was commented in section 2.2.2 it is possible to optimize the TPR and/or TNR by optimizing the parameters of the classifier. This optimization was performed via Particle Swarm Optimization (PSO). PSO is a stochastic evolutionary computation technique used in optimization, which was initially inspired in the social behavior of individuals (called

Table 2.9 Clinical variables description by group and their corresponding symbols and sampling frequency (Samp. Freq.). Discrete variables are shown in bold faces.

Biochemical	WBC	White Blood cells Count (10 <sup>6</sup> /microL)
	ALC	Absolute Lymphocyte Count (10 <sup>6</sup> /microL)
	AMC	Absolute Monocyte Count(10 <sup>6</sup> /microL)
	AEC	Absolute Eosinophil Count(10 <sup>6</sup> /microL)
	HGB	Hemoglobin (g/dL)
	PLT	Platelets (10 <sup>3</sup> /microL)
	ALB	Albumin (g/L)
	AST	Aspartate Aminotransferase (U/L)
	ALT	Alanine Aminotransferase (U/L)
	ALP	Alkaline phosphatase (U/L)
	CR	Creatinine (mg/dL)
	LDH	Lactate Dehydrogenase (U/L)
	ESR	Erythrocyte Sedimentation Rate (mm/hour)
	CRP	C-Reactive Protein (mg/L)
	GG	Gamma Globulin (g/L)
	IgG	Immunoglobulin G (g/L)
	IgA	Immunoglobulin A (g/L)
	IgM	Immunoglobulin M (g/L)
B2M	Beta 2 Microglobulin (mg/L)	
Cu	Copper (mEq/L)	
SF	Serum Ferritine (ng/mL)	
Immuno-histochemical Tests	<b>CD20</b>	B-lymphocyte antigen CD20 test: Positive or Negative
	<b>Ki67</b>	Ki-67 cellular marker for proliferation: Positive or Negative
	<b>EBV</b>	Ebstein-Barr Virus presence: Positive or Negative
HL Specific	OS	Overall survival from diagnosis to death (days)
	<b>Stage</b>	Ann Arbor staging: I, II , III and IV
	<b>SS</b>	Signs and Symptoms: fever, weight loss, anomalous night sweats
	ALA	Affected Lymphs Areas
	<b>LMM</b>	Large Mediastinal Mass: more than 1/3 of the thoracic diameter
	<b>ELI</b>	Extraganlionic Involvement
Treatment	<b>Bulky</b>	Mediastinal mass more than 10 cm
	<b>CHEMO</b>	Chemotherapy treatment
Personal	<b>RT</b>	Radiotherapy treatment
	AGE	Age
	<b>SEX</b>	Sex

particles) in nature, such as bird flocking and fish schooling. The algorithm consists of the following:

1. A space of admissible solution  $\mathbf{M}$ , is defined:

$$l_j \leq x_{ji} \leq u_j, \quad 1 \leq j \leq n, \quad l \leq i \leq n_{size},$$

where  $l_j, u_j$  are the lower and upper limits for the  $j$ -th coordinate for each optimization model. In PSO terminology, each model is called a particle, and is represented by a vector whose length is the number of model parameters of the optimization problem. Each particle has its own position in the search space. The particle velocity represents the parameter perturbations needed for these particles to move around in the search space and explore solutions of the inverse problem.

2. PSO updates the positions,  $\mathbf{x}_i(k)$  and velocities,  $\mathbf{v}_i(k)$  of each particle in the swarm in each iteration, according to 3 main components:

- The inertia term, which consists of the old velocity of the particle,  $\mathbf{v}_i(k)$ , weighted by a real constant,  $\omega$ , called inertia.
- The so-called social term, which is the difference between the global best position found so far in the entire swarm (called  $\mathbf{g}(k)$ ), and the particle's current position ( $\mathbf{x}_i(k)$ ).
- The so-called cognitive term, which is the difference between the particle's best position found so far (called  $\mathbf{I}_i(k)$ , the local best) and the particle's current position ( $\mathbf{x}_i(k)$ ):

$$\mathbf{v}_i(k+1) = \omega \mathbf{v}_i(k) + \phi_1 (\mathbf{g}(k) - \mathbf{x}_i(k)) + \phi_2 (\mathbf{I}_i^k - \mathbf{x}_i(k)),$$

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) + \mathbf{v}_i(k+1),$$

$$\phi_1 = r_1 a_g, \phi_2 = r_2 a_l, r_1, r_2 \in U(0, 1), \omega, a_g, a_l \in \mathbb{R},$$

$r_1$  and  $r_2$  are vectors of random numbers uniformly distributed in  $(0, 1)$ , to weight the global and local acceleration constants,  $a_g$  and  $a_l$ .  $(\omega, a_g, a_l)$  are the PSO parameters to be tuned in order to achieve convergence. PSO has been chosen for this purpose because its convergence has been analyzed using stochastic stability analysis. Consequently, the tuning of the PSO parameters can be done automatically, based on these stability results. Particularly, the RR-PSO (Fernández-Martínez and García Gonzalo, 2012) and CP-PSO (Fernández-Martínez and



García Gonzalo, 2009) versions are used due to their higher exploratory properties that allowed us to escape from local minima.

In this case, the particles  $\mathbf{x}_i$  (model parameters) are the weights of the k-NN distance-based classifier. These weights were optimized from their prior values given by the HVDM metric, that is, the inverse of two times the prior variability of the prognostic variables (see section 1.3.6 for further details), to balance confusion matrix. The cost function was defined as follows:

$$c(\mathbf{x}_i) = \omega_1 FP(\mathbf{x}_i) + \omega_2 FN(\mathbf{x}_i)$$

where  $\omega_1$  and  $\omega_2$  serve to weight the relative importance of the false positives and false negatives depending on  $\mathbf{x}_i$ . If  $\omega_1 = \omega_2 = 1$  then both terms have equal importance.

#### 2.4.4 HL results

Treatment response in HL is a difficult prediction problem. Aside from plasma EBV DNA (Gandhi et al., 2006), there is no predictive biomarker to predict the patient's response to the corresponding treatment with a reliable accuracy.

The first modeling decision was to transform the analysis of treatment response into a binary classification problem (two-class problem) that admits a more reliable and stable solution than the corresponding value regression problem, that is, it is easier to predict if a patient is in complete or partial remission than predicting the value of the biological variables related to this fact. Besides, the prediction in binary classification problems allows risk assessment through the analysis of the confusion matrix and the Receiving Operating Characteristic (ROC) curve. The comparison was composed of two main steps. In the first step (1. CR and PR Vs. PD), we established the differences between patients who experienced partial or complete remission (CR and PR, positive class) from those in which the disease progressed without any relevant change (PD, negative class). Then, a second comparison (2. CR Vs. PR) was used to establish the differences between CR (positive class) and PR (negative class) patients.

The best result was obtained by filtering out those variables having a sampling frequency lower than 30%, and imputing the rest. Besides, MPD (Maximum Percentile Distance) provided the shortest list of variables with the highest predictive accuracy. Table 2.10 shows the confusion matrix rates (TPR, TNR, FPR, FNR) for both comparisons, together with the False Discovery Rate (FDR) and the LOOCV predictive accuracy (Acc). No weight optimization was performed in this case, that is, the weights corresponded to the inverse of the prior variability of the prognostic variables (see section 1.3.6 for further details of how is calculated the weights).

Table 2.11 Mean values of the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), and weights ( $\omega_b$ ) for the optimum NN-classifier without weights optimization.

Comparisons	Variables	TP	TN	FP	FN	$\omega_b$
CR and PR (+)						
Vs.	SF	266.4	<b>3288.0</b>	452.4	3231.3	0.0005
PD (-)						
CR (+)	SF	249.9	<b>2401.0</b>	405.5	2131	0.0005
Vs.	ALT	<b>23.7</b>	18.0	44.2	74.4	0.0092
PR (-)	ALP	116.8	<b>376.0</b>	163.5	608.4	0.0017

Signs (+) and (-) represent positive and negatives groups respectively. Bold faces indicate the highest value for each prognostic variable. Normal bounds for the decision correspond to the TP and TN groups.  $\omega_b$  are the weights used in the classifier for data variability normalization (before weight optimization).

Table 2.10 Best results for all the comparisons obtained without weights optimization.

Comparisons	Base	MPD rate	TPR (%)	TNR (%)	FPR (%)	FNR (%)	FDR (%)	Acc (%)
CR and PR (+)								
Vs.	SF	75.2264	98.43	22.22	77.78	1.57	2.72	95.82
PD (-)								
CR (+)	SF	57.7157						
Vs.	ALT	41.3166	97.89	11.76	88.24	2.11	6.07	92.13
PR (-)	ALP	38.9228						

The algorithm used for all the comparisons was the same: filtering 30% of sampling frequency, imputing and MPD as feature selection method. Rate is the maximum percentile distance rate, TPR is the True Positive Rate, TNR is the True Negative Rate, FPR is the False Positive Rate, FNR is the False Negative Rate and Acc is the final accuracy of the prediction. Signs (+) and (-) represent respectively positive and negatives groups respectively.

Table 2.11 shows the mean values of the three prognostic variables for the different groups of the confusion matrix and the weights ( $\omega_b$ ) used to define the distance criterion in the classifier.

Optimization of the weights of the classifier via Particle Swarm Optimization (PSO) was performed to improve the true negative rate (or Specificity), that is, increasing TNR while the overall accuracy is also improved (TPR is not affected). Table 2.12 shows the TPR, TNR, FPR, FNR, FDR and predictive accuracy (Acc) obtained after weight optimization. TN rates were improved around 10% in comparisons 1, while in comparison 2 TP rate was improved around 1%. The overall accuracy was improved in all the cases around 1%.

Table 2.12 Best results for the comparisons obtained after weights optimization.

Comparisons	Base	MPD rate	TPR (%)	TNR (%)	FPR (%)	FNR (%)	FDR (%)	Acc (%)
CR and PR (+)								
Vs.	SF	75.2264	98.43	33.33	66.67	1.57	2.34	96.1977
PD (-)								
CR (+)	SF	57.7157						
Vs.	ALT	41.3166	99.58	11.76	88.24	0.42	5.98	93.7008
PR (-)	ALP	38.9228						

Rate is the maximum percentile distance rate, TPR is the True Positive Rate, TNR is the True Negative Rate, FPR is the False Positive Rate, FNR is the False Negative Rate and Acc is the final accuracy of the prediction. Signs (+) and (-) represent positive and negatives groups respectively.

Table 2.13 shows the mean values for TP, TN, FP, FN and the optimized weights for the prognostic variables ( $\omega_a$ ). It can be observed that values of the weights increased after optimization for all the prognostic variables. Therefore, it is possible to improve the quality of the prediction and minimize risk on the decisions, by optimizing the weights that are initially provided by the distance criterion.

Table 2.13 Mean values of the true positives, true negatives, false positives and false negatives and optimized weights  $\omega_a$  of the optimum NN classifier after weight optimization.

Comparisons	Variables	TP	TN	FP	FN	$\omega_a$
CR and PR (+)						
Vs.	SF	275.4	<b>2796.7</b>	225.5	2669.5	0.0020
PD (-)						
CR (+)	SF	276.7	<b>2401.0</b>	405.5	3330.0	0.0026
Vs.	ALT	<b>24.3</b>	18.0	44.2	140.0	0.0663
PR (-)	ALP	123.2	<b>376.0</b>	163.5	1059.0	0.0051

Signs (+) and (-) represent the positive and negatives groups, respectively. Boldfaces indicate the highest value for each prognostic variable. Normal bounds for the decision correspond to the TP and TN groups.

## 2.4.5 Conclusions for HL treatment response prediction

Overall, the results of this study show that the combined use of these prognostic variables, SF, ALT and ALP, in a simple classifier allows predicting first-line treatment response in HL patients with high accuracy and confirms a close relationship between treatment response in HL, inflammation, iron overload and liver and bone damage.

Serum ferritin has been frequently used as a surrogate marker for systemic iron stores, but may be also elevated in specific circumstances without excess iron stores, such as in inflammation, correlating closely to the activity of malignant lymphomas. However, to our knowledge, serum ferritin levels have not been yet related to the treatment response of HL patients.

Serum activity levels of ALT enzyme are routinely used as a biomarker of liver injury caused by drug toxicity, infection, alcohol and steatosis. Levels greater than 500 U/L occur most often in people with hepatic diseases, such as viral hepatitis, ischemic liver injury (shock liver), toxin-induced liver damage and tumor infiltration of liver.

The alkaline phosphatase test (ALP) is used to detect liver disease or bone disorders. In conditions affecting the liver, damaged liver cells release increased amounts of ALP into the blood. In non-Hodgkin lymphomas, ALP is increased in patients with bone marrow affection (Kittivorapart and Chinthammitr, 2011), thus reaching stage IV and worse prognosis. However, in a patient with fever of unknown origin (FUO), highly elevated alkaline phosphatase and normal/slightly elevated serum transaminase levels suggest the possibility of lymphoma (Brensilver and Kaplan, 1975; Brinckmeyer et al., 1982; Cunha, 2007).

To conclude, detecting those HL patients who do not respond to the treatment at early stages may help improve their treatment. This study proposed a new prognostic analysis method, based on mathematical models that identify three simple prognostic variables currently gathered at diagnosis that may help detect with high accuracy those HL patients with bad prognosis without any additional cost.

# Chapter 3

## Application to genetic data

### 3.1 Introduction

Genetic information is located in the DNA as a sequence of nucleotides. A gen is a part of the DNA that contains the necessary information for the synthesis of proteins, which is a critical process in the human body. Genes are not continuous and include both non-coding and coding regions for synthesis of proteins. The typical samples from the DNA are commonly extracted from blood, tissues or fluids. Thanks to the the development of high-throughput technologies for sequencing in genetic and genomic analyses, that sequence of nucleotides may be stored in a data set within a computer. Moreover, gene expressions can be analyzed through hybridization microarrays or RNA sequencing, which is a much cheaper way of analyzing genetic data.

Genetic data, particularly, gene expression data, are commonly used to compare two or more sets of patients (typically healthy control VS unhealthy patients) in order to figure out what (genes) is causing those differences. Those comparisons could be used to predict a certain disease occurrence (diagnosis optimization), to make the difference between two or more treatments (treatment optimization) or to evaluate the survival of a set of patients (prognosis optimization). These kind of problems will be addressed as the general denomination of phenotype prediction problems.

Genetic data has a very underdetermined character, since the number of samples/patients is always much lower than the number of genes. We do not have a unique solution to the inverse problem, therefore, reduction of dimension algorithms become a key element in the problem solution.

In this chapter we applied our methodology to address two different problems using gene expression data. Firstly, we identify and validate a specific gene cluster that is predictive of fatigue risk in prostate cancer patients treated with radiotherapy. This research work

was performed in collaboration with the National Institute of Nursing Research, National Institute of Health, Bethesda, Maryland, USA and Biomodels, LLC, Watertown, MA, USA. As a result a manuscript named "Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy-Related Fatigue in Patients with Prostate Cancer" was published in the journal "Cancer Informatics". Secondly we modeled a data expression microarray related to Chronic Lymphocytic Leukemia, predicting the occurrence of the main mutations, which are closely related with the survival of the patients. The results were included in a paper called "Genomic Data Integration in Chronic Lymphocytic Leukemia" and it is currently under review in the journal "Journal of Gene Medicine".

As a continuation of the research work on the Cancer treatment-related fatigue, we perform some statistical analysis using the methodology explained herein to a data set related to mitochondrial activity. The result was a publication named "Relationship of Mitochondrial Enzymes to Fatigue Intensity in Men With Prostate Cancer Receiving External Beam Radiation Therapy" in the journal "Biological research for nursing" (Filler et al., 2015).

As in the previous chapter, there are three main parts. Firstly we present the common methodology applied in both practical cases. Secondly we introduce the cancer related fatigue prediction problem and present the results and conclusions. Finally, we proceed in the same way with the Genomic data integration in Chronic Lymphocytic Leukemia.

## **3.2 Methodology applied to both practical cases using genetic data**

The common methodology applied to both cases has three main steps: 1) Obtain the gene discriminatory power. 2) Select the genes according to the discriminatory power. 3) Create the correlation networks between the selected genes. Figure 3.1 shows the flowchart describing these steps.

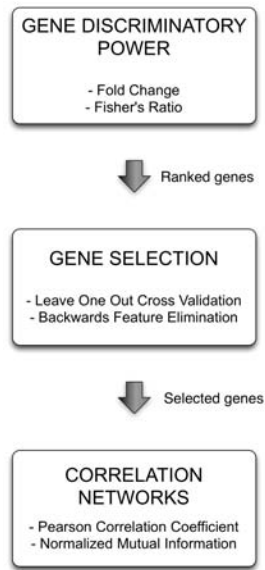


Fig. 3.1 Flow diagram for the prediction model. The methodology is composed of 3 steps: 1) Obtain the gene discriminatory power. 2) Select the genes according to the discriminatory power. 3) Create the correlation networks between the selected genes.

### 3.2.1 Gene discriminatory power

It is crucial important to be able to establish the discriminatory power of a gene in phenotype prediction problems. In section 1.3.6 we have presented the main feature selection methods for clinical and genetic data. A gene is said to be highly discriminatory if several conditions are met such as high Fisher's ratio, high Fold Change, low Entropy, high Percentile distance and high SAM ratio. In this case we used a combination between Fold Change (FC) and Fisher's ratio (FR). We first ranked genes according to their discriminatory power. We preselect the most differentially expressed genes above a certain absolute FC value and then we finally rank the genes according to their FR. The reason to first preselect with FC is because low dispersions in both classes can provide high FR values when in fact the centers of both distributions in expressions are very close (differences in means very small). Therefore, by preselecting differentially expressed genes above a certain absolute FC value we can avoid to have high FR values due to the low dispersions.

The FR values of the ranked prognostic variables draw a curve that could be interpreted as a singular value of a linear forward operator characterized by a matrix  $F \in M_{m \times n}(\mathbb{R})$ .  $F$  has a singular value decomposition  $F = U\Sigma V^T$ , where  $U \in M_{m \times m}(\mathbb{R})$ ,  $V \in M_{n \times n}(\mathbb{R})$  are orthogonal and  $\Sigma$  blocky diagonal. Besides,  $\text{rank}(F) = \text{rank}(\Sigma) = r$ .

$$\Sigma = \left( \begin{array}{ccc|c} \alpha_1 & & & 0_1 \\ & \ddots & & \\ & & \alpha_r & \\ \hline & & & 0_2 \\ & & & 0_3 \end{array} \right), \quad \alpha_i > 0.$$

Given a model  $\mathbf{x}$ , its prediction is

$$\begin{aligned} F\mathbf{x} &= U\Sigma V^T\mathbf{x} = \\ &= \sum_{k=1}^r \alpha_k \mathbf{u}_k \mathbf{v}_k^T \mathbf{x} = \\ &= \sum_{k=1}^r \alpha_k \mathbf{d}_k \end{aligned}$$

where  $\mathbf{d}_k = \mathbf{u}_k \mathbf{v}_k^T \mathbf{x}$  is the prediction due to  $k$ -th spectral term  $\mathbf{b}_k = \mathbf{u}_k \mathbf{v}_k^T$ . Therefore  $F\mathbf{x} = \sum_{k=1}^r \alpha_k \mathbf{d}_k$ . The contribution of the singular value  $\alpha_k$  to the energy of the data prediction  $\mathbf{d}$  is:

$$e_k = \frac{\alpha_k \|\mathbf{d}_k\|_2}{\|\mathbf{d}\|_2} 100.$$

Something similar can be said with the Fisher's Ratio curve, that can be interpreted as a measure of the prior discriminatory power of a gene. In the cancer treatment-related fatigue prediction problem we showed in figure 3.4 this kind of curve, where genes with the highest FR were the most important biological "eigenvectors" for the discrimination. The posterior discriminatory power is given by the predictive accuracy of the ranked lists of genes, see for example figure 3.6 for the cancer treatment-related fatigue problem. We can observe that adding genes with lower discriminatory power as defined by their FR does not imply an increase of the predictive accuracy, that is, the posterior predictive accuracy is not monotonous increasing by adding more genes to the discrimination. This is a simple way of reducing the high underdetermined character of any phenotype prediction problem.

### 3.2.2 Gene selection

Similarly to what we did with clinical data, we applied a Nearest Neighbor based algorithm to establish the accuracy of the different ranked sets of genes using Leave-One-Out-Cross-Validation (LOOCV) experiment. The combination of this procedure with a backwards feature elimination algorithm produced the shortest list of high discriminatory gene and served to validate the prognostic value of these gene signatures over the existing dataset by cross-validation (see section 1.3.6 for further details). This procedure serves to eliminate redundant or irrelevant genes to yield the most precise set of genes with the greatest predictive accuracy. The linear separability of the phenotype in the reduced set of genes could be



checked by performing principal component analysis (PCA) of the dataset expressed in this small-scale signature and projecting these samples in the corresponding 2D PCA space. Then, the problem approximates a linear separable behavior by reducing the dimension to the list of most discriminatory genes, if populations can be linearly separated by a given hyper-plane.

### 3.2.3 Correlation networks

Finally we built correlation networks using the selected genes to understand how the expression of the most discriminatory genes is controlled in each case. Correlation networks were generated using the approach presented in Lastra et al. (2011) but with two different coefficients measuring the dependency between genes:

- Pearson correlation coefficient (Pearson, 1895): It measures the linear correlation of two random variables.

$$p_{ij} = \frac{\text{cov}(g_i, g_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \quad (3.1)$$

where  $\text{cov}(g_i, g_j)$  is the covariance between the expressions of two genes  $g_i, g_j$  considered as random variables and  $\sigma_i^2, \sigma_j^2$  is the variance of the expression in gene  $i$  and  $j$  respectively.  $p_{ij}$  is zero when the variables are uncorrelated, that is, linearly independent, and varies between -1 (total negative correlation between expressions) and 1 (total positive correlation). This metric is not useful when the relationship between the variables is nonlinear. Nevertheless, we will show numerically that the classification problem approximates a linear separable behavior when the dimension is reduced to the most discriminator variables. Therefore, when the analysis is restricted to these variables, it makes perfect sense and serves to find the trade-offs between them (uncertainty of the corresponding prediction problem).

- Normalized Mutual Information (Strehl and Ghosh, 2003): The mutual information of two random variables is a measure of mutual dependence of both variables. In our case we have used the normalized mutual information, which is similar to a correlation coefficient:

$$NM_{ij} = \frac{I(g_i, g_j)}{\sqrt{H(g_i)H(g_j)}} \quad (3.2)$$

where  $I(g_i, g_j)$  is the mutual information content and  $H(g_i)$  the entropy of gene  $i$  calculated based in the ordering of its expression with respect to the class assignment. The mutual information  $I(g_i, g_j)$  content is calculated as follows:

$$I(g_i, g_j) = H(g_i) + H(g_j) - H(g_i \cup g_j)$$

being  $H(g_i \cup g_j)$  the joint entropy. The normalized mutual information can be interpreted as a correlation coefficient based exclusively in the diversity (entropy) in  $g_i$  and  $g_j$ . It varies between 0 (totally independent) and 1 (totally dependent):

$$NM_{ij} = 0 \leftrightarrow H(g_i \cup g_j) = H(g_i) + H(g_j)$$

Therefore, the normalized mutual information is null when one variable does not reduce the uncertainty about the knowledge of the other, that is, they are independent descriptors.

Once we have calculated these coefficients, the Kruskal's algorithm (Kruskal, 1956) is used to find the minimum-spanning-tree between the selected genes and building the correlation network, using as head the gene with the most discriminatory power. Two main gene categories can be identified in correlation networks: headers, which are the genes located in the top of the network and have higher discriminatory power, and helpers, which are the genes in the lower parts of the network that provide high frequency details for the discrimination. Moreover, correlation networks serve to analyze inter-relationships between genes, that impact the expression of other genes, and therefore their function. Finally, gene ontology is performed to cover the altered and disease pathways. For that purpose we used the GeneAnalytics tool provided by the Weizmann Institute of Science (Stelzer et al., 2009).

### **3.3 Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy - Related Fatigue in Patients with Prostate Cancer**

In this research work we applied the methodology explained in section 3.2 to a genetic data obtained from a microarray expression dataset where patients were diagnosed with non-metastatic prostate cancer and scheduled to receive radiotherapy treatment. In this case the correlation network step was not performed. The problem were to identify the smallest subset of genes that predict the cancer treatment-related fatigue before radiotherapy treatment was carried out. This work was published in the journal "Cancer Informatics" (see Appendix A.3).

### **3.3.1 Introduction to the cancer treatment-related fatigue prediction problem**

Fatigue is the most common, troublesome, and costly side effect of many cancer treatment regimens. Not only does it impact patients directly, but it also has significant repercussions on both direct and indirect health economic outcomes (Carlotto et al., 2013). Cancer treatment-related fatigue (CTRF) is defined as a "subjective sense of tiredness" that persists over time, interferes with activities of daily living, and is not relieved by adequate rest (Minton et al., 2008; Mock, 2003).

CTRF, like other regimen-related toxicities, does not occur in every patient, but rather in a subpopulation of at-risk individuals. In the context of individualizing care, the ability to predict CTRF risk has the potential to help guide treatment choices for patients and providers. However, as it becomes increasingly clear that CTRF is strongly related to a series of underlying genetically controlled biological events, the utility of identifying a group of genes that impact patients' risk of the condition seems compelling. In the current study, we evaluated our methodology to identify a group of genes that predicted CTRF in men being irradiated for prostate cancer. This proof-of-concept investigation not only demonstrated the utility of the analysis, but also confirmed the observation that focal radiation therapy is capable of inducing gene expression changes in peripheral white blood cell RNA (Sonis et al., 2007).

In this case we firstly applied the methodology explained in section 3.2 to a training data in order to identify/select the smallest and most precise set of CTRF-associated genes, and then check the legitimacy of the predictive accuracy based on the training set with a validation blind set. The validation was performed as follows:

1. We first considered the most predictive gene cluster, a group consisting of the 14 most discriminatory genes deduced from the training set. The samples of the training set expressed in the reduced base and their phenotype information were used to define the distance of the classifier.
2. Second, the values of these discriminatory genes in the validation samples were read from the validation dataset. For each sample of the validation set, its predicted class was established using the k-NN based algorithm explained in 1.3.6, using the 14 different most discriminatory reduced sets of genes that were defined by the training dataset.
3. The first step was repeated to generate 14 different reduced bases, which yielded 14 different class predictions for each sample in the validation set: 14 different Biomedical

Robots. The final estimated class was then made by consensus or majority voting classifiers (see section 1.3.5). A posterior probability was given to the class prediction, defined as the ratio of the number of votes assigned to the predicted class and the total number of voters.

### 3.3.2 CRTF gene expression data

The microarray expression dataset consists of men who were 18 years or older, diagnosed with non-metastatic prostate cancer with or without a history of prostatectomy, and scheduled to receive External Beam Radiation Treatment (EBRT) with or without concurrent androgen deprivation therapy (ADT). A total of 44 men with non-metastatic prostate cancer were studied, 27 of them were used as training set and 17 as validation blind set.

To assess fatigue in cancer therapy the 13-item Functional Assessment of Cancer Therapy-Fatigue (FACT-F) score was used. FACT-F is scored from 0-52, the higher the score, the lower the fatigue symptoms. A greater than three-point decrease in the FACT-F score is considered to be a minimally important change that is clinically relevant (Yost et al., 2011). To discretize the phenotypic characterization of the study participants, subjects were categorized into high-fatigue (HF) or low-fatigue (LF) groups based on their change in FACT-F scores from baseline to completion of EBRT. HF subjects had a decrease of three or more points in FACT-F scores, and those who had less than a three-point decrease in FACT-F scores between both time points were categorized in the LF group. Questionnaires were completed at baseline (prior to EBRT) and at completion of EBRT (day 38-42 after EBRT initiation). To avoid extraneous influences on their responses, subjects completed the questionnaires in an outpatient setting before clinical procedures were provided.

The biological sample collection, RNA extraction, and microarray experiments were extracted from peripheral blood at baseline and on the last day of EBRT, immediately after FACT-F was performed.

### 3.3.3 CRTF results

As presented above through the methodology explained in section 3.2, we try to identify/select the smallest and most precise set of CRTF-associated genes in a training set and then check the consistency of the results in a validation blind set. The training model was developed from the array outputs of 27 subjects; 18 were HF and 9 were LF. Each patient sample contained 604,258 different probes. The minimum and maximum gene expressions were 21 and 62,088 respectively. As shown in Figure 3.2, it was impossible to visually distinguish

HF and LF microarray outputs in heat map format using decibels as units of measure ( $\log_2$  of gene expression).

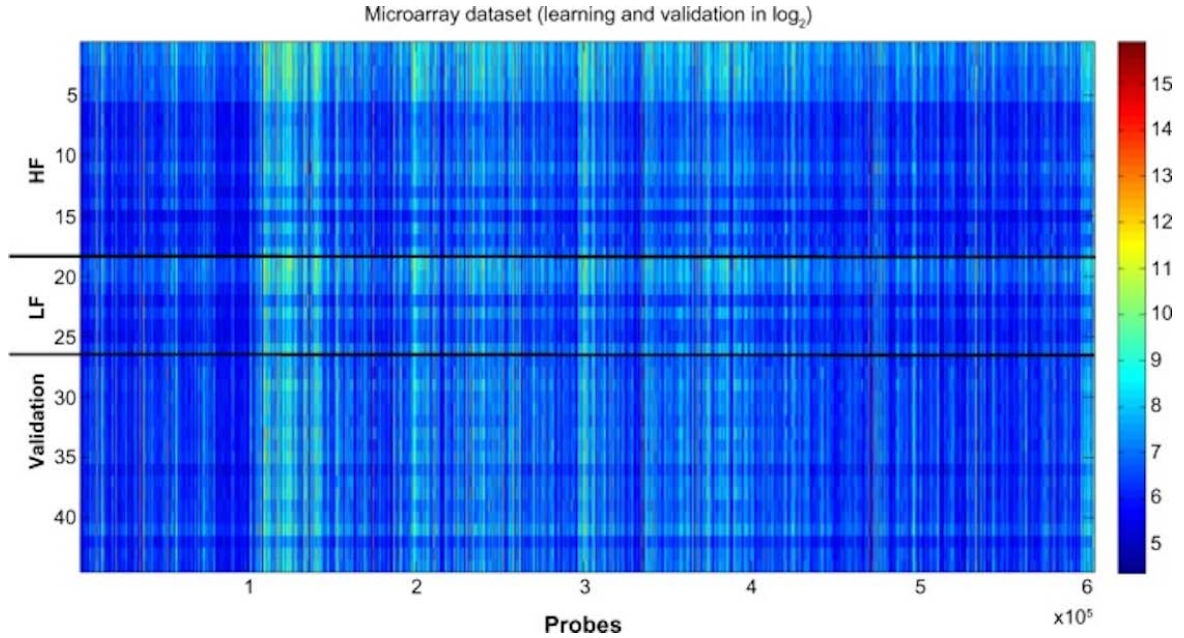


Fig. 3.2 Data visualization in decibels ( $\log_2$  of the expression). HF is composed of 18 samples, LF 9 samples and Validation 17 samples.

The similarities between the HF and LF groups in the learning dataset were confirmed by further histogram analysis of gene expression. Figure 3.3 shows that the corresponding statistical distributions of gene expressions in both groups were close to lognormal, with the main differences between both phenotypes occurring around the mode of both histograms (expressions around  $2^4$  and  $2^6$ ).

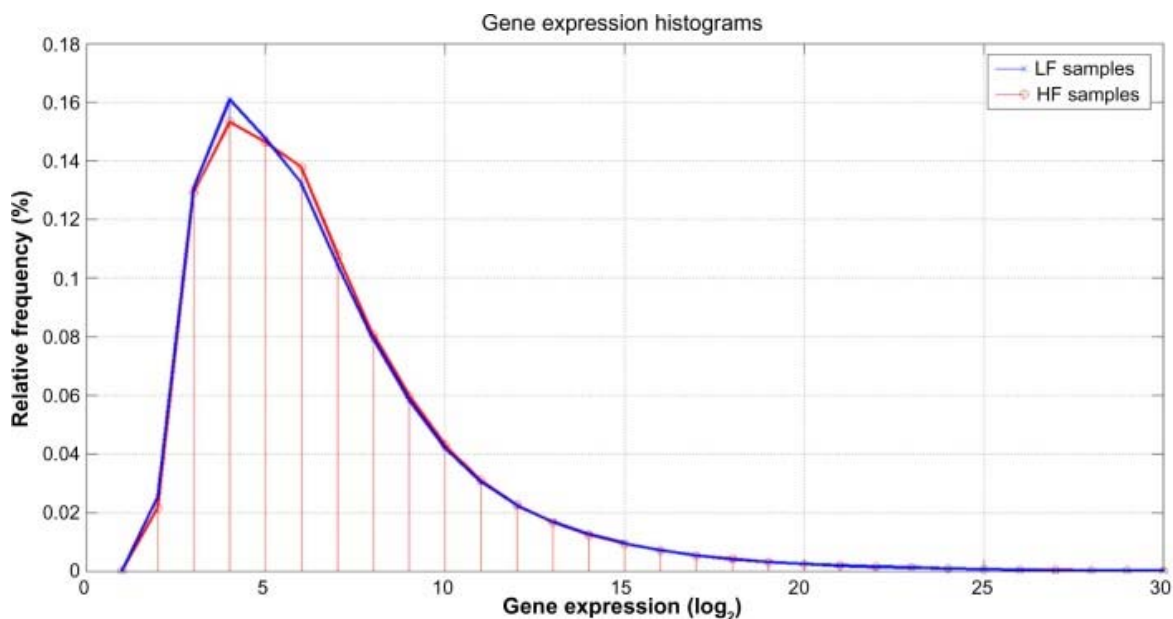


Fig. 3.3 Gene expression histograms in log<sub>2</sub> scale for the Low Fatigue and High Fatigue subjects.

A final list of 575 highly discriminatory genes according to expression was noted and defined by the intersection between those genes that were differentially expressed (located in the 0.05% and 99.5% tails of the fold-change ratio cumulative distribution) and which had a FR higher than 0.25 (figure 3.4). Genes with the highest FR were the most important biological eigenvectors for the discrimination, as it happens, for the Fourier analysis of a digital signal and its decomposition into different harmonics. In this case, the FR curve decreases very steeply, in such that only with the first set of genes (14 to 35 genes in this case), the highest discriminative accuracy of the learning data set can be achieved. Adding genes with lowest discriminatory power indiscriminately does not improve the LOOCV predictive accuracy. The BFE method (see section 1.3.6 for further details) is used to determine the amount of details that is needed.

Additionally, figure 3.5 shows the FC-FR plot for genes in the learning dataset with FC lower than - 0.52 and higher than 0.67. These values (of gene under- and over-expression) corresponded, respectively, to the 0.05% and 99.5% tails of the FC distribution. It can be observed that the highest FR was 2.12, and genes with the highest FC did not coincide with those exhibiting the highest FR.

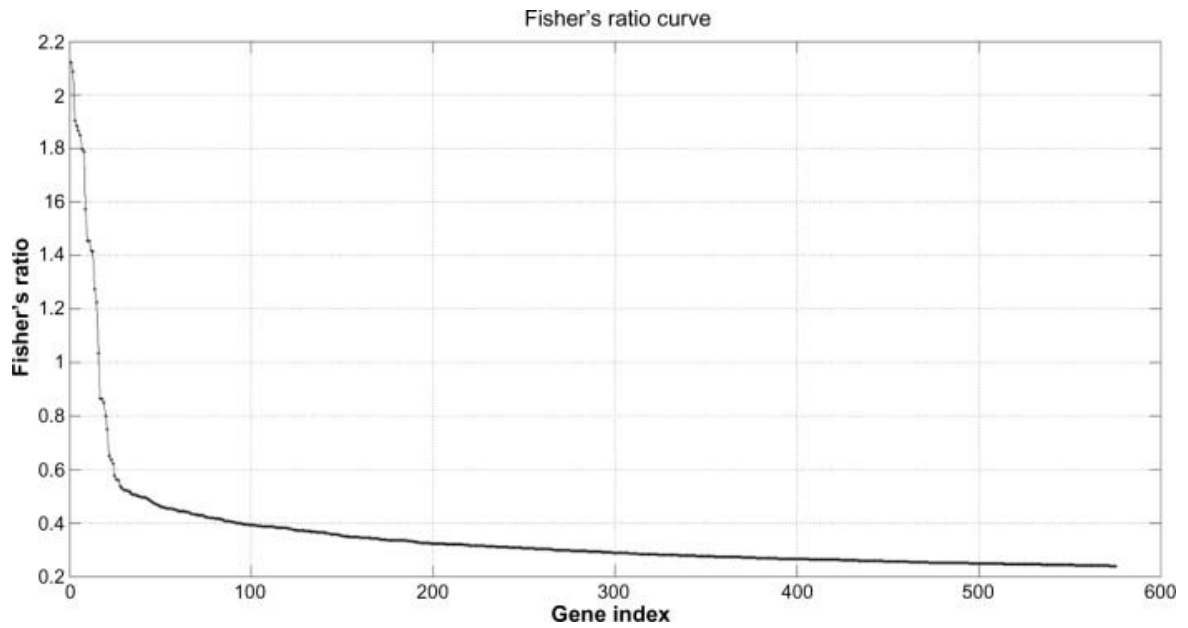


Fig. 3.4 Fisher's ratio curve for the Low Fatigue-High Fatigue phenotype discrimination.

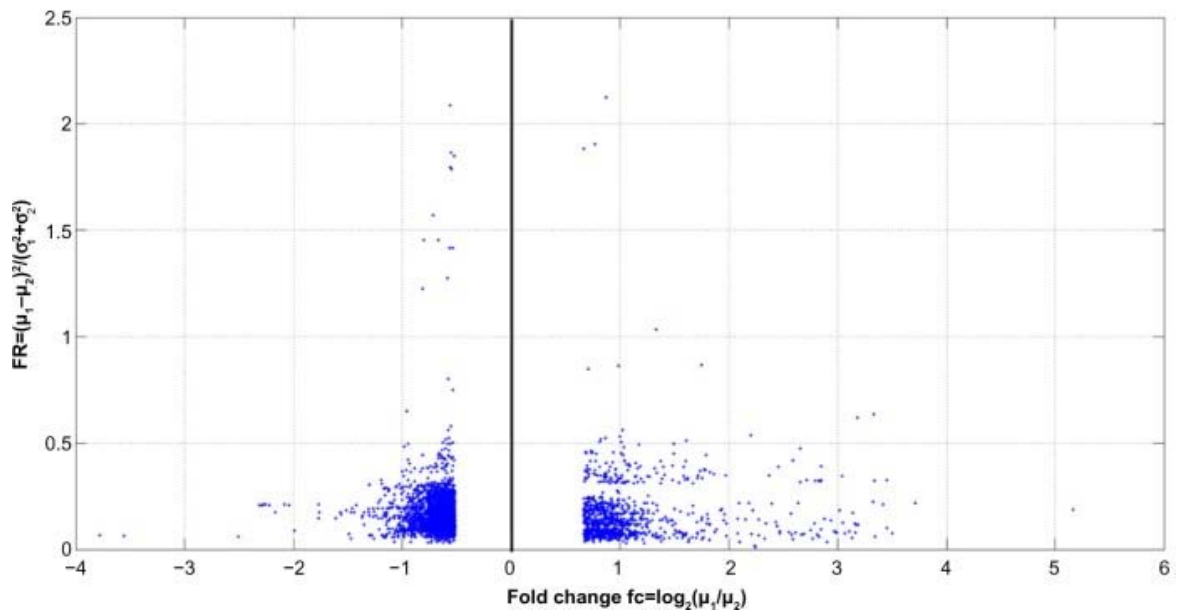


Fig. 3.5 Fold change-Fisher's ratio plot of genes in the learning dataset with absolute fold change greater than 0.52 that corresponds to the 0.005 and 99.5% tails of the fold change distribution. In this case the Fisher's ratio plays a similar role than  $-\log(P \text{ value})$  for the volcano plot analysis (Cui and Churchill, 2003).

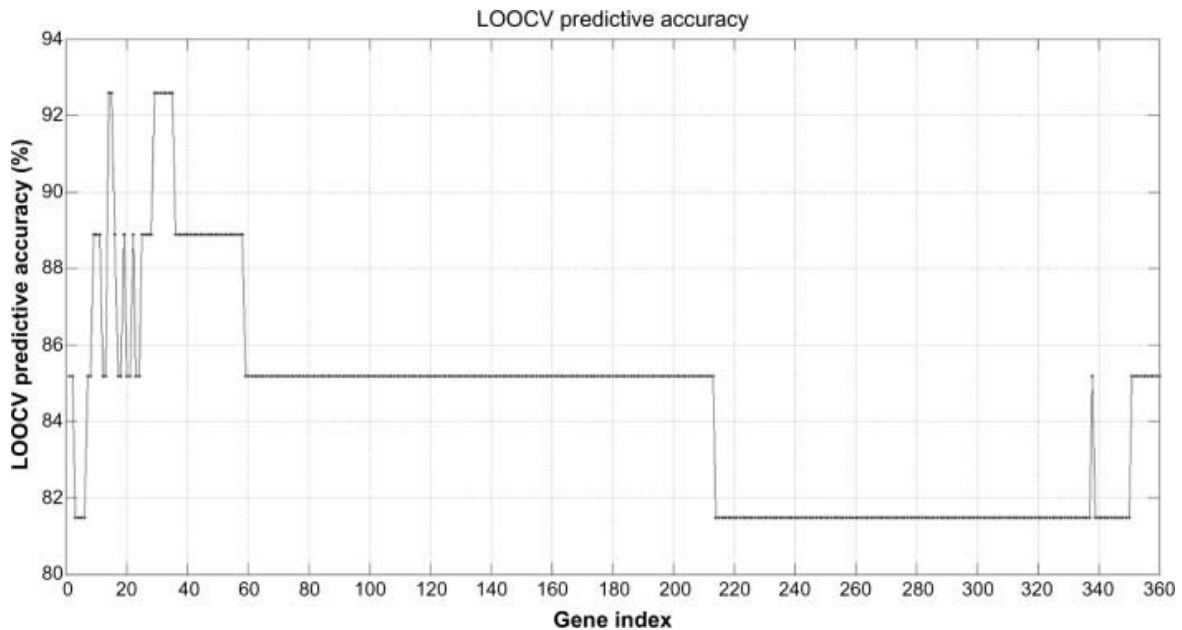


Fig. 3.6 Leave-One-Out-Cross-Validation (LOOCV) learning predictive accuracy of the first 360 gene sets with the highest discriminatory power. The shortest list with the highest accuracy (92.6%) contains only the first 14 genes.

Figure 3.6 shows the predictive accuracy curve of the different gene lists, established using the backward feature elimination algorithm. The shortest list with the highest accuracy (92.6%) was composed by the first 14 genes with the highest FR. The lists with the first 15, and 29 to 35 most discriminatory genes also provide the same maximum accuracy. As the data suggest, continuously adding genes with lower discriminatory power as defined by their FR failed to increase the accuracy of discrimination.

When a histogram was used to assess the first 360 most discriminatory genes found by our analysis, we noted a shift of the mode of distribution for the LF patients to higher expressions (29-210) with respect to the HF case (26-27), suggesting that HF patients show mostly lower expressions of these genes that we hypothesized were responsible for this phenotypic discrimination (figure. 3.7). Compared to Figure 3.3, a higher discrimination in the modes of the LF/HF phenotypes can be observed: the mode of HF samples is shifted to lowest values (approximately 64 instead of 512).



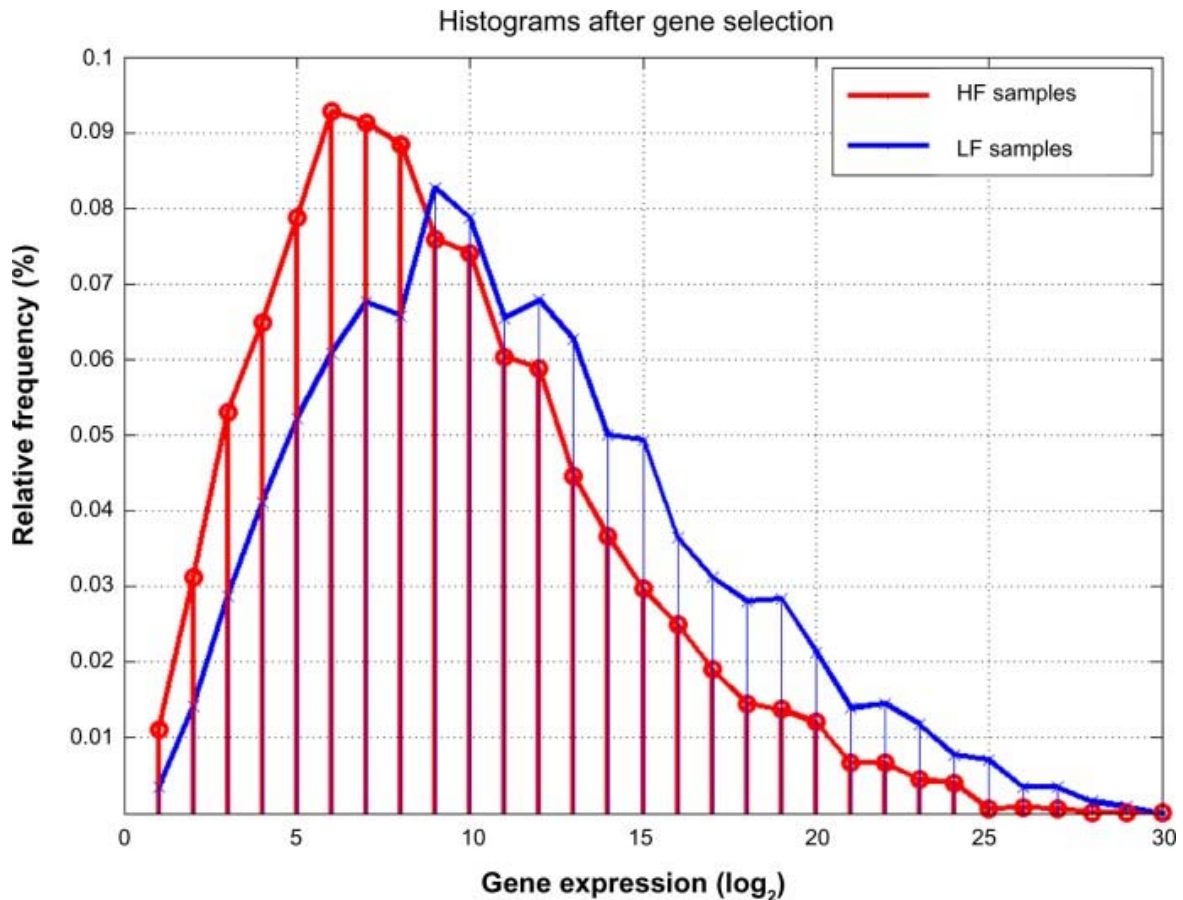


Fig. 3.7 Histograms (in log<sub>2</sub> scale) for the Low Fatigue (LF) and High Fatigue (HF) patients, of the first 360 most discriminatory genes.

Figure 3.8 shows the PCA plots (unsupervised method) of the learning dataset expressed in the base of the most 14 (figure 3.8A) and 35 (figure 3.8B) discriminatory genes having the highest predictive accuracy. The following can be observed:

- The LF/HF phenotype discrimination became easier to linearly separate in these reduced sets of genes, confirming the fact that the classification problem simplifies when reducing the dimension to the most discriminatory set of genes. Both plots have a similar structure. The LF samples lie between samples P1A and xrt28A, which is genetically close to the region of the HF samples.
- Also, sample xrt25A, which belongs to the LF category, is surrounded by HF samples. This sample might be a biological or behavioral outlier.
- The HF samples lie between samples xrtp2A and 13A. Sample xrt20A also seems to mark a transition between LF and HF samples toward the west of the plot.

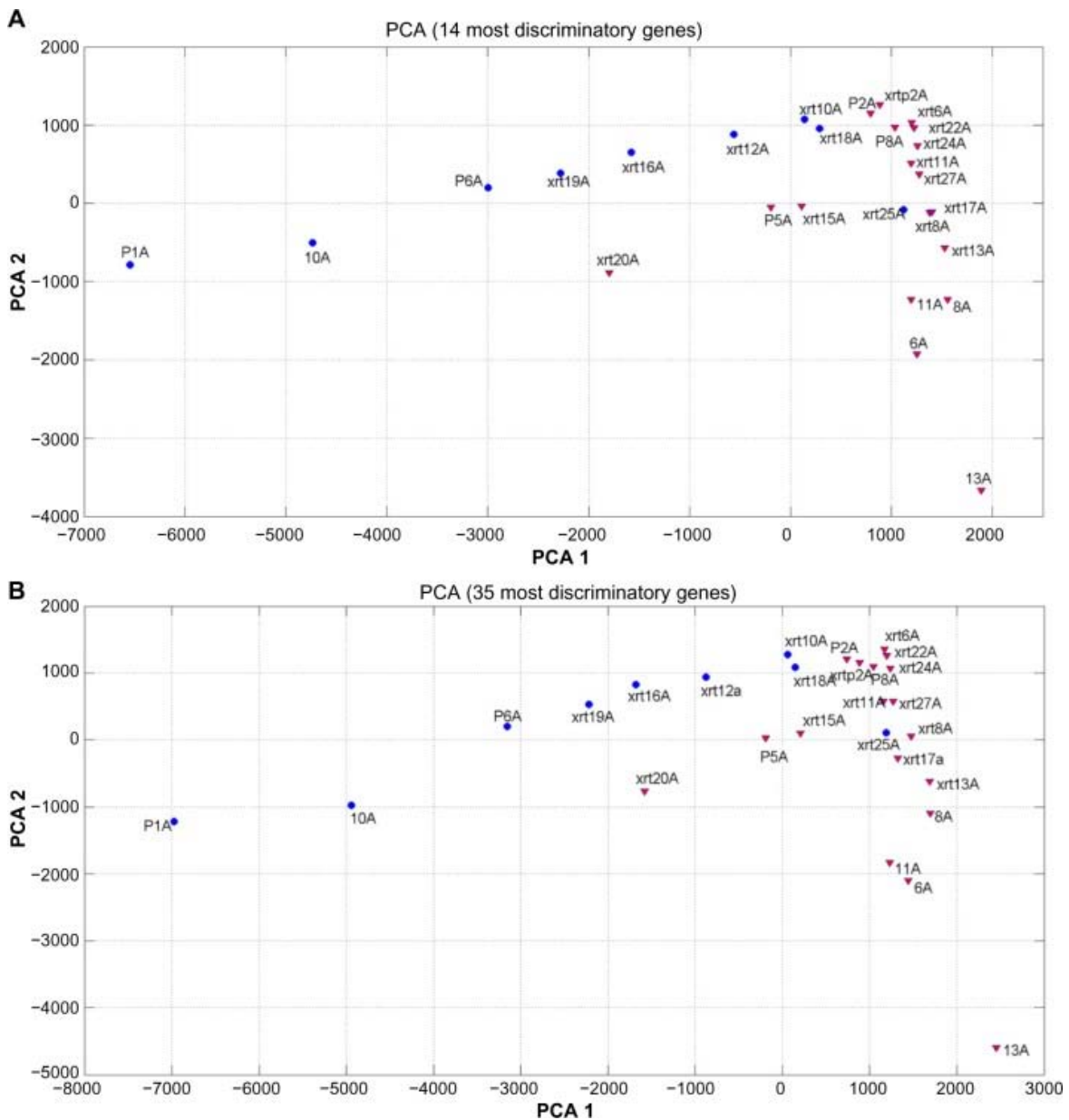


Fig. 3.8 **(A)** PCA plot for the learning set in the reduced base of the 14 most discriminatory genes. **(B)** PCA plot for the learning set in the reduced base of the 35 most discriminatory genes. A linear separability with a similar structure can be observed in both cases. Low Fatigue samples lie between P1A and xrt18A. Xrt25A might be a biological or behavioral outlier. High Fatigue (HF) samples lie between 13A and xrtp2A. Xrt20A marks the HF limit towards the west of the plot. Additional data are needed to perfectly delineate this PCA plot.

The algorithm provided 13 successes out of 17 validation samples. Three of the four misclassified samples belonged to the LF group (false positives, patients were predicted to be HF) and one to the HF (false negative, patient predicted to be LF). These samples are outliers

with respect to this classifier, because their expressions in the reduced base of genes are closer to the HF and LF groups, respectively (Tables 3.1, 3.2, and figure 3.9). Interestingly, the 14 different predictions for these misclassified samples coincide, that is, the probability of these samples belonging to their predicted class according to the consensus criterion is 1. This fact also strengthens the argument that these samples are biological or behavioral outliers, that is, their class assignment based on the change in their FACT-F scores was ambiguous.

Table 3.1 Mean values for the 14 most discriminatory genes.

Learning		Validation	
HF	LF	HF	LF
114	<b>388</b>	117	<b>401</b>
152	<b>644</b>	143	<b>546</b>
302	<b>1455</b>	326	<b>1569</b>
343	<b>1659</b>	364	<b>1535</b>
185	<b>861</b>	196	<b>841</b>
149	<b>611</b>	127	<b>460</b>
<b>585</b>	128	<b>381</b>	194
243	<b>1182</b>	252	<b>1049</b>
<b>689</b>	111	<b>536</b>	235
<b>160</b>	65	75	<b>126</b>
247	<b>1225</b>	275	<b>1187</b>
<b>223</b>	80	73	<b>171</b>
269	<b>1329</b>	331	<b>1573</b>
<b>1200</b>	281	<b>1083</b>	485

Bold values indicate the highest mean expression values in the learning and validation datasets for HF and LF classes.

Table 3.2 Misclassified samples.

S1 (xrt14)	S2 (xrt36)	S3 (xrt39)	S4 (xrt33)
57	<b>129</b>	87	<b>342</b>
78	<b>257</b>	105	<b>492</b>
136	<b>327</b>	201	<b>1354</b>
122	<b>309</b>	183	<b>1514</b>
79	<b>180</b>	125	<b>765</b>
92	<b>126</b>	168	<b>341</b>
42	<b>44</b>	54	<b>946</b>
103	<b>175</b>	184	<b>1045</b>
<b>41</b>	34	49	<b>1430</b>
62	<b>178</b>	<b>258</b>	52
77	<b>234</b>	183	<b>1142</b>
97	<b>286</b>	<b>374</b>	82
146	<b>239</b>	232	<b>1388</b>
162	<b>167</b>	137	<b>2518</b>

Expressions for the 14 most discriminatory probes. Samples S1, S2 and S3 were predicted to be High Fatigue and S4 to be Low Fatigue. The expression values for S1, S2 and S3 were closer to the mean expression of the High Fatigue group in the learning phase. Conversely, the expression values for S4 is closer to the Low Fatigue group. S1, S2 and S3 might define a new group of Low Fatigue with very small expressions (lower than the corresponding expressions observed among High Fatigue subjects) in this reduced base of 14 genes.

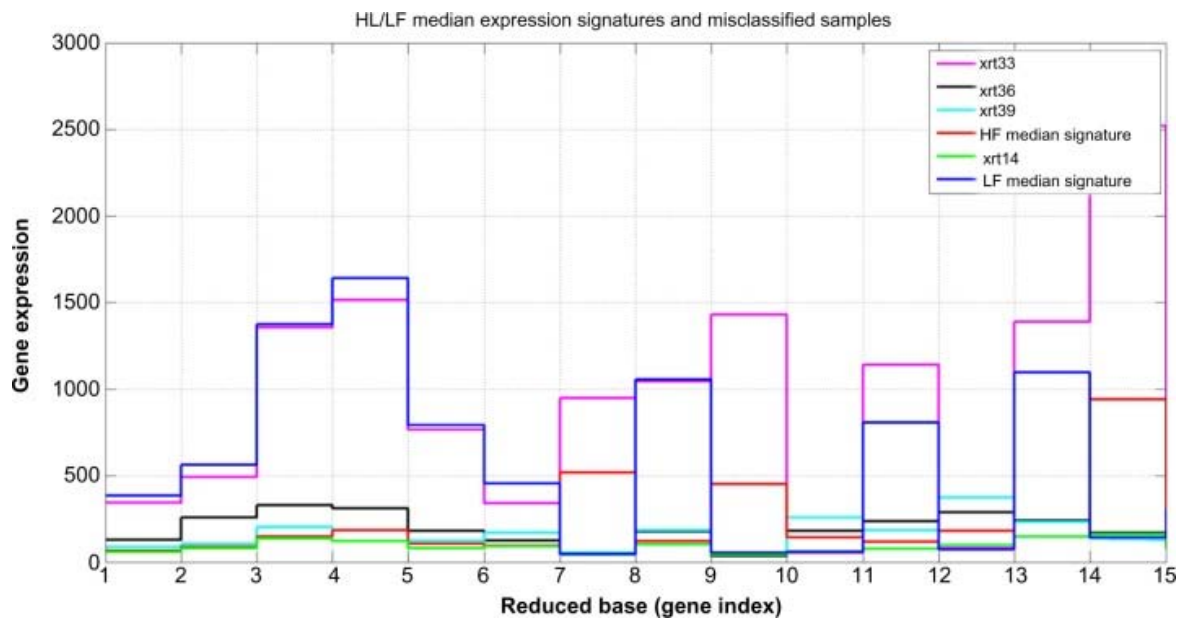


Fig. 3.9 High Fatigue (HF)/Low Fatigue (LF) median expression signatures and misclassified samples at validation. It can be observed that sample xrt33 is closer to LF median signature, while xrt14, xrt36 and xrt39 are closer to the HF median signature (values for the expressions are given in tables 2 and 3).

### 3.3.4 CRTF conclusions

EBRT is a highly utilized treatment option for many forms of cancer. While it is efficacious in many cases, its toxicity profile is significant and common, but not ubiquitous. Consequently, the ability to predict toxicities of EBRT has long been of interest. With better understanding of the pathobiology of radiation injury, using genomics as the basis for toxicity risk prediction has been the focus of active research.

We proposed that the risk of a complex disease, such as CRTF, could well be more easily defined by identifying groups of simultaneously expressed, synergistically functioning genes. Our finding that the gene cluster so identified was able to predict CRTF risk with an accuracy of  $> 75\%$  suggests that the approach has validity.

The process of selecting the most predictive cluster of genes revealed informative considerations. The genes with the highest FC did not coincide with those exhibiting the highest FR because the means of both distributions were different, hence their tails did not overlap. So, in this method we concluded that FR was a better feature selection method than FC. While, in the case of FC analysis, noisy genes are typically penalized by the FR selection method because of an increase of their variance; the noise might be amplified by the FC ratio. Genes

with the highest FR and FC have the biggest discriminatory power and are assumed to be involved in the genesis of fatigue.

Interestingly, the histogram analysis of the first 360 genes that most discriminated between HF and LF subjects was informative in that the shift of the mode of distribution showed lower expressions of these genes among HF subjects. It seems possible that it is this distributional shift that ultimately is responsible for discriminating the fatigue phenotype in this population.

We were unable to correctly predict four samples, based on our phenotypic approach, since the consensus provides the opposite class in all the cases. These classified samples were close to the border of separation between both fatigue classes (figure 3.8). There are three possibilities: (1) these samples are behavioral outliers, (2) the phenotypic approach needs further review and improvement, especially dealing with samples that are bordering the cut-off scores set for fatigue grouping, and (3) possible use of more sophisticated algorithms (black box neural networks) to classify the samples may be needed, which could run the risk of losing the clarity in the interpretation.

We recognize that this study was limited by its small sample size. Nonetheless, the fact that the analysis was successful in predicting LF/HF in an unrelated population with reasonable accuracy suggests that increasing the number of subjects in the training population would likely improve the predictive model's ability. Nevertheless, this analysis confirms that it is possible to separate both classes of the LF/HF phenotype by reducing the dimension to the most discriminatory genes, provided by their FR.

The importance of predicting toxicity or adverse event risk associated with cancer treatment regimens cannot be understated as the clinical implications in personalizing cancer therapy and prospectively attenuating toxicity risk are significant. Furthermore, this type of information provides patients and their care-givers more specific knowledge upon which to make treatment decisions. A future manuscript will be devoted to the gene attribution analysis of the cancer treatment-related fatigue biomarkers and pathways (in preparation).

### **3.4 Genomic data integration in Chronic Lymphocytic Leukemia**

In this work we applied our methodology explained in section 3.2 using both publicly available genetic data obtained from a microarray expression dataset and sequencing dataset to figure out how the main mutations defined by the sequencing data affect gene expression by finding small-scale signatures to predict those mutations. This work is currently under review in the "Journal of Gene Medicine" (see Appendix A.4).

### 3.4.1 Introduction to Genomic data Integration in Chronic Lymphocytic Leukemia

B-cell chronic lymphocytic leukemia (CLL) is a complex heterogeneous disease characterized by the accumulation of malignant B-cells in blood and lymphoid organs (Rodriguez-Vicente et al., 2013). Clinical diagnosis of CLL is based on the demonstration of an abnormal population of B lymphocytes in the blood, bone marrow, or tissues that display an unusual but characteristic pattern of molecules on the cell surface (CD5 and CD23 clusters of differentiation).

DNA analysis distinguishes two major types of CLL with different survival times (Hamblin et al., 1999). This distinction is based on lymphocyte maturity, as discerned by the immunoglobulin variable-region heavy chain (IgVH) gene mutation status. High-risk patients (with poor survival) have an immature cell pattern with few mutations in the IgVH gene region, whereas low risk patients show considerable mutations in the antibody gene region indicating mature lymphocytes. Since the determination of the IgVH mutation status is very labor-intensive and expensive, alternative markers have been investigated to better prognosticate disease progression.

Gene expression profiles were also used to understand the genesis and progression of CLL. Subsequently, whole-genome sequencing has identified four major genomic aberrations in cells that are strongly associated with the disease behavior and prognostically independent of IgVH mutational status (Döhner et al., 2000). More recently, whole-genome sequencing identified NOTCH1 and SF3B1 as the most frequently mutated genes that were predictive of CLL prognosis (Puente et al., 2011).

Given the low incidence of NOTCH1 (9%) and SF3B1 (8%) mutations, it seemed unlikely to us that CLL progression could be solely ascribed to the two. We therefore sought to identify shared/synergistic mechanisms among the three most common mutations (IgVH, NOTCH1 and SF3B1) which might better predict and explain disease progression and behavior.

### 3.4.2 CLL gene expression data

We used a publicly accessible microarray dataset consisting of 48807 probes were derived from 163 patients with a diagnosis of CLL (Ferreira et al., 2014). The expression data were originally presented in logarithmic scale ( $\log_2$ ) after the corresponding RMA preprocess. Of the original cohort of 163 patients, 92 had mutated IgVH, which was associated with a favorable prognosis, while IgVH was not mutated in the remainder ( $n=71$ ) and prognosticated an unfavorable outcome. The exome sequencing data is described by Quesada et al. (2012), who identified 1246 mutations resulting in protein coding changes. Six genes appeared to

be most frequently mutated (>5%): NOTCH1, SF3B1, NOP16, CHD2, ATM and LRP1B. Amongst the 163 samples we evaluated, NOTCH1 and SF3B1 mutational status were determined for 117 patients. Of these, 106 were unmutated for NOTCH1 and 107 were unmutated for SF3B1.

### 3.4.3 CLL results

In this research work we applied the methodology described in section 3.2 to elucidate how the main mutations affect gene expression by finding small-scale signatures to predict the IgVH, NOTCH1 and SF3B1 mutations (genomic data integration). We subsequently applied our method to define and understanding the biological pathways and correlation networks that are involved in the disease development with the potential goal of identifying new druggable targets.

#### **IgVH mutational status**

We determined the best set of genes that discriminates IgVH mutational status based on microarray expression and the class information defined by the IgVH phenotype using 92 mutated and 71 unmutated samples.

The shortest list with the highest predictive accuracy (93.3%) was composed by 13 first probes: LPL (2 probes), CRY1, LOC100128252 (2 probes), SPG20 (2 probes), ZBTB20, NRIP1 (2 probes), ZAP-70, LDOC1 and COBLL1. Table 3.3 shows the list of these genes, their associated FR, the mean ( $\mu_1$ ,  $\mu_2$ ) and the standard deviation ( $\sigma_1$ ,  $\sigma_2$ ) for each group, and the LOOCV accuracy (Acc(%)). FR was applied to the log2 of the expressions.

Figure 3.10A shows the Pearson Correlation (PC) network of the most discriminatory genes of the IgVH mutational status. The Normalized Mutual Information (NMI) correlation network is shown in figure 3.10B.



Table 3.3 IgVH mutational status prediction

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
LPL	40	70	380	272	4.6	87.1
LPL	26	33	146	102	3.7	86.5
CRY1	62	125	352	298	3.1	90.2
LOC100128252	29	43	224	194	3.0	90.2
LOC100128252	30	42	220	172	3.0	89.6
SPG20	24	35	111	85	2.9	91.4
ZBTB20	1943	505	982	417	2.8	91.4
NRIP1	275	183	63	81	2.7	91.4
SPG20	30	53	148	126	2.6	91.4
ZAP70	103	151	273	140	2.4	92.6
LDOC1	20	19	50	27	2.3	92.6
COBLL1	186	107	85	100	2.3	92.6
NRIP1	85	60	24	24	2.1	93.3

List of the 13 most discriminatory genes list with the highest predictive accuracy (93.3%), ordered by decreasing Fisher's ratio.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in class 1, (mutated IgVH), and  $\mu_2$  and  $\sigma_2$  for the unmutated group. FR (log) stands for the logarithmic Fisher's ratio.

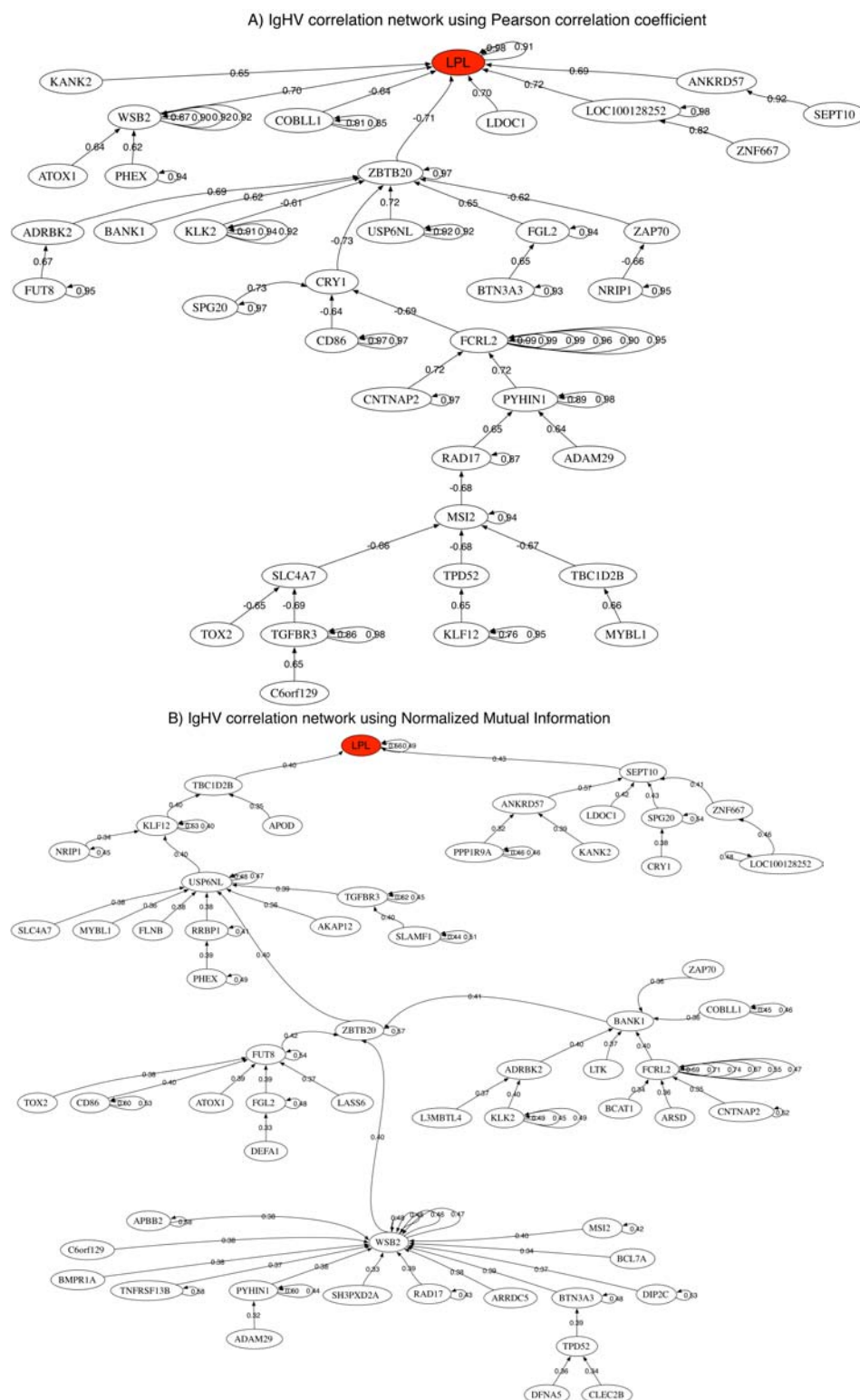


Fig. 3.10 Correlation network of the most discriminatory genes for the IgVH mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

### NOTCH1 mutational status

We recognized the challenge of analyzing those genes for which the NOTCH1 mutation impacted expression given the highly unbalanced sample mix (106 of 117 samples did not show the NOTCH1 mutation).

The shortest list with the highest predictive accuracy (95.7%) was composed by 60 probes with FR between 4.6 and 1.4 (see Table 3.4). The first five probes of this list corresponded to MSI2. Also using the two first probes of MSI2, the NOTCH1 mutation is predicted with 94.9% of accuracy. All MSI2 probes had lower expression in NOTCH1-mutation negative patients. One probe of the LPL gene appeared in eighth position in this list. Therefore the incremental accuracy from probe 5 to 60 was minimal (0.8%). That means the genes from the 6<sup>th</sup> position to the 60<sup>th</sup> serve to add high frequency details in the discrimination, as it has been pointed in our work commented on section 3.3.

Figure 3.11A shows the Pearson Correlation network of the most discriminatory genes of the NOTCH1 mutation in which three main networks associated to MSI2 through WSB2, ACSL5 and CNTNAP2 are apparent. The Normalized Mutual Information network (Figure 3.11B) demonstrates a main connection through NCK2.

Table 3.4 NOTCH1 mutational status prediction.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
MSI2	157	74	43	26	4.6	93.2
MSI2	238	123	62	49	4.1	94.9
MSI2	73	25	31	16	3.0	91.5
MSI2	283	149	92	61	2.8	90.6
MSI2	58	19	32	15	2.7	92.3
C10orf137	193	86	392	135	2.4	90.6
LAG3	236	155	77	103	2.4	90.6
LPL	357	250	170	254	2.3	92.3
NCK2	838	219	1560	529	2.2	93.2
CNTNAP2	66	96	667	799	2.1	92.3
ST3GAL1	38	11	85	36	2.1	90.6
CCDC24	109	73	48	44	2.0	92.3
LTK	216	96	103	132	2.0	90.6
FLNB	59	30	33	17	1.9	94.0
ZNF333	38	5	57	16	1.9	92.3
PREPL	190	62	329	108	1.9	93.2
C19orf28	120	37	217	80	1.9	93.2

Table 3.4 NOTCH1 mutational status prediction.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
C1orf38	365	148	189	109	1.8	91.5
LTK	107	52	52	64	1.8	91.5
SPG20	182	150	71	106	1.8	92.3
SAP30L	74	38	111	32	1.8	94.0
MYST1	248	37	322	60	1.7	93.2
C10orf137	99	41	187	66	1.7	94.9
ATP6V0B	831	198	596	183	1.7	91.5
LPL	130	89	75	99	1.7	92.3
SLC4A7	47	39	150	120	1.7	90.6
LOC100128252	161	126	112	156	1.7	89.7
HNRNPR	57	22	110	48	1.7	89.7
REEP5	41	18	80	39	1.6	90.6
SRSF1	110	60	175	52	1.6	94.0
GNPNAT1	37	8	64	24	1.6	94.0
SHPRH	270	64	383	83	1.6	94.0
CNTNAP2	101	140	804	1105	1.6	94.9
PHF2	119	44	175	60	1.6	92.3
FCRL1	234	180	525	308	1.6	93.2
WSB2	804	329	489	258	1.6	93.2
ATP6V0B	624	145	448	134	1.6	94.9
LYL1	87	31	140	47	1.5	94.9
ACSL5	230	85	332	106	1.5	94.9
STX17	50	21	75	25	1.5	94.0
SPG20	125	98	55	74	1.5	94.0
NHEJ1	29	7	37	8	1.5	94.0
ZNF248	48	25	89	45	1.5	93.2
MPST	55	20	35	10	1.5	93.2
CDK13	69	42	132	75	1.5	93.2
TRMT1	58	17	86	30	1.5	92.3
PI4K2A	224	101	115	84	1.5	93.2
ELOVL5	254	97	504	188	1.5	93.2
FAM30A	588	900	1535	1495	1.5	93.2
PTDSS1	129	21	190	44	1.5	94.0
PLGLB1	74	47	152	103	1.5	94.0

Table 3.4 NOTCH1 mutational status prediction.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
C5orf53	51	22	125	74	1.5	94.0
PSMD7	608	175	414	141	1.5	94.9
NASP	117	26	176	52	1.5	94.0
ATP6V0B	768	170	566	172	1.5	94.9
WDR36	108	36	164	43	1.4	94.9
LTN1	511	52	645	99	1.4	94.9
GAL3ST3	22	2	19	2	1.4	94.9
PDE7A	102	67	214	120	1.4	94.9
CAPRIN2	1098	345	1511	368	1.4	95.7

List of the 60 most discriminatory genes to predict the NOTCH1 mutation list with the highest predictive accuracy (95.7%), ordered by decreasing Fisher's ratio. Class 1 corresponds to samples with mutated NOTCH1 and class 2 corresponds to those with unmutated NOTCH1.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in class 1 (mutated NOTCH1), and  $\mu_2$  and  $\sigma_2$  for the unmutated group. FR (log) stands for the logarithmic Fisher's ratio.

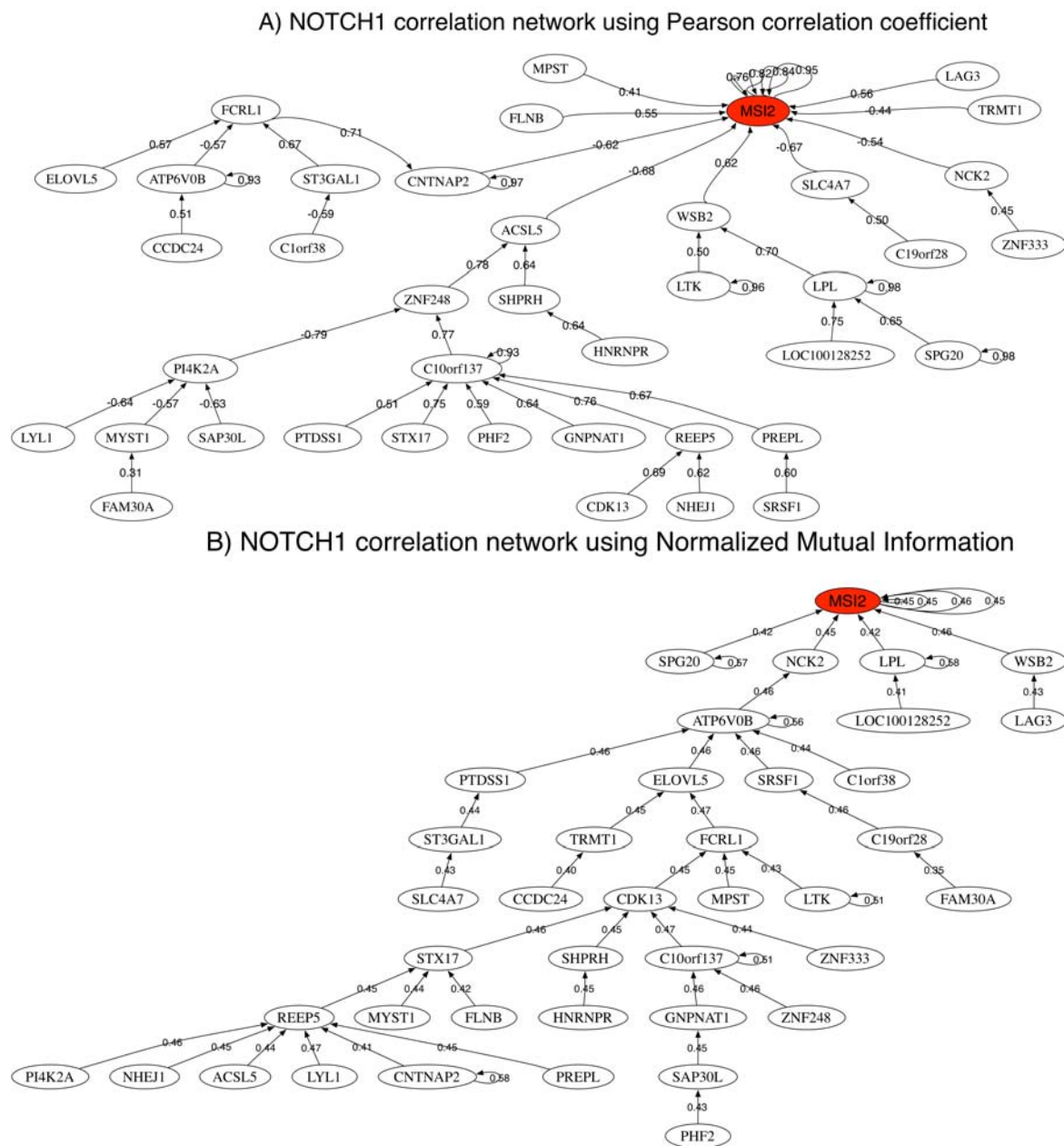


Fig. 3.11 Correlation network of the most discriminatory genes for the NOTCH1 mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

### SF3B1 mutational status

SF3B1 gene (Splicing Factor 3b, Subunit 1) is located in chromosome 2. Its importance in CLL has been analyzed by Wan and Wu (2013); Wang et al. (2011). As with NOTCH1, the

Table 3.5 SF3B1 mutational status prediction.

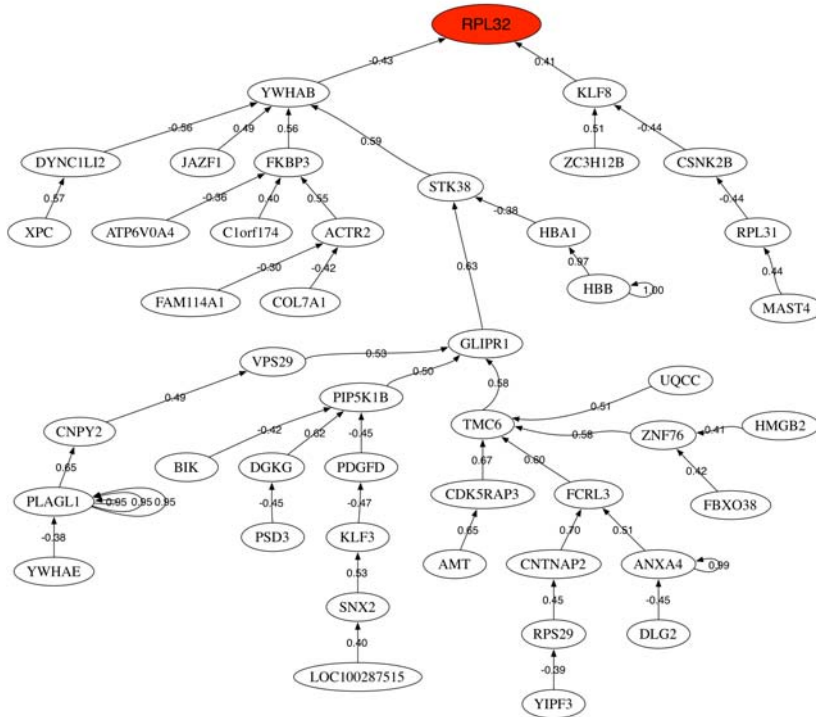
Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
RPL32	859	228	513	115	2.6	94.0
KLF8	131	45	59	30	2.4	94.0
PDGFD	85	34	42	20	2.2	95.7
PLAGL1	171	87	336	118	2.2	94.0
KLF3	40	29	239	221	2.2	94.0
UQCC	27	7	41	7	2.1	94.9
HBA1	3650	2978	755	2218	2.1	96.6
CNPY2	206	73	317	70	2.1	97.4
TMC6	322	74	546	155	2.0	97.4
CSNK2B	71	37	141	38	2.0	97.4
PLAGL1	282	135	507	174	2.0	97.4
PIP5K1B	55	32	212	200	1.9	98.3
DGKG	44	16	115	70	1.9	97.4
HBB	12044	6627	2783	5082	1.9	98.3
PLAGL1	138	83	252	92	1.9	98.3
ZNF76	34	8	61	20	1.8	98.3
AMT	48	8	97	41	1.8	97.4
STK38	206	108	368	156	1.8	97.4
HBB	8359	5278	1777	3669	1.8	97.4
ACTR2	3113	266	3789	506	1.8	97.4
GLIPR1	115	107	359	261	1.7	97.4
MAST4	136	89	59	60	1.7	99.1

List of most discriminatory genes (22) to predict the SF3B1 mutation, ordered by decreasing Fisher's ratio with an accuracy of 99.1%. Class 1 corresponds to samples with mutated SF3B1, and class 2 corresponds to those with unmutated SF3B1.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in class 1 (mutated SF3B1), while  $\mu_2$  and  $\sigma_2$  do for the unmutated group. FR (log) stands for the logarithmic Fisher's ratio.

SF3B1 classification problem was also highly unbalanced, since 107 CLL samples (out of 117) did not show the mutation.

The shortest list with the highest predictive accuracy (99.1%) was composed of 22 probes with FR's between 2.6 and 1.7. The most discriminatory gene was RPL32 (table 3.5). Figure 3.12A shows the Pearson Correlation network of the most discriminatory genes of the SF3B1 mutation. In general correlations between discriminatory genes are low. Two main networks were noted to be associated to the most discriminatory gene RPL32, through YWHAB and KLF8. Conversely, the correlation network using the Normalized Mutual Information (figure 3.12B) demonstrated a single network associated with CNPY2-STK38.

A) SF3B1 correlation network using Pearson correlation coefficient



B) SF3B1 correlation network using Normalized Mutual Information

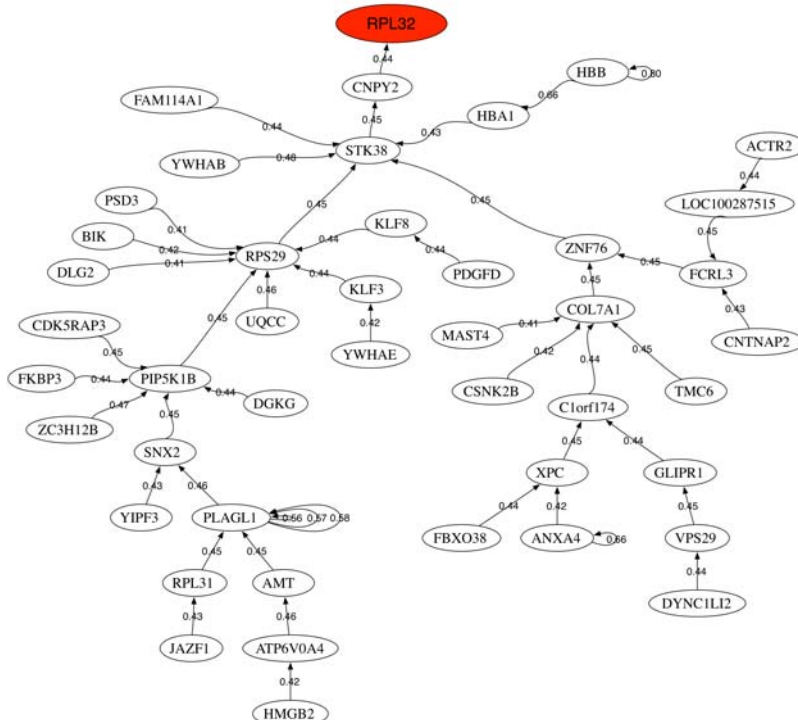


Fig. 3.12 Correlation network of the most discriminatory genes for the SF3B1 mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.



### Gene intersections for IgVH, NOTCH and SF3B1 mutations

We analyzed the intersection between the most discriminatory genes for IgVH, NOTCH1, and SF3B1 mutations as defined by FR and FC analyses. We consolidated both lists. The shortest lists found by FR and FC for each mutation and then performed pairwise intersections to establish shared genes. Figure 3.13 shows the result for these intersections. The intersection with the greater number of genes is NOTCH1-SF3B1 (19 genes), followed by IgVH-NOTCH1 (11 genes) and IgVH-SF3B1 with only 5 genes. Only four genes were common to all mutations: IGHG1, MYBL1, NRIP1 and RGS13.

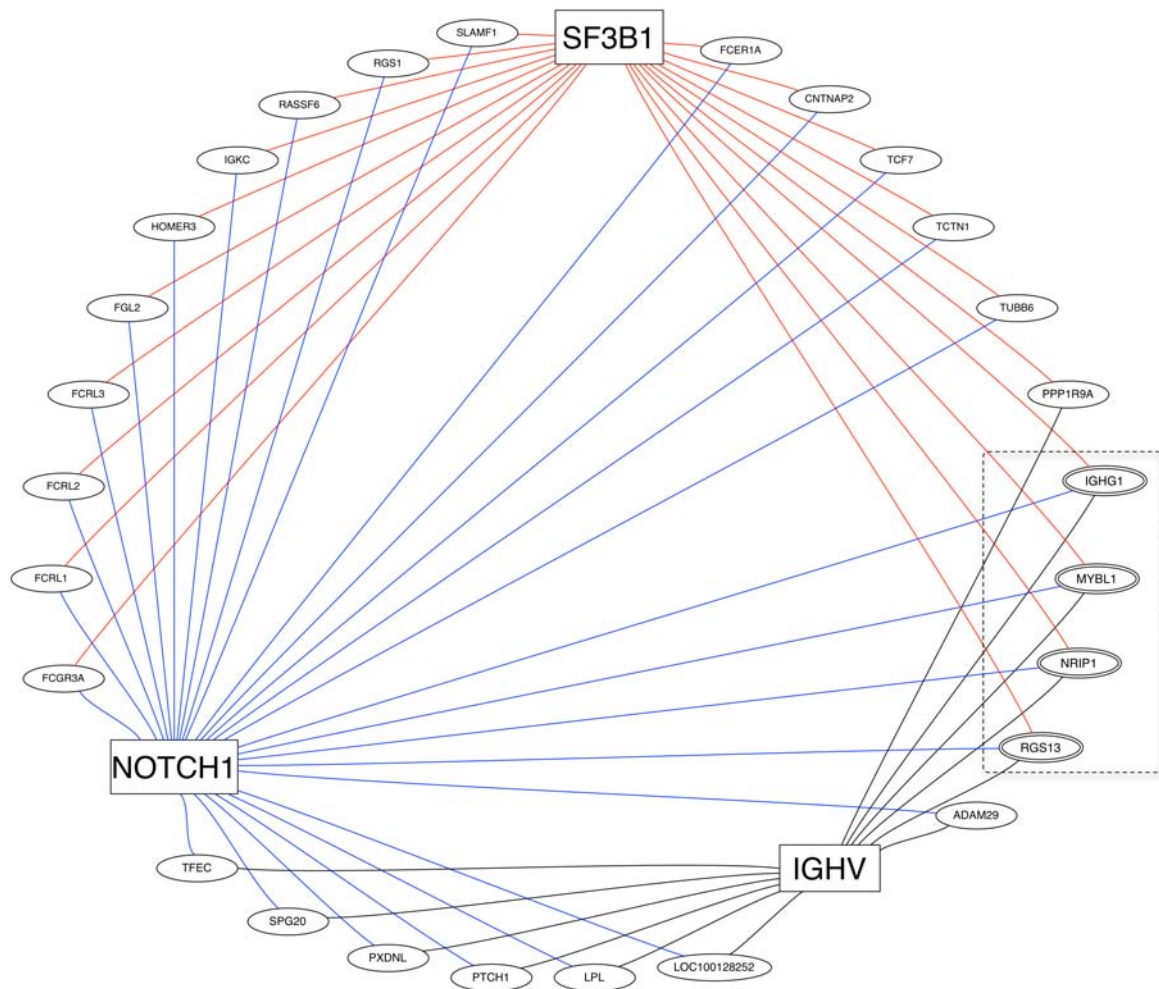


Fig. 3.13 Intersection among the most discriminatory genes of the IgVH, NOTCH1 and SF3B1 mutations. The three main mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with three edges (surrounded by a dot rectangle) are common to these three main mutations.

### 3.4.4 CLL conclusions

We showed the genomic data integration in CLL patients, by linking together microarray expression data and their IgVH, NOTCH1 and SF3B1 mutational status. Our methodological approach could define hierarchical gene relationships among CLL patients expressing these 3 different mutations and establishing the predictive accuracy of gene clusters relative to each mutation. Besides, our methodology served to depict the gene clusters that are most strongly associated with the expression of each selective mutation (networks of synergistically working genes), and their relationship between mutation expressions with a particular clinical outcome (survival). The biological significance of the findings for each of the mutational statuses can be found on the original manuscript (see appendix A).

The aim of this retrospective analysis was to provide a deeper understanding on the effects of the different mutations in the CLL disease progression, hoping that these findings will be used clinically in the near future with the development of new drugs. A future verification of these findings with other independent cohorts could lead to a better design of the therapeutic targets.

### 3.4.5 Additional results for NOP16 mutational status

NOP16 is the third mutation by percentage of occurrence (6.84%) in our cohort. Other authors have identified POT1 as the third most mutated gene using a more restricted dataset (Ramsay et al., 2013). Besides, NOP16 (NOP16 nucleolar protein) is an interesting target, since it is transcriptionally regulated by c-Myc, a gene that plays an important role in cell cycle progression, apoptosis and cellular transformation. Also, NOP16 is upregulated in breast cancer, being its over-expression associated to poor patient survival (Butt et al., 2008).

The NOP16 mutation has been predicted with an accuracy of 100% using the list of 26 most discriminatory genes provided by the FR (Table 3.6). Interestingly, the predictive accuracy obtained with only the two first genes of this list (SLC39A4 and WARS) is very high (97.4%).

Figure 3.14 shows the intersections between the lists of most discriminatory genes provided by the Fisher's ratio and fold change lists in each case. The intersections are as follows:

1. The intersection between NOP16 and NOTCH1 contains 6 genes: IGHG1, IGKC, IGKV3D-11, PXXDNL, RASF6 and RSG13.
2. The intersection between NOP16 and IgVH contains 4 genes: IGHG1, PXXDNL, SEPT10 and RSG13.

Table 3.6 NOP16 mutational status prediction.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
SLC39A4	32	5	48	13	2.5	88.9
WARS	122	63	70	46	1.4	97.4
CORIN	14	1	16	1	1.4	95.7
BRWD1	179	68	121	56	1.4	95.7
KLHL8	207	60	295	73	1.3	94.0
SIRT6	16	2	19	2	1.3	96.6
TCOF1	123	33	160	39	1.3	95.7
DCX	17	2	15	1	1.3	95.7
DSE	29	4	24	3	1.2	94.9
NONO	2685	175	2407	229	1.2	94.9
SLC1A7	20	2	18	3	1.2	96.6
BAD	51	11	65	14	1.2	95.7
SNORA16B	26	6	21	4	1.2	95.7
OR51F1	18	2	16	2	1.2	96.6
C9orf57	17	2	15	2	1.2	97.4
ABHD2	25	4	21	4	1.1	97.4
KIAA0907	871	364	1219	455	1.1	97.4
EDN3	15	1	17	1	1.1	97.4
UNC5B	29	4	25	4	1.1	97.4
OR1J4	18	1	16	2	1.1	97.4
PROZ	22	2	19	4	1.1	98.3
SEMA6A	14	1	15	2	1.1	98.3
MECR	31	10	38	7	1.1	99.1
GNA14	15	1	14	1	1.0	99.1
OPN5	14	1	16	2	1.0	99.1
CYP4Z2P	13	1	15	1	1.0	100.0

List of the 26 most discriminatory genes ordered by decreasing Fisher's ratio.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in class 1, (mutated NOP16), and  $\mu_2$  and  $\sigma_2$  for the unmutated group. FR (log) stands for the logarithmic Fisher's ratio, and Acc is the LOOCV predictive accuracy. The maximum accuracy (100%) is obtained with the first 26 most discriminatory genes. Also the list composed by the two first genes (SLC39A4 and WARS) provides a predictive accuracy of 97.4%.

3. The intersection between NOP16 and SF3B1 also contains 4 genes: IGHG1, IGKC, RASF6 and RSG13

Therefore, the longest intersection of NOP16 is with NOTCH1 mutation and only two genes belong to the intersection of the 4 mutations: IGHG1 and RSG13. IGHG1 (Immunoglobulin Heavy Constant Gamma 1) has been already related to hypogammaglobulinemia and B-cell chronic lymphocytic leukemia. This gene also plays a major role in antigen binding. RGS13 (Regulator of G-protein signaling 13) encodes a protein that is a member of the regulator of G protein signaling (RGS) family. Down-regulation of RGS13 has been observed in mantle cell lymphoma (Islam et al., 2003). In the present case RGS13 is upregulated in the group with mutated NOP16. RGS13 over expression inhibited CXCL12-evoked Ca(2+) mobilization, Akt phosphorylation and chemotaxis (Bansal et al., 2008). Also it has been also shown that p53 negatively regulates RGS13 protein expression in immune cells (Iwaki et al., 2011).

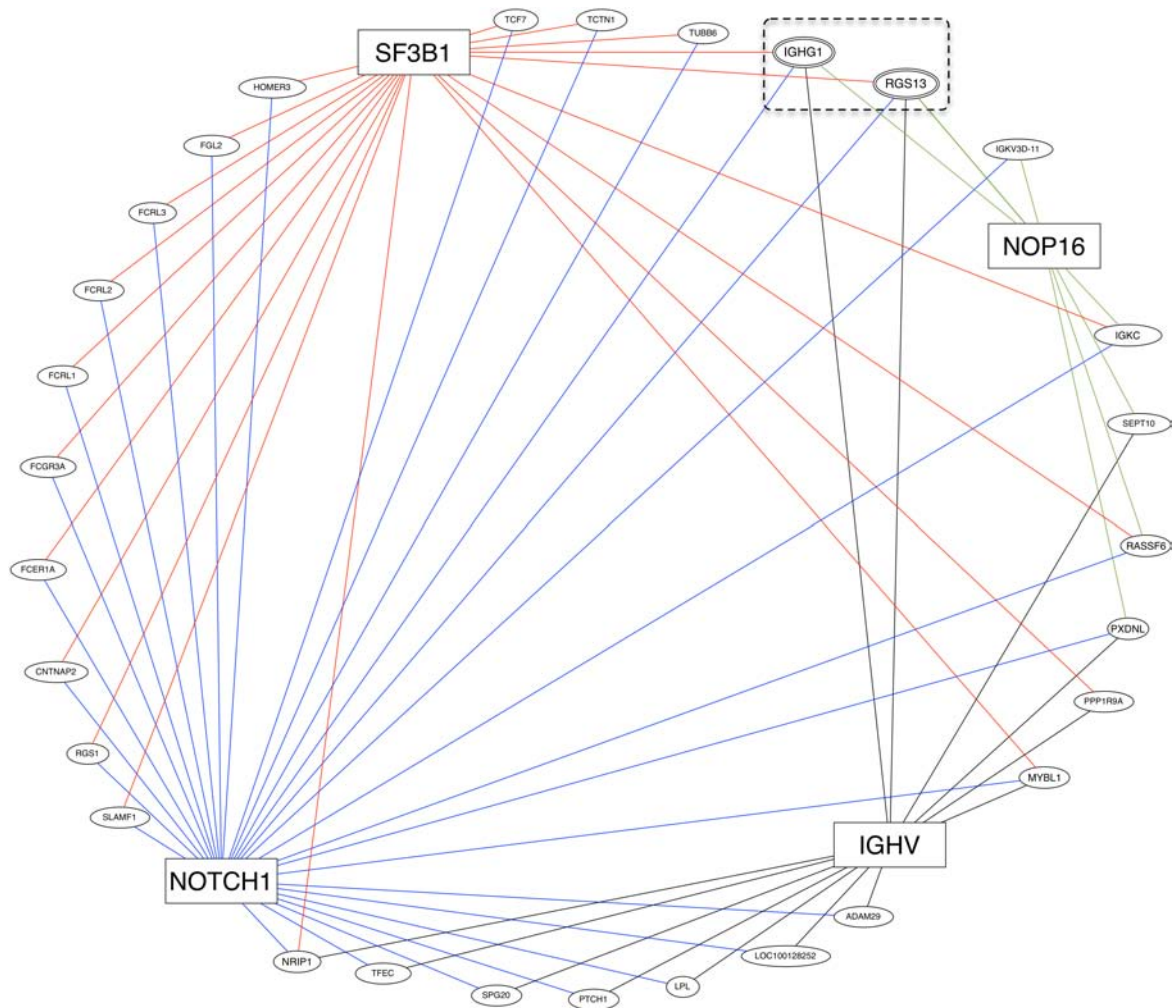


Fig. 3.14 Intersection among the most discriminatory genes of the IgVH, NOTCH1, SF3B1 and NOP16 mutations. The four mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with four edges (surrounded by a dot rectangle) are common to these four mutations.



# Chapter 4

## Sensitivity analysis

### 4.1 Introduction

Hitherto, we have applied our methodology to the main different kind of biomedical data, showing that we can solve diverse biomedical problems precisely and effectively, and using limited resources. In this section we check the robustness of the methodology against the main sources of noise and how the most common biomedical data preprocessing techniques affect it. We have tested it using genetic data, particularly microarray datasets. The result of the noise analysis using genetic data could be extended to other types of data since the noise we manage is present in every type of data regardless of their origin. However, the preprocessing techniques we managed are common to microarray expression data. We focused in these preprocessing techniques since they are very well-known and commonly applied. Preprocessing techniques in clinical data, due to their heterogeneity, is an extensive topic out of scope in this dissertation. Both works are reflected in two manuscripts: "Sensitivity analysis of gene ranking methods in phenotype prediction" currently under review in the "Journal of Biomedical Medicine" (see Appendix A.5) and "Impact of microarray preprocessing techniques in unraveling biological pathways", accepted for publication in the "Journal of Computational Biology" (see Appendix A.6).

The chapter is structured in two main parts. Firstly we theoretically analyzed the effect of noise in phenotype prediction problems. Via synthetic modeling, we performed the sensitivity analysis for the main gene ranking methods applied in our methodology to different types of noise. We then studied the predictive accuracy of our biomedical robot in synthetic data and in three different datasets related to cancer, rare and neurodegenerative diseases to better understand the translational aspects of our findings. Secondly, we analyze the impact of the main microarray preprocessing techniques on the analysis of biological pathways in the prediction of cancer treatment-related fatigue performed in section 3.3. We compared the

Robust Microarray Averaging (RMA) and the Affymetrix's MAS5 method with the results that are obtained working with raw data.

## 4.2 Sensitivity analysis of gene ranking methods in phenotype prediction

In this section we first theoretically analyzed the effect of noise in phenotype prediction problems by casting them into abstract optimization problems. To accomplish this, we first show that noise in data can be expressed as a modeling error that partially falsifies the set of discriminatory probes that are phenotype-related, and therefore the biological pathways that are involved. Secondly, the sensitivity to different kind of noise (in expression and class assignment) for the following gene ranking methods explained in section 1.3.6: Fold Change (FC), Fisher's Ratio (FR), Maximum Percentile Distance (MPD) and Entropy (EN); compared to well-established Significance Analysis of Microarrays (SAM) is performed via synthetic microarray modeling. This analysis has shown that in general terms FR is the most robust method in terms of precision closely followed by SAM. Besides, both methods provided the smallest subsets of genes with the highest discriminatory power. The effect of noise increases the number of genetic probes that are needed to slightly improve the predictive accuracy. This is a very important result concerning parsimony principle. Therefore, an optimum method to find the biological pathways in translational problems will consist of ranking the differential expressed genes decreasingly by their corresponding FR. Additionally, to avoid variable distributions with very low variances but with means/medians very close which would derive in high FR's, a first preselection with a low cut off using FC could be performed, as we will see in section 4.3.

The results of these analyses are confirmed using three different datasets concerning the study of cancer (Chronic Lymphocytic Leukemia), rare diseases (Inclusion Body Myositis) and neurodegenerative diseases (Amyotrophic Lateral Sclerosis). We found that FR and SAM provide the highest predictive accuracies with the smallest number of genes, exploiting the principle of parsimony. Besides, we show their corresponding biological found with an expanded list of genes whose discriminatory power has been established via FR. In these three cases, the effect of viral infections in the corresponding pathways is clear. The results of this analysis is important to optimize the use of these methods in translational medicine, particularly in the biological understanding of different diseases and in drug optimization problems.



### 4.2.1 The effect of noise in phenotype prediction

One of the main obstacles in the analysis of genomic data is the absence of a conceptual model that relates the different probes to the class prediction (phenotype). Therefore, we need to model these complex relationships. For this reason, similarly what we defined in section 1.3.1 and equation (1.1), a classifier  $L^*(\mathbf{g})$  has to be constructed and it is defined as an application between the set of genetic signatures  $\mathbf{g}$  and the set of classes  $C = \{c_1, c_2, \dots, c_n\}$ : in which the phenotype is divided:

$$L^*(\mathbf{g}) : \mathbf{g} \in \mathbb{R}^s \rightarrow C = \{c_1, c_2, \dots, c_n\}. \quad (4.1)$$

For this specific problem and following equations (1.2), (1.3) and (1.4) the optimization problem of finding the subset of genetic signatures  $\mathbf{g}$  that maximizes the learning accuracy, giving a subset of samples  $\mathbf{T}$  (training data set) whose class vector is known,  $\mathbf{c}^{obs}$ , can be written as follows:

$$\mathbf{g} : O(\bar{\mathbf{g}}) = \min_{\mathbf{g} \in \mathbb{R}^s} O(\mathbf{g}), \quad (4.2)$$

$$O(\mathbf{g}) = \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p \quad (4.3)$$

$$\mathbf{L}^*(\mathbf{g}) = (L^*(\mathbf{g}_1), \dots, L^*(\mathbf{g}_i), \dots, L^*(\mathbf{g}_m)), \quad (4.4)$$

As we explained in chapter 3, one of the main numerical difficulties in learning is the high dimensionality of the genomic data since the number of monitored probes (or genes) is much greater than the number of samples (or patients). This fact provokes that the phenotype prediction in the learning stage will have a very high underdetermined character. Therefore, several gene lists with similar predictive accuracy might exist. All these high predictive lists are expected to be involved in the genetic pathways that explain the phenotype. In practice, the predictive accuracy of a genetic signature,  $O(\mathbf{g})$ , is performed via cross-validation. This knowledge could be very important for early diagnosis and treatment optimization.

The presence of noise in the genomic data will impact the classification and obviously the pathway analysis resulting from this procedure. There are at least two main sources of noise in phenotype prediction problems as we detailed in section 1.3.3: Noise in the feature data (gene expression), and noise in class assignment. Consequently the perturbed and noise-free cost functions,  $O^p(\mathbf{g})$  and  $O^f(\mathbf{g})$  will never achieve their corresponding minima for the same genetic signatures  $\mathbf{g}$ . Therefore, the impact of noise in the optimum genetic signature is a fail in the generalization of new incoming samples. For that reason, it is also desirable to inspect the genetic signatures having a lower predictive accuracy than the optimum.

To alleviate the high underdetermined character of genomic-phenotype prediction problems, feature selection methods are used to reduce the dimensionality of the genetic data. The problem of determining the genes that separate two (or more) classes corresponding to given phenotypes has been traditionally been addressed by filter, wrapper and embedded methods (Saeys et al., 2007). In the case of filter methods, the gene selection and the classifier for phenotype prediction are independent (uncoupled). Wrapper and embedded techniques are most sophisticated approaches where the gene selection is the solution of an optimization problem; therefore selection and classification are coupled. Wrapper and embedded methods usually involve the use of neural network, support vector machines, decision trees and global optimization algorithms. Filter methods rank different genes according to different measures of their discriminatory power in phenotype prediction problems. The fact that wrapper and embedded methods involve optimization also implies that an uncertainty analysis of the feature selection problem is involved. For that reason we find that filter methods are more interesting.

#### 4.2.2 Gene selection ranking methods and noise

To determine the stability and robustness of the mentioned ranking algorithms in mitigating microarray-generated noise, we compared them using a synthetic dataset and publicly available datasets associated with Chronic Lymphocytic Leukemia, Inclusion Body Myositis, and Amyotrophic Lateral Sclerosis. At a translational level, the aim of this analysis is to establish an optimum way to find the most discriminatory genes in a phenotype prediction and the biological pathways that are involved.

A variety of analyses have been performed to study the sensitivity of some of these methods to noise in the expression data (Dinu et al., 2007; Jeffery et al., 2006; Kooperberg et al., 2002; Larsson et al., 2005). However, so far the robustness against different kind of noises for all these ranking methods has not been addressed.

For that purpose, we used a synthetic dataset where three different types of noise were introduced: additive Gaussian noise, lognormal noise and noise in the class assignment. The Gaussian noise has been introduced through a random number generator following a normal distribution  $n_j \rightarrow N(0, r_k E_j^t)$  for each gene, being  $r_k$  the noise level, and  $E_j^t$  is the noise-free expression of the gene  $j$ . Therefore, the noisy expression corresponding to the gene  $j$  would be:

$$E_j^p = E_j^t + n_j. \quad (4.5)$$

The lognormal noise has been obtained by adding Gaussian noise to the logarithms of the expression:

$$\ln_j = \log_2 s_j \rightarrow N(0, r_k \log_2 E_j^t). \quad (4.6)$$

Therefore, the lognormal noise has a scaling effect, since:

$$\begin{aligned} \log_2 E_j^p &= \log_2 E_j^t + \ln_j, \\ E_j^p &= s_j E_j^t. \end{aligned} \quad (4.7)$$

In the case of class assignment noise, a given number of samples are misclassified. The class assignment and lognormal noises belong to the category of non-Gaussian noise. The synthetic dataset was built with a predefined number of differentially expressed genes. We subsequently introduced different levels of noise: 1 to 6% for Gaussian and log-Gaussian noises and 10 to 40% for the class assignment noise.

To check the performance of the different ranking methods we used the Precision metric:

$$Precision = \frac{|\{DE_{genes}\} \cap \{Selected_{genes}\}|}{|\{Selected_{genes}\}|} \quad (4.8)$$

where  $\{DE_{genes}\}$  is the set of the differentially expressed genes and  $\{Selected_{genes}\}$  the set of genes selected by the ranking algorithm.

### 4.2.3 The synthetic and diseases datasets

A flow diagram for the methodology used in this paper is shown in figure 4.1. The synthetic datasets was created to compare the various filtering methods against a known dataset and then, based on these findings, create a hierarchy which defines the effectiveness of the ranking methods against different kind of noise and to understand how to find optimally the biological pathways in disease datasets.

The synthetic dataset was built simulating a real dataset related to Chronic Lymphocytic Leukemia (see section 3.4.2 for further details) using the OC-plus package available for The Comprehensive R Archive Network (Pawitan and Ploner, 2015). The original data was compound of 163 samples and 48807 probes. We have chosen this dataset for building the synthetic dataset because it has a good sample size and the class is well balanced. The experiment was set up as follows:

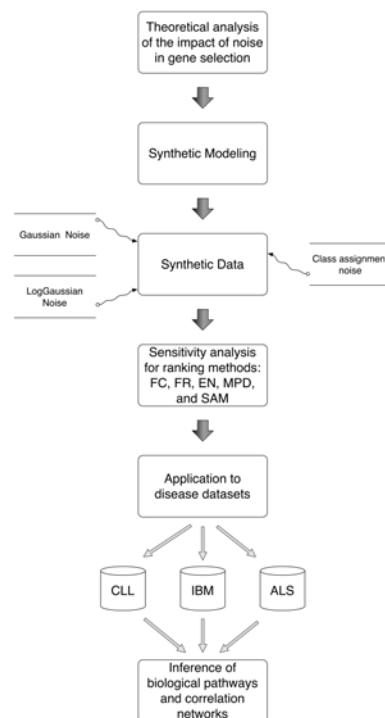


Fig. 4.1 Flow diagram of the noise analysis methodology

- The class of the synthetic dataset was the same as the one observed for the IgVH status (Ferreira et al., 2014): 92 samples had mutated IgVH, while in the other 71 samples IgVH was not mutated.
- The noise-free synthetic data set (expression) was generated using as main parameters  $D = 2$  and  $P_0 = 0.47$  where  $D$  is the effect size for differentially expressed genes expressed in units of the gene-specific standard deviation and  $P_0$  is the proportion of differentially expressed genes. This simulation made 229 genes be differentially expressed which we will try to recover via the different gene-ranking methods. These genes are supposed in the synthetic dataset to optimally differentiate the known IgVH status.

Furthermore, we have modeled different real microarray datasets to confirm these findings:

- B-cell Chronic Lymphocytic Leukemia (CLL) dataset composed by 163 samples and 48807 probes (Ferreira et al., 2014). CLL is a complex and molecular heterogeneous disease which is the most common adult Leukemia in western countries. DNA analyses served to distinguish two major types of CLL with different survival times based on the maturity of the lymphocytes, as discerned by the Immunoglobulin Heavy chain

Variable-region (IgVH) gene mutation status. 92 samples had the IgVH gene mutated versus 71 samples with worse prognosis. The aim of this analysis is to find the pathways that are associated with bad prognosis in CLL patients (see section 3.4.2 for further details about this dataset).

- **Inclusion Body Myositis (IBM):** microarray studies (with 22283 probes) were performed on muscle biopsy specimens from 34 patients with inclusion body myositis and 11 samples without neuromuscular disease (Greenberg et al., 2005). IBM is a muscle disease characterized by chronic, progressive muscle inflammation accompanied by muscle weakness. The aim of this analysis is to find the pathways that are associated to the development of IBM with respect to healthy controls.
- **Amyotrophic Lateral Sclerosis (ALS)** dataset composed by 85 samples (57 samples are ALS cases and 28 healthy controls) and 54675 probes (Lincecum et al., 2010). ALS is a fatal neurodegenerative disease characterized by progressive loss of motor neurons. These authors have shown that the co-stimulatory pathway is upregulated in the blood of a high percentage of human patients with ALS (56%). The aim of this analysis is to define the genes that are associated with a diagnosis of ALS, the possible causes and the biological pathways that are involved.

These datasets are representative of 3 different types of diseases: cancer, rare and neurodegenerative diseases. Besides, they have a reasonable sample size and a good balance between both classes in each case. Although all the microarray datasets treated herein are post processed via the RMA algorithm that performs an estimation and correction of the noise (Irizarry et al., 2003), noise is still present due to the complexity of the data acquisition. Because the genes which are differentially expressed in real datasets are unknown, we applied the methodology explained in section 1.3.6 to select the smallest subset of high discriminatory probes.

#### 4.2.4 Results using synthetic dataset

In order to compare the performance we calculated the precision for each method, considering the set of 229 genes that were differentially expressed in the synthetic dataset. Table 4.1 provides the precision for all the ranking methods mentioned above for different noise types and levels. Table 4.2 shows the LOOCV mean accuracy and the number of selected genes in each method. What is more, we have also calculated the empirical Cumulative Distribution Functions (CDF) of the positions of the differentially expressed genes captured by each method. For the sake of clearness we only used the first 1000 gene positions. A perfect CDF

would be a straight line reaching the value of 1 at position 229. Figure 4.2, 4.3 and 4.4 shows these CDF curves for each type of noise and noise level.

Table 4.1 Synthetic modeling precision. Precision for each of the noise types at different noise levels.

		1%	2%	3%	4%	5%	6%
GAUSSIAN	FR	<b>1.00</b>	<b>0.97</b>	<b>0.86</b>	<b>0.72</b>	<b>0.65</b>	<b>0.55</b>
	FC	0.64	0.64	0.61	0.56	0.53	0.46
	EN	0.85	0.75	0.68	0.55	0.5	0.43
	MPD	0.28	0.31	0.32	0.34	0.34	0.34
	SAM	0.94	0.91	0.80	0.67	0.60	0.51
LOG-GAUSSIAN	FR	<b>0.84</b>	<b>0.62</b>	<b>0.41</b>	0.26	0.21	<b>0.16</b>
	FC	0.60	0.54	0.38	<b>0.27</b>	<b>0.23</b>	<b>0.16</b>
	EN	0.67	0.45	0.31	0.18	0.12	0.10
	MPD	0.32	0.36	0.28	0.24	0.19	0.14
	SAM	0.79	0.57	0.38	0.25	0.20	0.15
		10%	15%	20%	25%	30%	35%
CLASS	FR	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>0.94</b>	<b>0.68</b>	<b>0.40</b>
	FC	0.53	0.52	0.41	0.29	0.25	0.16
	EN	0.87	0.88	0.82	0.77	0.50	0.32
	MPD	0.27	0.26	0.22	0.19	0.18	0.12
	SAM	0.94	0.94	0.93	0.88	0.64	0.37

Table 4.2 Synthetic modeling accuracy. Mean LOOCV predictive accuracy for each of the noise types at different noise levels.

		1%	2%	3%	4%	5%	6%
GAUSSIAN	FR	100.00 / 8	<b>100.00 / 5</b>	<b>100.00 / 6</b>	100.00 / 5	100.00 / 13	<b>100.00 / 8</b>
	FC	100.00 / 9	100.00 / 12	100.00 / 12	100.00 / 9	<b>100.00 / 9</b>	100.00 / 12
	EN	100.00 / 17	100.00 / 11	100.00 / 8	100.00 / 12	100.00 / 19	100.00 / 28
	MPD	100.00 / 21	100.00 / 19	100.00 / 23	100.00 / 17	100.00 / 17	100.00 / 22
	SAM	<b>100.00 / 6</b>	<b>100.00 / 5</b>	<b>100.00 / 6</b>	<b>100.00 / 4</b>	100.00 / 14	<b>100.00 / 8</b>
LOG-GAUSSIAN	FR	100.00 / 6	100.00 / 22	100.00 / 47	<b>100.00 / 29</b>	100.00 / 37	<b>100.00 / 88</b>
	FC	100.00 / 9	100.00 / 16	100.00 / 48	<b>100.00 / 29</b>	100.00 / 37	100.00 / 119
	EN	<b>100.00 / 4</b>	<b>100.00 / 14</b>	<b>100.00 / 24</b>	100.00 / 38	100.00 / 45	100.00 / 132
	MPD	100.00 / 22	100.00 / 23	100.00 / 111	100.00 / 37	100.00 / 46	100.00 / 128
	SAM	100.00 / 8	100.00 / 18	100.00 / 47	<b>100.00 / 29</b>	<b>100.00 / 28</b>	100.00 / 90
CLASS		10%	15%	20%	25%	30%	35%
	FR	90.18 / 3	85.28 / 14	<b>83.44 / 2</b>	<b>76.07 / 4</b>	<b>73.62 / 2</b>	69.94 / 213
	FC	90.18 / 10	84.66 / 8	80.98 / 188	76.07 / 52	72.39 / 183	<b>69.94 / 85</b>
	EN	90.18 / 25	85.28 / 28	81.60 / 18	75.46 / 2	73.62 / 3	71.17 / 4
	MPD	90.80 / 121	85.89 / 180	80.98 / 29	75.46 / 23	71.17 / 33	66.87 / 46
SAM	<b>90.80 / 5</b>	<b>85.28 / 4</b>	<b>83.44 / 2</b>	<b>76.07 / 4</b>	<b>73.62 / 2</b>	69.33 / 5	

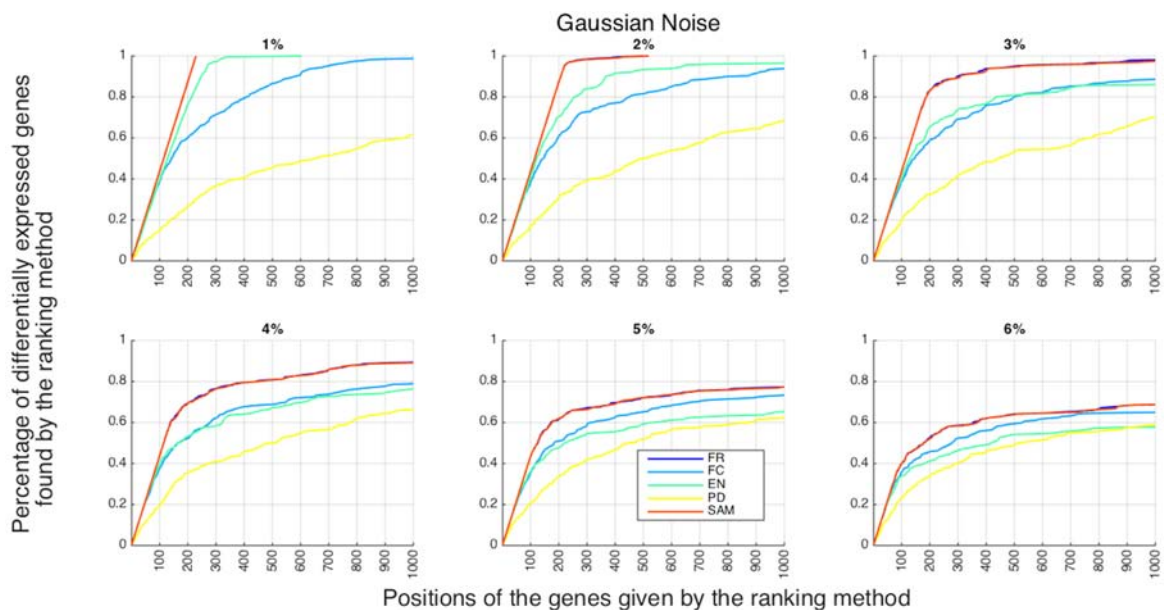


Fig. 4.2 Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for Gaussian noise.

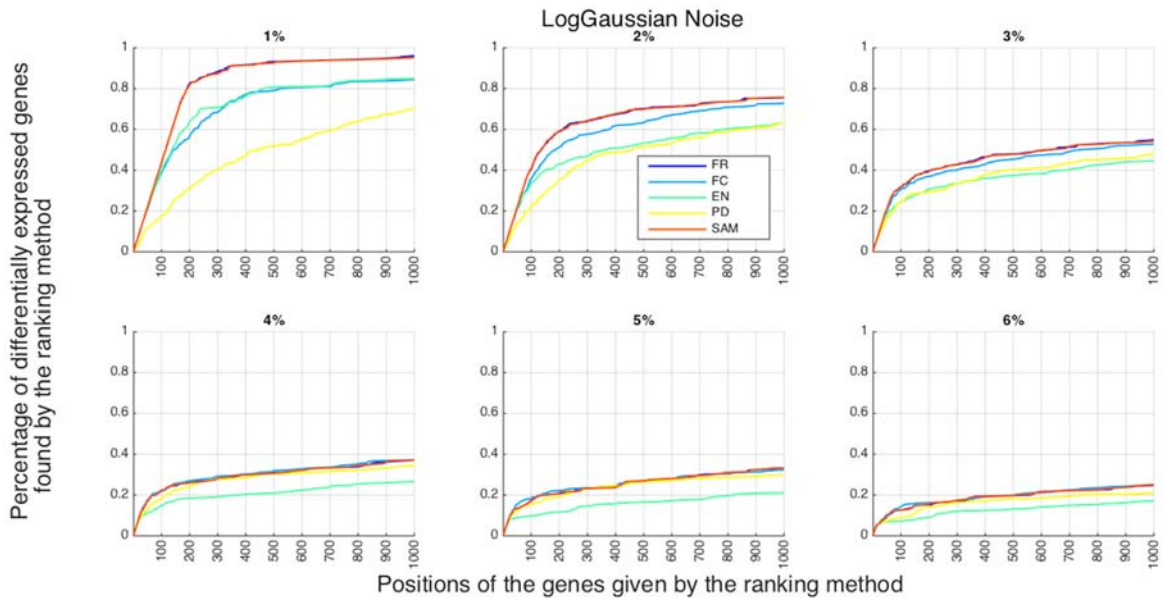


Fig. 4.3 Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for log-Gaussian noise.

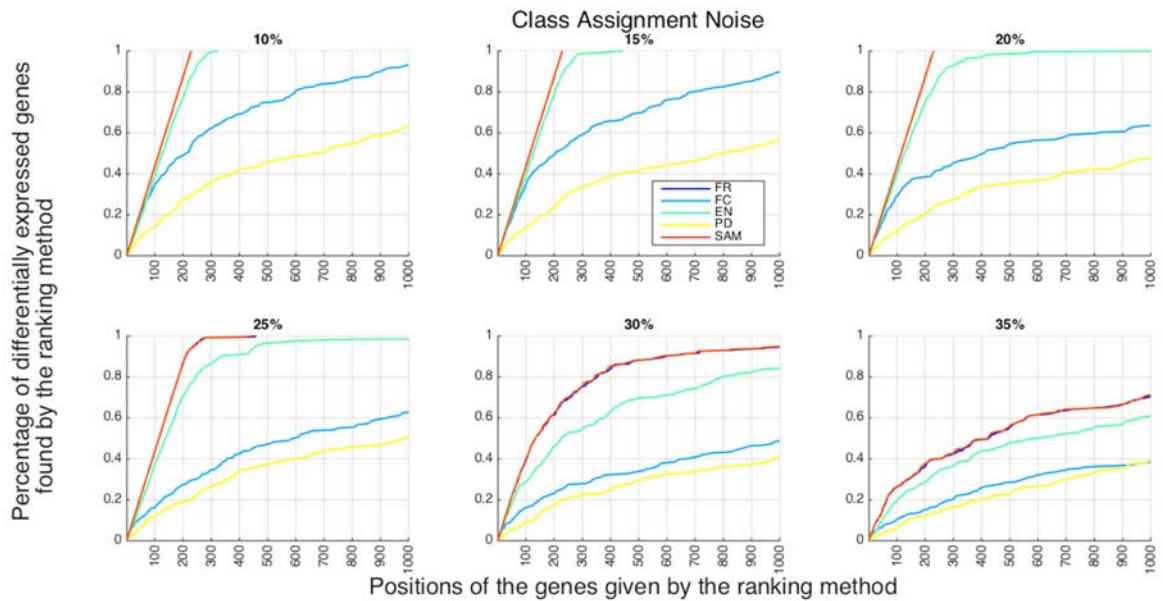


Fig. 4.4 Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for class assignment noise.

It can be observed the following:

- The precision decreases for all the methods as the noise level increases (refer to Table 4.1). The FR provides the best precision score for all the noise types and levels. These



differences decrease very fast with the noise level in the case of lognormal noise. The precision figures for SAM, in some cases, are very close to FR. In the case of class assignment noise, FR keeps precision levels up to 90% for 10 to 25% of noise, showing a very good robustness against this type of noise. This result has an important translational impact in real datasets to find the biological pathways that are involved in the disease development.

- The differences in the LOOCV mean accuracy (table 4.2) is not so clear and all methods provide similar results for the three types of noise at the different levels in the expenses of increasing the number of probes needed to improve the LOOCV predictive accuracy. In the case of Gaussian noise, SAM and FR show very similar results obtaining 100% of predictive accuracy with a much more reduced set of selected probes. Regarding lognormal noise, entropy seems to be the best for lower level of noises, while SAM and FR behave better when the noise level increases. FR and SAM are the best methods with a very little difference between them in the case of class assignment noise. These conclusions can also be clearly observed in the CDF curves (figures 4.2 to 4.4).
- We have also combined the Gaussian noise and the Log-Gaussian noise with the noise in the class assignment obtaining similar results. Adding the class assignment noise to a noisy dataset (for both Gaussian and Log-Gaussian noises) affects much more in finding the differentially expressed genes since the Precision decreases drastically (see table 4.3). What is interesting is that the FC seems to work better in terms of precision for a combination of class assignment and log-Gaussian noise. In terms of predictive accuracy more genes are needed to have a high predictive accuracy when class assignment noise is present (see table 4.4). In this case, FR and SAM provide the best results. Furthermore, it is possible to observe that for high levels of noise we can achieve high predictive accuracy with null precision at the expenses of adding a lot of genes to the predictive genetic signature. In this case, the biological pathways are clearly falsified.

In conclusion, noise in class assignment affects the selection of the important discriminatory genes in phenotype prediction problems more than noise in the expression data. This result emphasizes the importance in translational medicine of having at disposal a correct class assignment of the samples, provided by the doctors. Moreover, the methodologies used for solving the phenotype prediction problems should be accordingly designed, since the strategy of finding the best result might be in this case suboptimal, because noise impact the results.

Table 4.3 Synthetic modeling precision with combined noise

		1% / 10%	2% / 15%	3% / 20%	4% / 25%	5% / 30%	6% / 35%
GAUSSIAN & CLASS	FR	<b>0.97</b>	<b>0.72</b>	<b>0.45</b>	<b>0.16</b>	0.07	<b>0.04</b>
	FC	0.45	0.39	0.35	0.15	<b>0.11</b>	0.08
	EN	0.82	0.55	0.36	0.10	0.04	0.04
	MPD	0.26	0.26	0.22	0.11	0.09	0.08
	SAM	0.91	0.68	0.42	0.16	0.07	0.04
LOG-GAUSSIAN & CLASS	FR	<b>0.71</b>	0.25	0.12	0.06	<b>0.03</b>	0.00
	FC	0.46	<b>0.28</b>	<b>0.16</b>	<b>0.08</b>	0.03	0.00
	EN	0.55	0.21	0.11	0.03	0.00	0.00
	MPD	0.25	0.22	0.14	0.07	0.03	0.01
	SAM	0.66	0.24	0.12	0.05	0.02	0.00

Precision for Gaussian and Log-Gaussian noises combined with Class assignment noise at different levels using the synthetic dataset.

Table 4.4 Synthetic modeling accuracy with combined noise

		1% / 10%	2% / 15%	3% / 20%	4% / 25%	5% / 30%	6% / 35%
GAUSSIAN & CLASS	FR	<b>87.12 / 5</b>	<b>82.21 / 10</b>	<b>78.53 / 6</b>	<b>85.28 / 45</b>	<b>95.71 / 197</b>	<b>97.55 / 218</b>
	FC	84.66 / 11	82.21 / 196	75.46 / 68	70.55 / 6	77.30 / 176	85.89 / 107
	EN	84.66 / 3	81.60 / 171	77.30 / 4	83.44 / 36	93.87 / 212	95.71 / 173
	MPD	85.28 / 130	82.21 / 50	76.07 / 61	66.87 / 17	73.01 / 215	79.14 / 229
	SAM	85.89 / 7	81.60 / 8	77.91 / 5	85.28 / 47	95.71 / 228	95.71 / 218
LOG-GAUSSIAN & CLASS	FR	<b>84.66 / 7</b>	<b>88.96 / 226</b>	<b>99.39 / 172</b>	<b>99.39 / 128</b>	<b>100.00 / 79</b>	<b>100.00 / 90</b>
	FC	85.28 / 8	85.28 / 205	94.48 / 196	97.55 / 204	100.00 / 109	100.00 / 90
	EN	84.66 / 10	87.12 / 221	99.39 / 226	96.93 / 226	100.00 / 214	99.39 / 224
	MPD	84.66 / 46	83.44 / 190	93.87 / 223	96.32 / 204	100.00 / 148	100.00 / 154
	SAM	<b>84.66 / 6</b>	<b>89.57 / 228</b>	<b>99.39 / 167</b>	<b>99.39 / 154</b>	<b>100.00 / 96</b>	<b>100.00 / 67</b>

LOOCV mean accuracy / Number of selected probes for Gaussian and Log-Gaussian noises combined with Class assignment noise at different levels using the synthetic dataset.

## 4.2.5 Results using disease datasets

Table 4.5 shows the mean accuracy and number of selected probes for each ranking method and dataset. For these three datasets we achieved accuracies higher than 90% with a very small subset of probes.

Table 4.5 Mean LOOCV accuracy / Number of selected probes for CLL, IBM, and ALS datasets.

	CLL	IBM	ALS
FR	93.25 / 6	97.06 / 2	94.12 / 12
FC	93.87 / 35	79.41 / 2	87.06 / 254
MPD	93.25 / 7	91.18 / 32	94.12 / 17
EN	94.48 / 99	79.41 / 6	88.24 / 114
<b>SAM</b>	<b>93.87 / 26</b>	<b>97.06 / 1</b>	<b>95.29 / 42</b>

In the case of CLL, the difference between all the methods is very small. The entropy method achieved 94% of accuracy with 99 probes. However SAM got almost 94% of accuracy with 26 probes and FR 93% with only 6 probes. High discriminatory genes of the IgVH phenotype include: LPL, CRY1, LOC100128252, SPG20, ZBTB20, NRIP1, ZAP-70, LDOC1, COBLL1 and NRIP1. The pathway analysis has revealed the importance of the Inflammatory Response, the PAK pathway and the ERK signaling super pathway that includes ERK signaling, ILK signaling, MAPK signaling, Molecular Mechanisms of cancer and Rho Family GTPases pathway. These pathways control Proliferation, Differentiation, Survival and Apoptosis. Also, other important pathways found were Allograft Rejection, the Inflammatory Response Pathway, CD28 Co-stimulation, TNF-alpha/NF-kB Signaling Pathway, Akt Signaling, PAK Pathway and TNF Signaling. The presence of some of these pathways opens the hypothesis of viral infection as a cause of CLL.

Regarding the IBM dataset, we found that SAM and FR were able to correctly predict 97% of the samples just with 2 and 1 probes respectively. Differences between SAM and FR and other methods are remarkable. The list of most discriminatory genes of the IBM phenotype include: HLA-C, HLA-B, TMSB10, S100A6, HLA-G, STAT1, TIMP1, HLA-F, IRF9, BID, MLLT11 and PSME2. Note the presence of different HLA-x genes of major histocompatibility. Particularly, the function of the gene HLA-B would explain alone the genesis of IBM: "HLA-B (major histocompatibility complex, class I, B) is a human gene that provides instructions for making a protein that plays a critical role in the immune system. HLA-B is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system to distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria". The analysis of biological pathways has revealed the importance of viral infections, mainly in IBM patients: Allograft Rejection, Influenza A, Class I MHC Mediated Antigen Processing and Presentation, Staphylococcus Aureus Infection, Interferon Signaling, Immune Response IFN Alpha/beta Signaling Pathway, Phagosome, Tuberculosis, Cell Adhesion Molecules (CAMs), Epstein-Barr Virus Infection,

and TNF Signaling. We can see several viral infections in this list. It is interesting to remark that 75% of the cases of viral myositis are due to *Staphylococcus Aureus* infection (Fayad et al., 2007).

Finally, in the case of ALS dataset, SAM reached an accuracy of 95% with 42 probes, while FR and MPD got a 94% with 12 and 17 probes respectively. High discriminatory genes of the ALS phenotype include: CASP1, ZNF787 and SETD7. The pathway analysis has revealed the importance of the GPCR Pathway, RhoA Signaling Pathway, EPHB Forward Signaling, EphrinA-EphR Signaling, EBV LMP1 Signaling, and Regulation of Microtubule Cytoskeleton. These pathways have different important signaling roles and suggest a possible link to the Epstein-Barr virus (EBV). The activation of Caspases plays a central role in cell apoptosis and activates interleukin-1, a cytokine involved in the processes such as inflammation. Caspases have been also associated to the pathogenesis of Huntington disease. Obviously, the complete exploitation of these results needs from the analysis of geneticists.

#### 4.2.6 Conclusions for noise analysis

We have experimentally showed that noise in expression data and class assignment partially falsifies the sets of discriminatory probes in phenotype prediction problems. Via synthetic modeling we have shown that FR and SAM are the most robust gene selection methods for different kind of noises. Besides, FR and SAM seem to exploit the parsimony principle, being able to find the smallest-scale high discriminatory gene signature. Nevertheless, SAM is much more computationally expensive than FR while the achieved results are similar. We have also found that noise in class assignment affect the predictive accuracy and the precision much more than noise in the expression data. Nevertheless, the No-Free-Lunch Theorem in search and optimization (Wolpert and Macready, 1997) states that all these algorithms are needed to understand the complex relationships hidden in the genetic datasets. Therefore, the prior knowledge provided by the doctors is of paramount importance in the search for solutions of the different diseases. From the translational point of view this analysis shows the importance of establishing the discriminatory power of the genes in phenotype prediction problems to correctly find the biological pathways that are involved. To accomplish this task in the most efficient way possible we suggest ranking the most differentially expressed genes according to their FR (or SAM ratio). Examples to cancer (CLL), rare (IBM) and neurodegenerative diseases (ALS) are also outlined in this paper obtaining very interesting conclusions that might imply an important role of several viral infections.

### 4.3 Impact of microarray preprocessing techniques in unraveling biological pathways

In this section we analyzed the precision in biological pathways analysis obtained with a raw dataset and the preprocessed datasets via Robust Multi-array Average (RMA) and Affymetrix Microarray Suit 5.0 algorithm (MAS5). For that purpose we use a the combination FC-FR ranking methods as explained in section 3.2, establishing the predictive accuracy via LOOCV (see section 1.3.6). One of the main complexities of this analysis is having at disposal synthetic data to perform it as we did with the sensitivity noise analysis. For that reason we decided to work with international standards, such as, the Affymetrix Latin Square Data for Expression Algorithm Assessment (Human Genome U133 Data Set Affymetrix (2015)), where 42 different control genes are spiked-in at known concentrations. This is commonly known as the Spike-In experiment. As a result, this study has two main parts: A) Analysis of the precision and accuracy of the ranking methods using a synthetic data set for both raw and preprocessed datasets. B) Analysis of the accuracy and biological pathways of the selected genes using the cancer related fatigue raw and preprocessed datasets.

In part A we used the Affymetrix Latin Square Data for Expression Algorithm Assessment. As we know the genes that are differentially expressed we first rank the genes through the combination FC-FR and then we analyzed the precision (see equation (4.8)) of the obtained ranking using raw and preprocessed data. Subsequently, we perform a gene selection to study the discrimination power of the selected genes in both cases (raw and preprocessed).

In part B we managed the cancer treatment-related fatigue dataset described in section 3.3. We carried out a similar analysis using both raw and preprocessed microarray data consisting of 44 men with non-metastatic prostate cancer, where 25 of them coursed high cancer related fatigue and 19 experimented low cancer related fatigue. We perform a gene selection based on the same ranking method as for the synthetic dataset and analyzed the biological pathways derived from the selection using the Gene-Analytics software (Stelzer et al., 2009). We have also built a correlation network using the Pearson correlation coefficient with the selected genes (see section 3.2 for further details about how are created the correlation networks). Then we compare the biological pathways and correlation networks derived from the selection with raw and preprocessed data. Also an independent validation dataset containing 17 samples is at disposal that is used to confirm the predictive power of these lists in each case. Obviously, the genes that are responsible for the disease development are unknown. The purpose of this analysis is to analyze the impact of the preprocessing techniques in understanding the biological pathways, keeping in mind the results that were found for the spike-in experiment

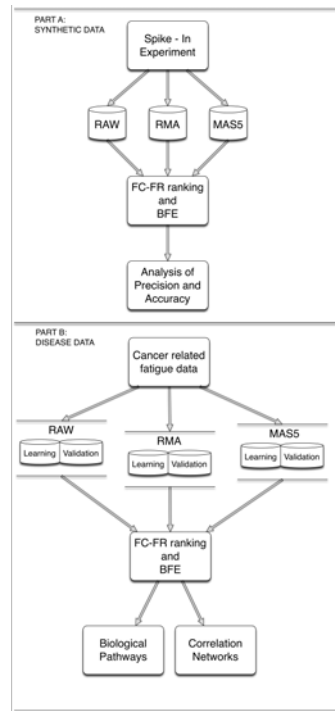


Fig. 4.5 Flow chart of the methodology

where the set of differentially expressed genes is known, and also the a priori knowledge in the chronic fatigue disease. A flow chart of this methodology is shown in figure 4.5.

### 4.3.1 Microarrays preprocessing techniques

In this section we provide a brief introduction about the microarray preprocessing techniques used to unravel the biological pathways in phenotype prediction problems.

Microarrays are manufactured using photo-lithographic techniques to attach hundreds of thousands of different oligonucleotide sequences on the surface of a glass slide. These oligonucleotides correspond to known DNA or RNA sequences that are arranged in different probe sets. Quantification of the levels of transcripts in a sample is performed via hybridization to the specific probes and measurement of the expression through fluorescence-based methods. Generally, raw data contains about 20 pairs of oligonucleotides for each DNA or RNA target (gene) known as probe set. The first component of these pairs is referred to as the Perfect Match (PM) probe. Each PM probe is paired with a Mismatch (MM) probe that is artificially created by changing the middle base with the intention of measuring non-specific binding. Typically, to define a measure of gene expression, probe intensities are summarized for each probe set into a single value.

Different studies have been performed to analyze the accuracy of these measurements and to correct the effect of noise in microarrays (Benito et al., 2004; Chen et al., 2011; Scherer, 2009). Two techniques of particular importance are MAS5 (Affymetrix, 2001) and RMA (Irizarry et al., 2003):

- **MAS5:** The Affymetrix Microarray Suite 5.0 (MAS5) algorithm uses both PM and MM probes to summarize gene expression. The MAS5 signal of a probe set  $i$  is defined as the anti-log of the Tuckey's biweight robust mean (Huber and Ronchetti, 2009) of the following values:

$$u_{ij} = \log(PM_{ij} - CT_{ij}), \quad j = 1, \dots, N \quad (4.9)$$

where

$$CT_{ij} = \begin{cases} MM_{ij} & \text{if } MM_{ij} < PM_{ij} \\ PM_{ij} - \varepsilon^2 & \text{if } MM_{ij} > PM_{ij} \end{cases}, \quad (4.10)$$

being  $N$  the number of probes in the probe set (or gene)  $i$  and  $\varepsilon^2$  a given positive amount that has to be individually adjusted for each probe set. Therefore, the robust Tuckey's mean of a probe set  $i$  is defined as:

$$\bar{u}_i = \frac{\sum_{j=1}^N \psi(u_{ij}; c) u_{ij}}{\sum_{j=1}^N \psi(u_{ij}; c)}, \quad (4.11)$$

where

$$\psi(x; c) = \begin{cases} x \left(1 - \frac{x^2}{c^2}\right)^2 & \text{for } |x| < c, \\ 0 & \text{for } |x| > c. \end{cases} \quad (4.12)$$

- **RMA:** Robust Multiarray Average (RMA), basically consists in three steps:

1. **Background correction** using the following additive probabilistic model:

$$PM_{ij} = s_{ij} + bg_{ij}, \quad (4.13)$$

where  $PM_{ij}$  is the Perfect Match of the probe  $j$  in gene  $i$ ,  $s_{ij}$  is the gene signal and it is supposed to follow an exponential distribution  $s_{ij} \sim Exp(\lambda_i)$ , and  $bg_{ij}$  is the background correction caused by the optical noise and non-specific binding and it

is supposed to follow a normal distribution  $bg_{ij} \sim N(\mu_i, \sigma_i^2)$ . This identification problem has three unknown parameters  $(\lambda_i, \mu_i, \sigma_i)$  and  $N$  different realizations for  $PM_{ij}$ , and can be typically solved by least squares and the maximum likelihood estimation.

2. **Normalization** across all arrays to make all distributions the same. This is typically performed by quantile normalization, that consists in normalizing the background corrected array to a common set of quantiles. This process is aimed at correcting for array biases and avoiding the effect of outliers. This process provided a set of normalized probe values  $sn_{ij}$ .
3. **Probe set summarizing**, where the final expression is calculated separately for each gene  $i$  using the following linear model in  $\log_2$  scale:

$$Y_{ij} = \mu_i + \alpha_{ij} + \varepsilon_{ij}, \quad (4.14)$$

where  $Y_{ij}$  are the background corrected, normalized, log transformed probe intensities ( $Y_{ij} = \log_2(sn_{ij})$ ),  $\mu_i$  is the log-expression level for gene  $i$ ,  $\alpha_{ij}$  is the probe affinity effect of probe  $j$  in the gene  $i$ , and  $\varepsilon_{ij}$  is the independent identically distributed error term with zero mean. The probe affinities  $\alpha_{ij}$  verifies  $\sum_{j=1}^N \alpha_j = 0$ . This linear model is solved using the median polish algorithm and provides the final summarized gene intensity value  $\mu_i$ , that is commonly used in phenotype prediction problems.

### 4.3.2 Results with the Spike-in experiment

In order to check the precision using both raw and preprocessed data we need a dataset where we know the genes that are differentially expressed. In such case we used the Affymetrix Latin Square Data for Expression Algorithm Assessment (Human Genome U133 Data Set) that consists of 3 technical replicates of 14 separate hybridization of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125pM to 512pM. The concentration in the first experiment composed by three replicas is 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512pM (the concentrations table can be consulted in appendix B). Each subsequent experiment and its three replicas rotates the spike-in concentrations by one group; i.e. experiment 2 and its three replicas begins with 0.125pM and ends at 0pM, on up to experiment 14 and its three replicas, which begins with 512pM and ends with 256pM. Further details can be consulted in Affymetrix (2015).



Table 4.6 Precision on the selection of the differential expressed genes using raw data or preprocessed data with RMA and MAS5. The data is the Affymetrix Latin Square Data for Expression Algorithm Assessment. The selection is performed between the first group and the rest to include all the differences between the spike-in concentrations.

Group comparison	RAW	RMA	MAS5
1 vs 2	7.14	<b>9.52</b>	4.76
1 vs 3	<b>26.19</b>	16.67	16.67
1 vs 4	<b>38.10</b>	11.90	14.29
1 vs 5	<b>28.57</b>	<b>28.57</b>	16.67
1 vs 6	26.19	<b>28.57</b>	<b>28.57</b>
1 vs 7	<b>40.48</b>	26.19	23.81
1 vs 8	<b>35.71</b>	21.43	30.95
1 vs 9	<b>40.48</b>	23.81	23.81
1 vs 10	<b>35.71</b>	19.05	21.43
1 vs 11	<b>38.10</b>	14.29	21.43
1 vs 12	<b>23.81</b>	16.67	9.52
1 vs 13	<b>23.81</b>	<b>23.81</b>	14.29
1 vs 14	7.14	4.76	<b>9.52</b>
Mean Precision	<b>28.57</b>	18.86	18.13

There are 42 differentially expressed probes and we selected the first 42 probes of the ranking. We compared the first group with the rest of groups to cover all the possible concentration comparisons. In the first comparison (group 1 Vs. group 2) the difference of concentration between all the differentially expressed probes was 0.125pM, in the second comparison (group 1 Vs. group 3) the difference was 0.5, on up to the 12 comparison (group 1 Vs. group 13), which was 256pM. Due to the rotation of the concentrations, the last comparison (group 1 Vs. 14) had again a difference of 0.125pM in concentration among all the differentially expressed probes.

Table 4.6 shows the Precision for each comparison using raw, RMA and MAS5 datasets, showing the mean Precision along the different comparisons. In almost all the comparisons we got better results in terms of precision working with RAW data than with preprocessed data. Also, the higher mean precision was obtained with RAW data.

We have also calculated the empirical Cumulative Distribution Functions (CDF) of the positions of the differentially expressed genes. A perfect CDF would be a straight line reaching the value of 1 at position 42, which is the number of differentially expressed genes. These curves serve to visualize how many genes we have to select in order to get all the differentially expressed genes. Figure 4.6 shows these CDF curves for each comparison and type of data. As the raw data obviously have more probes (248.152 for raw data and 22.300 for preprocessed data, see Affymetrix (2015)), the positions given by the ranking

method are divided by a correction factor:  $C = nR/nP$  where  $nR$  is the number of raw probes equal to 248.152 and  $nP$  is the number of preprocessed probes equal to 22.300. Therefore,  $C = 11.13$ .

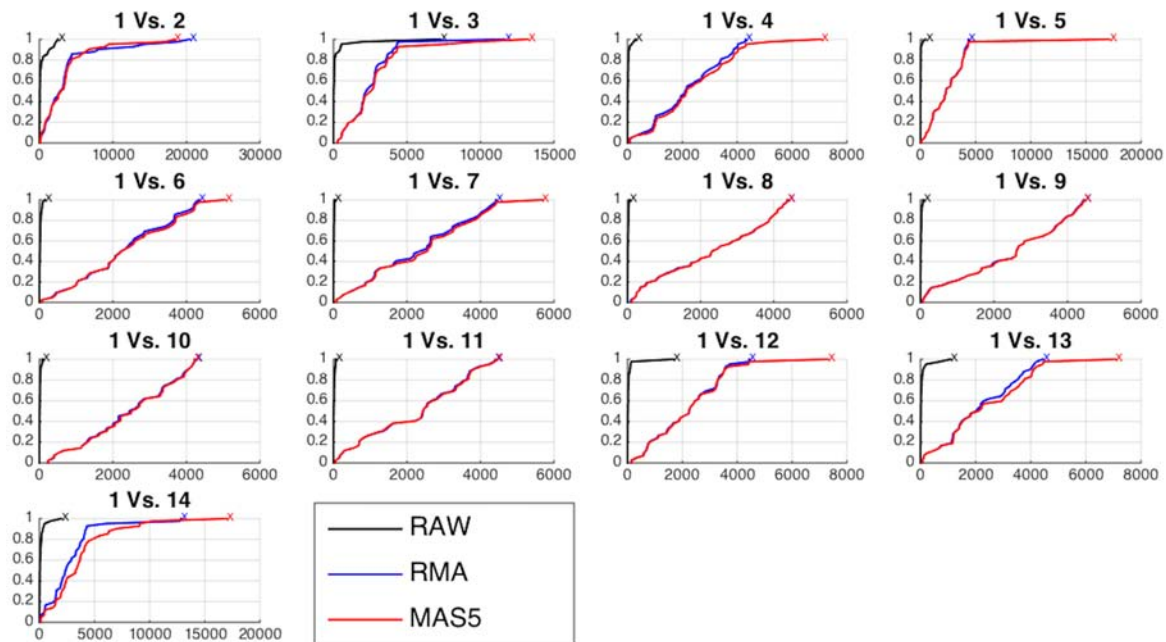


Fig. 4.6 Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes ranked by the FC/FR methods for each comparison and different types of data.

In this figure the X-axis represents the positions of the probes given by the ranking method and the Y-axis represents the percentage of differentially expressed genes that were located. Therefore in the first comparison we were able to find all the of differentially expressed genes (42) selecting less than 5000 ( $0.5E4$  in the X-axis of the graphic) while working with preprocessed data we need almost all the probes ( $2.23E4$ ). In all the comparisons, we were able to find all the differentially expressed genes selecting rather less number of probes with raw data than with preprocessed data.

### 4.3.3 Results for the cancer related fatigue dataset

Table 4.7 shows the LOOCV accuracy of first 50 most discriminatory probes in each case. The highest predictive accuracy we obtained a 92.59% of accuracy with only the first 3 probes. However, using RMA and MAS5 we achieved a 100% with 6 and 44 probes respectively.

We managed a raw microarray dataset which have 604,258 probes and the preprocessed dataset which have 54,675 different probes in both cases, RMA and MAS5. Obviously

the dimensionality of the raw data set is 11.05 times higher than the preprocessed datasets, that is, using the raw data, the probe sets have not been summarized in one gene like in the preprocessed data. For that reason the repetition of a probe in the raw data indicates the importance of the gene that corresponds to this gene. This is the case of TUBB2A, HLA-DQA1, TUBB3, HLA-DQB1, and BTNL3. It can be observed that RMA also find these genes within the most discriminatory set, but not using MAS5.

Additionally, a blind validation of these results has been performed using the set of 17 subjects, independent of the training set, originally used to assess the validity of the learned predictive model (see section 3.3). The results of this blind validation using raw data was 76.47% accurate, while using MAS5 and RMA, the accuracy dropped to 58.82 and 64.7% respectively. This result is very important and shows that RMA and MAS5 increase the accuracy in the learning process at the price of decreasing the accuracy in blind validation. Therefore, this implies that the biological pathways associated with the predictive genes found using raw data are more meaningful, and both preprocessing techniques (RMA and MAS5) highly impact the biological pathway analysis and the corresponding phenotype prediction problem.

The raw data generated predictive genes associated with pathways mainly related to pathogenic infections (HLA-DQX genes) and also oligomerization of connexins into connexons (TUBB2A and TUBB3) involved in intercellular signals and metabolic communication between communicating cells in a tissue (Koval, 2006). These are crucial mechanism in the development of many human diseases (Kellsell et al., 2001).

The main pathways associated with predictive genes generated by RMA are mitotic prometaphase (BIRC5, CLIP1, STAG2, TUBB3) that control the nuclear membrane breaking apart into numerous membrane vesicles, cytoskeleton remodeling neurofilaments (EEPK1, KRT6A, TUBB2A and TUBB3) and also mitotic meta-phase and ana-phase (BIRC5, CLIP1, TUBB2A and TUBB3). The beta-tubulin gene family control the tubulin protein super-family of globular proteins. Beta-tubulins polymerize into micro-tubules which is a major component of the cytoskeleton formation. Micro-tubules function in many essential cellular processes, including mitosis. For instance, tubulin-binding drugs serve to kill cancerous cells by inhibiting micro-tubule dynamics that are required for DNA segregation and cell division. The main pathways associated with predictive genes generated by MAS5 are GADD45 Pathway, EGFR1 Signaling Pathway, Interferon Type I related to the MAP3KX genes.

We also provide the correlation networks, using the Pearson correlation coefficient metric, for the 50 most discriminatory genes for each dataset. Figure 4.7, 4.8 and 4.9 shows the correlation graphs for raw, RMA and MAS5 respectively. In the case of raw data we can observe one main tree connecting the tubulin genes to the major histocompatibility complex

gene and other genes that serve to expand the tree. RMA privileges the connection between the beta-tubuline genes and two probes (241238\_at and 1566585\_at) whose gene name is unknown. MAS5 privileges the role of SOCS3. This gene encodes a member of the STAT-induced STAT inhibitor (SSI), also known as suppressor of cytokine signaling (SOCS), family. SSI family members are cytokine-inducible negative regulators of cytokine signaling. The expression of SOCS3 gene is induced by various cytokines, including IL6, IL10, and interferon (IFN)-gamma (Masuhara et al., 1997; Minamoto et al., 1997).

Table 4.7 Probe/Gene name and Accuracy (Acc %) of the selected probes for raw data and preprocessed data with RMA and MAS5

RAW		RMA		MAS5	
Probe/Gene	Acc(%)	Probe/Gene	Acc(%)	Probe/Gene	Acc(%)
<b>TUBB2A</b>	85.19	<b>TUBB2A</b>	88.89	<b>SOCS3</b>	85.19
<b>HLA-DQA1</b>	96.3	<b>C11orf1</b>	88.89	<b>TMEM194A</b>	92.59
<b>TUBB2A</b>	<b>92.59</b>	<b>PPOX</b>	96.3	<b>1561478_at</b>	92.59
TUBB2A	92.59	<b>TTC23</b>	92.59	<b>CIB3</b>	96.3
TUBB2A	88.89	<b>NRIP3</b>	96.3	<b>ESYT2</b>	92.59
TUBB2A	85.19	<b>SCAMP4</b>	<b>100</b>	<b>ABHD1</b>	92.59
TUBB2A	85.19	HLA-DQA1	100	<b>JTB</b>	92.59
HLA-DQA1	88.89	234253_at	100	<b>1556412_at</b>	92.59
TUBB2A	88.89	223313_s_at	96.3	<b>207371_at</b>	96.3
TUBB2A	88.89	BTNL3	100	<b>LOC100131756</b>	92.59
BTNL3	88.89	YSK4	96.3	<b>CDK6</b>	92.59
TUBB2A	88.89	236963_at	100	<b>ALS2CR8</b>	96.3
HLA-DQA1	92.59	ZCCHC2	100	<b>SEL1L2</b>	96.3
TUBB2A	88.89	DSG3	100	<b>FLJ35220</b>	96.3
TUBB3	88.89	TMEFF2	100	<b>215626_at</b>	96.3
HLA-DQB1	85.19	1566585_at	100	<b>SPAM1</b>	96.3
HLA-DQB1_LOC101060835	85.19	231141_at	100	<b>FTCD</b>	96.3
HLA-DQA1	88.89	SPATA20	100	<b>1570285_at</b>	96.3
IMMP1L	85.19	CSN1S2A	100	<b>216795_at</b>	96.3
BTNL3	85.19	RAB11FIP3	100	<b>MAP3K2</b>	96.3
240231_at	85.19	239587_at	100	<b>MTSS1L</b>	96.3
ZFPL1	85.19	RIMS3	100	<b>GMEB1</b>	96.3
GNRHR2	85.19	234548_at	100	<b>SOCS7</b>	96.3
DR1	88.89	C20orf103	100	<b>GNA12</b>	96.3
DOCK11	88.89	AGR2	100	<b>244274_at</b>	96.3
HLA-DQB1	88.89	SAT1	100	<b>PLP2</b>	96.3
FMR1	88.89	RGS18	100	<b>ATG9B</b>	96.3
ACAP2	85.19	1570044_at	100	<b>1564056_at</b>	96.3
HLA-DQB1	85.19	TUBB3	100	<b>PCCB</b>	96.3
ZEB1_LOC10096668	85.19	HDLBP	100	<b>239370_at</b>	96.3
FLJ32790	85.19	1560087_a_at	100	<b>ANK1</b>	96.3
LOC100505812	88.89	AVL9	100	<b>SCAND2</b>	96.3
DENND4C	88.89	241238_at	100	<b>1564872_at</b>	96.3
PREPL	88.89	PHLDB3	100	<b>SMAD2</b>	96.3
LOC100505812	85.19	PIGK	100	<b>CMTM3</b>	96.3
FAM63B	88.89	F11	100	<b>INSR</b>	96.3
LYSMD3	85.19	C1orf21	100	<b>PSG1</b>	96.3
RP11-727A23.11_OTTHUMG00000183952	85.19	IL9	100	<b>1560169_at</b>	96.3
HIPK3	85.19	229733_s_at	100	<b>MAP3K1</b>	96.3
POLR2J4	85.19	241776_at	100	<b>KCNRG</b>	96.3
PHF17	85.19	WDR27	100	<b>DOCK7</b>	96.3
SP3	85.19	D21S2091E	100	<b>1560995_s_at</b>	96.3
MRGBP	85.19	239632_at	100	<b>WNT5A</b>	96.3
NAPIL1	85.19	HGSNAT	100	<b>1562673_at</b>	<b>100</b>
FAM126A	85.19	242839_at	100	<b>GSK3B</b>	100
EPS15P1	85.19	KCTD4	100	<b>NCKIPSD</b>	100
SMCR8	85.19	MECOM	100	215439_x_at	100
HLA-DQA1_LOC100509457	85.19	LOC257152	100	<b>CDHR3</b>	96.3
ZMYM2	85.19	MLH3	100	<b>PCGEM1</b>	96.3
EIF1AX_LOC101060318	85.19	DDX60	100	<b>GNG13</b>	96.3

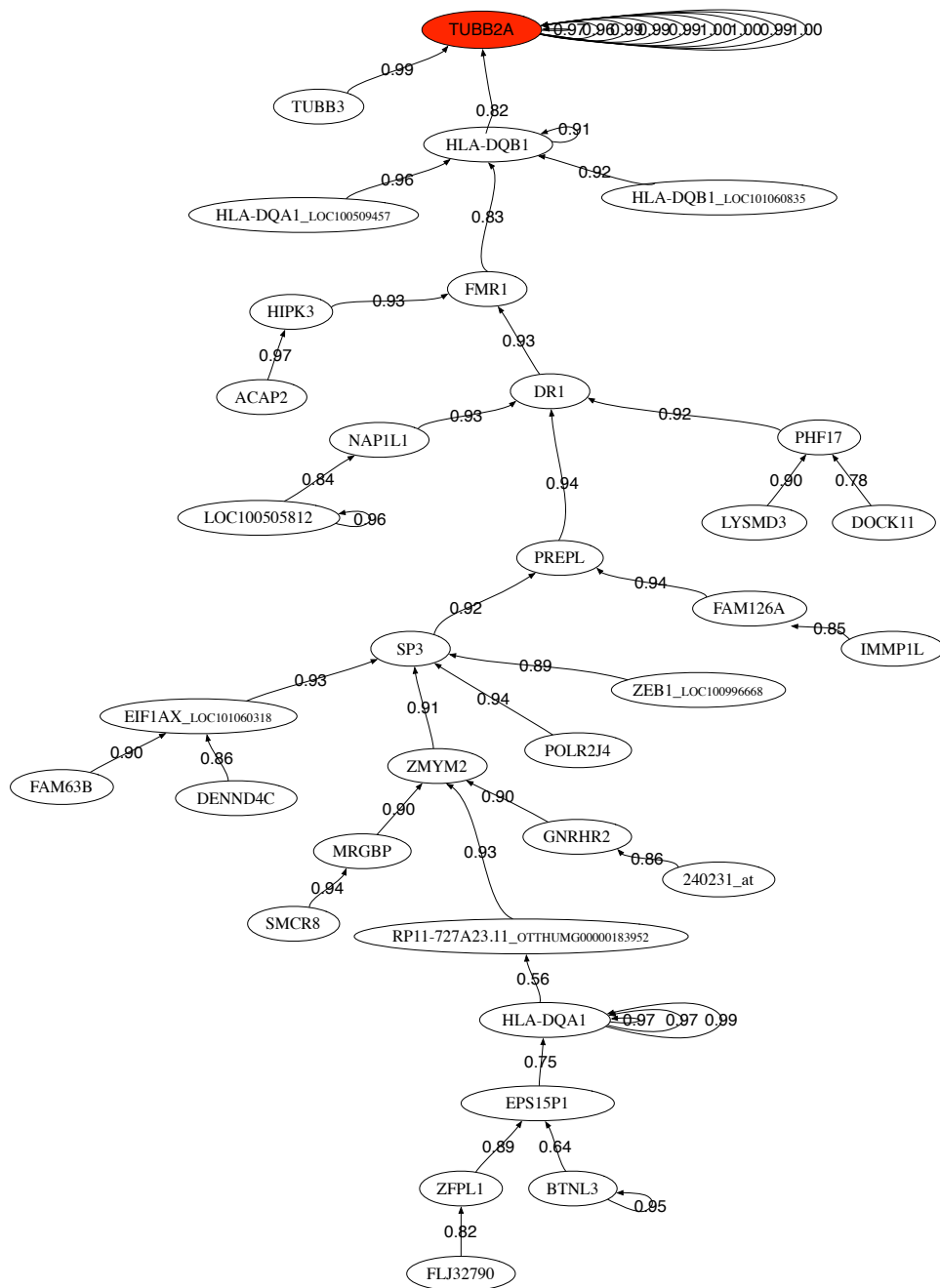


Fig. 4.7 Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using raw data.

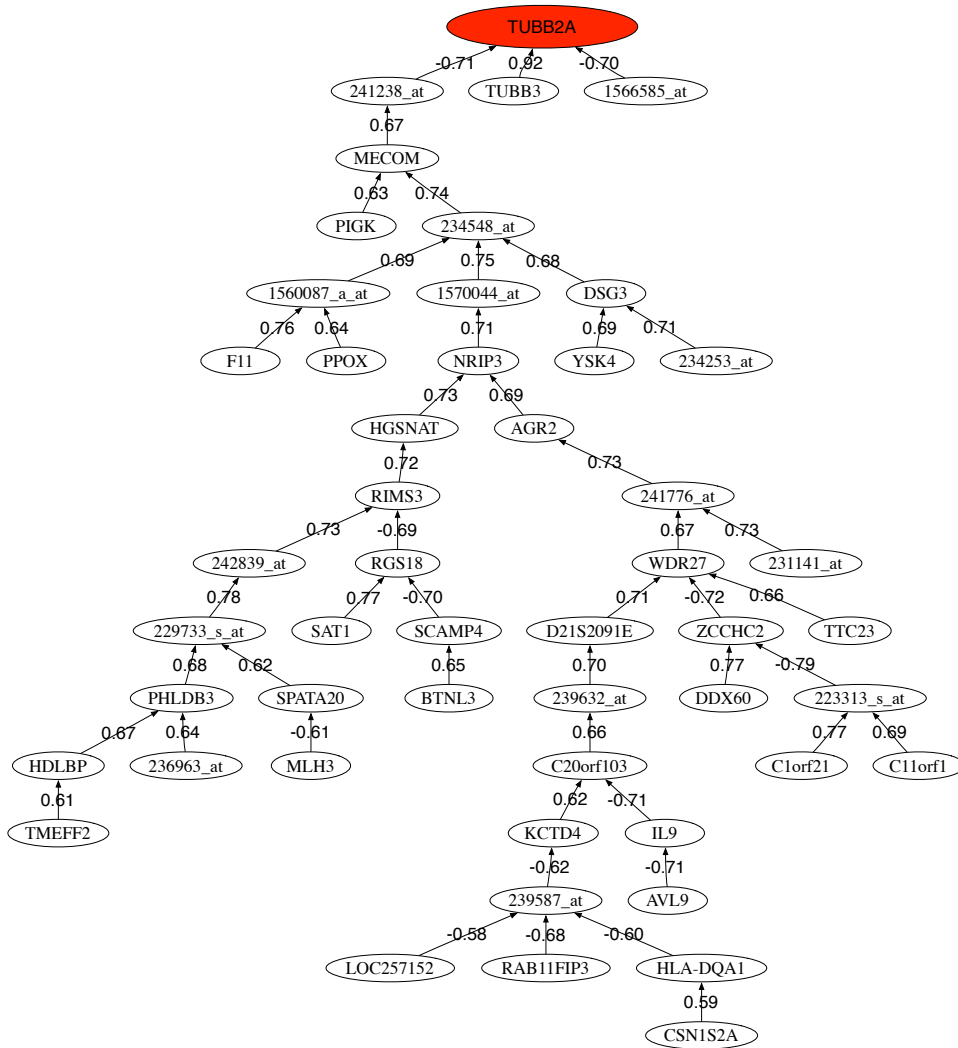


Fig. 4.8 Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using preprocessed data with RMA.

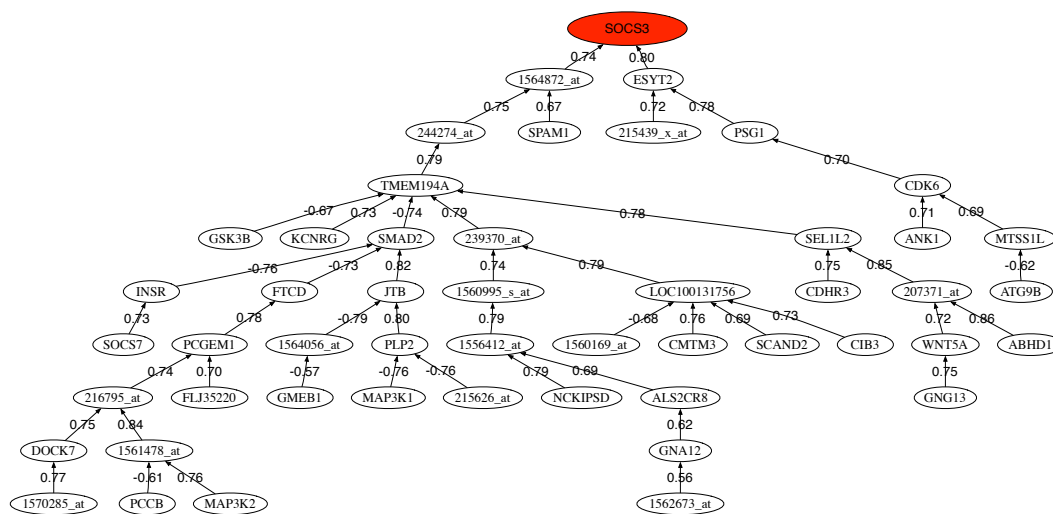


Fig. 4.9 Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using preprocessed data with RMA.

## 4.4 Conclusions of the microarrays preprocessing techniques impact analysis

We analyzed the impact of the main preprocessing microarrays techniques (MAS5 and RMA) in identifying the biological pathways that are associated with discriminatory genes that can accurately predict the cancer treatment-related fatigue phenotype. We found that in the case of the Affymetrix synthetic dataset, the mean precision along all the comparisons was higher using raw data than using preprocessed data. This difference is even more remarkable in the CDF curves for all the comparisons. We were able to find all the differentially expressed genes selecting rather less number of genes with raw data than with preprocessed data.

Regarding the cancer related fatigue dataset, working with RMA and MAS5 datasets we got better accuracy results than using raw data. However, in the blind validation, working with raw data allowed us to generalize better than using preprocessed data (RMA and MAS5). Besides the pathways analysis and the correlation networks were significantly different between raw, RMA and MAS5. This would explain why some genetic signatures found in real practice fail to predict unseen samples. Consequently it can be concluded that interpreting results from predictive gene profiles generated by RMA and MAS5 should be done with caution. This is an important conclusion with a high translational impact that should be confirmed in other disease datasets. A retrospective analysis of different cancer datasets will be performed in future research.



# Chapter 5

## Design and application of biomedical robots to phenotype prediction problems

### 5.1 Introduction

So far we have applied our methodology to different kind of biomedical data, and checked the robustness of the proposed methods against different kind of noises and the impact of the main preprocessing techniques in microarray expression data. Now we apply the methodology, generating several biomedical robots, in order to address a specific biomedical problem: phenotype prediction in genomics, with the intention of giving a translational approach.

Despite all of its promises, clinical translation of genomics findings has been tempered by analytical limitations, the requirement for extensive numbers of subjects, and cost. To help address these issues and following the scheme 1.1 we have generated different sets of biomedical robots and applied them to different phenotype prediction problems. In this case, the difference between the biomedical robots will lie in the ranking algorithms (described in section 1.3.6) we take for selecting the genes. The final prediction will be performed via consensus strategy. Obviously there are different ways of designing the biomedical robots, but all of them must be based in sampling the uncertainty space of any phenotype prediction problem. A future research will be devoted to explore other designs.

Aside from specifically addressing the interpretation of genomic data, strength of the method is its ability to synchronously include non-genomic inputs (epigenetics, demographic variables, etc.) as a component of a comprehensive analysis. For a clinical perspective, phenotype prediction problem applies to linking a set of genes to a specific disease or condition. As we do in section 4.2, firstly we performed the sensitivity analysis to noise of

the different sets of biomedical robots using synthetic microarrays perturbed by different kind of noises in expression and class assignment. Subsequently, we provide specific applications of the methodology to the microarray datasets we managed in section 4.2: predicting IgHV mutation in patients with Chronic Lymphocytic Leukemia, predicting Inclusion Body Myositis-Polymyositis and predicting Amyotrophic Lateral Sclerosis; inferring the pathways and the correlation networks in each case.

The result of this research work was a manuscript titled "Design and application of biomedical robots to phenotype prediction problems" currently accepted for publication in the "Journal of Computational Biology" (see Appendix A.7).

## 5.2 Noise Sensitivity Analysis of Biomedical Robots

We have generated different synthetic data sets using three types of noise: additive Gaussian noise, lognormal noise, and noise in class assignment (see section 4.2.2). Then we set up several biomedical robots using the ranking algorithms we described in section 1.3.6.

Table 5.1 shows the results obtained for the sensitivity analysis.  $\delta$  represents the level of noise imputed for each type of noise, *Acc* the mean LOOCV accuracy, *P* the Precision (see equation (4.8)) established using the set of genes constructed with the union of all the genes found by the robots, and *#R* the number of robots used in the consensus strategy.

Table 5.1 Noise results

$\delta(\%)$	Class Assignment			Gaussian			Log Gaussian		
	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>
1	98.77	1.00	98	100.00	1.00	98	100.00	1.00	98
3	96.93	1.00	98	100.00	1.00	98	100.00	0.74	98
5	94.48	1.00	98	100.00	1.00	98	100.00	0.35	98
10	90.18	1.00	98	100.00	0.60	98	100.00	0.14	10
15	87.12	1.00	3	100.00	0.33	98	99.39	0.05	37
20	80.98	1.00	1	100.00	0.22	11	100.00	0.03	43
25	77.30	1.00	10	99.39	0.13	81	98.77	0.04	98
30	73.62	0.92	43	99.39	0.14	7	100.00	0.03	14

$\delta$  the percentage of noise introduced, *Acc* the mean LOOCV predictive accuracy, *P* the precision of the selection using the union of all the genes found by the robots and *#R* the number of robots applied in the consensus strategy

The results can be summarized as follows:

- The Precision  $P$  remains stable when noise in class assignment is increased. This result is very interesting since the biomedical robots are able to find the differentially expressed genes when the noise in class assignment is introduced. In the case of Gaussian noise the precision is very high for noise levels less than 5%. The worst result was obtained when multiplicative noise is added to the expressions. The fact that the precision gradually decreases when noise in the expression increases, implies that some of the biological pathways that are inferred might be partially falsified. Therefore, any filtering step that it is usually performed in the microarray data will have important consequences with respect to the pathway analysis, as we shown in section 4.3.
- The mean predictive accuracy ( $Acc$ ) systematically decreases when a higher level of the noise is added to the class assignment vector, and is very stable when Gaussian and non-Gaussian noises are added to the expression data, meaning that the biomedical robots are robust in terms of accuracy with respect to the presence of noise in the expressions. This result also suggests that noise acts as regularization with respect to the accuracy in the prediction as it has been theoretically proved by Fernández-Martínez et al. (2014a,b) in inverse problems. It can be also concluded that if the biomedical robots are unable to improve the accuracy of the best prediction, the dataset could have some wrong class assignment that prevents achieving a perfect classification. Other possibility is that parameterization of the samples is incorrect, that in the present case would mean that none of the genes that have been measured bring enough information to achieve a good phenotype discrimination.

### 5.3 Predicting IgHV mutation with biomedical robots

In this first example we had at disposal a microarray data set consisting of 163 samples and 48807 probes (see section 3.4.2). The best robot predicted the IgVH mutational status with 93.25% accuracy using small-scale signature composed by 13 genes: LPL (2 probes), CRY1, LOC100128252 (2 probes), SPG20 (2 probes), ZBTB20, NRIP1, ZAP-70, LDOC1, COBLL1 and NRIP1. Although in section 4.2.5 the best result was 94.48% using Entropy and 99 probes (see table 4.5), we considered a better result the one achieved by the Fisher's ratio (93.25%), since it took only 6 probes to perform the prediction.

Table 5.2 shows the results of applying the methodology of biomedical robots to this problem. In this case the highest prediction accuracy obtained by the set of biomedical robots equal the accuracy provided by the best robot within the set (93.25%). This happened with 11 samples that are identified in the PCA space in two dimensions (figure 5.1) using the genetic signature of these 13 genes. It can be observed how the classification in this

reduced set of genes approximates a linear separable behavior while using all the genetic information that we have at disposal the classification is nonlinear. Therefore, as an important conclusion we can affirm that reducing the dimension to the set of discriminatory genes helps to linearize the phenotype classification problem. Figure 5.2 also shows the correlation network, using the Pearson correlation coefficient, of the most discriminatory genes of the CLL-IgVH mutational status found in this analysis. This analysis will serve us to understand how the most discriminatory genes regulate the expression of other genes involved in different biological pathways. The head of graph is the gene that has the highest discriminatory power LPL. It can be observed one main network associated to ZBTB20. Finally, the pathway analysis can be consulted in section 4.2.5.

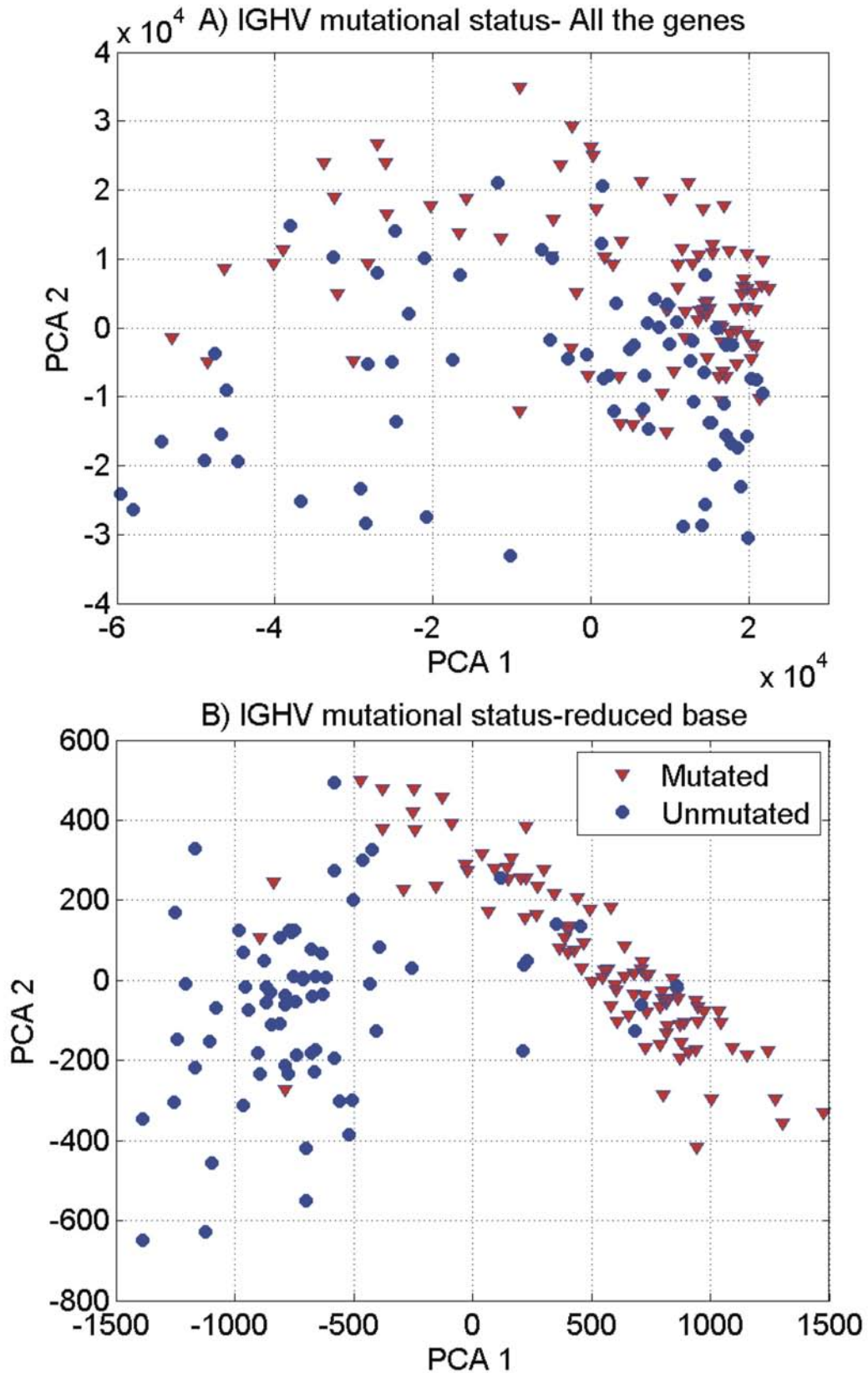


Fig. 5.1 IgVH classification in CLL: A) Considering all the genes of the microarray, the classification problem is nonlinear. B) Using the most discriminatory genes (13 probes) the classification problem approximates a linear separable behavior.

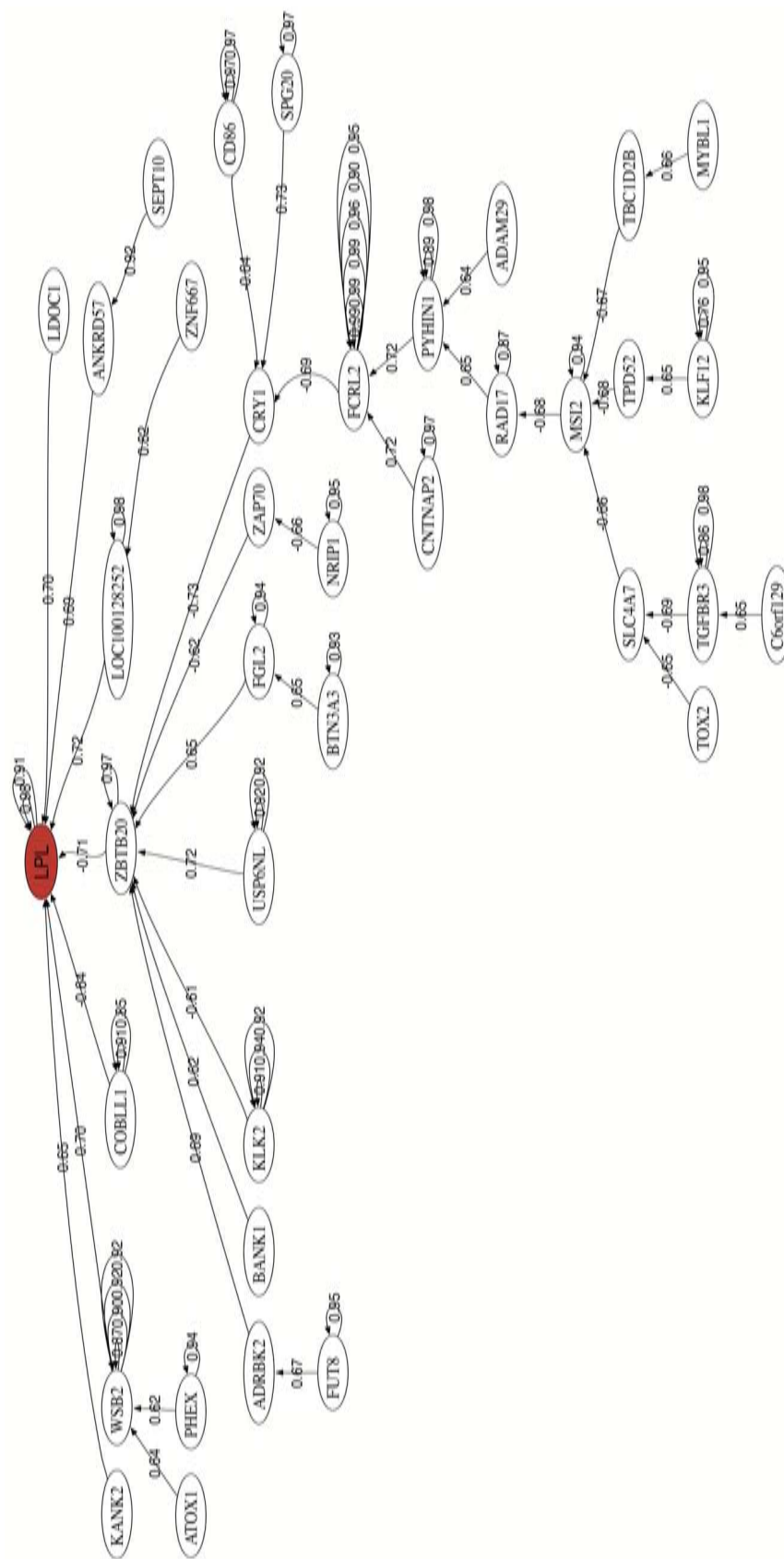


Fig. 5.2 Correlation network for IgVH mutational status in Chronic Lymphocytic Leukemia.

## 5.4 Predicting Inclusion Body Myositis and Polymyositis with Biomedical Robots

Myositis means muscle inflammation, and can be caused by infection, injury, certain medicines, exercise, and chronic disease. Some of the chronic, or persistent, forms are idiopathic inflammatory myopathies whose cause is unknown. We have modeled the Inclusion Body Myositis /Polymyositis (IBM/PM) dataset published by Greenberg et al. (2005). The data consisted in the microarray analysis of 23 patients with IBM, 6 with PM and 11 samples corresponding to healthy controls. The best robot performed the classification of the IBM+PM vs control obtaining a predictive accuracy of 97.5% using a reduced base with only 17 probes. The genes belonging to the highest predictive small-scale genetic signature are HLA-C (3 probes), HLA-B (4 probes), TMSB10, S100A6, HLA-G, STAT1, TIMP1, HLA-F, IRF9, BID, MLLT11 and PSME2. It can be observed the presence of different HLA-x genes of the major histocompatibility. Particularly the function of the gene HLA-B would explain alone the genesis of IBM: "HLA-B (major histocompatibility complex, class I, B) is a human gene that provides instructions for making a protein that plays a critical role in the immune system. HLA-B is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria".

Table 5.2 shows the results using the biomedical robots methodology. In this case we are able to hit the 100% of the samples with two robots, improving the results of the best robot. The analysis of biological pathways has revealed the importance of viral infections, mainly in IBM patients: Allograft Rejection, Influenza A, Class I MHC Mediated Antigen Processing and Presentation, Staphylococcus Aureus Infection, Interferon Signaling, Immune Response IFN Alpha/beta Signaling Pathway, Phagosome, Tuberculosis, Cell Adhesion Molecules (CAMs), Epstein-Barr Virus Infection, and TNF Signaling. Several viral infections appeared in this list.

Figure 5.3 shows the correlation network of the most discriminatory genes found in this analysis. It can be observed the presence of one main dense network involving different HLA-X genes. Among its related pathways are ERK Signaling and Apoptosis Pathway. GO annotations related to this gene include calcium ion binding and cysteine-type peptidase activity.





Figure 5.4 A) shows the PCA projection for the IBM+PM versus control samples using the optimum reduced base. It can be observed that the separability is almost perfect and only one PM sample that is close to the control samples might be misclassified. This graphic also explains that this basis set is not optimum to perform the classification of IBM vs PM. This separability can be achieved with 100% accuracy using a reduced base composed by the following genes: RHOBTB2, MT1P2, FBXL8, HIF3A, C17orf101, RPL12, RBM19, MT1G, WT1-AS, HEXIM1, NQO2, ENOSF1, ADRM1, EIF5A, CSF2RA, CPLX3 /// LMAN1L, C10orf95, NFIC, POLR2J2. The main pathways involved in the IBM vs PM phenotype differentiation can be consulted in 4.2.5. Figure 5.4 B) shows the PCA graphic of the IBM vs PM classification, and how this separability can be achieved. This methodology can be extrapolated to the analysis of other rare diseases in the search of orphan drugs. This project has been named FINISTERRAE and it is under development within the group of inverse problems, optimization and machine learning of the mathematics department of the university of Oviedo.

Table 5.2 CLL, IBM &amp; PM and ALS results

CLL			IBM & PM			ALS		
<i>Acc</i> (%)	<i>tol</i>	<i>#R</i>	<i>Acc</i> (%)	<i>tol</i>	<i>#R</i>	<i>Acc</i> (%)	<i>tol</i>	<i>#R</i>
92.64	85.89	488	87.50	82.50	223	84.71	83.53	547
92.64	86.50	487	87.50	85.00	159	85.88	84.71	441
92.64	89.57	486	90.00	87.50	138	87.06	85.88	241
92.64	90.18	479	90.00	90.00	71	88.24	87.06	197
92.64	90.80	446	92.50	92.50	32	90.59	88.24	134
92.64	91.41	373	100.00	95.00	2	91.76	89.41	96
93.25	92.02	255	97.50	97.50	1	90.59	90.59	54
93.25	92.64	120				92.94	91.76	32
93.25	93.25	22				95.29	92.94	20
						94.12	94.12	10
						95.29	95.29	6
						96.47	96.47	1

## 5.5 Predicting Amyotrophic Lateral Sclerosis with Biomedical Robots

Amyotrophic Lateral Sclerosis (ALS) is a motor neuron disease that characterized by stiff muscles, muscle twitching, and gradually worsening weakness. Between 5 and 10% of the cases are inherited from a relative, and for the rest of cases, the cause is still unknown (NINDS, 2013). It is a progressive disease that the average survival from onset to death is three to four years, in which most of them die from a respiratory failure. There is no cure yet.

We reinterpreted the dataset published by Lincecum et al. (2010) consisting of 57 ALS cases and 28 healthy controls. The best result yields an accuracy of 96.5% with small scale signature involving the following genes: CASP1, ZNF787 and SETD7. Table 5.2 shows the results of applying this methodology to this problem. The biomedical robots in this case did not improve this prediction. The pathway analysis can be consulted in section 4.2.5.

Figure 5.5 shows the correlation network of the most discriminatory genes found in this analysis. The head of the network is the CASP1 that is connected to MAP2K5, through ZNF3 and LUC7. MAP2K5 acts as a scaffold for the formation of a pathway that seems to play a critical role in protecting cells from stress-induced apoptosis, neuronal survival,

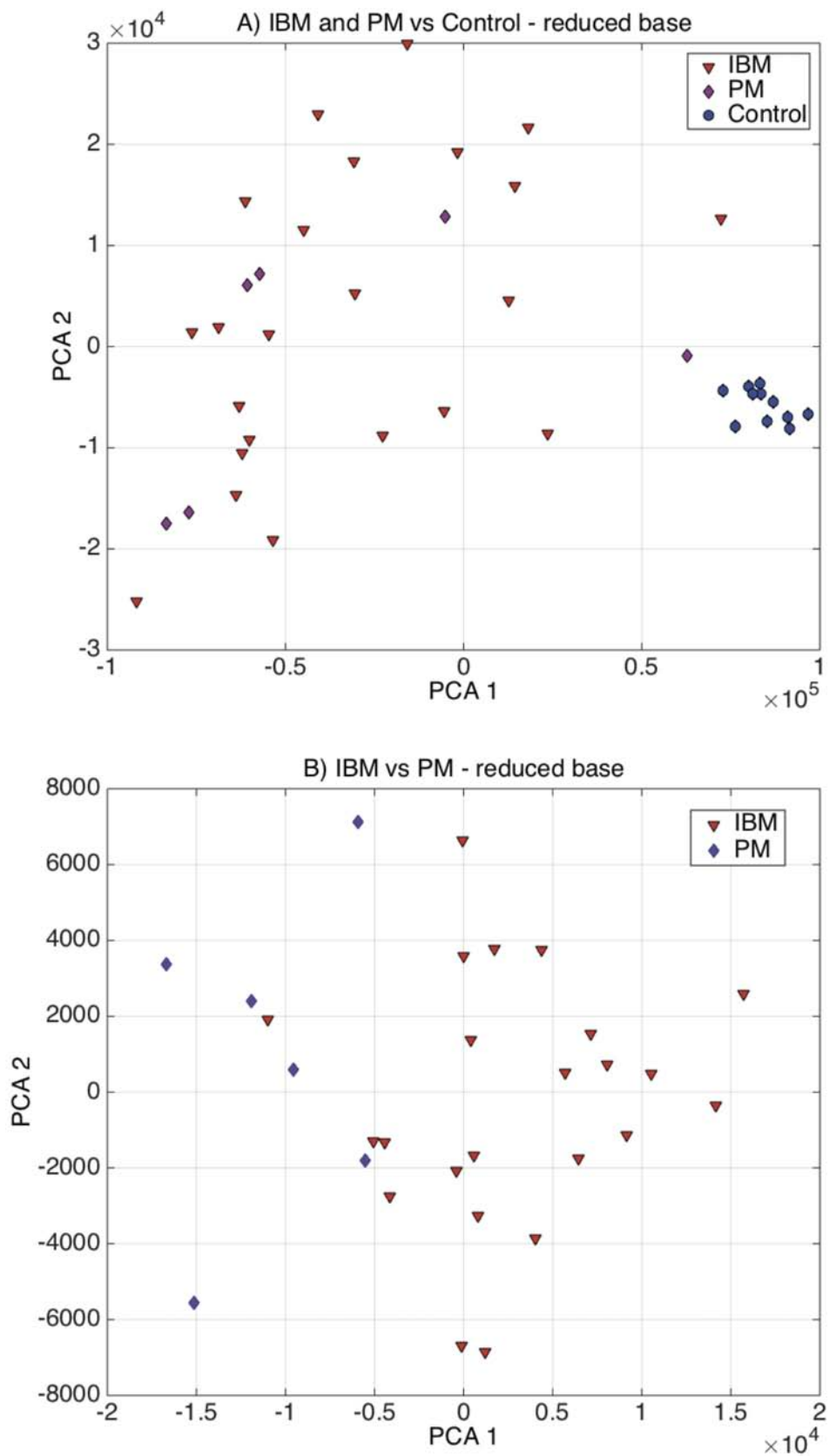


Fig. 5.4 Classification of IBM, PM and Control: A) PCA graphic for IBM+PM versus control samples. B) PCA graphic for IBM versus PM.

cardiac development and angiogenesis. Also DCAF8 has been associated to neuropathies. Figure 5.6 shows the PCA graphic for the ALS vs control samples.

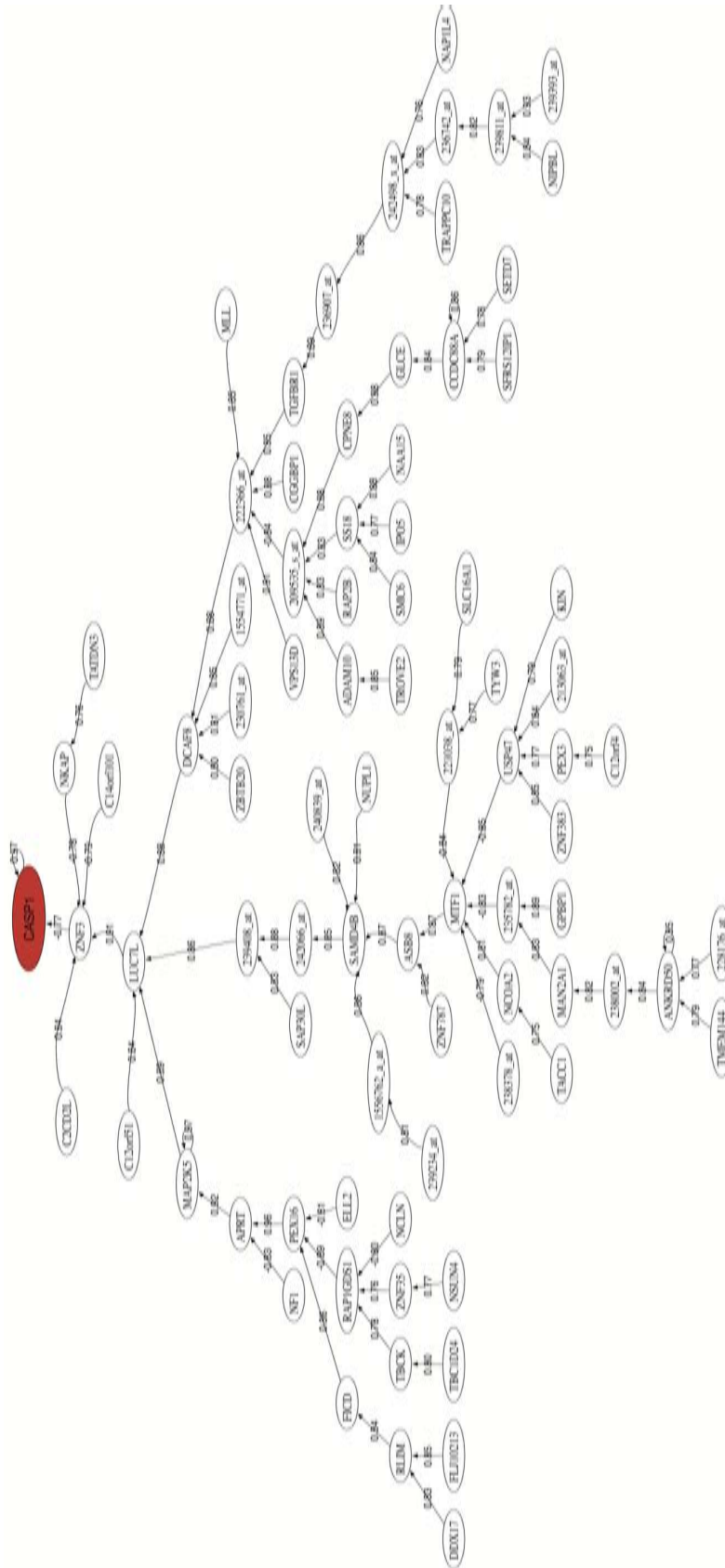


Fig. 5.5 Correlation network for Amyotrophic Lateral Sclerosis. Probe names are used when gene names are unknown.

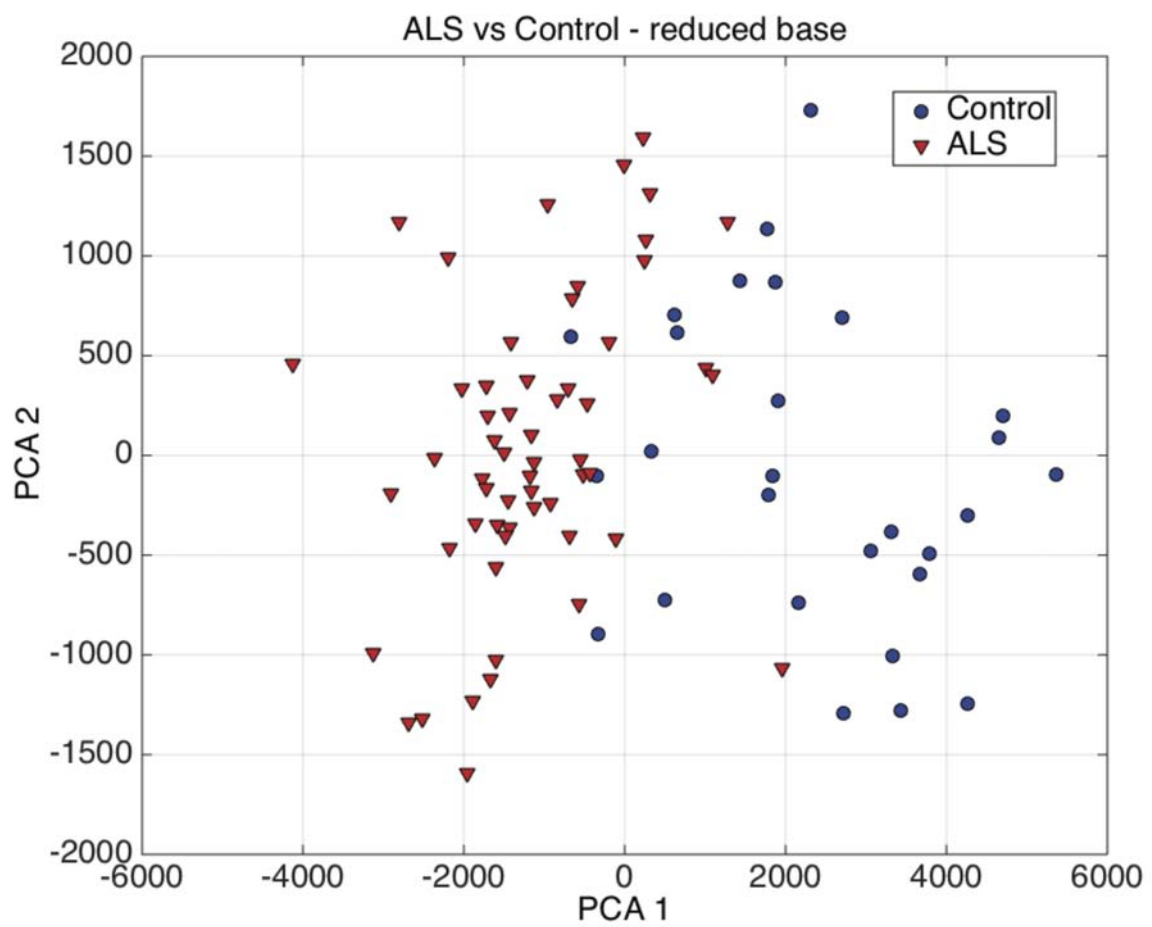


Fig. 5.6 PCA graphic for ALS versus control samples

# Chapter 6

## Conclusions and future research

In this dissertation we introduced the novel concept, analysis and design of biomedical robots, defined as the ensemble of methodologies and bioinformatics algorithms, derived from applied mathematics, statistics and computer science that are able of dynamically analyzing high dimensional data, discovering knowledge, generating new biomedical working hypothesis, and supporting medical decision making with its corresponding uncertainty assessment. This methodology is based in exploring the uncertainty space of any biomedical classification problem, and using the structure of the uncertainty space to adopt decisions and inferring knowledge.

The first complexity that we have to face is that in most of the biomedical problems there is no physical relationship available relating input and output variables, therefore, the functional accounting for the forward problem in these cases is a priori unknown. We have decided to approach them as nonlinear classification problems, since the classifier and the variables that serve to achieve an optimum prediction are a priori unknown. Due to the impact of the decisions that are going to be made, we decided to avoid the use of black-box methodologies that provide solution without the medical doctor's understanding. Medical doctors and biomedical researches are interested in both, the final prediction results, and the different sets of prognostic variables that allowed to optimally solve the corresponding prediction problem. These sets of prognostic variables belong to the "uncertainty space" of the prediction problem. Thereby, to approach the nonlinear character of the classification problem we have decoupled the feature selection problem from the final prediction and risk analysis assessment. Additionally, we have numerically showed (via Principal Component Analysis, PCA) that when all the variables parametrizing the samples are considered, the corresponding classification problem is nonlinear separable, that is, it is not possible to define in the feature space a set of hyperplanes that optimally separates the samples of each class. Nevertheless, it is possible to discard irrelevant variables that do not provide any useful

information for the discrimination, and introduce ambiguity in the classification. We also proved that when the relevant prognostic variables are correctly selected, the classification problem approximates a linear separable behavior. This simple approach allows identifying the behavioral outliers, that are the samples incorrectly classified.

The second complexity we had to face was that due to economic constraints the number of samples is finite (hundreds of samples) and the number of prognostic variables (genes or genetic probes) is much higher (hundred of thousands). This fact confers to these problems an ill-posed character due to their high degree of underdetermination. To reduce the dimensionality of the feature space we used different filter/ranking methods: Fisher's Ratio, Fold Change, Entropy, Maximum Percentile Distance and Significance Microarray Analysis. These methods served to rank the variables according to their discriminatory power. Besides, in the case of clinical data the feature data had to be previously imputed to fully take advantage of the strengths of this methodology. The use of a distance-based classifier combined with iterative feature elimination allowed to determine the small-scale list of ranked prognostic variables for the prediction. These lists are of paramount importance for cheaper and faster diagnosis, prognosis and treatment optimization. Moreover, using an expanded list of high discriminatory features, correlation networks (Pearson correlation coefficient and Normalized Mutual Information) were established using the minimum spanning tree through the Kruskal algorithm. These correlation networks served to visualize the existing univariate relationship between the main prognostic variables. In the case of genetic data the correlation networks can be used to classify genes into two main categories: Headers and Helpers. Header genes control most of the accuracy in the phenotype prediction, and therefore are those which are supposed to be highly correlated to the disease progression. Conversely, helper genes are those that provide high frequency details in the discrimination. The morphology of these networks (elongated or horizontally very dense) would explain the genetic/molecular complexity that has to be faced. Additionally, the methodology is completed by the ontology analysis of the most discriminatory genes through the GeneAnalytics platform and related software made at disposal by the Weizmann Institute of Science.

Biomedical data is also notorious for containing noise that could affect to the inference process and the mechanistic conclusions deduced from the analysis. An important part of this PhD is composed by the sensitivity analysis to noise in both feature selection and classification problems. For that purpose we formally defined two main sources of noise: noise in the feature data and in the class assignment. We checked the robustness of the methods applied in the methodology against these sources of noise, experimentally showing that noise in expression data and class assignment partially falsifies the sets of discriminatory probes in phenotype prediction problems, that is, in presence of noise the set of variables



---

with the highest predictive accuracy will never perfectly coincide with the set(s) of variables that explains the disease. Via synthetic modeling we have shown that Fisher's Ratio and Significance Analysis of Microarrays are the most robust gene selection methods, exploiting the parsimony principle, and being able of finding small-scale high discriminatory gene signatures. We have also proved that noise in class assignment affect the predictive accuracy and the precision more than noise in the expression data. Moreover, we showed that the combination of both types of noise (expression and class assignment) affects drastically the finding of differentially expressed genes, and therefore, the involved biological pathways.

We showed the importance of establishing the discriminatory power of the genes in phenotype prediction problems to correctly find the biological pathways that are involved. Nevertheless, not only noise affects to the unraveling of biological pathways but also the preprocessing techniques that are commonly employed in genetic expression data to filter and damping acquisition noise, before their analysis. We evaluated how the most common genetic expression data preprocessing techniques, RMA and MAS5, impact the finding of biological pathways associated to phenotype prediction problems. We showed that the pathways analysis and the correlation networks were significantly different depending if we used raw or preprocessed data. Through an international well-known synthetic data (spiked-in experiment), we proved that the mean precision was higher, and we were also able to locate all the differentially-expressed genes selecting rather less number of genes using raw data instead of preprocessed data. Besides, RMA provided better results than MAS5 in general terms. Working with the radiotherapy related fatigue dataset, the performance of preprocessing techniques are better than working with raw data in training data sets. However, in validation data, raw datasets allow us to generalize better than using preprocessed data. Besides the pathways analysis and the correlation networks were significantly different between them. Consequently, it can be concluded that preprocessed data "falsify" the biological pathway analysis that is performed in phenotype prediction problems, and interpreting results from predictive gene profiles generated by RMA and mainly via MAS5 should be done with caution. This is an important conclusion with a high translational impact that should be confirmed in other disease datasets and can importantly impact in drug discovery.

The presence of noise combined with the high underdetermined character of the biomedical problems provokes an uncertainty that should be managed properly. Consequently, it is necessary to sample the "uncertainty space" looking for other prediction models, using different set of variables, with lower predictive accuracy than the optimum, and carry out a final prediction with its uncertainty assessment using all the prediction models via a consensus strategy. Under all of these circumstances, we presented the biomedical robots methodology as a simple and robust solution in response to those challenges. We have shown

the application of this novel concept to 3 different illnesses: Chronic Lymphocytic Leukemia, Inclusion Body Myositis - Polymyositis and Amyotrophic Lateral Sclerosis, proving that it is possible to infer at the same time, both, high discriminatory small-scale signatures and the description of the biological pathways involved. The pathway analyses revealed in the three cases a possible link to viral infections and served to identify actionable genes and drug targets. Additionally, the methodology has been also applied and tested to both clinical and genetic data modeling.

In the case of clinical data we were able to select the optimum prognostic variables that provide simple biomedical discriminatory rules for diagnosis, prognosis and treatment optimization for two different problems: Chronic Lymphocytic Leukemia and Hodgkin Lymphoma. We found that clinical data has an heterogeneity character where variables are expressed in different measures with different bounds, and they are not usually available in all the samples. This fact makes clinical data preprocessing techniques the key for finding a solution. Taking this into account we presented a methodology that consists in three steps: 1) Data preprocessing, 2) Feature selection and 3) Risk assessment. We provided the corresponding risk assessment using ROC curves, and improving it optimizing the confusion matrix given by the classification problem (this ROC optimization was only performed in the case of Hodgkin Lymphoma).

In the case of Chronic Lymphocytic Leukemia we established the relevance of clinical variables that are not widely used as prognostic factor for the need of chemotherapy treatment and autoimmune disease development, optimizing thus both prognosis and treatment. The need of chemotherapy treatment seems to be related to the amount of malignant leukemia cells that are measured by the different leucocytes counts. The results of autoimmune disease development prediction showed the importance of variables associated with the characteristics of platelets, reticulocytes and natural killers, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia. The prognostic significance of the selected variables might probably reflect the relevance of some clinical aspects of this disease that are more important for prognosis than it is currently thought. These variables are obtained at diagnosis, and their use would not increase the cost or complexity of the diagnosis in CLL patients. We also showed an additional results for survival analysis showing that the selected variables for patients that did not survive during the first year are quite different from those that survive until the 3rd or 5th year since the diagnosis. Remarkably, results suggest that the adequate identification and treatment of these complications may play a more important role in the survival.

Regarding the Hodgkin Lymphoma we detected those patients who do not respond to the treatment at early stages identifying prognostic variables currently gathered at diagnosis

---

that may help to detect, with high accuracy, those patients with bad prognosis without any additional cost. Thereby we could improve their treatment. The results of this study show that the combined use of these selected prognostic variables gathered at diagnosis, allows predicting first-line treatment response with high accuracy and confirms a close relationship between treatment response, inflammation, iron overload and liver and bone damage.

We have also applied our methodology using genetic data, proving that we were able to manage the underdetermined character of this type of data. Genetic data has a high underdetermined character, since the number of samples/patients is always much lower than the number of genes. We do not have a unique solution to the inverse problem, therefore, reduction of dimension algorithms become a key element in the problem solution. We applied our methodology to address two different problems using gene expression data. Firstly, we identify and validate a specific gene cluster that is predictive of fatigue risk in prostate cancer patients treated with radiotherapy. Secondly we modeled a data expression microarray related to Chronic Lymphocytic Leukemia, predicting the occurrence of the main mutations, which are closely related with the survival of the patients.

In the first case we can predict radiotherapy-related fatigue in men with non-metastatic prostate cancer, by reducing the dimension to the most discriminatory genes. We proposed that the risk of a complex disease, such as radiotherapy-related fatigue, could well be more easily defined by identifying groups of simultaneously expressed, synergistically functioning genes. Our finding that the gene cluster so identified was able to predict radiotherapy-related fatigue risk with an accuracy of  $> 75\%$  suggests that the approach has validity. Applicability of this novel methodology to detect other treatment-related toxicities in other cancer populations would be worthwhile to pursue. The importance of predicting toxicity or adverse event risk associated with cancer treatment regimens cannot be understated as the clinical implications in personalizing cancer therapy and prospectively attenuating toxicity risk are significant. Furthermore, this type of information provides patients and their care-givers more specific knowledge upon which to make treatment decisions.

In the second case we figured out how the main mutations in Chronic Lymphocytic Leukemia affect gene expression by finding small-scale signatures. Our methodological approach could define hierarchical gene relationships among patients with Chronic Lymphocytic Leukemia expressing the different main mutations (IgHV, NOTCH1 and SF3B1) and establishing the predictive accuracy of gene clusters relative to each mutation. Besides, our methodology served to depict the gene clusters that are most strongly associated with the expression of each selective mutation (networks of synergistically working genes), and their relationship between mutation expressions with a particular clinical outcome (survival). The results allowed to define and understand the biological pathways and correlation networks

that are involved in the disease development with the potential goal of identifying new druggable targets. We also analyzed the intersection between the most discriminatory genes for IgVH, NOTCH1, and SF3B1 mutations, showing that only four genes were common to all mutations: IGHG1, MYBL1, NRIP1 and RGS13. Additionally, we included the analysis of the next more important mutation: NOP16; showing that the intersection got reduced to 2 genes: IGHG1 and RGS13 when we include NOP16. IGHG1 has been already related to hypogammaglobulinemia and B-cell chronic lymphocytic leukemia. RGS13 is related to mantle cell lymphoma.

Finally, the methodology shown in this dissertation is not computationally very expensive, since all the simulations shown in this research work were done with a personal computer in real time (several minutes).

Future research will be devoted to:

1. Developing new simpler and faster algorithms that allow us reducing drastically the dimension when dealing with big data.
2. Applying and exploring new approaches in the consensus strategy.
3. Verification of the findings in the practical cases with other independent cohorts that could lead to a better design of the therapeutic targets.
4. Developing new preprocessing techniques or improving existing ones that allow a correct generalization and do not impact in the pathway analysis.
5. Exploring preprocessing techniques in clinical data that become a key element in solving clinical-related problems.
6. Extending the application of the proposed methodology to other biomedical problems, such as, the estimation of surgical risk using different types of biomedical data (data fusion), or the Finisterrae Project that consists in the analysis of rare, neurodegenerative diseases, and cancer.
7. Exploring other machine learning techniques, such as, extreme learning machines, proximal algorithms and deep learning, with their uncertainty analysis.
8. Analysis of the genomic aberrations in CLL using the methodology that has been developed for understanding mutations.
9. Design of fast and cheap methods of early diagnosis for very aggressive cancers (breast, lung, colorectal, pancreas, etc). Particularly it is important predicting the metastasis or the recurrence at diagnosis.

# References

- Affymetrix (2001). Microarray suite user guide, version 5. <http://www.affymetrix.com/support/technical/manuals.affx>.
- Affymetrix (2015). Latin square data for expression algorithm assessment. [http://www.affymetrix.com/support/technical/sample\\_data/datasets.affx](http://www.affymetrix.com/support/technical/sample_data/datasets.affx).
- Alix-Panabieres, C. and Pantel, K. (2013). Circulating tumor cells: liquid biopsy of cancer. *Clin Chem*, 59(1):110–118.
- Aster, R., Borchers, B., and Thurber, C. (2012). *Parameter Estimation and Inverse Problems*. Elsevier, second edition.
- Bansal, G., DiVietro, J. A., Kuehn, H. S., Rao, S., Nocka, K. H., Gilfillan, A. M., and Druey, K. M. (2008). Rgs13 controls g protein-coupled receptor-evoked responses of human mast cells. *J Immunol*, 181(11):7882–7890.
- Bellazzi, R., Diomidous, M., Sarkar, I. N., Takabayashi, K., Ziegler, A., and McCray, A. T. (2011). Data analysis and data mining: Current issues in biomedical informatics. *Methods of information in medicine*, 50(6):536–544.
- Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114.
- Binet, J., Auquier, A., Dighiero, G., Chastang, C., Piguët, H., Goasguen, J., Vaugier, G., Potron, G., Colona, P., Oberling, F., et al. (1981). A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer*, 48(1):198–206.
- Brensilver, H. L. and Kaplan, M. M. (1975). Significance of elevated liver alkaline phosphatase in serum. *Gastroenterology*, 68(6):1556–1562.
- Brinckmeyer, L. M., Skovsgaard, T., Thiede, T., Vesterager, L., and Nissen, N. I. (1982). The liver in hodgkin's disease—i. clinico-pathological relations. *Eur J Cancer Clin Oncol*, 18(5):421–428.
- Butt, A. J., Sergio, C. M., Inman, C. K., Anderson, L. R., McNeil, C. M., Russell, A. J., Nusch, M., Preiss, T., Biankin, A. V., Sutherland, R. L., and Musgrove, E. A. (2008). The estrogen and c-myc target gene hspc111 is over-expressed in breast cancer and associated with poor patient outcome. *Breast Cancer Res*, 10(2):R28.

- Carlotto, A., Hogsett, V. L., Maiorini, E. M., Razulis, J. G., and Sonis, S. T. (2013). The economic burden of toxicities associated with cancer treatment: review of the literature and analysis of nausea and vomiting, diarrhoea, oral mucositis and fatigue. *Pharmacoeconomics*, 31(9):753–766.
- Chawla, N., Bowyer, K., Hall, L., and WP, K. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE*, 6(2):e17238.
- Cheson, B. D. (2008). New staging and response criteria for non-hodgkin lymphoma and hodgkin lymphoma. *Radiol Clin North Am*, 46(2):213–223.
- Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cdna microarray experiments. *Genome Biol*, 4(4):210.
- Cunha, B. A. (2007). Fever of unknown origin (fuo): diagnostic importance of serum ferritin levels. *Scand J Infect Dis*, 39(6-7):651–652.
- Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by sam-gs. *BMC Bioinformatics*, 8:242.
- Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., Döhner, K., Bentz, M., and Lichter, P. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. *New England Journal of Medicine*, 343(26):1910–1916.
- Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36.
- Fayad, L. M., Carrino, J. A., and Fishman, E. K. (2007). Musculoskeletal infection: role of ct in the emergency department. *Radiographics*, 27(6):1723–1736.
- Fernández-Martínez, J., Fernández-Muñiz, M., and Tompkins, M. (2012). On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics*, 77(1):1–15.
- Fernández-Martínez, J., Fernández-Muñiz, Z., Pallero, J., and Pedruelo-González, L. (2013). From bayes to tarantola: New insights to understand uncertainty in inverse problems. *Journal of Applied Geophysics*, 98:62–72.
- Fernández-Martínez, J. and García Gonzalo, E. (2009). The pso family: deduction, stochastic analysis and comparison. *Swarm Intell*, 3(4):245–73.
- Fernández-Martínez, J. and García Gonzalo, E. (2012). Stochastic stability and numerical analysis of two novel algorithms of the pso family: Pp-gpso and rr-gpso. *Int J Artif Intell Tools*, 21(3):1240011.
- Fernández-Martínez, J., Pallero, J., and Fernández Muñiz, Z. (2014a). The effect of noise and tikhonov’s regularization in inverse problems. part i: The linear case. *Journal of Applied Geophysics*, 108:176 – 185.

- Fernández-Martínez, J., Pallero, J., and Fernández Muñoz, Z. (2014b). The effect of noise and tikhonov's regularization in inverse problems. part ii: The nonlinear case. *Journal of Applied Geophysics*, 108:186 – 193.
- Fernández-Martínez, J. L. and García Gonzalo, E. (2008). The generalized pso: A new door to pso evolution. *Journal of Artificial Evolution and Applications*, 2008.
- Ferreira, P. G., Jares, P., Rico, D., Gomez-Lopez, G., Martinez-Trillos, A., Villamor, N., Ecker, S., Gonzalez-Perez, A., Knowles, D. G., Monlong, J., Johnson, R., Quesada, V., Djebali, S., Papasaikas, P., Lopez-Guerra, M., Colomer, D., Royo, C., Cazorla, M., Pinyol, M., Clot, G., Aymerich, M., Rozman, M., Kulis, M., Tamborero, D., Gouin, A., Blanc, J., Gut, M., Gut, I., Puente, X. S., Pisano, D. G., Martin-Subero, J. I., Lopez-Bigas, N., Lopez-Guillermo, A., Valencia, A., Lopez-Otin, C., Campo, E., and Guigo, R. (2014). Transcriptome characterization by rna sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res*, 24(2):212–226.
- Filler, K., Lyon, D., McCain, N., Bennett, J. J., Fernandez-Martinez, J. L., deAndres Galiana, E. J., Elswick, R. K. J., Lukkahatai, N., and Saligan, L. (2015). Relationship of mitochondrial enzymes to fatigue intensity in men with prostate cancer receiving external beam radiation therapy. *Biol Res Nurs*.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188.
- Futschik, M. E., Sullivan, M., Reeve, A., and Kasabov, N. (2003). Prediction of clinical behaviour and treatment for cancers. *Appl Bioinformatics*, 2(3 Suppl):S53–8.
- Gandhi, M. K., Lambley, E., Burrows, J., Dua, U., Elliott, S., Shaw, P. J., Prince, H. M., Wolf, M., Clarke, K., Underhill, C., Mills, T., Mollie, P., Gill, D., Marlton, P., Seymour, J. F., and Khanna, R. (2006). Plasma epstein-barr virus (ebv) dna is a biomarker for ebv-positive hodgkin's lymphoma. *Clin Cancer Res*, 12(2):460–464.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.
- Gonzalez-Rodriguez, A. P., Contesti, J., Huergo-Zapico, L., Lopez-Soto, A., Fernandez-Guizan, A., Acebes-Huerta, A., Gonzalez-Huerta, A. J., Gonzalez, E., Fernandez-Alvarez, C., and Gonzalez, S. (2010). Prognostic significance of cd8 and cd4 t cells in chronic lymphocytic leukemia. *Leuk Lymphoma*, 51(10):1829–1836.
- Greenberg, S. A., Bradshaw, E. M., Pinkus, J. L., Pinkus, G. S., Burleson, T., Due, B., Bregoli, L., O'Connor, K. C., and Amato, A. A. (2005). Plasma cells in muscle in inclusion body myositis and polymyositis. *Neurology*, 65(11):1782–1787.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422.
- Hadamard, J. (1902). Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52.

- Hallek, M., Cheson, B. D., Catovsky, D., Caligaris-Cappio, F., Dighiero, G., Dohner, H., Hillmen, P., Keating, M. J., Montserrat, E., Rai, K. R., and Kipps, T. J. (2008). Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the international workshop on chronic lymphocytic leukemia updating the national cancer institute-working group 1996 guidelines. *Blood*, 111(12):5446–5456.
- Hallek, M., Wanders, L., Ostwald, M., Busch, R., Senekowitsch, R., Stern, S., Schick, H. D., Kuhn-Hallek, I., and Emmerich, B. (1996). Serum beta(2)-microglobulin and serum thymidine kinase are independent predictors of progression-free survival in chronic lymphocytic leukemia and immunocytoma. *Leuk Lymphoma*, 22(5-6):439–447.
- Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G., and Stevenson, F. K. (1999). Unmutated ig v(h) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*, 94(6):1848–1854.
- Hasenclever, D., Diehl, V., Armitage, J. O., Assouline, D., Björkholm, M., Brusamolino, E., Canellos, G. P., Carde, P., Crowther, D., Cunningham, D., Eghbali, H., Ferm, C., Fisher, R. I., Glick, J. H., Glimelius, B., Gobbi, P. G., Holte, H., Horning, S. J., Lister, T. A., Longo, D. L., Mandelli, F., Polliack, A., Proctor, S. J., Specht, L., Sweetenham, J. W., and Hudson, G. V. (1998). A prognostic score for advanced hodgkin's disease. *New England Journal of Medicine*, 339(21):1506–1514. PMID: 9819449.
- He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328.
- Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics, Second Edition*. New York: Wiley.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.
- Islam, T. C., Asplund, A. C., Lindvall, J. M., Nygren, L., Liden, J., Kimby, E., Christensson, B., Smith, C. I. E., and Sander, B. (2003). High level of cannabinoid receptor 1, absence of regulator of g protein signalling 13 and differential expression of cyclin d1 in mantle cell lymphoma. *Leukemia*, 17(9):1880–1890.
- Iwaki, S., Lu, Y., Xie, Z., and Druey, K. M. (2011). p53 negatively regulates rgs13 protein expression in immune cells. *J Biol Chem*, 286(25):22219–22226.
- Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:359.
- Josting, A. (2010). Prognostic factors in hodgkin lymphoma. *Expert Rev Hematol*, 3(5):583–592.



- Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.
- Kellsell, D. P., Dunlop, J., and Hodgins, M. B. (2001). Human diseases: clues to cracking the connexin code? *Trends Cell Biol*, 11(1):2–6.
- Kennedy, J. and Eberhart, R. (1995). *Particle Swarm Optimization*, volume 4. PICNN.
- Kittivorapart, J. and Chinthammitr, Y. (2011). Incidence and risk factors of bone marrow involvement by non-hodgkin lymphoma. *J Med Assoc Thai*, 94 Suppl 1:S239–45.
- Kooperberg, C., Fazzio, T. G., Delrow, J. J., and Tsukiyama, T. (2002). Improved background correction for spotted dna microarrays. *J Comput Biol*, 9(1):55–66.
- Koval, M. (2006). Pathways and control of connexin oligomerization. *Trends Cell Biol*, 16(3):159–166.
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50.
- Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634.
- Larsson, O., Wahlestedt, C., and Timmons, J. A. (2005). Considerations when using the significance analysis of microarrays (sam) algorithm. *BMC Bioinformatics*, 6:129.
- Lastra, G., Luaces, O., Quevedo, J., and Bahamonde, A. (2011). Graphical feature selection for multilabel classification tasks. In Gama, J., Bradley, E., and Hollmén, J., editors, *Advances in Intelligent Data Analysis X*, volume 7014 of *Lecture Notes in Computer Science*, pages 246–257. Springer Berlin / Heidelberg.
- Lincecum, J. M., Vieira, F. G., Wang, M. Z., Thompson, K., De Zutter, G. S., Kidd, J., Moreno, A., Sanchez, R., Carrion, I. J., Levine, B. A., Al-Nakhala, B. M., Sullivan, S. M., Gill, A., and Perrin, S. (2010). From transcriptome analysis to therapeutic anti-cd40l treatment in the sod1 model of amyotrophic lateral sclerosis. *Nat Genet*, 42(5):392–399.
- Liu, X., Wu, J., and Z, Z. (2006). Exploratory under-sampling for class-imbalance learning. In *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 965–969.
- Masuhara, M., Sakamoto, H., Matsumoto, A., Suzuki, R., Yasukawa, H., Mitsui, K., Wakioka, T., Tanimura, S., Sasaki, A., Misawa, H., Yokouchi, M., Ohtsubo, M., and Yoshimura, A. (1997). Cloning and characterization of novel cis family genes. *Biochem Biophys Res Commun*, 239(2):439–446.
- Minamoto, S., Ikegame, K., Ueno, K., Narazaki, M., Naka, T., Yamamoto, H., Matsumoto, T., Saito, H., Hosoe, S., and Kishimoto, T. (1997). Cloning and functional analysis of new members of stat induced stat inhibitor (ssi) family: Ssi-2 and ssi-3. *Biochem Biophys Res Commun*, 237(1):79–83.
- Minton, O., Richardson, A., Sharpe, M., Hotopf, M., and Stone, P. (2008). A systematic review and meta-analysis of the pharmacological treatment of cancer-related fatigue. *J Natl Cancer Inst*, 100(16):1155–1166.

- Mock, V. (2003). Clinical excellence through evidence-based practice: fatigue management as a model. *Oncol Nurs Forum*, 30(5):787–796.
- Moore, J. H. and Hill, D. P. (2015). Epistasis analysis using artificial intelligence. *Methods Mol Biol*, 1253:327–346.
- NINDS (2013). *Motor Neuron Diseases Fact Sheet*. National Institute of Neurological Disorders and Stroke.
- Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization*, 1:123–231.
- Pawitan, Y. and Ploner, A. (2015). Ocplus: Operating characteristics plus sample size and local fdr for microarray experiments. *R package*.
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.
- Provencio, M., Espana, P., Millan, I., Yebra, M., Sanchez, A. C., de la Torre, A., Bonilla, F., Regueiro, C. A., and de Letona, J. M. L. (2004). Prognostic factors in hodgkin’s disease. *Leuk Lymphoma*, 45(6):1133–1139.
- Puente, X. S., Pinyol, M., Quesada, V., Conde, L., Ordonez, G. R., Villamor, N., Escaramis, G., Jares, P., Bea, S., Gonzalez-Diaz, M., Bassaganyas, L., Baumann, T., Juan, M., Lopez-Guerra, M., Colomer, D., Tubio, J. M. C., Lopez, C., Navarro, A., Tornador, C., Aymerich, M., Rozman, M., Hernandez, J. M., Puente, D. A., Freije, J. M. P., Velasco, G., Gutierrez-Fernandez, A., Costa, D., Carrio, A., Guijarro, S., Enjuanes, A., Hernandez, L., Yague, J., Nicolas, P., Romeo-Casabona, C. M., Himmelbauer, H., Castillo, E., Dohm, J. C., de Sanjose, S., Piris, M. A., de Alava, E., San Miguel, J., Royo, R., Gelpi, J. L., Torrents, D., Orozco, M., Pisano, D. G., Valencia, A., Guigo, R., Bayes, M., Heath, S., Gut, M., Klatt, P., Marshall, J., Raine, K., Stebbings, L. A., Futreal, P. A., Stratton, M. R., Campbell, P. J., Gut, I., Lopez-Guillermo, A., Estivill, X., Montserrat, E., Lopez-Otin, C., and Campo, E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 475(7354):101–105.
- Quesada, V., Conde, L., Villamor, N., Ordonez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Bea, S., Pinyol, M., Martinez-Trillos, A., Lopez-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M., Rozman, M., Delgado, J., Gine, E., Hernandez, J. M., Gonzalez-Diaz, M., Puente, D. A., Velasco, G., Freije, J. M. P., Tubio, J. M. C., Royo, R., Gelpi, J. L., Orozco, M., Pisano, D. G., Zamora, J., Vazquez, M., Valencia, A., Himmelbauer, H., Bayes, M., Heath, S., Gut, M., Gut, I., Estivill, X., Lopez-Guillermo, A., Puente, X. S., Campo, E., and Lopez-Otin, C. (2012). Exome sequencing identifies recurrent mutations of the splicing factor sf3b1 gene in chronic lymphocytic leukemia. *Nat Genet*, 44(1):47–52.
- Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.
- Rai, K. R., Sawitsky, A., Cronkite, E. P., Chanana, A. D., Levy, R. N., and Pasternack, B. S. (1975). Clinical staging of chronic lymphocytic leukemia. *Blood*, 46(2):219–234.

- Ramsay, A. J., Quesada, V., Foronda, M., Conde, L., Martinez-Trillos, A., Villamor, N., Rodriguez, D., Kwarciak, A., Garabaya, C., Gallardo, M., Lopez-Guerra, M., Lopez-Guillermo, A., Puente, X. S., Blasco, M. A., Campo, E., and Lopez-Otin, C. (2013). Pot1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet*, 45(5):526–530.
- Rodriguez-Vicente, A. E., Diaz, M. G., and Hernandez-Rivas, J. M. (2013). Chronic lymphocytic leukemia: a clinical and molecular heterogenous disease. *Cancer Genet*, 206(3):49–62.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*, 65(6):386–408.
- Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P., and Davis, R. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *PNAS*, 20(93):10614–10619.
- Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. John Wiley and Sons.
- Schreck, S., Friebel, D., Buettner, M., Distel, L., Grabenbauer, G., Young, L. S., and Niedobitek, G. (2009). Prognostic impact of tumour-infiltrating th2 and regulatory t cells in classical hodgkin lymphoma. *Hematol Oncol*, 27(1):31–39.
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(379–423, 623).
- Smolewski, P., Robak, T., Krykowski, E., Blasinska-Morawiec, M., Niewiadomska, H., Pluzanska, A., Chmielowska, E., and Zambrano, O. (2000). Prognostic factors in hodgkin's disease: multivariate analysis of 327 patients from a single institution. *Clin Cancer Res*, 6(3):1150–1160.
- Sonis, S., Haddad, R., Posner, M., Watkins, B., Fey, E., Morgan, T. V., Mookanamparambil, L., and Ramoni, M. (2007). Gene expression changes in peripheral blood cells provide insight into the biological mechanisms associated with regimen-related toxicities in patients being treated for head and neck cancers. *Oral Oncol*, 43(3):289–300.
- Stelzer, G., Inger, A., Olender, T., Iny-Stein, T., Dalah, I., Harel, A., Safran, M., and D, L. (2009). Genedecks: paralog hunting and gene-set distillation with genecards annotation. *OMICS*, 13(6):477–87.
- Strehl, A. and Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.
- Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. By JOHN A. SWETS. Lawrence Erlbaum Associates.
- Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):659–665.

- Troyanskaya, O. G., Cantor, M., Sherlock, G., Brown, P. O., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.
- Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*, 98(9):5116–21.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M.,

- Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.
- Wan, Y. and Wu, C. J. (2013). Sf3b1 mutations in chronic lymphocytic leukemia. *Blood*, 121(23):4627–4634.
- Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., Zhang, L., Zhang, W., Vartanov, A. R., Fernandes, S. M., Goldstein, N. R., Folco, E. G., Cibulskis, K., Tesar, B., Sievers, Q. L., Shefler, E., Gabriel, S., Hacohen, N., Reed, R., Meyerson, M., Golub, T. R., Lander, E. S., Neuberg, D., Brown, J. R., Getz, G., and Wu, C. J. (2011). Sf3b1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med*, 365(26):2497–2506.
- Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34.
- Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82.
- Yang, F. and Mao, K. (2011). Robust feature selection for microarray data based on multi-criterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):1080–1092.
- Yost, K. J., Eton, D. T., Garcia, S. F., and Cella, D. (2011). Minimally important differences were estimated for six patient-reported outcomes measurement information system-cancer scales in advanced-stage cancer patients. *J Clin Epidemiol*, 64(5):507–516.
- Zander, T., Wiedenmann, S., and Wolf, J. (2002). Prognostic factors in hodgkin’s lymphoma. *Ann Oncol*, 13 Suppl 1:67–74.



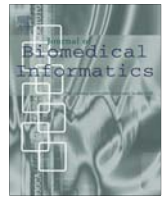
# **Appendix A**

## **Publications**

### **A.1 Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems**

Published in the Journal of Biomedical Informatics.

DOI: <http://dx.doi.org/10.1016/j.jbi.2016.02.017>



## Analysis of clinical prognostic variables for Chronic Lymphocytic Leukemia decision-making problems



Enrique J. deAndrés-Galiana<sup>a,d,1</sup>, Juan L. Fernández-Martínez<sup>a,1,\*</sup>, Oscar Luaces<sup>d</sup>, Juan J. del Coz<sup>d</sup>, Leticia Huergo-Zapico<sup>c</sup>, Andrea Acebes-Huerta<sup>c</sup>, Segundo González<sup>c</sup>, Ana P. González-Rodríguez<sup>b</sup>

<sup>a</sup> Department of Mathematics, University of Oviedo, Spain

<sup>b</sup> Hematology Department, Hospital Central de Asturias, Oviedo, Spain

<sup>c</sup> Instituto Universitario Oncológico del Principado de Asturias (IUOPA), University of Oviedo, Spain

<sup>d</sup> Artificial Intelligence Center, University of Oviedo, Spain

### ARTICLE INFO

#### Article history:

Received 14 September 2015

Revised 22 February 2016

Accepted 29 February 2016

Available online 5 March 2016

#### Keywords:

Chronic Lymphocytic Leukemia  
Chemotherapy Treatment  
Autoimmune disease development  
Machine learning

### ABSTRACT

**Introduction:** Chronic Lymphocytic Leukemia (CLL) is a disease with highly heterogeneous clinical course. A key goal is the prediction of patients with high risk of disease progression, which could benefit from an earlier or more intense treatment. In this work we introduce a simple methodology based on machine learning methods to help physicians in their decision making in different problems related to CLL.

**Material and methods:** Clinical data belongs to a retrospective study of a cohort of 265 Caucasians who were diagnosed with CLL between 1997 and 2007 in Hospital Cabueñes (Asturias, Spain). Different machine learning methods were applied to find the shortest list of most discriminatory prognostic variables to predict the need of Chemotherapy Treatment and the development of an Autoimmune Disease. **Results:** Autoimmune disease occurrence was predicted with very high accuracy (>90%). Autoimmune disease development is currently an unpredictable severe complication of CLL. Chemotherapy Treatment has been predicted with a lower accuracy (80%). Risk analysis showed that the number of false positives and false negatives are well balanced.

**Conclusions:** Our study highlights the importance of prognostic variables associated with the characteristics of platelets, reticulocytes and natural killers, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia for autoimmune disease development, and also, the relevance of some clinical variables related with the immune characteristics of CLL patients that are not taking into account by current prognostic markers for predicting the need of chemotherapy. Because of its simplicity, this methodology could be implemented in spreadsheets.

© 2016 Elsevier Inc. All rights reserved.

### 1. Introduction

Chronic Lymphocytic Leukemia (CLL) is the most common adult Leukemia in western countries, and it is characterized by the accumulation of malignant B-cells in blood and lymphoid organs. The clinical course of CLL is highly heterogeneous since the survival of some patients is only slightly affected by the disease, whereas other patients have a progressive disease associated with infectious and autoimmune complications. These progressive patients have poor prognosis, but they could benefit from an earlier or more intense chemotherapeutic treatment. It has been reported that many poor prognostic factors (including CD38, ZAP-70,

$\beta$ 2-microglobulin, IgVH mutation status and deletions of 11q23 or 17p53) may help to identify high-risk patients at early stages [1–6]. Most of these prognostic factors focus on the analysis of the characteristics of malignant leukemia cells. Additionally, the characteristics of the immune system of CLL patients, such as the number of CD8 and CD4-T cells at diagnosis, may also predict the progression of the disease [6]. Nevertheless, due to their high cost and complexity some of these prognostic factors are not used in most hospitals on regular basis. To overcome this problem in the clinical practice staging systems using few, simple, cheap and accessible clinical variables have been popularized. The Rai staging system [7] and the Binet classification [8] are useful to predict the prognosis of CLL patients, to stratify them, and to achieve comparisons for interpreting specific treatment results. Staging systems stratify subsets of patients who have significant differences in the overall survival but they fail to identify patients who have a high

\* Corresponding author at: Jesús Arias de Velasco s/n, 33005 Oviedo, Spain. Tel.: +34 985 103 199.

E-mail address: [jlfm@uniovi.es](mailto:jlfm@uniovi.es) (J.L. Fernández-Martínez).

<sup>1</sup> Both authors have contributed equally to this study.



risk of progression in early stages of the disease. Additionally, no current prognostic factors exist to predict the development of some severe complications such as the development of Autoimmune Diseases (AD), or the need for chemotherapy. Consequently, the identification of currently available clinical variables to assess the medical decisions in these CLL-related diagnosis problems is a key goal in the management of this disease. The development of AD or the need of CT is not known at diagnosis. So far, only with the evolution of the patient during the 5 years follow up, medical doctors can answer these questions. Therefore, the interest of the methodology presented herein consists in being able of predicting both CLL related problems at diagnosis. Particularly, AD problem was very hard to predict, and up to our knowledge no previous research was successful to explain this phenomenon using biochemical variables.

In this paper we show whether machine learning methods and clinical data obtained from a large population of well-studied CLL patients [6] can be efficiently applied to address these CLL diagnosis problems in medical practice by capturing the hidden implicit relationships between the clinical variables and the corresponding class of the different patients that have been established by medical experts. The use of machine learning techniques [9] in clinical medicine [10] and in cancer prediction and prognosis [11] is not new, and it has the advantage of treating more general prediction problems than survival analysis (usually treated through the Kaplan-Meier estimator) as supervised classification problems that admit more stable solutions than the corresponding regression problems.

The machine learning methodologies that are proposed in this paper are simple in their design and serve to provide to the physicians a simple and robust decision-making support system. Other more complex algorithms could be used, but the goal of this work is to obtain a simple decision rule and not to compare different learning algorithms. This manuscript is structured in three main parts. Firstly we provide an exhaustive explanation of the methods. Secondly, we present the results obtained for the two clinical CLL-related problems addressed herein: need of Chemotherapy Treatment (CT) and Autoimmune Disease development (AD). Finally, we provide coherent explanations and discussion of the findings.

## 2. Material and methods

A cohort of two hundred sixty-five Caucasians who were diagnosed in the Cabueñas Hospital (Gijón, Spain) with CLL between 1997 and 2009 were enrolled in this study. The population distribution by gender and age is the following: 154 are males and 111 are females, with ages ranging from 42 to 92, and 47 to 94 years old respectively. Clinical characteristics of patients including time for diagnosis to first treatment, need of Chemotherapy Treatment and appearance of autoimmune complications were also taken into account in this study. Additionally, thirty-six different clinical and biological variables were measured at diagnosis of the disease. Table 1 shows the variables description used in this study. Some variables reflect the malignant characteristic of leukemia cells; others measure the immunological characteristics of CLL patients, and some may be associated with the presence or development of autoimmune complications (autoimmune haemolytic anemia and immune-thrombocytopenia). Finally, some of the variables are demographic and biochemical. Most of them have a sampling frequency higher than 80%, however, the reticulocyte count (RET) and ZAP-70 are the ones that show the lowest sampling frequency. Particularly, ZAP-70 is only sampled in 21.9% of the patients (58 out of 265), showing that this popular CLL prognostic factor is not always available in medical practice. Although some of these variables were not at disposal at diagnosis (LD for instance), they have been used for analysis purposes. We provide the database as [supplementary material](#) (see “CLL.xls”).

**Table 1**

Clinical variables description by group and their corresponding symbols and sampling frequency (Samp. Freq.). Discrete variables are shown in bold faces.

Group	Variable name	Samp. freq. (%)
Biochemical	ALB – Albumin (g/L)	98.49
	ALC – Absolute Lymphocyte Count (cells/microL)	100.00
	ALP – Alkaline phosphatase (U/L)	95.47
	B2 M – Beta 2 Microglobulin (mg/L)	93.58
	BU – Bilirubin (mg/dL)	96.23
	CR – Creatinine (mg/dL)	99.62
	GOT – Glutamic-Oxaloacetic Transaminase (U/L)	98.11
	GPT – Glutamic-Pyruvic Transaminase (U/L)	99.25
	HGB – Hemoglobin (g/dL)	100.00
	IgA – Immunoglobulin A (g/L)	96.60
	IgG – Immunoglobulin G (g/L)	96.60
	IgM – Immunoglobulin M (g/L)	96.60
	K – Potassium (mEq/L)	90.94
	LDH – Lactate Dehydrogenase (U/L)	96.98
	MBC – Monoclonal B cell Count (cells/microL)	90.94
	MCV – Mean Corpuscular Volume (fl)	100.00
	NA (mEq/L)- Sodium	90.57
	NCC – Natural killer Cell Count (cells/microL)	90.94
	PLT – Platelets (cells/microL)	100.00
	RET – Reticulocyte count (cells/microL)	75.47
	SNC – Segmented Neutrophils Count (cells/microL)	100.00
	T8C – CD8 T cell Count (cells/microL)	86.42
	TLC – Total Lymphocyte Count, CD8 + CD4 (cells/microL)	96.60
	UA – Uric acid (mg/dL)	97.36
	UR – Urea (mg/dL)	99.25
	WBC – White Blood cells Count (cells/microL)	100.00
CLL Specific	<b>CD38</b> – CD38 positive	81.51
	<b>COOMBS</b> – Coombs test	94.34
	<b>LD</b> – Time for duplication of the number of lymphocytes	96.98
	<b>MOR</b> – Morphology	98.49
	<b>MP</b> – Monoclonal Peak	98.87
	<b>NLymph</b> – Number of affected lymph nodes	99.62
	<b>SMG</b> – Splenomegaly	99.62
	ZAP70 – Zeta-chain-associated protein kinase 70 (%)	21.89
Personal	AGE – Age	100.00
	<b>SEX</b> – Sex	100.00

The problems to be solved in this manuscript are the prediction of the need for Chemotherapy Treatment (CT) and the development of Autoimmune Disease (AD). Both classification problems are binary (two class classification problem). In our methodology we have explored the minimum-size list of prognostic variables (named as reduced base) having the highest predictive accuracy using different feature selection methods. The selected prognostic variables will be subsequently used for diagnosis and prognosis.

Fig. 1 shows the flowchart of the methodology, that includes 4 different steps:

### 2.1. Data preprocessing

Data preprocessing is applied to improve the quality of data used for performing feature selection, prediction and optimization. It includes two main sub steps that can be applied or not depending on their impact on the prediction:

- **Filtering:** All the features that were sampled less than a certain sampling frequency are removed. The filtering cut offs used were 30%, 40% and 50%.
- **Imputation:** This technique consists in interpolating all the missing values using a Nearest-Neighbor algorithm [12]. Given a partially-informed sample (with missing values) the algorithm finds the closest sample within the set of fully-informed

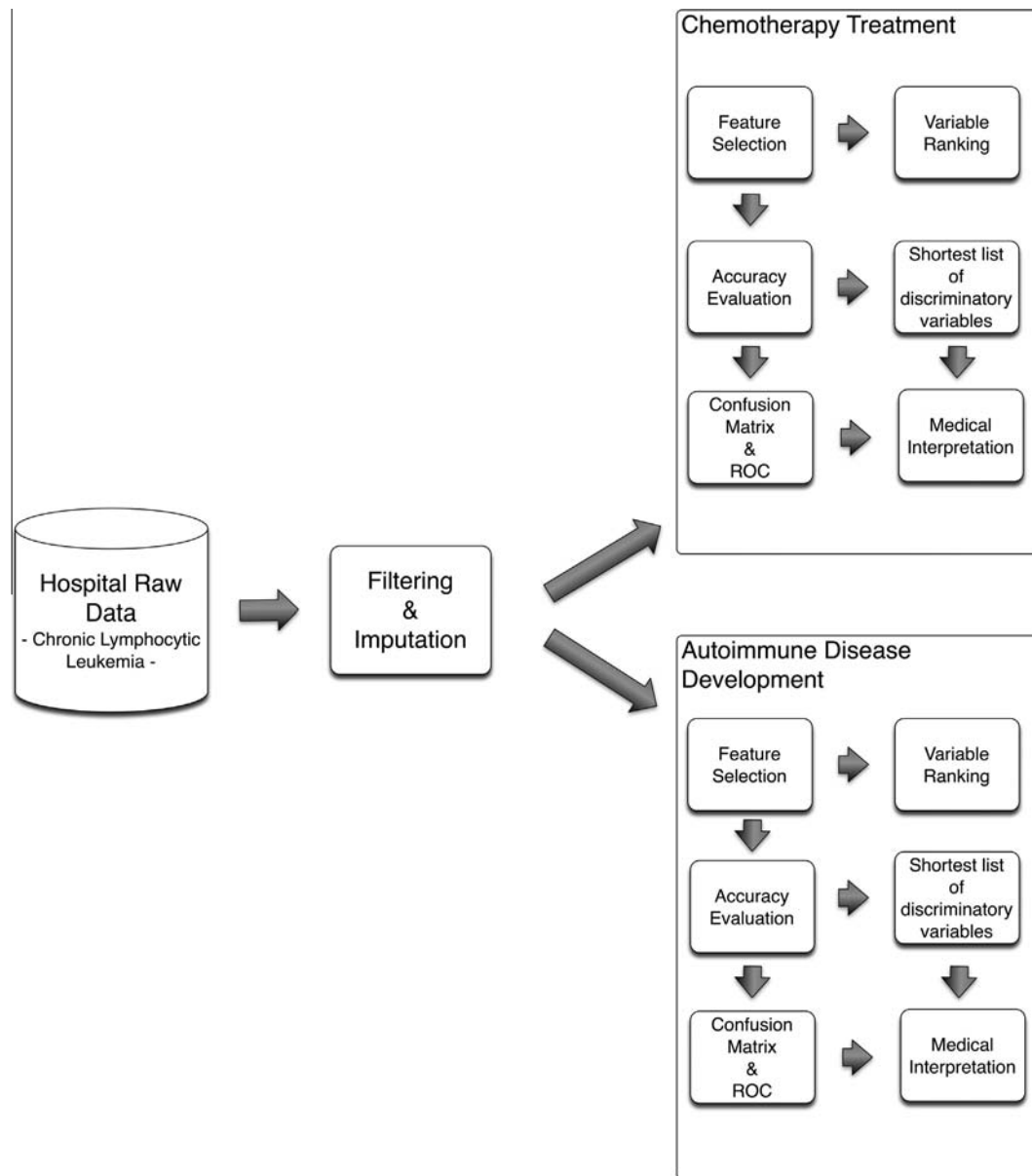


Fig. 1. Methodology flowchart.

samples and gives the values of the missing variables in this closest sample to the imputed sample. The similarity between samples is measured using the standard Euclidean dot product in  $N$ -dimensional vector spaces, where  $N$  is the number of fully-informed variables. This way of interpolation has the advantage of not introducing additional outliers that are not originally present in the dataset before imputation. Although the success of the different imputed algorithms might be data-driven, imputing the data improved the accuracy in the predictions and did not alter the prognostic variables that were involved providing shorter lists with higher discriminatory power.

## 2.2. Feature selection methods

**Maximum Fisher's ratio [13,14]:** The Fisher's ratio of an attribute  $j$ , in a two-class problem,  $c_1$ ,  $c_2$ , is defined as follows:

$$GFR_j(c_1, c_2) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2},$$

where,  $\mu_{j1}$ ,  $\mu_{j2}$  are measures of the center of the distribution (means) of gene  $j$  in classes 1 and 2, and  $\sigma_{j1}^2$ ,  $\sigma_{j2}^2$  are measures of the dispersion (variance) within these classes. This method looks for prognostic variables that separate the classes further apart and are very homogeneous within classes (low intra class variance).

**Minimum class Entropy [15,16]:** Entropy is a measure of the number of specific ways in which a system may be rearranged, and it is often considered a measure of disorder, or progression toward thermodynamic equilibrium. In the case of a binary classification problem, the entropy of each attribute is defined as follows:

$$E_j(c_1, c_2) = - \sum_{k=1}^2 \sum_{j=1}^{N_c} p_{kj} \log_2 p_{kj},$$

where  $N_c$  are the number of bins used to describe the probability distribution of attribute  $j$  in class  $k$ , and  $p_{kj}$  is the probability that this attribute takes the center class value  $x_{kj}$ . The algorithm to compute the entropy is based in ordering the variables according to their value and calculating the mismatch to the class vector. A perfect ordering occurs when the values correspond perfectly to the class vector. Variables with higher ordering (or lower entropy) are therefore the most discriminatory.

**Maximum Percentile Distance:** This feature selection method selects the attributes with higher distances between the corresponding cumulative probability functions (percentile array) within each class, defined for attribute  $j$  as follows:

$$d_j(c_1, c_2) = \frac{\|\mathbf{p}_{j1} - \mathbf{p}_{j2}\|_2}{\max(\|\mathbf{p}_{j1}\|_2, \|\mathbf{p}_{j2}\|_2)},$$

where  $\mathbf{p}_{ji}$  stands for the percentile vector  $j$  in class  $i$ , and  $\|\mathbf{p}_{ji}\|_2$  its Euclidean norm. Percentiles vary from 5 to 95 to avoid the possible effect of outliers [17]. This method can be considered as a generalization of a Mann–Whitney selection test, which is only based in the median (percentile 50).

The main reason for choosing these methods is due their clear interpretation, low computational cost, and the possibility of being applied to both, discrete and continuous variables. A survey about FS methods can be consulted in [18].

### 2.3. Accuracy evaluation

Once the most discriminatory variables are determined and ranked in decreasing order by their discriminatory power, the aim is to determine the shortest (having the smallest number of variables) list of prognostic variables with the highest predictive accuracy. The algorithm to find the minimum-size list of features is the Backwards Feature Elimination (BFE), which is similar to the Recursive Feature Elimination [19]. Feature elimination tries to unravel the existence of redundant or irrelevant features to yield the smallest set of prognostic variables that provide the greatest possible classification accuracy. Redundant features are those that provide no additional information than the currently selected features, while irrelevant features provide no useful information in any context.

The algorithm of BFE works as follows:

1. Beginning by the tail of the ranked list of prognostic variables, the algorithm iteratively generates increasingly shorter lists by eliminating one prognostic variable at a time, calculating their classification accuracy.
2. Finally, the list with the optimum accuracy and minimum size is therefore selected.

This way of proceeding is based on the following idea: prognostic variables with higher discriminatory ratios span low frequency features of the classification, while variables with lowest discriminatory ratios account for the details in the discrimination (high frequency features). This method determines the minimum amount of high frequency details that are needed to optimally discriminate between classes.

The predictive accuracy estimation is based on a Leave One Out Cross-Validation experiment (LOOCV), using the average distance of the reduced set of features to each training class set. The goal of cross-validation is to estimate how accurately a predictive model (classifier) will perform in practice. LOOCV involves using a single sample from the original dataset as the validation data (sample test), and the remaining samples as training data. The class assignment is based in a nearest-neighbor classifier in the reduced base, that is, the class with the minimum distance in

the reduced base to the sample test is assigned to the sample test. The average LOOCV predictive accuracy is calculated by iterating over all the samples using as metric the Euclidean distance between the corresponding normalized variables. For that purpose the weights used to normalize the variables are the inverse of two times the prior variability (standard deviation) of the prognostic variables. These weights serve to scale the different kinds of measurements into approximately the same range in order to give to each variable a similar influence on the overall distance measurement. The distance between a new sample  $\mathbf{s}_{new}$  and the average signature  $\mathbf{m}_j$  in class  $j$  is:

$$d(\mathbf{s}_{new}, \mathbf{m}_j) = \|W(\mathbf{s}_{new} - \mathbf{m}_j)\|_2,$$

with  $W$  is a diagonal matrix with  $W(k, k) = \frac{1}{2std(v_k)}$ , where  $std(v_k)$  is the standard deviation of the  $k$ th discriminatory prognostic variable.

In this procedure the feature selection method is executed only once using all training samples before estimating the accuracy by means of a leave-one-out procedure. For each new sample the classifier computes the average distance to the training samples of each class, being  $d_1$  the average distance to class 1, and  $d_2$  the average distance to class 2.

Based on these distances the probability of a new sample  $\mathbf{s}_{new}$  to be in class 1 can be written as:

$$P(\mathbf{s}_{new} \in c_1) = \frac{d_2}{d_1 + d_2}.$$

The procedure to decide the class assignment is as follows:

$$\mathbf{s}_{new} \in c_1 \text{ if } P(\mathbf{s}_{new} \in c_1) > p_{th} = 0.5.$$

Otherwise,  $\mathbf{s}_{new} \in c_2$ . The threshold probability ( $p_{th}$ ) can be considered as a continuous variable to establish the Receiver Operator Characteristic (ROC) curve for this classifier [20]. Finally, the reduced base might be tested over different randomly chosen training and testing dataset, and averaging the results over a set of independent simulations.

Although this simple classifier seems to be similar to a nearest neighbor algorithm (k-NN), it is not obviously the same, since neither the centroid definition of the distributions, nor the way of adopting the decisions coincide. Besides, we have testing k-NN nearest neighbor classifiers without success. Notice that in this process, the feature selection method is executed only once using all training samples, before estimating the accuracy by means of a leave-one-out procedure. Our goal is to study the effectiveness of feature selection methods in finding the groups of prognosis variables with higher predictive accuracy of these two CLL-related problems. Also, if the attribute selection process was performed each time the classifier was executed (i.e. in each of the folds of the leave-one-out), different sets of attributes would be obtained, thus, it would more difficult to assess the goodness of any concrete group of prognosis variables. The only way will be performing frequency analysis of the selected prognostic variables and applying BFE to this set of variables ranked by decreasing order of their posterior frequency. Besides, since the accuracy is established by Leave-One-Out Cross Validation (LOOCV) the selected attributes within each fold of the LOOCV would not be so different from selecting them using the whole dataset, considering that the training set of each of fold in a LOOCV is composed by all the samples but one. These facts have been confirmed through numerical experimentation.

### 2.4. ROC curves and risk assessment

In the previous step, maximizing the predictive accuracy according to the LOOCV criterion allowed to determine the best

reduced-base of prognostic variables. However, it is also important to analyze the structure of the confusion matrix, obtained from the set of predictions of the training set using the LOOCV method. The confusion matrix is composed by: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These concepts depend on how the classification problem is set up. From the confusion matrix we can calculate different rates that are very useful to understand the risk in the prediction:

- True Positive Rate or Sensitivity (TPR): measures the proportion of actual positives that are correctly predicted as such.
- True Negative Rate or Specificity (SPC): measures the proportion of negatives that are correctly predicted as such.
- Positive Predicted Value (PPV): is the proportion of positives values that are true positives.
- False Positive Rate (FPR): fraction of false positives out of the total actual negatives.
- False Negative Rate (FNR): fraction of false negatives out of the total actual positives.
- False Discovery Rate (FDR): fraction of false positives out of the total actual positives.

Based in these rates it is possible to construct a receiver operating characteristic curve (or ROC curve), which is a graphical plot that illustrates the performance of a binary classifier as a function of one parameter (the cut-off probability in this case). The curve is created by plotting the true positive rate or sensitivity (TPR) against the false positive rate (FPR) or fall-out. A perfect classifier has as ROC curve the step function at the origin. ROC analysis is related to cost/benefit analysis of diagnostic decision making (see for instance [17]).

The selected attributes are used to provide simple biomedical discriminatory rules for diagnosis and prognosis since for each classification problem we provide the bounds for the four groups of the confusion matrix. This knowledge can be used by the physicians in their decision-making process. Additionally to the LOOCV results, we also provide the mean accuracy obtained for 100 random holdouts 75/25 (75% for training and 25% for testing). In any case, and independently of how the predictive accuracy is established, it is crucially important to understand that there exist different combinations of prognostic variables with similar predictive accuracy whose knowledge might be useful to understand the genesis of the problem from a medical point of view. The existence of these different lists is related to the uncertainty analysis of the solutions in any decision-making problem [21,22].

Finally, the aim of this paper is not to compare different machine learning methods, but to introduce a simple methodology to select the shortest list of prognostic variables that could be easily interpreted by medical doctors to perform prognostic predictions with their corresponding risk assessment. However, we have compared this distance based nearest-neighbor algorithm to more sophisticated learning methods and the results did not improve or were clearly worse. The success of the methodology is not based on the sophistication of the classifier but on selecting the most discriminatory variables in each case and building the classifier based on these variables. By doing that it has been shown that the classification problem becomes linearly separable [23].

The methodology presented herein is easy to understand, since we avoid the use of black-box methodologies that provide estimations without MD's understanding, and has been successfully applied to predict response to treatment in Hodgkin lymphoma [17] using clinical data, and also in the prediction of risk of radiotherapy-related fatigue in prostate cancer patients using high dimensional expression data [24].

### 3. Results and discussion

#### 3.1. Chemotherapy Treatment assessment

As it was already mentioned CLL has a highly variable clinical course. Some patients have an indolent disease and they do not require CT. Other patients who present a progressive disease may require an intense CT. The identification of those patients at early stages of the disease with a high risk of rapid disease progression may help to significantly improve their prognosis. Thus, we try to establish the prognostic variables and criteria to assess the need for CT, assuming that the clinical decisions on the 71 (out of 259, therefore there are 6 missing values since the total cohort is 265) patients that have received CT were correct. The criteria for initiating CT were established in 2008 by the International Workshop on Chronic Lymphocytic Leukemia [25]. Particularly the presence of constitutional symptoms, such as, unintentional weight loss of 10% or more within the previous 6 month and significant fatigue or fevers or night sweats without other evidence of infection.

The Fisher's ratio method provided the minimum-size set of prognostic variables with the highest accuracy of 80.3%: B2M, WBC, ALC and MBC. Fig. 2 shows the ROC curve and the Recall (or True Positive Rate – TPR) against Precision (or Positive Predicted Value – PPV) curves for several probability thresholds in the CT classification problem. The optimum result ( $p_{th} = 0.47$ ) shows that 63.4% (TPR) of the patients that need CT and 86.7% (True Negative Rate or Specificity – SPC) of the patients that do not need CT were correctly predicted. Besides, with that probability threshold we got a Precision (or Positive Predicted Value – PPV) of 64.3%. Nevertheless, other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery Rate (FDR) was 36.62%. The confusion matrix is shown below:

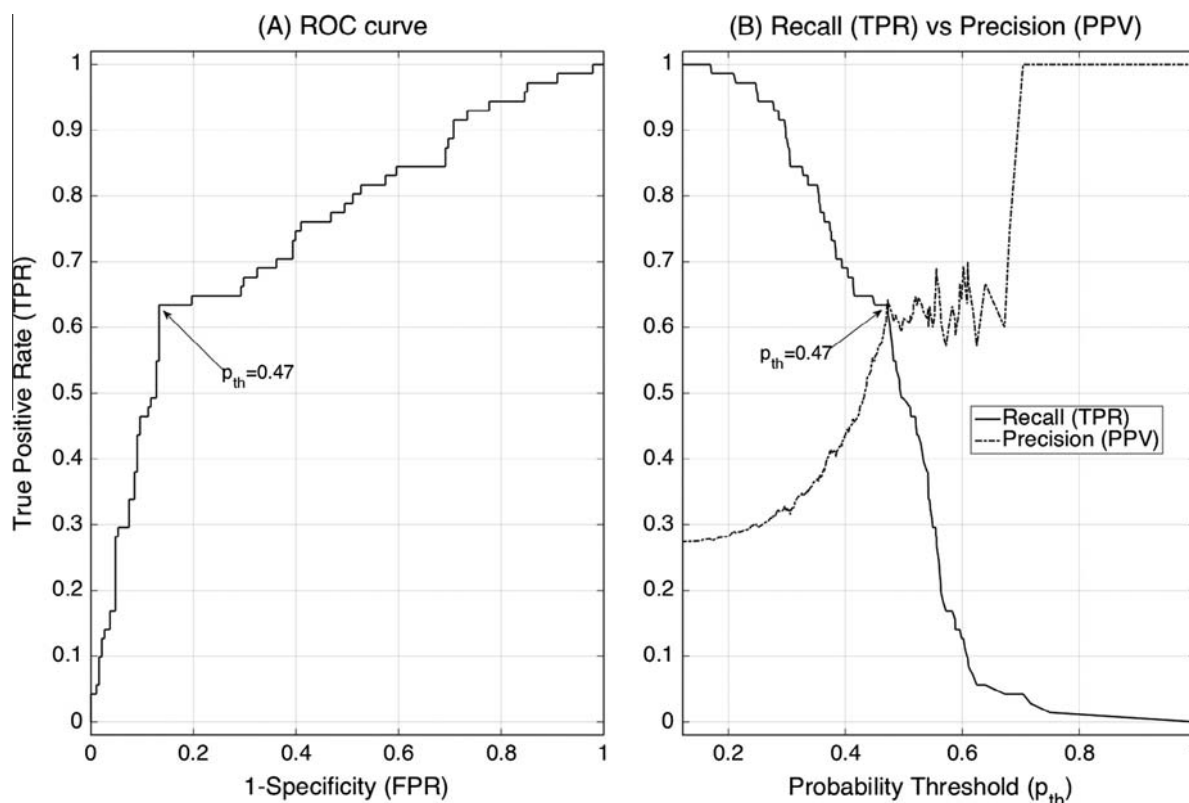
$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 45 & 25 \\ 26 & 163 \end{pmatrix}.$$

The True Positives (TP) are formed by the group of patients that need CT (+) and are correctly predicted, and the True Negatives (TN) are formed by the groups of patients that do not need CT (–) and are correctly predicted. Thus, False Positives (FP) are the patients that do not need CT (–) and are not correctly predicted and False Negative (FN) are the patients that need CT (+) and are not correctly predicted.

Besides, we have performed a two sample T-test and a Mann–Whitney U-test to clarify the differences between the TP and TN groups in the selected variables. The null hypothesis was rejected for all the prognostic variables. Therefore their statistical distributions should be considered to be different and the differences to be significant (see Appendix).

Additionally the Maximum Percentile Distance method also found a subset of variables with lower accuracy (78%): MBC, ALC, ZAP70, WBC and B2M. Moreover, the results using the minimum class Entropy were quite similar (76.8%): ZAP70, BU, WBC, ALC and MBC.

CT is recommended in patients with advanced and progressive disease. Thus, the amount of malignant leukemia cells that it is measured by the different counts of leucocytes; particularly WBC (White Blood Cells count), ALC (Absolute Lymphocyte Count) and MBC (Monoclonal B Cell Count) are key clinical parameters. Nevertheless, these variables are not currently used to select patients who may benefit from CT. On the other hand, AGE, B2M and ZAP70 are traditional clinical parameters that have demonstrated their prognostic importance independently of the clinical stage. Our results also indicated the great prognostic significance of other variables that are mainly related with the characteristics of the



**Fig. 2.** (A) ROC curve. (B) Sensitivity (or True Positive Rate – TPR) and Precision (or Positive Predicted Value – PPV) for Chemotherapy Treatment. The optimum result (TPR = 63.4 and PPV = 64.3) is obtained for  $p_{th} = 0.47$ .

immune system and are not currently used as prognostic markers in this disease. The fact that the prediction accuracy is barely above 80% means that these variables only contain partial information to establish the need of CT and/or to incorrect medical decisions that might input noise in the class assignment.

Table 2 shows the median/mean signatures for the 4 groups of the confusion matrix for the main decision variables found by this methodology. We can observe that there exists a significant distance between the mean signatures of the TP and TN groups, being the median/mean signatures in all the decision variables much higher in the TP group. Moreover, the distance between the median and the mean values of the decision variable distributions is much higher in the TP and in the FP groups, meaning a higher variability in these groups:

- The normal value of B2M is less than 2 mg/L [26]. Levels of B2M can be elevated in multiple myeloma and lymphoma. Besides, elevated values (>4 mg/L) are known to be an indicator of poor prognosis and survival [27]. In our case B2M is higher than this cut-off value (4.24) for the patients in the TP group.

**Table 2**

Chemotherapy Treatment. This table shows the list of most discriminatory continuous variables with a predictive accuracy of 80.3%. Median and mean values (median/mean) of the prognostic variables for the different groups of the confusion matrix are also given. Variables with (K) are expressed in kilo units. Bold faces indicate the highest value for each prognostic variable in the TP and TN groups. Bounds for the decision correspond to the TP and TN groups.

Variables	TP (median/mean)	TN	FP	FN
B2M	<b>3.9/4.24</b>	2.06/2.15	4.37/4.58	2.0/2.18
WBC (K)	<b>34.1/61.8</b>	14.3/16.8	18.3/28.3	14.2/15.5
ALC (K)	<b>24.7/47.6</b>	9.0/11.2	12.4/21.8	8.5/10.4
MBC (K)	<b>21.7/40.3</b>	6.1/8.4	10.1/18.4	6.9/7.8

- For the second decision variable, the normal value of WBC in the blood is 4.5–10.0 Kcells/microL. In our case the patients of the TN group have a mean WBC value (16.8 Kcells/microL) that exceeds four times the minimum normal value. Also the patients in the TP group show even higher mean WBC values (61.8 Kcells/microL).
- The reference range for the ALC is 4.5–11.0 Kcells/microL. It can be also observed that the ALC mean value in the TP group (47.6 Kcells/microL) exceeds 4 times the maximum normal value.
- Finally, the MBC is also very high (40.3 Kcells/microL) in the TP group compared to the TN group (8.4 Kcells/microL). The definition of CLL implies having a rate of CLL-phenotype B-cell lymphocytes higher than 5 Kcells/microL.

This analysis shows the typical profile of CLL patients with need of CT. The same tendencies are observed for the corresponding median values.

With respect to the analysis of the classification errors, the mean signatures of the FN group (patients that need CT and are incorrectly predicted) are very close to the mean signatures of the TN group. These patients will never be correctly predicted according to this classifier. The mean and median signatures of the FP group have the following singularities:

1. The mean B2M value (4.58 mg/L) is even higher than the corresponding B2M mean value in the TP group (4.24 mg/L). The same is observed for the median values.
2. Their mean WBC, ALC and MBC values are closer to the corresponding mean values of the TN group, exceeding in all the cases the mean values of the TN group. These differences are smaller in the case of the median values. These patients could be detected using only these three variables, not considering the value of B2M in these patients that is distorting the prediction.

Furthermore, to understand the ambiguity in the CT prediction, it should be taken into account that the criteria used to establish the need of CT [25] sometimes have not correlation with the biological data. The reason is that some patients are diagnosed in early stages of the disease when a low burden tumor mass has been detected but they have a very fast progression which implies the need of CT.

### 3.2. Autoimmune disease development

An Autoimmune Disease (AD) occurs when an adaptive immune response is mounted against self-antigen. In CLL, an autoimmune response against red blood cells (known as autoimmune haemolytic anemia), and an autoimmune response against platelets (known as immune thrombocytopenia) are severe complication of this disease. To the best of our knowledge no prognostic factors capable to predict the presence or development of an autoimmune disease in CLL patients have been currently disclosed. In our cohort only 16 patients (out of 263, therefore there are 2 missing values since the total cohort is 265) have shown autoimmune disorders. Therefore this classification problem, independently of the data sampling, is intrinsically highly unbalanced.

The shortest list of prognostic variables with the highest accuracy (97.3%) was found by the Fisher's ratio method and includes 13 clinical variables: PLT, RET, ALB, HGB, BU, UR, MCV, NCC, K, WBC, LDH, ALC and MBC. Furthermore, considering only the first nine attributes the predictive accuracy was 95.4%. Besides, only the two first attributes provided a predictive accuracy of 91%. Fig. 3 shows the ROC and the Recall (or True Positive Rate – TPR) against Precision (or Positive Predicted Value – PPV) curves throughout all possible probability thresholds for the AD classification problem. The optimum result ( $p_{th} = 0.5$ ) shows that 62.5%

(TPR) of the patients that have AD and 99.6% (True Negative Rate or Specificity – SPC) of the patients that do not have AD are correctly predicted. Moreover, over that probability threshold we get a Precision (or Positive Predicted Value – PPV) of 90.1%. However, other probability thresholds could be adopted depending on the Recall/Specificity balance, and therefore on the PPV as well. The False Discovery Rate (FDR) in this case is 9.1%. The confusion matrix is the following one:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 6 & 246 \end{pmatrix}.$$

The True Positives (TP) group is formed in this case by the patients that present AD (+) and are correctly predicted and True Negatives (TN) correspond to the patients that do not have AD (–) and are correctly predicted. Similarly, the False Positives (FP) are the patients that do not have AD (–) and are not correctly predicted and the False Negatives (FN) correspond to the patients that present AD (+) and are not correctly predicted. As in the previous section, we have performed the T-test and the Mann–Whitney U-test to analyze the differences between the TP and TN groups. The null hypothesis was rejected for all selected variables, except for K and LDH in the T-test (see Appendix).

Additionally the percentile distance method also found a subset of variables with 95.1% accuracy composed only by one prognostic variable: NCC. Entropy method also found a subset of 4 prognostic variables with 94.3% accuracy: TLC, T8C, NCC and MBC. PLT and RET, that were ranked in the first positions by the Fisher's Ratio, were found by the Entropy method in the fifth and sixth positions (TLC, T8C, NCC, MBC, RET and PLT), but the accuracy of this final subset was 93.2%.

PLT and RET appears in the first two positions of the FR list. They are responsible for most of the discriminatory power of the

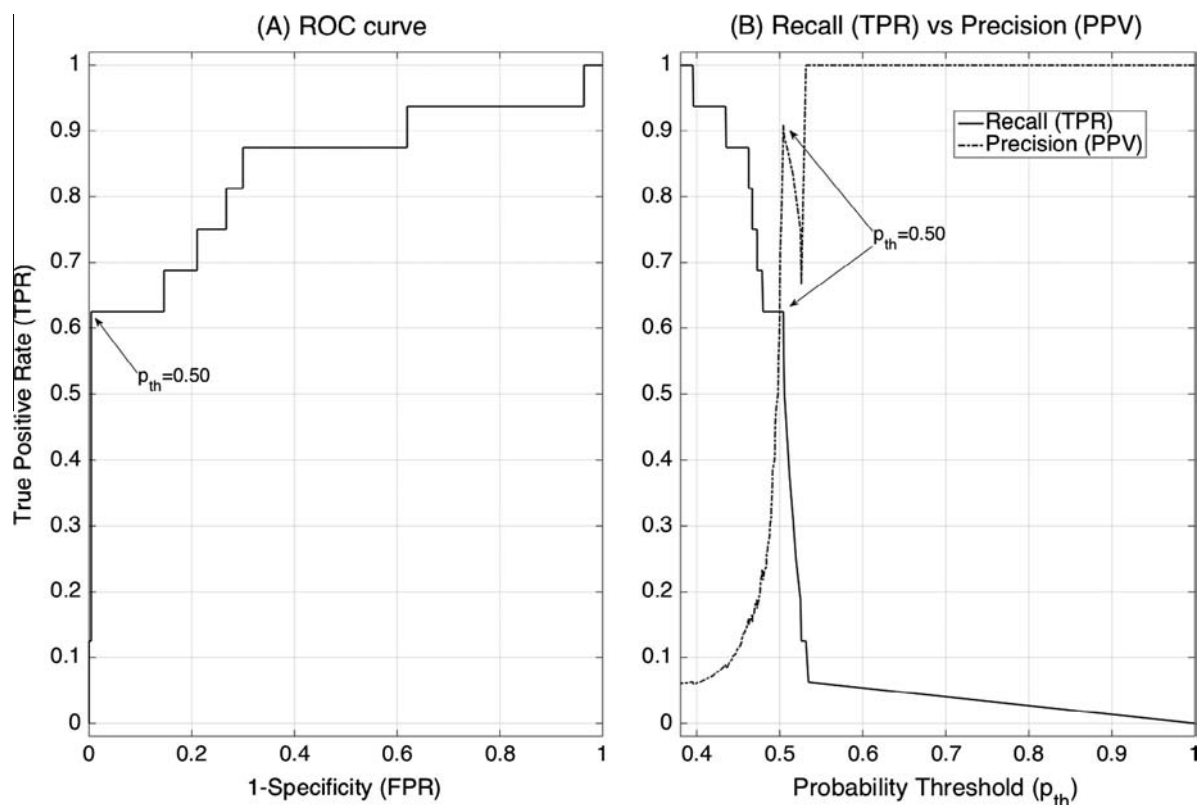


Fig. 3. (A) ROC curve. (B) Sensitivity (or True Positive Rate – TPR) and Precision (or Positive Predicted Value – PPV) for Autoimmune Disease occurrence. The optimum result (TPR = 62.5 and PPV = 90.1) is obtained for  $p_{th} = 0.5$ .

reduced base of features and the rest of variables span high frequency details in the classification. They also appear in the first positions of the list using Entropy method. It seems they could have an important role in the development of an autoimmune disease. Table 3 shows the medians and means for the 13 prognostic variables for the 4 groups of the confusion matrix. The differences between the means in TP and TN groups decrease with the Fisher's ratio. Prognostic variables with lower Fisher's ratios (secondary variables) also contribute to improve the discrimination. Except for the main variable, PLT, and the secondary variables HGB and K, the mean and median values are higher in the group with autoimmune disease (TP). The analysis of the two main prognostic variables shows that patients that develop AD and are correctly predicted (TP) have much lower medians and means PLT values (97.7/95.0 Kcells/microL). The normal platelet count lays in the range 150–450 Kcells/microL, being the average 237 Kcells/microL in men, and 266 in women. On the other hand, the reticulocyte count (RET) in the TP group almost doubles (136 Kcells/microL) the average RET count in patients with no AD (70 Kcells/microL). Median values also show similar tendencies.

The False Positives (FP group) is composed in this case only by 1 sample, whose signature is closer for all the 13 variables to the TP group, except for PLT, RET that are somewhere in between the median/mean values for TP and TN. This fact points out the difficulty of classifying this sample, and it can be concluded that it could be a 'biological' outlier. On the other hand, the FN group is composed by 6 samples. The mean PLT count (147 Kcells/microL) of the FN group lies between the mean value for the TP (95 Kcells/microL) and TN (202.2 Kcells/microL) groups. The RET count is however closer to the TN group showing a tendency to very low median values (54.4 Kcells/microL).

The percentile distance method found a subset of variables with 95.1% accuracy composed only by the Natural killer Cell Count

(NCC). The mean NCC value in the TP group (2251 cells/microL) is higher than in the TN (741 cells/microL) and FN (393 cells/microL) groups. Natural killer cells provide rapid responses to virally infected cells and respond to tumor formation. Therefore, this result suggests a possible link between AD development, viral infection and tumor progression. The percentile method also gives a great importance to IgM due to the higher values in the group of patients without AD (TN group with a mean of 1.12 g/L) with respect to the TP group (mean value of 0.36 g/L). This result is important since IgM is the first antibody to appear in response to initial exposure to antigens [28] and lower levels of this immunoglobulin is related to selective immunoglobulin M deficiency, which in turn is also related with autoimmune disorders like celiac disease or systemic lupus erythematosus [29].

Overall, these results show the importance of variables associated with the characteristics of platelets and red cells, which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia, such as PLT, HGB, MCV and RET. Other variables depend on the presence of autoantibodies (COOMBS) or products or symptoms derived from the lysis of blood cells (BU, LDH and SMG). Moreover, some variables associated with the immunological characteristics of patients, such as IgM, IgG, IgA, TLC, NCC and T8C, constitute a relevant subset of variables that may predict an autoimmune disease occurrence. The association of these variables with an autoimmune disease is not unexpected based on the biology of CLL, but we would like to highlight that no prognostic factors or system may currently predict the development of an autoimmune disease in the clinical practice. To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an autoimmune disease.

**Table 3**

Autoimmune disease development. This table shows the list of most discriminatory continuous variables with a predictive accuracy of 97.3%. Median and mean values of the prognostic variables for the different groups of the confusion matrix are also given. FP is composed only by 1 sample (median and mean coincides). Variables with (K) are expressed in kilo units. Bold faces indicate the highest value for each prognostic variable in the TP and TN groups. Bounds for the decision correspond to the TP and TN groups.

Variables	TP (median/mean)	TN	FP	FN
PLT (K)	97.7/95.0	<b>191/202.2</b>	138	163 /147.2
RET (K)	<b>128.0/135.7</b>	67.2/69.8	101.3	54.4 /71.8
ALB	<b>42.0/40.4</b>	38.0/37.4	41.1	39 /39.7
HGB	14.0/11.5	<b>14/13.6</b>	13.6	14/13
BU	<b>1.0/1.1</b>	1/0.6	0.6	1.0 /0.76
UR	<b>52.0/64.1</b>	43/46.7	49	42 /43.7
MCV	<b>93. 0/98.1</b>	90/89.6	88.9	87 /86.7
NCC	<b>966/2251</b>	576/741	1657	338 /393.4
K	4.0/4.09	<b>4.0/4.33</b>	4.0	4.0 /4.33
WBC (K)	<b>23.1/56.0</b>	15.4/24.7	23.6	13.5 /13.9
LDH	<b>360/398.1</b>	325/343.4	288	333/333
ALC (K)	<b>16.1/42.2</b>	10.1/ 17.8	18.4	8.5/6.7
MBC (K)	<b>10.2/36.3</b>	7.3/14.2	14.7	5.2/4.6

**Table 4**

Summary table. Summary table shows the Sensitivity or True Positive Rate (TPR) and Specificity or True Negative Rate (SPC) together with the mean accuracy (Acc.) for both experiments leave one out (LOO) and 100 repetitions of a hold-out 75/25 (HO, 75% for training and 25% for testing); and the positive and negative case description of each problem. Bold faces indicate the prognostic variables that have been discussed in the text.

Problem	Variables	TPR/SPC	LOO Acc.	HO-100 Acc.
CT(+) vs. No CT (-)	<b>B2M WBC ALC MBC</b>	63.4%/86.7%	80.30%	76.10%
AD (+) vs. No AD	<b>PLT RET</b> ALB HGB BU UR MCV <b>NCC</b> K WBC LDH ALC MBC	62.5%/99.6%	97.30%	92.80%

### 3.3. Summary of the results

Finally, Table 4 summarizes the main results found for both classification problems (CT and AD): the optimum reduced set of features, the LOOCV accuracy, the hold out (HO) mean accuracy over 100 different random simulations using 75% and 25% of samples for training and testing, the Sensitivity or True Positive Rate (TPR), and the Specificity or True Negative Rate (SPC) statistics. TPR and SPC values are important due to the impact on the patients of the decision taken by physicians.

It is possible to observe that:

1. The median accuracy of the predictions is quite stable with respect to the LOOCV accuracy.
2. The TPR/SPC statistics are optimally balanced in all the problems. The TPR/SPC statistics might be the target of a different optimization for the weights of the linear classifier depending on the risk that is given by the medical doctors to the False Positives (FP) and False Negatives (FN) diagnostic in each classification problem. This approach has been adopted to predict response to treatment in Hodgkin Lymphoma [17].

#### 4. Conclusions

Different prognostic factors are presented in this paper to predict two clinically important classification problems for CLL patients: Chemotherapy Treatment assessment and autoimmune disease development.

From the machine learning point of view, working imputed data produced better results in reliability (accuracy) than working with raw data. Fisher's ratio and percentile distance are the feature selection methods that produced the best biomarkers in terms of medical interpretability. The minimum-size of variables is established using BFE. The class prediction is based on a simple classifier, and its accuracy is determined by LOOCV experiment. The results show that the accuracies are rather high and the difference between both experiments LOOCV and 100 repetitions of a Hold Out (75/25) is quite low, which highlights the robustness of the methodology. In addition, risk assessment ROC curves are provided for each problem and show a good balance between False Positives and False Negatives.

From a medical point of view, machine learning methods allow the identification of clinical variables obtained at diagnosis of CLL patients, which may predict the development of AD and the need of CT. These variables are obtained at diagnosis of CLL patients on a regular basis, and consequently, their use does not increase the cost or complexity of the diagnosis in CLL patients. The need of CT seems to be related to the amount of malignant leukemia cells that are measured by the different leucocytes counts.

The best prognostic variables to predict the need of CT were B2M, WBC, ALC and MBC. Although the results concerning these prognostic variables are well known in other plasma disorders, this analysis served to conclude that these variables only carry partial information to adopt this important decision, that most of the times, is taken based on criteria that have not correlation with the biological data. To the best of our knowledge this is the first description so far that a group of clinical variables obtained at diagnosis of CLL patients may predict an occurrence of an AD, which is a severe and currently unpredictable complication of this disease. These results show the importance of variables associated with the characteristics of platelets, reticulocytes and natural killers (PLT, RET and NCC), which are the main targets of the autoimmune haemolytic anemia and immune thrombocytopenia. Additionally, machine learning methods focus on the relevance of some variables, such as the immunological ones, which may have an important impact on the prognosis of CLL patients, but they are not currently used by hematologists. Particularly, this analysis has shown that the low sampling frequency of RET and ZAP-70 could be troubling given their predictive significance in all the problems that have been treated: RET is a key factor for predicting AD, while ZAP-70 seems to be important for predicting the need of CT.

In conclusion, machine learning methods allow an accurate prediction of risk in CLL related problems. Additionally, they may establish the relevance of clinical variables that are not widely used as prognostic factor in this disease. The prognostic significance of these variables may probably reflect the relevance of some clinical aspects of this disease that are more important for prognosis than it is currently thought. This bioinformatics system can be easily applied in medical practice and updated along time through a simple computer program or excel spreadsheet (see [supplementary material file "CLL\\_predictor.xls"](#)).

#### Ethics statement

This study was approved by the Ethics Committee of Clinical Investigation of Principado de Asturias (date: 21th of January of 2009; n°1/2009). All the patients signed an informed consent to

participate in this study with the approval of the Ethics Committee. All studies were performed in accordance with the ethical standards of the Declaration of Helsinki and informed consent was obtained from all patients and controls.

#### Conflict of interest

No conflict of interest exists for the authors of this paper.

#### Authors' contributions

JLFM and EJAG prepared the data, designed the machine learning methodology, carried out the experiment, analyzed and interpreted the results and drafted the manuscript. OLR and JJC revised the design of the methodology critically, analyzed the results and drafted the manuscript. LHZ and AAH participated in the acquisition of the data, analyzed and interpreted the results and drafted the manuscript. SG and APG participated in the acquisition of the data, analyzed and interpreted the results, established main clinical conclusions and drafted the manuscript. All authors read and approved the final manuscript.

#### Acknowledgments

Enrique J. de Andrés was supported by the Spanish Ministerio de Economía y Competitividad (grant TIN2011-23558). The medical analysis was supported by the Fondo de Investigaciones Sanitarias. FEDER European Union (Instituto Carlos III-grant PI12/01280). No other financial support has been received to perform this retrospective analysis. We would like to acknowledge Dr. Stephen T. Sonis for his constructive review and suggestions that served to improve the translational approach shown in this manuscript.

#### Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jbi.2016.02.017>.

#### References

- [1] M. Hallek, L. Wanders, M. Ostwald, et al., Serum beta(2)-microglobulin and serum thymidine kinase are independent predictors of progression-free survival in chronic lymphocytic leukemia and immunocytoma, *Leuk. Lymphoma* 22 (5–6) (1996) 439–447.
- [2] T. Zenz, D. Mertens, H. Dohner, S. Stilgenbauer, Molecular diagnostics in chronic lymphocytic leukemia—pathogenetic and clinical implications, *Leuk. Lymphoma* 49 (5) (2008) 864–873.
- [3] T.J. Hamblin, Z. Davis, A. Gardiner, D.G. Oscier, F.K. Stevenson, Unmutated ig v (h) genes are associated with a more aggressive form of chronic lymphocytic leukemia, *Blood* 94 (6) (1999) 1848–1854.
- [4] M. Crespo, F. Bosch, N. Villamor, et al., Zap-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia, *N. Engl. J. Med.* 348 (18) (2003) 1764–1775.
- [5] R.N. Damle, T. Wasil, F. Fais, et al., IgV gene mutation status and cd38 expression as novel prognostic indicators in chronic lymphocytic leukemia, *Blood* 94 (6) (1999) 1840–1847.
- [6] A.P. Gonzalez-Rodriguez, J. Contesti, L. Huergo-Zapico, et al., Prognostic significance of CD8 and CD4 T cells in chronic lymphocytic leukemia, *Leuk. Lymphoma* 51 (10) (2010) 1829–1836.
- [7] K.R. Rai, A. Sawitsky, E.P. Cronkite, A.D. Chanana, R.N. Levy, B.S. Pasternack, Clinical staging of chronic lymphocytic leukemia, *Blood* 46 (2) (1975) 219–234.
- [8] J.L. Binet, A. Auquier, G. Dighiero, et al., A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis, *Cancer* 48 (1) (1981) 198–206.
- [9] J.D. Olden, J.J. Lawler, N.L. Poff, Machine learning methods without tears: a primer for ecologists, *Quart. Rev. Biol.* 83 (2) (2008) 171–193.
- [10] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, *Int. J. Med. Inform.* 77 (2) (2008) 81–97.
- [11] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, *Cancer Inform.* 2 (2006) 59–77.



- [12] O.G. Troyanskaya, M. Cantor, G. Sherlock, et al., Missing value estimation methods for DNA microarrays, *Bioinformatics* 17 (6) (2001) 520–525.
- [13] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7 (7) (1936) 179–188.
- [14] F. Yang, K. Mao, Robust feature selection for microarray data based on multicriterion fusion, *IEEE/ACM Trans. Comput. Biol. Bioinf.* 8 (4) (2011) 1080–1092.
- [15] C. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 27 (379–423) (1948) 623.
- [16] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1993.
- [17] E.J. deAndrés-Galiana, J.L. Fernández-Martínez, O. Luaces, et al., On the prediction of Hodgkin Lymphoma treatment response, *Clin. Transl. Oncol.* 17 (8) (2015) 612–619.
- [18] Y. Saeys, In Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics* 23 (19) (2007) 2507–2517.
- [19] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Mach. Learn.* 46 (1–3) (2002) 389–422.
- [20] J.A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum Associates, Mahwah, N.J., 1996.
- [21] J.L. Fernández-Martínez, Z. Fernández-Muñiz, M.J. Tompkins, On the topography of the cost functional in linear and nonlinear inverse problems, *Geophysics* 77 (1) (2012) 1–15.
- [22] J.L. Fernández-Martínez, Z. Fernández Muñiz, G. Pallero, L.M. Pedruelo González, From Thomas Bayes to Albert Tarantola. New insights to understand uncertainty in inverse problems from a deterministic point of view, *J. Appl. Geophys.* 98 (2013) 62–72.
- [23] J.L. Fernández-Martínez, E.J. deAndrés-Galiana, S. Sonis, Design of biomedical robots for the analysis of cancer, neurodegenerative and rare diseases, in: *Proceedings of the International Conference on Man-Machine Interactions*, vol. 391(4), 2015, pp. 29–44.
- [24] L. Saligan, J.L. Fernández-Martínez, E.J. deAndrés-Galiana, S. Sonis, Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer, *Cancer Inform.* 13 (141–152) (2014) 12.
- [25] M. Hallek, B.D. Cheson, D. Catovsky, et al., Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the international workshop on chronic lymphocytic leukemia updating the national cancer institute-working group 1996 guidelines, *Blood* 111 (12) (2008) 5446–5456.
- [26] M. Pignone, D. Nicoll, S.J. McPhee, *Pocket guide to diagnostic tests*, fourth ed., McGraw-Hill, New York, 2004.
- [27] N.C. Munshi, D.L. Longo, K.C. Anderson, Plasma cell disorders, in: J. Loscalzo, D. L. Longo, A.S. Fauci, L.K. Dennis, S.L. Hauser (Eds.), *Harrison's Principles of Internal Medicine*, McGraw-Hill Professional, 2011.
- [28] Houghton Mifflin Company, *Immunoglobulin M. The American Heritage Dictionary of the English Language*, Fourth ed., 2004.
- [29] L. Yel, S. Ramanuja, S. Gupta, Clinical and immunological features in IgM deficiency, *Int. Arch. Allergy Immunol.* 150 (3) (2009) 291–298.

## **A.2 On the prediction of Hodgkin Lymphoma treatment response**

Published in the Clinical & Translational Oncology.

DOI: <http://dx.doi.org/10.1007/s12094-015-1285-z>

## On the prediction of Hodgkin lymphoma treatment response

E. J. deAndrés-Galiana<sup>1,2</sup> · J. L. Fernández-Martínez<sup>1</sup> · O. Luaces<sup>2</sup> · J. J. del Coz<sup>2</sup> · R. Fernández<sup>3</sup> · J. Solano<sup>4</sup> · E. A. Nogués<sup>4</sup> · Y. Zanabilli<sup>5</sup> · J. M. Alonso<sup>6</sup> · A. R. Payer<sup>4</sup> · J. M. Vicente<sup>7</sup> · J. Medina<sup>5</sup> · F. Taboada<sup>8</sup> · M. Vargas<sup>9</sup> · C. Alarcón<sup>5</sup> · M. Morán<sup>5</sup> · A. González-Ordóñez<sup>5</sup> · M. A. Palicio<sup>9</sup> · S. Ortiz<sup>9</sup> · C. Chamorro<sup>10</sup> · S. Gonzalez<sup>11</sup> · A. P. González-Rodríguez<sup>4</sup>

Received: 11 December 2014 / Accepted: 20 March 2015  
© Federación de Sociedades Españolas de Oncología (FESEO) 2015

### Abstract

**Purpose** The cure rate in Hodgkin lymphoma is high, but the response along with treatment is still unpredictable and highly variable among patients. Detecting those patients who do not respond to treatment at early stages could bring improvements in their treatment. This research tries to identify the main biological prognostic variables currently gathered at diagnosis and design a simple machine learning methodology to help physicians improve the treatment response assessment.

**Methods** We carried out a retrospective analysis of the response to treatment of a cohort of 263 Caucasians who were diagnosed with Hodgkin lymphoma in Asturias (Spain). For that purpose, we used a list of 35 clinical and biological variables that are currently measured at diagnosis before any treatment begins. To establish the list of most discriminatory prognostic variables for treatment

response, we designed a machine learning approach based on two different feature selection methods (Fisher's ratio and maximum percentile distance) and backwards recursive feature elimination using a nearest-neighbor classifier (k-NN). The weights of the k-NN classifier were optimized using different terms of the confusion matrix (true- and false-positive rates) to minimize risk in the decisions.

**Results and conclusions** We found that the optimum strategy to predict treatment response in Hodgkin lymphoma consists in solving two different binary classification problems, discriminating first if the patient is in progressive disease; if not, then discerning among complete and partial remission. Serum ferritin turned to be the most discriminatory variable in predicting treatment response, followed by alanine aminotransferase and alkaline phosphatase. The importance of these prognostic variables suggests a close relationship between inflammation, iron overload, liver damage and the extension of the disease.

E. J. deAndrés-Galiana and J. L. Fernández-Martínez contributed equally to this study.

**Electronic supplementary material** The online version of this article (doi:10.1007/s12094-015-1285-z) contains supplementary material, which is available to authorized users.

✉ J. L. Fernández-Martínez  
jl\_fm@uniovi.es

- <sup>1</sup> Department of Mathematics, University of Oviedo, Oviedo, Spain
- <sup>2</sup> Artificial Intelligence Center, University of Oviedo, Oviedo, Spain
- <sup>3</sup> Hematology Department, Hospital de Cabueñes, Gijón, Spain
- <sup>4</sup> Hematology Department, Hospital Universitario Central de Asturias, Oviedo, Spain
- <sup>5</sup> Hematology Department, Hospital San Agustín, Avilés, Spain

**Keywords** Hodgkin lymphoma · Treatment response · Machine learning · Serum ferritin (SF) · Alanine aminotransferase (ALT) · Alkaline phosphatase (ALP)

- <sup>6</sup> Hematology Department, Hospital Valle del Nalon, Langreo, Spain
- <sup>7</sup> Hematology Department, Hospital de Mieres, Mieres, Spain
- <sup>8</sup> Hematology Department, Hospital Cangas de Narcea, Cangas de Narcea, Spain
- <sup>9</sup> Hematology Department, Hospital de Jarrio, Jarrio, Spain
- <sup>10</sup> Hematology Department, Hospital de Arriendas, Arriendas, Spain
- <sup>11</sup> Instituto Universitario Oncológico del Principado de Asturias (IUOPA), University of Oviedo, Oviedo, Spain

## Purpose

Lymphoma is the most common blood cancer and comprises two types: Hodgkin lymphoma (HL) and non-Hodgkin lymphoma. HL is characterized by the presence of the so-called malignant Reed–Sternberg cells, surrounded by an inflammatory infiltrate consisting of lymphocytes, neutrophils, eosinophils, plasma cells, macrophages and fibroblasts, constituting a model of interaction of tumor cells with their microenvironment. Components of inflammatory background are associated with classical HL: the presence of tumor-infiltrating lymphocytes is a negative prognostic factor for survival in these patients [1]. This kind of cancer is most commonly diagnosed in young adults between the ages of 15 and 35 years and in older adults over 50 years. The cure rate in HL patients is high, but the response along the treatment is still unpredictable and varies from patient to patient. Besides, a small minority is resistant or relapses before treatment. Detecting those patients with a poor prognosis at early stages (diagnosis) could bring improvements in their treatment and prognosis.

There was an international effort to identify the prognostic factors to accurately predict the development and treatment of HL, mainly in patients with advanced stage. The adverse prognostic factors identified were: male older than 45 years, stage IV disease, hemoglobin lower than 10.5 g/dl, lymphocyte count lower than 600/ $\mu$ l (or less than 8 %), albumin lower than 4.0 g/dl and white blood count greater than 15,000/ $\mu$ l [2, 3]. Other studies also took into account mixed cellularity or lymphocyte-depleted histologies, the presence of B symptoms or high erythrocyte sedimentation rate and bulky disease as adverse prognostic factors [4, 5]. Moreover, disease extensions measured by computed tomography (CT) and early response to treatment measured by positron emission tomography (PET) have demonstrated a powerful prognostic ability [6, 7].

Several research works highlighted the importance of the identification of prognostic variables to predict patients who will suffer relapse and the adaptation of treatments to individual risks [8–11]. Particularly, the result of treatment optimization provoked some criteria modification, with the disappearance of some factors that were considered to be of poor prognosis and with the proposal of new ones that allowed establishing groups with differing risks of relapse and different treatments.

In this manuscript, we inferred prognostic variables for HL treatment response using clinical data and machine learning techniques in a retrospective study of a cohort of 263 Caucasians. For this purpose, we designed a methodology to find the shortest list of clinical variables providing the highest predictive accuracy for Hodgkin lymphoma first-line treatment response (at diagnosis). We found that the best way of addressing this problem is to proceed in two steps: comparing

first the complete/partial remission hypothesis against progressive disease hypothesis, and secondly differentiating between complete and partial remission in case it proceeds. Serum ferritin (SF) turned to be the most important prognostic variable, achieving cross-validation predictive accuracies higher than 90 %. Ferritin concentrations increase drastically in the presence of an infection or cancer [12]. Our study also showed the importance of alanine aminotransferase (ALT) and alkaline phosphatase (ALP). The normal ranges for these three prognostic factors are provided in Table S1 (see Supplementary Material). The importance of these variables in the treatment response suggests a close relationship to iron overload, liver damage and bone affection. An adequate staging of newly diagnosed patients using this methodology will enable optimal treatment planning, which is particularly important in health care to find an optimum balance between treatment efficacy and drug toxicity.

## Methods

The present research work is a retrospective study of a cohort of 263 Caucasians who were diagnosed with classical Hodgkin lymphoma in Asturias (Spain) and enrolled in this study between 2002 and 2012. This study was approved by the institutional review boards of the different hospitals involved and performed in accordance with the Helsinki Declaration of 1975. Besides, this study was approved by the Ethics Committee of the Principado de Asturias (date: 17th of January; Project 6th number 13).

Staging definitions from the German Hodgkin Study Group (GHSG) were evaluated in this analysis. All patients were treated with the ABVD (doxorubicin, bleomycin, vinblastine, dacarbazine) regimen: 125 patients (47 %) received involved field radiotherapy, 91 with early stage and 34 with advanced stage disease. Response to therapy was evaluated by physical and radiographic evaluation, including computed tomography (CT) and the follow-up of the patients. In the last 5 years, PET scan was also included to assess treatment response. The treatment response was divided into three categories according to international standards [13]: 237 of the patients were in complete remission (CR), 17 in partial remission (PR) and only in 9 cases the disease progressed without any relevant change. This last category was named as progressive disease (PD). Table 1 describes the main characteristics of the patients: age, sex, stage at diagnosis, percentage of early favorable and early unfavorable and percentage of advanced disease depending on Hasenclever Prognostic Score.

Progression-free survival (PFS) was calculated from the date of diagnosis to the date of progression, relapse or death by of any cause. Overall survival (OS) was calculated from the date of diagnosis to the date of death from any cause or last follow-up. Overall and progression-free

**Table 1** Main characteristics of the patients (number of patients/percentage), including Hasenclever International Prognostic Score (IPS)

Age
Median: 37
Males range: 9–82
Females range: 10–83
Sex
Males: 171/65 %
Females: 92/35 %
Stage at diagnosis
Stage I: 42/16 %
Stage II: 92/35 %
Stage III: 82/31 %
Stage IV: 47/18 %
Early disease: 113/43 %
Favorable: 57/22 %
Not favorable: 56/21 %
Advanced disease: 150/57 %
IPS ≤2: 81/31 %
IPS >2: 69/26 %

survival distribution curves were estimated using the product-limit method of Kaplan–Meier. The median PFS and OS for the entire group were, respectively, 150 and 160 months. The probabilities of PFS and OS at 7 years were 57 and 76 %, correspondingly.

Thirty-five clinical and biological variables were measured at diagnosis and before treatment. These variables were classified into five groups: biochemical, immunohistochemical, Hodgkin lymphoma specific, treatment specific and host information. Table 2 shows the description of all these variables, boldfacing those that take discrete predefined values. Most of the variables had a sampling frequency higher than 90 %. However, others were scarcely sampled, such as CRP (14 %), immunoglobulins and Ki67 (20 %). The need of imputing/filtering those variables has turned out to be a very important step in the modeling process.

The problem addressed in this manuscript consists in building an efficient and simple machine methodology to predict HL first-line treatment response with the highest predictive accuracy, and at the same time minimizing risk in the decisions. For that purpose we have used the response criteria defined by Cheson et al. [13]. Patients were divided into three categories: complete remission (CR), defined as the disappearance of all evidence of disease; partial remission (PR) defined as regression of measurable disease and no new sites; and progressive disease (PD) defined as any new lesion or increase by 50 % of previously involved sites. Four different classification problems were performed to find the optimum way of separating these three classes.

**Table 2** Variable description gathered into five groups

Biochemical	
WBC	White blood cells count (10 <sup>6</sup> /μL)
ALC	Absolute lymphocyte count (10 <sup>6</sup> /μL)
AMC	Absolute monocyte count (10 <sup>6</sup> /μL)
AEC	Absolute eosinophil count (10 <sup>6</sup> /μL)
HGB	Hemoglobin (g/dL)
PLT	Platelets (10 <sup>3</sup> /μL)
ALB	Albumin (g/L)
AST	Aspartate aminotransferase (U/L)
ALT	Alanine aminotransferase (U/L)
ALP	Alkaline phosphatase (U/L)
CR	Creatinine (mg/dL)
LDH	Lactate dehydrogenase (U/L)
ESR	Erythrocyte sedimentation rate (mm/h)
CRP	C-reactive protein (mg/L)
GG	Gamma globulin (g/L)
IgG	Immunoglobulin G (g/L)
IgA	Immunoglobulin A (g/L)
IgM	Immunoglobulin M (g/L)
B2M	Beta-2 microglobulin (mg/L)
Cu	Copper (mEq/L)
SF	Serum ferritin (ng/mL)
Immuno-histochemical tests	
<b>CD20</b>	<b>B-lymphocyte antigen CD20 test: positive or negative</b>
<b>Ki67</b>	<b>Ki-67 cellular marker for proliferation: positive or negative</b>
<b>EBV</b>	<b>Epstein–Barr virus presence: positive or negative</b>
HL specific	
OS	Overall survival from diagnosis to death (days)
<b>Stage</b>	<b>Ann Arbor staging: I, II, III and IV</b>
<b>SS</b>	<b>Signs and symptoms: A = no SS, B = fever, weight loss, anomalous night sweats</b>
ALA	Affected lymph areas
<b>LMM</b>	<b>Large mediastinal mass: more than 1/3 of the thoracic diameter</b>
<b>ELI</b>	<b>Extraganglionic involvement</b>
<b>Bulky</b>	Mediastinal mass more than 10 cm
Treatment	
<b>CHEMO</b>	<b>Chemotherapy treatment</b>
<b>RT</b>	<b>Radiotherapy treatment</b>
Personal	
AGE	Age
<b>SEX</b>	<b>Sex</b>

Discrete variables are boldfaced

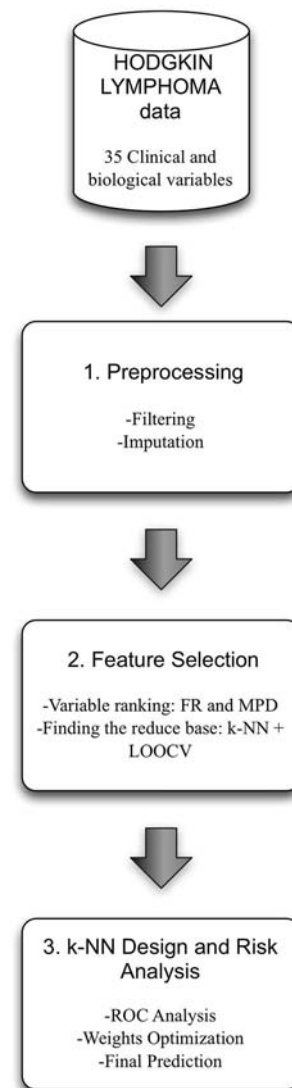
The machine learning methodology is explained in Appendix 1 (see Supplementary Material) and is composed of three main steps: (1) pre-processing, (2) feature selection and (3) k-NN design and risk analysis.

Basically, the learning method consists, according to the parsimony principle, in finding the shortest subset of most discriminatory clinical variables (also called the reduced base of prognostic variables) to predict treatment response in HL patients. The clinical and biological variables are first ranked according to two different filter methods: Fisher's ratio (FR) and maximum percentile distance (MPD). In a second step, the predictive accuracy of the different ranked lists of prognostic variables is established by leave-one-out-cross-validation (LOOCV) experiment using a simple k-nearest-neighbor (k-NN) classifier (Appendix 1 in Supplementary Material). This methodology has been successfully applied to predict risk of radiotherapy-related fatigue in prostate cancer patients using high-dimensional expression data [14]. In this case, the challenge is not related to the dimension of the dataset, but to the heterogeneous degree of sampling of the different clinical variables. Besides, in this case the methodology incorporates the weight optimization of the k-NN classifier according to the receiver operating characteristic (ROC) curve to improve risk decision-making, that is, to provide a very high predictive accuracy with an optimum balance between the different rates of the confusion matrix (the true-positive and false-positive rates defining the corresponding ROC curve). Figure 1 shows a flow diagram explaining the methodology.

Finally, we would like to point out that the aim of this work is not to numerically compare different machine learning methods, but to introduce simple algorithms to select the shortest list of prognostic variables that could be easily interpreted by medical doctors, to improve the patient prognostic in HL treatment response, with its corresponding risk assessment. Particularly, we tried to avoid the use of black boxes that provide estimations without medical doctors' understanding. As a matter of fact, this methodology can be easily implemented in any platform such as a spreadsheet (see Supplementary Material—HL treatmentResponse\_Predictor.xls-, and the corresponding explanation provided in Appendix 2). That said, the classifier that is proposed in this paper outperformed other more sophisticated classifiers that are proposed in the machine learning literature, highlighting the importance of selecting the correct prognostic variables.

## Results

Treatment response in HL is a difficult prediction problem. Aside from plasma EBV DNA [15], there is no predictive biomarker to predict the patient's response to the corresponding treatment with a reliable accuracy. This classification problem is intrinsically highly unbalanced, mainly due to the discrete sampling of the samples (number of patients) and also because a high percentage of the patients are cured from this kind of malignancy.



**Fig. 1** Flow diagram for HL treatment response prediction model. The methodology is composed of three steps: 1 filtering and imputing data, 2 feature selection and 3 k-NN design and risk analysis. In each box, the different substeps are also detailed

The first modeling decision was to transform the analysis of treatment response into a binary classification problem (two-class problem) that admits a more reliable and stable solution than the corresponding value regression problem, that is, it is easier to predict if a patient is in complete or partial remission than predicting the value of the biological variables related to this fact. Besides, the prediction in binary classification problems allows for risk assessment through the analysis of the confusion matrix and the receiving operating characteristic (ROC) curve. The confusion matrix consists of four different groups: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), whose definition depends on how the classification problem has been set up. From the confusion

matrix, different rates can be calculated to understand the risk on the prediction:

1. True-positive rate or sensitivity (TPR): measures the proportion of actual positives that are correctly predicted as such.
2. True-negative rate or specificity (SPC): measures the proportion of negatives that are correctly predicted as such.
3. False-positive rate (FPR): fraction of false positives out of the total actual negatives.
4. False-negative rate (FNR): fraction of false negatives out of the total actual positives.
5. False discovery rate (FDR): fraction of false positives out of the total actual positives.

These rates could be used by the physicians in their decision-making process. A perfect classifier would have 100 % sensitivity and specificity.

The following comparisons were performed:

1. CR vs. PR+PD,
2. CR+PR vs. PD and CR vs. PR,
3. CR vs. PR vs. PD.

Comments for the prognostic variables in comparisons 1 and 3 are given in Appendix 3 (see Supplementary Material), since we have obtained worse results. The most

effective comparison was the second one and it was composed of two main steps. In the first step (2.1 CR+PR vs. PD), we established the differences between patients who experienced partial or complete remission (CR+PR, positive class) from those in which the disease progressed without any relevant change (PD, negative class). Then, a second comparison (2.2 CR vs. PR) was used to establish the differences between CR (positive class) and PR (negative class) patients.

The best result was obtained by filtering out those variables having a sampling frequency lower than 30 %, and imputing the rest. Besides, MPD (maximum percentile distance) provided the shortest list of variables with the highest predictive accuracy.

Table 3 shows the confusion matrix rates (TPR, TNR, FPR, FNR) for all the binary classifications (comparisons 1 and 2), together with the false discovery rate (FDR) and the LOOCV predictive accuracy (ACC). No weight optimization was performed in this case, that is, the weights corresponded to the inverse of the prior variability of the prognostic variables (see Appendix 1 Supplementary Material). Table 4 shows the mean values of the three prognostic variables for the different groups of the confusion matrix and the weights ( $W$ ) used to define the distance criterion in the nearest-neighbor classifier.

**Table 3** Best results for all the comparisons obtained without weight optimization

#	Comparisons	Base	MPD rate	TPR (%)	TNR (%)	FPR (%)	FNR (%)	FDR (%)	Acc (%)
1	CR (+) vs. PR and PD (-)	SF	67.2818	97.89	19.23	80.77	2.11	8.30	90.11
		GPT	41.0086						
2.1	CR and PR (+) vs. PD (-)	SF	75.2264	98.43	22.22	77.78	1.57	2.72	95.82
2.2	CR (+) vs. PR (-)	SF	57.7157	97.89	11.76	88.24	2.11	6.07	92.13
		ALT	41.3166						
		ALP	38.9228						

The algorithm used for all the comparisons was the same: filtering 30 % of sampling frequency, imputing and MPD as feature selection method. Rate is the maximum percentile distance rate, *TPR* true-positive rate, *TNR* true-negative rate, *FPR* false-positive rate, *FNR* false-negative rate and *Acc* final accuracy of the prediction. Signs (+) and (-) represent the positive and negatives groups, respectively

**Table 4** Mean values (for all the comparisons) of the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and weights ( $\omega_b$ ) for the optimum NN classifier (without weight optimization)

#	Comparisons	Variables	TP	TN	FP	FN	$\omega_b$
1	CR (+) vs. PR and PD (-)	SF	246.7	<b>2480.2</b>	391.8	2280.6	0.0005
		GPT	23.9	<b>52.0</b>	38.0	65.2	0.0239
2.1	CR and PR (+) vs. PD (-)	SF	266.4	<b>3288.0</b>	452.4	3231.3	0.0005
2.2	CR (+) vs. PR (-)	SF	249.9	<b>2401.0</b>	405.5	2131	0.0005
		ALT	<b>23.7</b>	18.0	44.2	74.4	0.0092
		ALP	116.8	<b>376.0</b>	163.5	608.4	0.0017

Signs (+) and (-) represent positive and negatives groups, respectively. Boldfaces indicate the highest value for each prognostic variable. Normal bounds for the decision correspond to the TP and TN groups.  $W_b$  are the weights used in the classifier for data variability normalization (before weight optimization)

### Comparison 2.1: CR+PR vs. PD

In this comparison serum ferritin was the single selected prognostic variable (using MPD as feature selection) with a predictive accuracy of 95.82 %. The SF mean value in the TN group (3288 ng/mL) was even higher than in the previous comparison, the SF value in the TP group being 266 ng/mL. Therefore, patients with progressive disease show a very clear inflammatory behavior as shown by the SF value at diagnosis. The TPR of this comparison is very high (98.43 %), and the TNR is higher than in the comparison 1 (22.22 %). The TP and FP maximum and minimum SF values are closer to normal SF values (see Table S1 of Supplementary Material). Conversely, the TN and FN corresponding SF signatures are extreme values.

### Comparison 2.2: CR vs. PR

The best subset of prognostic variables for this case was found by MPD and was composed of SF, ALT and ALP, providing 92 % of LOOCV predictive accuracy. The TPR is very high (97.89 %) and the TNR is quite low (11.76 %), that is, the difference between partial and complete remission is very hard to tell, and the classifier tends to assign the complete remission class in most of the cases. There is a big gap between SF mean levels of both TP (249 ng/mL) and TN (2401 ng/mL) groups. Moreover, FP (405 ng/mL) and FN (2131 ng/mL) mean SF values are similar to the mean SF values of the TN and TP groups, respectively. The same happens with ALP; there is also a big difference between TP (116.8 U/L) and TN (376 U/L) values. The mean values in the FP (163.5 U/L) and FN (608.4 U/L) groups are also close to the TP and TN groups, which make those samples very difficult to correctly predict using this k-NN classifier. SF and ALP have higher mean values in the TN group than in the TP group. However, in the case of ALT, the mean value in the TP group (23.7 U/L) is higher than in the TN group (18 U/L). Moreover, the difference between these two groups is very

low. The ALT mean value in the FN group (74.4 U/L) is closer to the TP group, instead of being closer to the TN group, as it should be expected. This is due to the presence of some PR patients with anomalously large ALT values.

### k-NN weight optimization

Optimization of the weights of the k-NN classifier via Particle Swarm Optimization (PSO) was performed to improve the true negative rate (or specificity), that is, increasing TNR while the overall accuracy is also improved (TPR is not affected). Details about PSO are given in Appendix 1 (see Supplementary Material).

Table 5 shows the TPR, TNR, FPR, FNR, FDR and predictive accuracy (Acc) obtained after weight optimization. TN rates were improved around 10 % in comparisons 2.1, while in comparison 2.2 TP rate was improved around 1 %. The overall accuracy was improved in all the cases around 1 %. Table 6 shows the mean values for TP, TN, FP, FN and the optimized weights for the prognostic variables ( $\omega_a$ ). It can be observed that values of the weights increased after optimization for all the prognostic variables. Therefore, it is possible to improve the quality of the prediction and minimize risk on the decisions, by optimizing the weights that are initially provided by the distance criterion.

### Conclusions

In this paper we presented an optimum strategy to predict treatment response in HL. Three main discriminatory prognostic variables were used in this analysis: serum ferritin, ALT and ALP.

Serum ferritin has been frequently used as a surrogate marker for systemic iron stores, but may be also elevated in specific circumstances without excess iron stores, such as in inflammation, correlating closely to the activity of malignant lymphomas. Serum ferritin levels have been reported to be elevated in HL patients, in particular in

**Table 5** Best results for all the comparisons obtained after weight optimization

#	Comparisons	Base	Rate	TPR (%)	TNR (%)	FPR (%)	FNR (%)	FDR (%)	Acc (%)
1	CR (+) vs. PR and PD (-)	SF	67.2818	98.73	23.08	76.92	1.27	7.87	91.2548
		GPT	41.0086						
2.1	CR and PR (+) vs. PD (-)	SF	75.2264	98.43	33.33	66.67	1.57	2.34	96.1977
2.2	CR (+) vs. PR (-)	SF	57.7157	99.58	11.76	88.24	0.42	5.98	93.7008
		ALT	41.3166						
		ALP	38.9228						

Rate is the maximum percentile distance rate, *TPR* true-positive rate, *TNR* true-negative rate, *FPR* false-positive rate, *FNR* false-negative rate and *Acc* final accuracy of the prediction. Signs (+) and (-) represent the positive and negatives groups, respectively



**Table 6** Mean values of the true positives, true negatives, false positives and false negatives and optimized weights ( $\omega_a$ ) of the optimum NN classifier after weight optimization, for all the comparisons

#	Comparisons	Base	TP	TN	FP	FN	$\omega_a$
1	CR (+) vs. PR and PD (-)	SF	254.7	<b>2309.3</b>	338.6	3007.0	0.0016
		GPT	24.4	<b>99.2</b>	23.2	55.3	0.0319
2.1	CR and PR (+) vs. PD (-)	SF	275.4	<b>2796.7</b>	225.5	2669.5	0.0020
2.2	CR (+) vs. PR (-)	SF	276.7	<b>2401.0</b>	405.5	3330.0	0.0026
		ALT	<b>24.3</b>	18.0	44.2	140.0	0.0663
		ALP	123.2	<b>376.0</b>	163.5	1059.0	0.0051

Signs (+) and (-) represent the positive and negatives groups, respectively. Boldfaces indicate the highest value for each prognostic variable. Normal bounds for the decision correspond to the TP and TN groups

advanced stages and during disease progression [16, 17]. Moreover, it has been proposed that the release of IL-6 stimulates the overproduction of hepcidin in the liver, which correlates with the iron restriction and contributes to anemia in HL [18]. In addition, the abundant microenvironment surrounding the neoplastic Hodgkin's and Reed-Sternberg cells may contribute to alterations in iron metabolism [19]. Besides, serum ferritin concentration closely follows the activity of malignant lymphomas [20]. Another research work [21] has shown that levels of serum ferritin higher than 500 ng/mL are an important marker for predicting poor survival outcomes for non-Hodgkin lymphoma. Nevertheless, and to our knowledge, serum ferritin levels have not been yet related to the treatment response of HL patients.

Serum activity levels of ALT enzyme are routinely used as a biomarker of liver injury caused by drug toxicity, infection, alcohol and steatosis. ALT plays a key role in the intermediary metabolism of glucose and amino acids and also participates in cellular nitrogen metabolism and liver gluconeogenesis. This cytosolic enzyme catalyzes the transfer of the  $\alpha$ -amino group from alanine to  $\alpha$ -ketoglutaric acid. Serum levels of ALT are normally low (10–40 U/L), but any type of liver cell injury may modestly increase the ALT levels. Levels greater than 500 U/L occur most often in people with hepatic diseases, such as viral hepatitis, ischemic liver injury (shock liver), toxin-induced liver damage and tumor infiltration of liver. Despite the association between greatly elevated ALT levels and hepatocellular diseases, the levels of ALT do not correlate with the extent of liver cell damage [22].

The alkaline phosphatase test (ALP) is used to detect liver disease or bone disorders. In conditions affecting the liver, damaged liver cells release increased amounts of ALP into the blood. Further, any condition that affects bone growth or causes increased activity of bone cells can affect ALP levels in the blood. In non-Hodgkin lymphomas, ALP is increased in patients with bone marrow disorders [23], thus reaching stage IV and worse prognosis. A recent study suggests that ALP together with gamma-

glutamyl transferase and albumin may define advanced stages of HL [24]. Moreover, bone affection is also associated with a high progression degree (HR: 1.96) [25]. However, in a patient with fever of unknown origin (FUO), highly elevated alkaline phosphatase and normal/slightly elevated serum transaminase levels suggest the possibility of lymphoma [26–28].

Overall, the results of this study show that the combined use of these prognostic variables, SF, ALT and ALP, in a simple classifier allows predicting first-line treatment response in HL patients with high accuracy and confirms a close relationship between treatment response in HL, inflammation, iron overload and liver and bone damage. Particularly, the combination of feature selection methods (maximum percentile distance), risk assessment analysis (ROC curve) and global optimization (PSO) provides biomarker discovery that is easily implemented in spreadsheet.

To conclude, detecting those HL patients who do not respond to the treatment at early stages may help improve their treatment. This study proposed a new prognostic analysis method, based on mathematical models that identify three simple prognostic variables currently gathered at diagnosis that may help detect with high accuracy those HL patients with bad prognosis without any additional cost.

**Acknowledgments** Enrique J. de Andrés was supported by the Spanish Ministerio de Economía y Competitividad (Grant TIN2011-23558), and the medical analysis was supported by the Fondo de Investigaciones Sanitarias (Instituto Carlos III-Grant PI12/01280). No other financial support has been received to perform this retrospective analysis.

**Conflict of interest** None.

## References

- Alvaro-Naranjo T, Lejeune M, Salvado-Usach MT, Bosch-Princep R, Reverter-Branchat G, Jaen-Martinez J, et al. Tumor-infiltrating cells as a prognostic factor in Hodgkin's lymphoma: a quantitative tissue microarray study in a large retrospective cohort of 267 patients. *Leuk Lymphoma*. 2005;46(11):1581–91.

2. Schreck S, Friebel D, Buettner M, Distel L, Grabenbauer G, Young LS, et al. Prognostic impact of tumour-infiltrating th2 and regulatory t cells in classical Hodgkin lymphoma. *Hematol Oncol*. 2009;27(1):31–9.
3. Hasenclever D, Diehl V, Armitage JO, Assouline D, Björkholm M, Brusamolino E, et al. A prognostic score for advanced Hodgkin's disease. *New Eng J Med*. 1998;339(21):1506–14.
4. Friedman S, Henry-Amar M, Cosset JM, Carde P, Hayat M, Dupouy N, et al. Evolution of erythrocyte sedimentation rate as predictor of early relapse in posttherapy early-stage Hodgkin's disease. *J Clin Oncol*. 1988;6(4):596–602.
5. Mauch P, Larson D, Osteen R, Silver B, Yeap B, Canellos G, et al. Prognostic factors for positive surgical staging in patients with Hodgkin's disease. *J Clin Oncol*. 1990;8(2):257–65.
6. Cheson BD. New staging and response criteria for non-Hodgkin lymphoma and Hodgkin lymphoma. *Radiol Clin North Am*. 2008;46(2):213–23.
7. Biggi A, Gallamini A, Chauvie S, Hutchings M, Kostakoglu L, Gregianin M, et al. International validation study for interim pet in abvd-treated, advanced-stage Hodgkin lymphoma: interpretation criteria and concordance rate among reviewers. *J Nucl Med*. 2013;54(5):683–90.
8. Smolewski P, Robak T, Krykowski E, Blasinska-Morawiec M, Niewiadomska H, Pluzanska A, et al. Prognostic factors in Hodgkin's disease: multivariate analysis of 327 patients from a single institution. *Clin Cancer Res*. 2000;6(3):1150–60.
9. Zander T, Wiedenmann S, Wolf J. Prognostic factors in Hodgkin's lymphoma. *Ann Oncol*. 2002;13(Suppl 1):67–74.
10. Josting A. Prognostic factors in Hodgkin lymphoma. *Expert Rev Hematol*. 2010;3(5):583–92.
11. Provencio M, Espana P, Millan I, Yebra M, Sanchez AC, de la Torre A, et al. Prognostic factors in Hodgkin's disease. *Leuk Lymphoma*. 2004;45(6):1133–9.
12. Ong DST, Wang L, Zhu Y, Ho B, Ding JL. The response of ferritin to lps and acute phase of pseudomonas infection. *J Endotoxin Res*. 2005;11(5):267–80.
13. Cheson BD. The international harmonization project for response criteria in lymphoma clinical trials. *Hematol Oncol Clin North Am*. 2007;21(5):841–54.
14. Saligan L, Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Inform*. 2014;13(141–152):12.
15. Gandhi MK, Lambley E, Burrows J, Dua U, Elliott S, Shaw PJ, et al. Plasma epstein-barr virus (EBV) DNA is a biomarker for EBV-positive Hodgkin's lymphoma. *Clin Cancer Res*. 2006;12(2):460–4.
16. Bezwoda WR, Derman DP, Bothwell TH, Baynes R, Hesdorffer C, MacPhail AP. Serum ferritin and Hodgkin's disease. *Scand J Haematol*. 1985;35:505–10.
17. Dörner MH, Abel U, Fritze D, Manke HG, Drings P. Serum ferritin in relation to the course of Hodgkin's disease. *Cancer*. 1983;52:2308–12.
18. Hohaus S, Massini G, Giachelia M, Vannata B, Bozzoli V, Cuccaro A, et al. Anemia in Hodgkin's lymphoma: the role of interleukin-6 and hepcidin. *J Clin Oncol*. 2010;28(15):2538–43.
19. Hohaus S, Giachelia M, Cuccaro A, Voso MT, Leone G. Iron in Hodgkin's lymphoma. *Crit Rev Oncog*. 2013;18(5):463–9.
20. Aulbert E, Steffens O. Serum ferritin—a tumor marker in malignant lymphomas? *Onkologie*. 1990;13(2):102–8.
21. Yoh KA, Lee HS, Park LC, Lee EM, Shin SH, Park DJ, et al. The prognostic significance of elevated levels of serum ferritin before chemotherapy in patients with non-Hodgkin lymphoma. *Clin Lymphoma Myeloma Leuk*. 2014;14(1):43–9.
22. Kaplan MM. Alanine aminotransferase levels: what's normal? *Ann Intern Med*. 2002;137(1):49–51.
23. Kittivorapart J, Chinthammitr Y. Incidence and risk factors of bone marrow involvement by non-Hodgkin lymphoma. *J Med Assoc Thai*. 2011;94(Suppl 1):S239–45.
24. Jamakovic M, Baljic R. Significance of copper level in serum and routine laboratory parameters in estimation of outspreading of Hodgkin's lymphoma. *Med Arch*. 2013;67(3):185–7.
25. El-Galaly TC, Hutchings M, Mylam KJ, Brown Pde N, Bukh A, Johnsen HE, et al. Impact of <sup>18</sup>F-fluorodeoxyglucose positron emission tomography/computed tomography staging in newly diagnosed classical Hodgkin lymphoma: fewer cases with stage I disease and more with skeletal involvement. *Leuk Lymphoma*. 2014;55(10):2349–55.
26. Brensilver HL, Kaplan MM. Significance of elevated liver alkaline phosphatase in serum. *Gastroenterology*. 1975;68(6):1556–62.
27. Brinckmeyer LM, Skovsgaard T, Thiede T, Vesterager L, Nissen NI. The liver in Hodgkin's disease. Clinico-pathological relations. *Eur J Cancer Clin Oncol*. 1982;18(5):421–8.
28. Cunha BA. Fever of unknown origin (FUO): diagnostic importance of serum ferritin levels. *Scand J Infect Dis*. 2007;39(6–7):651–2.

## **A.3 Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy - Related Fatigue in Patients with Prostate Cancer**

Published in the journal Cancer Informatics.

DOI: <http://dx.doi.org/10.4137/CIN.S19745>

## Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy-Related Fatigue in Patients with Prostate Cancer

Leorey N. Saligan<sup>1</sup>, Juan Luis Fernández-Martínez<sup>2</sup>, Enrique J. deAndrés-Galiana<sup>2</sup> and Stephen Sonis<sup>3</sup>

<sup>1</sup>National Institute of Nursing Research, National Institutes of Health, Bethesda, Maryland, USA. <sup>2</sup>Universidad de Oviedo, Spain. <sup>3</sup>Biomodels, LLC, Watertown, MA, USA.

### ABSTRACT

**BACKGROUND:** Fatigue is a common side effect of cancer (CA) treatment. We used a novel analytical method to identify and validate a specific gene cluster that is predictive of fatigue risk in prostate cancer patients (PCP) treated with radiotherapy (RT).

**METHODS:** A total of 44 PCP were categorized into high-fatigue (HF) and low-fatigue (LF) cohorts based on fatigue score change from baseline to RT completion. Fold-change differential and Fisher's linear discriminant analyses (LDA) from 27 subjects with gene expression data at baseline and RT completion generated a reduced base of most discriminatory genes (learning phase). A nearest-neighbor risk (k-NN) prediction model was developed based on small-scale prognostic signatures. The predictive model validity was tested in another 17 subjects using baseline gene expression data (validation phase).

**RESULT:** The model generated in the learning phase predicted HF classification at RT completion in the validation phase with 76.5% accuracy.

**CONCLUSION:** The results suggest that a novel analytical algorithm that incorporates fold-change differential analysis, LDA, and a k-NN may have applicability in predicting regimen-related toxicity in cancer patients with high reliability, if we take into account these results and the limited amount of data that we had at disposal. It is expected that the accuracy will be improved by increasing data sampling in the learning phase.

**KEYWORDS:** cancer-related fatigue, radiation therapy, prostate cancer, Fisher's linear discriminant analysis (LDA), k-NN backward recursive feature elimination

**CITATION:** Saligan et al. Supervised Classification by Filter Methods and Recursive Feature Elimination Predicts Risk of Radiotherapy-Related Fatigue in Patients with Prostate Cancer. *Cancer Informatics* 2014;13:141–152 doi: 10.4137/CIN.S19745.

**RECEIVED:** August 27, 2014. **RESUBMITTED:** October 23, 2014. **ACCEPTED FOR PUBLICATION:** October 23, 2014.

**ACADEMIC EDITOR:** JT Efrid, Editor in Chief

**TYPE:** Methodology

**FUNDING:** This study was supported by the Division of Intramural Research of the National Institute of Nursing Research of the NIH, Bethesda, Maryland, and Biomodels, LLC, Watertown, MA. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.

**COPYRIGHT:** © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

**CORRESPONDENCE:** ssonis@biomodels.com

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

### Introduction

Fatigue is the most common, troublesome, and costly side effect of many cancer (CA) treatment regimens. Not only does it impact patients directly, but it also has significant repercussions on both direct and indirect health economic outcomes.<sup>1</sup> CA treatment-related fatigue (CTRF) is defined as a “subjective sense of tiredness” that persists over time, interferes with activities of daily living, and is not relieved by adequate rest.<sup>2,3</sup> The majority of CTRF studies are

associated with chemotherapy regimens; however, fatigue during and after external beam radiation therapy (RT) is common, increasing in severity during treatment and persisting after RT has been completed.<sup>4</sup> CTRF has been reported to be the most distressing symptom reported by patients with non-metastatic prostate CA who receive RT with the greatest negative impact on daily activity, physical well-being/function, and relationships with significant others.<sup>5</sup> The trajectory of CTRF is still being defined. During RT, fatigue



intensification peaks at midpoint, declines after completion of RT,<sup>6</sup> and becomes chronic in a subpopulation of patients. The pathobiology of CTRF, like other toxicities, is complex and is probably attributable to a cascade of events resulting in radiation-induced pro-inflammatory cytokine production, hypothalamic–pituitary–adrenal (HPA) activation dysfunction, and neuromuscular function abnormalities.<sup>7,8</sup>

CTRF, like other regimen-related toxicities, does not occur in every patient, but rather in a subpopulation of at-risk individuals. In the context of individualizing care, the ability to predict CTRF risk has the potential to help guide treatment choices for patients and providers. There have been a number of attempts to predict CTRF. For example, one study reported that elevated pre-treatment fatigue, anxiety, and a specific breast cancer diagnosis (eg, ductal carcinoma in situ, invasive lobular carcinoma) predicted CTRF during RT in early stage breast cancer.<sup>9</sup> Another study found dyspnea, pain, lack of appetite, feeling drowsy, feeling sad, and feeling irritable to be forecasters of CTRF among hematology–oncology patients.<sup>10</sup>

However, as it becomes increasingly clear that CTRF is strongly related to a series of underlying genetically controlled biological events, the utility of identifying a group of genes that impact patients' risk of the condition seems compelling. We hypothesized that radiation-associated fatigue risk, like other regimen-related toxicities, is determined not by a single gene, but rather a synergistically functioning group of genes. This theory is supported by the finding that clusters of SNPs, discovered by Bayesian network analysis, have been reported to be associated with CTRF risk in patients being treated with cycled chemotherapy for breast and colorectal cancers.<sup>11,12</sup> In the current study, we evaluated an alternative analytical method in which genes were identified using a series of hierarchical filters and nearest-neighbor (NN) analysis to identify a group of genes that predicted CTRF in men being irradiated for prostate cancer. This proof-of-concept investigation not only demonstrated the utility of the analysis, but also confirmed the observation that focal radiation therapy is capable of inducing gene expression changes in peripheral white blood cell RNA.<sup>13</sup>

## Methods

**Patients.** This study (NCT00852111) was approved by the Institutional Review Board of the National Institutes of Health (NIH), Bethesda, Maryland, USA. The study involving human participants is in compliance with the Declaration of Helsinki. Men who were 18 years or older, diagnosed with non-metastatic prostate cancer with or without a history of prostatectomy, and scheduled to receive EBRT with or without concurrent androgen deprivation therapy (ADT), were enrolled. Men with progressive disease causing significant fatigue, those with psychiatric disease within the past five years, uncorrected hypothyroidism and anemia, taking sedatives, steroids, and non-steroidal anti-inflammatory

agents, and those with second malignancies, were excluded. Patients were recruited at the Magnuson Clinical Research Center, NIH, between May 2009 and September 2011. Subjects signed written informed consents prior to study participation.

**Fatigue assessment instruments.** Clinical and demographic data (eg, age, race, stage of prostate cancer, EBRT dose, type of EBRT technique used, and laboratory values) were obtained from chart review. Questionnaires were completed at baseline (prior to RT) and at completion of RT (day 38–42 after EBRT initiation). To avoid extraneous influences on their responses, subjects completed the questionnaires in an outpatient setting before clinical procedures were provided.

The 13-item Functional Assessment of Cancer Therapy–Fatigue (FACT-F), a frequently used, validated, reliable, stand-alone measure of fatigue in cancer therapy with coefficient alphas in the mid-90s, was used.<sup>14</sup> FACT-F is scored from 0–52, the higher the score, the lower the fatigue symptoms. A greater than three-point decrease in the FACT-F score is considered to be a minimally important change that is clinically relevant.<sup>15</sup> To optimize the phenotypic characterization of the study participants, subjects were categorized into high-fatigue (HF) or low-fatigue (LF) groups based on their change in FACT-F scores from baseline to completion of EBRT. HF subjects had a decrease of three or more points in FACT-F scores, and those who had less than a three-point decrease in FACT-F scores between both time points were categorized in the LF group. Depressive symptoms were also assessed using the 21-item Hamilton Depression Rating Scale (HAM-D), a clinician-administered questionnaire with good psychometric properties.<sup>16</sup>

**Biological sample collection, RNA extraction, and microarray experiments.** Peripheral blood (2.5 mL) was collected at baseline and on the last day of RT, immediately after FACT-F was administered, from each subject using PAXgene™ Blood RNA tubes (Qiagen, Frederick, Maryland, USA) containing red blood cell lysis buffer and a RNA-stabilizing solution and stored at –80 °C until RNA extraction. Total RNA was extracted using the PAXgene™ Blood RNA system (Qiagen, Frederick, Maryland, USA) according to manufacturer's instructions. The quantity of total RNA was measured by a spectrophotometer at an optical density of 260 nm. RNA quality was assessed using the RNA 6000 Nano LabChip® on a Bioanalyzer Agilent 2100 (Agilent Technologies, Palo Alto, CA, USA). RNA purification, cDNA and cRNA synthesis, amplification, hybridization, scanning, and data analyses were conducted by one laboratory technician following standard protocols as previously described.<sup>17</sup> Affymetrix microarray chips (HG-U133 Plus 2.0, Santa Clara, California, USA) were used for gene expression analysis. The Affymetrix HG-U133 Plus 2.0 microarray chip is comprised of 47,000 transcripts, including 38,000 well-characterized human genes.

Affymetrix GeneChip Command Console (AGCC, 3.0 V) was used to scan the images for data acquisition. Affymetrix raw data were acquired using comparison expression analysis of GCOS Software to yield CHP files according to the user instructions. Peripheral blood has been previously utilized to describe gene expression signature that predicted radiation-related toxicities.<sup>18</sup>

Ingenuity Pathway analysis (Ingenuity® Systems, www.ingenuity.com, Redwood City, California, USA) identified the functional networks of the differentially expressed probe sets from Ingenuity's Knowledge Base. Right-tailed Fisher's exact test was used to calculate the *P*-values determining the probability that each biological function and/or disease assigned to these networks is due to chance alone. The one-tailed analysis was used to reduce the random chances of over-representation of focused genes in the relevant pathways.<sup>19</sup>

**Statistical rationale.** Descriptive analyses were used to assess the demographic characteristics of the sample. Paired *t*-tests were used to compare fatigue scores and clinical variables between time points. To facilitate the identification of a group of synergistically functioning genes that were associated with CTRF risk, we used an approach that optimized an initial supervised component with a subsequent statistically driven hierarchical ranking. Using microarray data from the training set of patients for which the presence or absence of CRTF was known, we identified the genes that most discriminated between individuals who developed CTRF from those who did not. Those genes were then ranked according to their discriminatory value (as defined by their Fisher's ratio [FR]), in which the predictive accuracy of the different-ordered reduced sets was determined using a backward recursive feature elimination algorithm (see flow diagram in Fig. 1 below). This procedure serves to eliminate redundant or irrelevant genes (features) to yield the most precise set of genes with the greatest predictive accuracy.

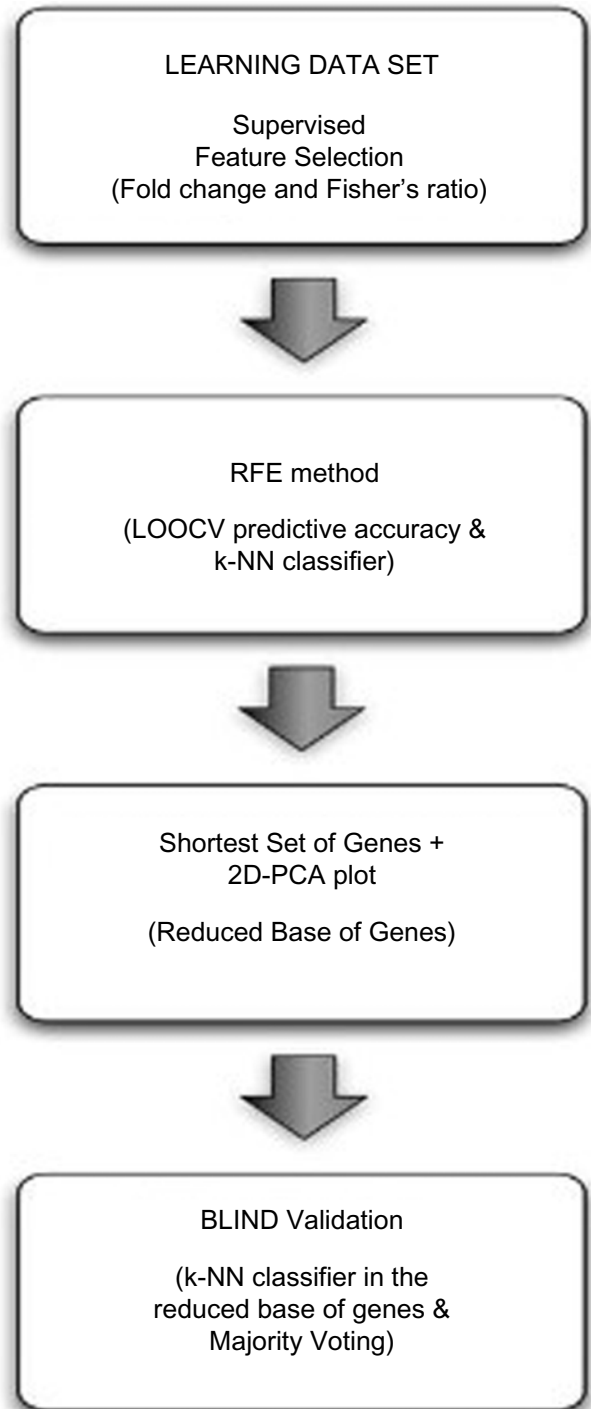
**Feature selection (gene ranking).** Feature selection identified genes with the highest fold change,<sup>20,32</sup>  $fc_j(c_1, c_2)$ , and FR,<sup>21</sup>  $FR_j(c_1, c_2)$ , using the phenotype information. The fold change and the FR for probe *j* in a binary classification problem are defined as follows:

$$fc_j(c_1, c_2) = \log_2 \frac{\mu_{j1}}{\mu_{j2}}, \quad (1)$$

$$FR_j(c_1, c_2) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \quad (2)$$

where  $\mu_{j1}, \mu_{j2}$  are measures of the center of the distribution (means) of gene *j* in classes 1 and 2, and  $\sigma_{j1}^2, \sigma_{j2}^2$  are measures of the dispersion (variance) within these classes.

The following relationship holds:



**Figure 1.** Flow diagram for the radiation-related fatigue prediction model. The methodology is composed of 4 steps: feature selection, backward recursive feature elimination, small-scale separability analysis and blind validation.

$$FR_j = k^2 > 1 \Leftrightarrow |\mu_{j1} - \mu_{j2}| = k \cdot \sqrt{\sigma_{j1}^2 + \sigma_{j2}^2}, \quad (3)$$

that is,

$$|\mu_{j1} - \mu_{j2}| = k \cdot \sigma_j^T, \quad (4)$$



where  $|\mu_{j1} - \mu_{j2}|$  is the distance between the centers of the classes, and  $\sigma_j^T = \sqrt{\sigma_{j1}^2 + \sigma_{j2}^2}$  is the total variance of the gene  $j$  in both classes.

The above relationship means that the centers of the distribution are further apart the distance,  $k \cdot \sigma_j^T$ . Also, taking into account that  $\mu_{j1} = 2^{fc_j} \cdot \mu_{j2}$ , then

$$|\mu_{j1} - \mu_{j2}| = k \cdot \sigma_j^T \Rightarrow \mu_{j2} = \frac{\sqrt{FR_j}}{|2^{fc_j} - 1|} \sigma_j^T. \quad (5)$$

This last relationship implies that given a gene characterized by its FR,  $FR_j$ , and fold-change value,  $fc_j$ , only the most discriminatory genes with means  $\mu_{j1}, \mu_{j2}$  and dispersions  $\sigma_{j1}, \sigma_{j2}$  in both classes are selected by this procedure.

**Identification/selection of the smallest and most precise set of CTRF-associated genes.** We used the following algorithm to select the smallest and most precise set of discriminatory genes for the LF/HF phenotype:

1. Genes identified by feature selection (see above) were ranked in decreasing order according to their FR value.
2. The predictive accuracy of the different sets was iteratively calculated after the sequential elimination of the genes with lowest FR. We termed this novel algorithm, a modification of the technique described by Guyon et al (2002),<sup>22</sup> "backward recursive feature elimination." It served to determine the number of helper genes (genes with the lowest FR) needed to maximize the Leave-One-Out-Cross-Validation (LOOCV) predictive accuracy,<sup>23</sup> in a procedure similar to the Fourier decomposition of a signal into a sum of harmonics of increasing frequency.<sup>24</sup> Genes with lower FR provide high frequency details for the discrimination. This procedure yielded the shortest gene set that predicted fatigue risk association with optimum accuracy (most precise). Other sets with similar and lower accuracy were also determined by this procedure and were of value, because these sets were also considered as noise buffers; as the classifier with the highest learning accuracy might not be the one that generalizes (predicts correctly unseen samples) better. This approach is appropriate and is especially helpful in designing small-scale signatures that were able to predict HF/LF with a high degree of accuracy.

The linear separability of the phenotype in the reduced set of genes that is determined in step 2 was checked by performing principal component analysis (PCA) of the learning dataset expressed in this small-scale signature and projecting these samples in the corresponding 2D PCA space. Then, the LF/HF phenotype becomes linearly separable by reducing the dimension to the list of most discriminatory genes, if both populations (HF and LF) can be linearly separated by a given hyper-plane.

3. The accuracy estimation was based on the LOOCV method, using the average Euclidean distance on the reduced set of features to each training class set. The goal of cross-validation was to estimate how accurately a predictive model (classifier) will perform in practice. This procedure, applied to the training dataset, is supervised because the phenotype information of the patients was needed to establish the predictive accuracy of each gene list. LOOCV implies using a single sample from the original dataset as the validation data (sample test), and the remaining samples as training data. This was repeated such that each sample in the dataset was used once, as a sample test. Each sample was characterized by a vector whose dimension was the number of genes that belonged to the reduced base that differentiated between HF and LF. The class with the minimum Euclidean distance was assigned to the sample test (NN classifier),<sup>25</sup> and the average accuracy was calculated by iterating over all the samples. For that purpose, all the samples were normalized according to their gene variability (each attribute or gene separately). In this way, all the genes had the same importance in the distance criterion. The distance between a sample and a phenotype class could have been defined in several ways, but the most robust one was using the median distance between the sample test and all the samples in the corresponding class.
4. The legitimacy of the predictive accuracy based on the training set was then tested with the validation set, using the above-mentioned predictive model. It is important to remark that the application of the prediction model, designed in steps 1 to 3, to the validation set was unsupervised. The final decision was made by consensus (majority voting) of the predictions made using the lists of most discriminatory genes.

## Results

**Demographic and clinical characteristics.** A total of 44 men with non-metastatic prostate cancer were studied. Subjects were primarily Caucasian (67%), had a mean age of  $65.2 \pm 6.7$  years and were not depressed based on Hamilton Depression Scale ( $1.1 \pm 2.2$ ) criteria. All subjects received a cumulative radiation dose of at least 68.4 Gy and more than 90% received a total dose of 75.6 Gy. Most (64%) of the subjects had a Gleason score of 7–8, and 71% had clinical T-stage below T3. The Gleason scoring and clinical staging are unique systems to classify the extent of the prostatic carcinoma.<sup>26,27</sup> There was no difference between the clinical and demographic features of subjects in the training and validation sets. In general, CTRF as indicated by a significant decrease in FACT-F scores from baseline ( $45.4 \pm 7.2$ ) to completion of EBRT ( $39.4 \pm 10.0$ ,  $P < 0.05$ ) was found. The characteristics of both study sets are shown in Table 1.

**Training model development.** The training model was developed from the array outputs of 27 subjects; 18 were HF

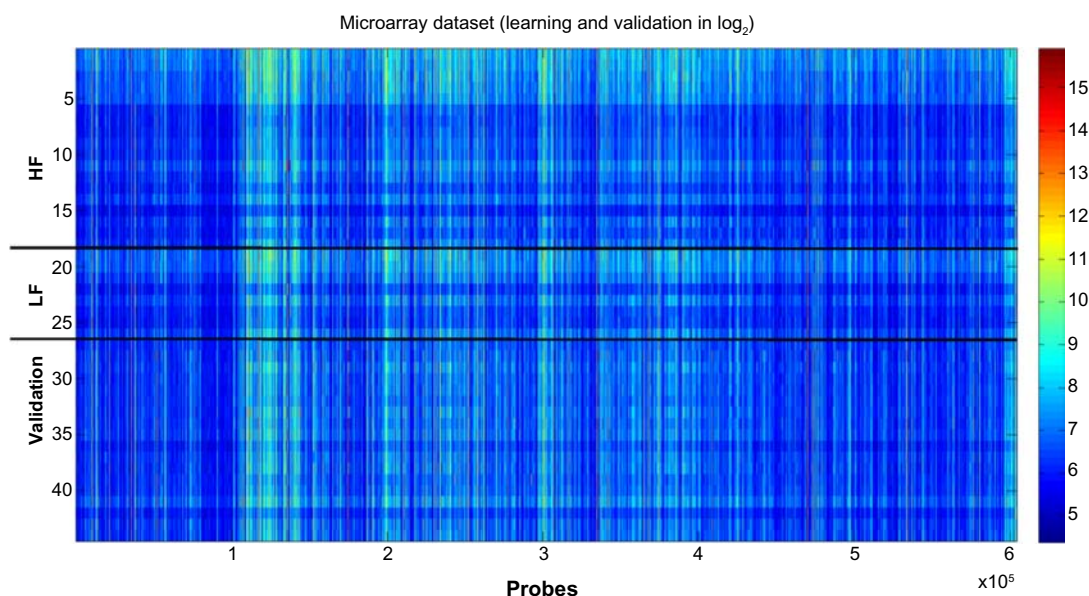
**Table 1.** Demographic characteristics of the sample.

	TRAINING						VALIDATION					
	HIGH FATIGUE (N = 18)			LOW FATIGUE (N = 9)			HIGH FATIGUE (N = 7)			LOW FATIGUE (N = 10)		
	MEAN (SD)	RANGE	N (%)	MEAN (SD)	RANGE	N (%)	MEAN (SD)	RANGE	N (%)	MEAN (SD)	RANGE	N (%)
Age in Years	64.6 (5.7)	53–73		65.2 (7.0)	55–74		66.7 (5.3)	58–73		66.5(7.0)	53–74	
Ethnicity n(%)												
Caucasian	18 (100)			7 (78)			2 (29)			5 (50)		
African-American				2 (22)			4 (57)			4 (40)		
Other							1 (14)			1 (10)		
Clinical T stage												
T1 (a-c)	4 (22)			2 (22)			2 (29)			2 (20)		
T2 (a-c)	10 (56)			7 (78)			3 (43)			7 (70)		
T3 (a-c)	4 (22)						2 (29)			1 (10)		
BMI	30.3 (4.5)	22–37		30.4 (2.7)	26–34		30.4 (6.3)	24–42		31.5 (5.5)	25–40	
FACT-F score												
Baseline	43.6 (8.4)	28–52		47.0 (5.6)	36–52		48.9 (5.8)	36–52		42.3 (7.7)	32–51	
Endpoint (day 42)	32.5 (8.1)	20–46		47.4 (4.4)	41–51		39.6 (8.0)	26–48		43.1 (8.1)	31–52	
HAM-D score												
Baseline	1.1 (2.2)	0–7		0.6 (0.9)	0–2		0.1 (0.4)	0–1		1.0 (1.3)	0–4	
Endpoint (day 42)	1.8 (2.2)	0–7		0.8 (0.7)	0–2		1.6 (2.2)	0–6		1.6 (1.4)	0–5	

**Abbreviations:** SD, standard deviation; BMI, body mass index; FACT-F, Functional Assessment of Cancer Therapy – Fatigue subscale; HAM-D, Hamilton - Depression.

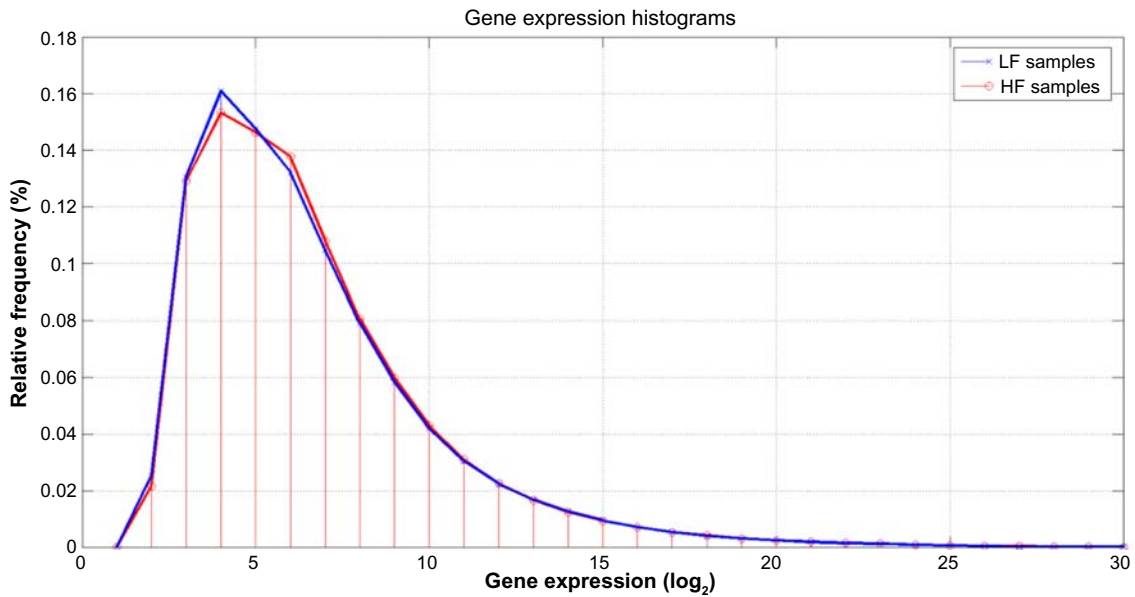
(mean FACT-F change =  $-11.8 \pm 6.8$ ) and 9 were LF (mean FACT-F change =  $0.8 \pm 3.3$ ). Each patient sample contained 604,258 different probes. The minimum and maximum gene expressions were 21 and 62,088, respectively.

As shown in Figure 2, it was impossible to visually distinguish HF and LF microarray outputs in heat map format using decibels as units of measure ( $\log_2$  of gene expression). The similarities between the HF and LF groups in the learning



**Figure 2.** Data visualization in decibels ( $\log_2$  of the expression). HF is composed of 18 samples, LF 9 samples and Validation 17 samples. The phenotype of the validation samples is not used for learning purposes. The expression varies from 21 to 62,088, that is, a fold change of 11, 53. No filtering is performed in the expression data, since the feature selection methods that are used are robust to the presence of outliers. Also, the gene selection is not only based on differential expression that might be affected by the presence of noise.





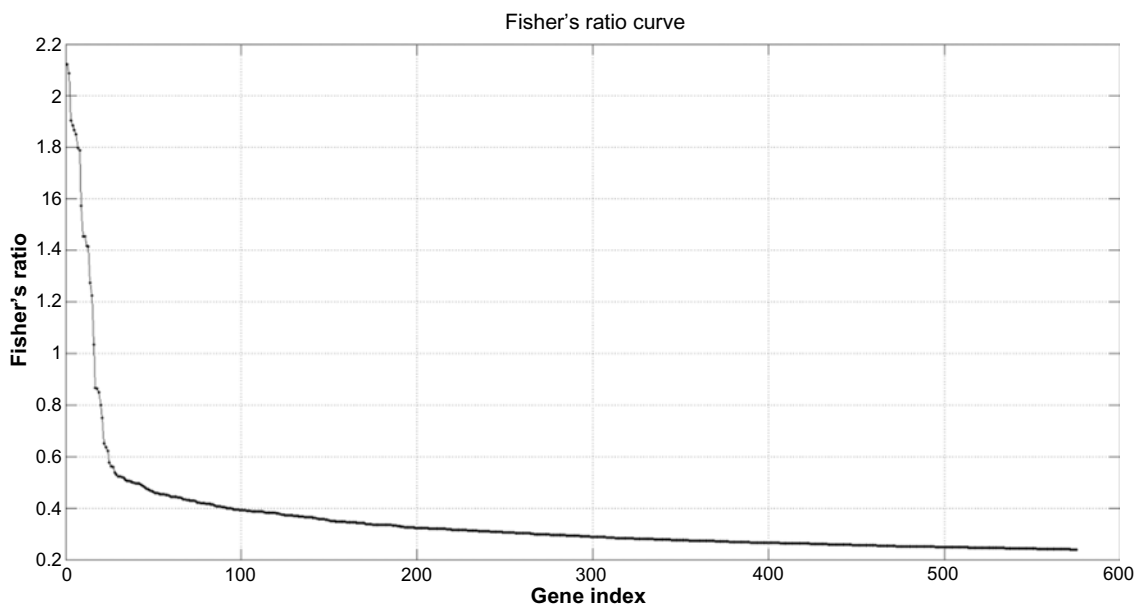
**Figure 3.** Gene expression histograms in  $\log_2$  scale for the Low Fatigue and High Fatigue subjects. Slight difference can be observed between them around the modes of the histograms ( $2^4$  to  $2^5$ ).

dataset were confirmed by further histogram analysis of gene expression. Figure 3 shows that the corresponding statistical distributions of gene expressions in both groups were close to lognormal, with the main differences between both phenotypes occurring around the mode of both histograms (expressions around  $2^4$  and  $2^6$ ).

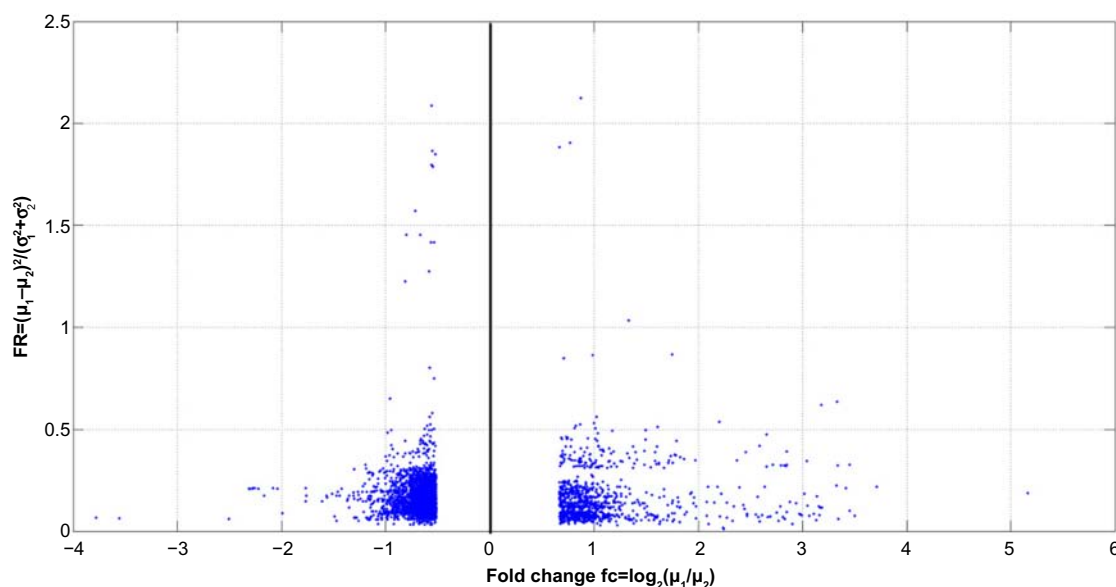
A final list of 575 highly discriminatory genes according to expression was noted and defined by the intersection

between those genes that were differentially expressed (located in the 0.05% and 99.5% tails of the fold-change ratio cumulative distribution) and which had a FR higher than 0.25 (Fig. 4).

Additionally, Figure 5 shows the fold change–FR plot for genes in the learning dataset with fold change lower than  $-0.52$  and higher than  $0.67$ . These values (of gene under- and over-expression) corresponded, respectively, to



**Figure 4.** Fisher's ratio curve for the Low Fatigue-High Fatigue phenotype discrimination. Genes with the highest Fisher's ratio were the most important biological eigenvectors for the phenotype discrimination, as it happens, for the Fourier analysis of a digital signal and its decomposition into different harmonics. In this case, the Fisher's ratio curve decreases very steeply, in such that only with the first set of genes (14 to 35 genes in this case) can the highest discriminative accuracy of the learning data set, can be achieved. Adding genes with lowest discriminatory power indiscriminately does not improve the LOOCV predictive accuracy. The backward RFE method is used to determine the amount of details that is needed.



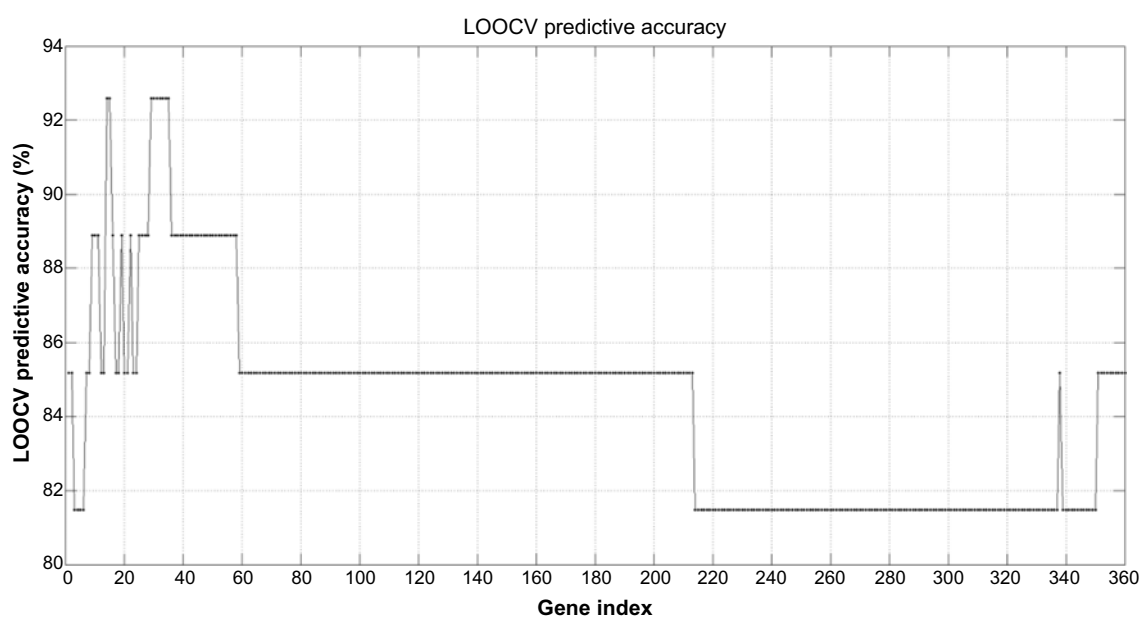
**Figure 5.** Fold change-Fisher's ratio plot of genes in the learning dataset with absolute fold change greater than 0.52 that corresponds to the 0.005 and 99.5% tails of the fold change distribution. In this case the Fisher's ratio plays a similar role than  $-\log(P \text{ value})$  for the volcano plot analysis.<sup>30</sup>

the 0.05% and 99.5% tails of the fold-change distribution. It can be observed that the highest FR was 2.12, and that genes with the highest fold change did not coincide with those exhibiting the highest FR.

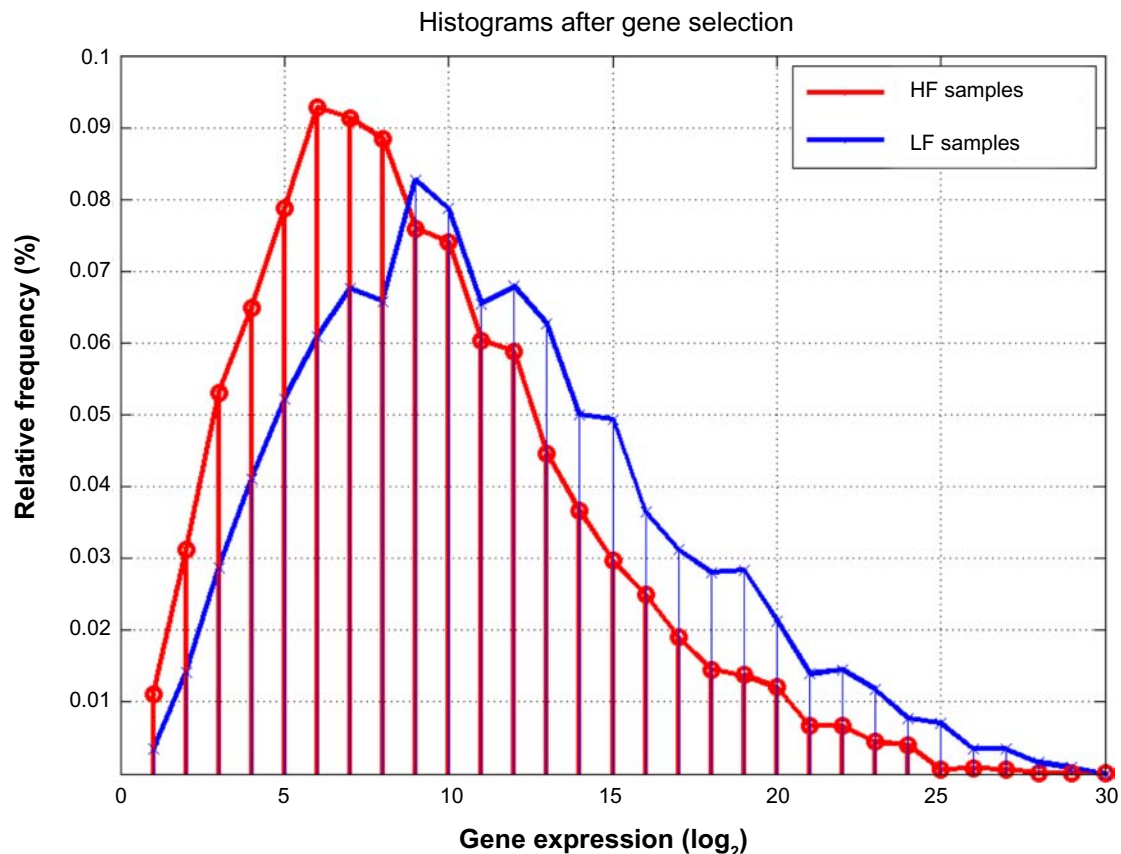
Figure 6 shows the predictive accuracy curve of the different gene lists, established using the backward feature elimination algorithm. The shortest list with the highest accuracy (92.6%) was composed by the first 14 genes with the highest FR. The lists with the first 15, and 29 to 35 most discriminatory genes also provide the same maximum

accuracy. As the data suggest, continuously adding genes with lower discriminatory power as defined by their FR failed to increase the accuracy of discrimination.

When a histogram was used to assess the first 360 most discriminatory genes found by our analysis, we noted a shift of the mode of distribution for the LF patients to higher expressions ( $2^9-2^{10}$ ) with respect to the HF case ( $2^6-2^7$ ), suggesting that HF patients show mostly lower expressions of these genes that we hypothesized were responsible for this phenotypic discrimination (Fig. 7).



**Figure 6.** Leave-One-Out-Cross-Validation (LOOCV) learning predictive accuracy of the first 360 gene sets with the highest discriminatory power. The shortest list with the highest accuracy (92.6%) contains only the first 14 genes. Other sets with similar accuracy adding additional helper genes also exist.



**Figure 7.** Histograms (in  $\log_2$  scale) for the Low Fatigue (LF) and High Fatigue (HF) patients, of the first 360 most discriminatory genes. Compared to Figure 3, a higher discrimination in the modes of the LF/HF phenotypes can be observed: the mode of HF samples is shifted to lowest values (approximately 64 instead of 512).

Figure 8 shows the PCA plots (unsupervised method) of the learning dataset expressed in the base of the most 14 (Fig. 8A) and 35 (Fig. 8B) discriminatory genes having the highest predictive accuracy. The following can be observed:

1. The LF/HF phenotype discrimination became linearly separable in these reduced sets of genes, confirming the fact that the classification problem simplifies when reducing the dimension to the most discriminatory set of genes. Both plots have a similar structure. The LF samples lie between samples P1A and xrt28A, which is genetically close to the region of the HF samples.
2. Also, sample xrt25A, which belongs to the LF category, is surrounded by HF samples. This sample might be a biological or behavioral outlier.
3. The HF samples lie between samples xrtp2A and 13A. Sample xrt20A also seems to mark a transition between LF and HF samples toward the west of the plot.

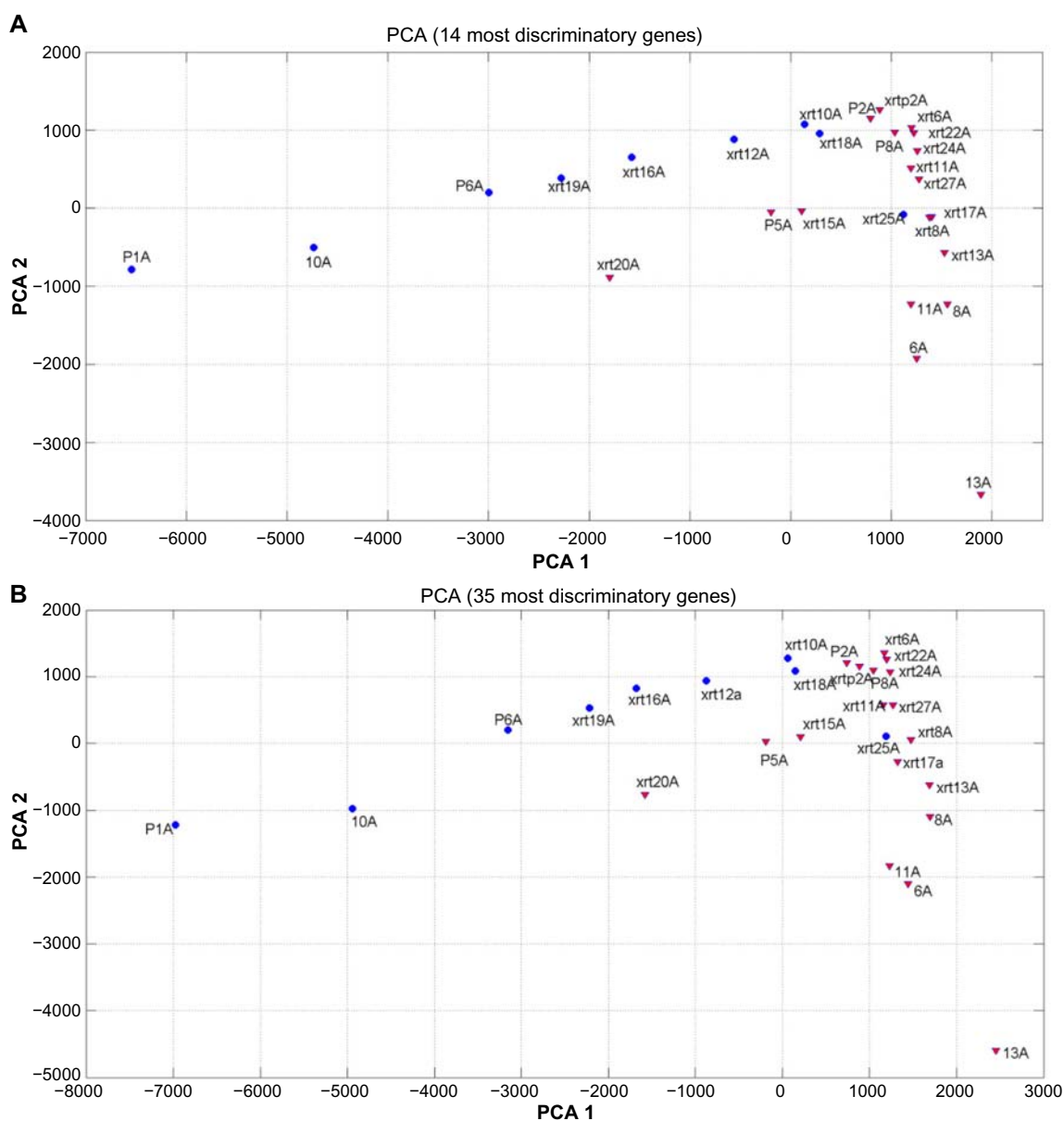
**Interpretative phenomenological analysis.** Interpretative phenomenological analysis (IPA) revealed that the 575 highly discriminatory genes were associated with the following canonical pathways: B cell development, autoimmune thyroid disease signaling, allograft rejection signaling,

graft-versus-host disease signaling, and Nur77 signaling in T lymphocytes. Further, the differentially expressed genes were associated with the following functional networks: cancer and neurological disease. Additional IPA was performed on the 360 most predictive genes (having a learning predictive accuracy higher than 81%), a part of the 575 highly discriminatory genes, and it revealed concordance of pathway attributions observed in the initial IPA. The top canonical pathways of the 360 most predictive genes remained to be related to B cell development, but it also revealed other focused pathways related to T helper cell differentiation and interferon signaling. The top functional networks of the 360 genes remained to be related to cancer, followed by neurological disease and psychological disorders, suggesting that the most predictive genes are related to behavior experienced by cancer patients.

**Validation.** Seventeen subjects, independent of the training set, were used to assess the validity of the learned predictive model. Seven were classified as HF (mean FACT-F change =  $-10.6 \pm 6.9$ ) and 10 were LF (mean FACT-F change =  $0.8 \pm 2.2$ ) subjects.

The prediction was based on majority voting, as follows:

1. We first considered the most predictive gene cluster, a group consisting of the 14 most discriminatory genes



**Figure 8.** (A) PCA plot for the learning set in the reduced base of the 14 most discriminatory genes. (B) PCA plot for the learning set in the reduced base of the 35 most discriminatory genes. A linear separability with a similar structure can be observed in both cases. Low Fatigue samples lie between P1A and xrt18A. Xrt25A might be a biological or behavioral outlier. High Fatigue (HF) samples lie between 13A and xrt2A. Xrt20A marks the HF limit towards the west of the plot. Additional data are needed to perfectly delineate this PCA plot.

deduced from the learning set, and the values of the expressions of these genes on both classes (LF and HF) represented in the training dataset. The samples of the training set expressed in the reduced base and their phenotype information were used to define the distance of the NN classifier used in this paper.

2. Second, the values of these discriminatory genes in the validation samples were read from the validation dataset. For each sample of the validation set, its predicted class was established using the k-NN algorithm, using the 14 different most discriminatory reduced sets of genes that were defined by the learning dataset. For instance, given the base composed by three first genes of the 14-size

reduced set of genes, the k-NN algorithm calculated the distance defined in three-dimensional space between each validation sample and the samples of the training dataset belonging to each phenotype class. The class with the minimum distance was then predicted for the validation sample. This was repeated for the 14 different reduced bases, which yielded 14 different class predictions for each sample in the validation set.

3. The final estimated class was then made by consensus or majority voting classifiers.<sup>28</sup> A posterior probability was given to the class prediction, defined as the ratio of the number of votes assigned to the predicted class and the total number of voters. For example, if a validation sample

**Table 2.** Mean values for the 14 most discriminatory genes.

HF IN LEARNING	LF IN LEARNING	HF IN VALIDATION	LF IN VALIDATION
114	<b>388</b>	117	<b>401</b>
152	<b>644</b>	143	<b>546</b>
302	<b>1455</b>	326	<b>1569</b>
343	<b>1659</b>	364	<b>1535</b>
185	<b>861</b>	196	<b>841</b>
149	<b>611</b>	127	<b>460</b>
<b>585</b>	128	<b>381</b>	194
243	<b>1182</b>	252	<b>1049</b>
<b>689</b>	111	<b>536</b>	235
<b>160</b>	65	75	<b>126</b>
247	<b>1225</b>	275	<b>1187</b>
<b>223</b>	80	73	<b>171</b>
269	<b>1329</b>	331	<b>1573</b>
<b>1200</b>	281	<b>1083</b>	485

**Notes:** Mean values of the 14 most discriminatory genes in the High Fatigue/Low Fatigue groups in the learning and the validation phases. Observe the coherence in values in both phases. Bold values indicate the highest mean expression values in the learning and validation datasets for HF and LF classes.

has 12 predictions in the LF class (and two in the HF class), the posterior probability to belong to LF will be 12/14.

The application of this algorithm provided 13 successes out of 17 validation samples. Three of the four misclassified samples belonged to the LF group (false positives, patients

**Table 3.** Misclassified samples.

S1 (XRT14)	S2 (XRT36)	S3 (XRT39)	S4 (XRT33)
57	129	87	342
78	257	105	492
136	327	201	1354
122	309	183	1514
79	180	125	765
92	126	168	341
42	44	54	946
103	175	184	1045
41	34	49	1430
62	178	258	52
77	234	183	1142
97	286	374	82
146	239	232	1388
162	167	137	2518

**Notes:** Misclassified samples. Expressions for the 14 most discriminatory probes. Samples S1, S2 and S3 were predicted to be High Fatigue and S4 to be Low Fatigue. The expression values for S1, S2 and S3 were closer to the mean expression of the High Fatigue group in the learning phase. Conversely, the expression values for S4 is closer to the Low Fatigue group. S1, S2 and S3 might define a new group of Low Fatigue with very small expressions (lower than the corresponding expressions observed among High Fatigue subjects) in this reduced base of 14 genes.

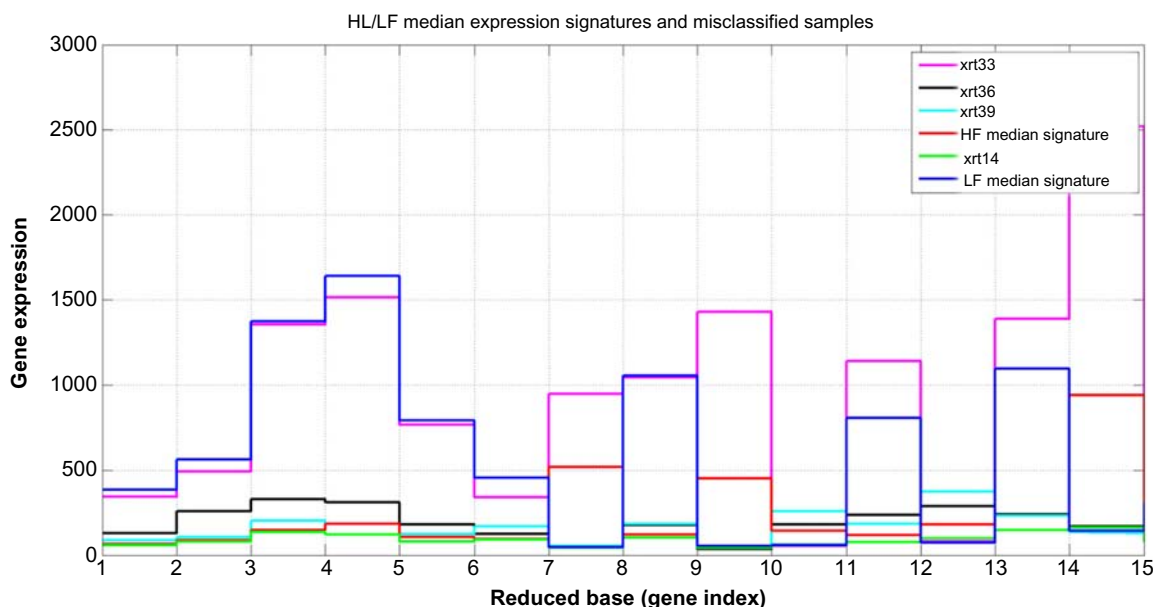
were predicted to be HF) and one to the HF (false negative, patient predicted to be LF). These samples are outliers with respect to this classifier, because their expressions in the reduced base of genes are closer to the HF and LF groups, respectively (Tables 2, 3, and Fig. 9). Interestingly, the 14 different predictions for these misclassified samples coincide, that is, the probability of these samples belonging to their predicted class according to the consensus criterion is 1. This fact also strengthens the argument that these samples are biological or behavioral outliers, that is, their class assignment based on the change in their FACT-F scores was ambiguous.

## Discussion

We have described a novel analytical algorithm to predict radiation-related fatigue. RT is a highly utilized treatment option for many forms of cancer. While it is efficacious in many cases, its toxicity profile is significant and common, but not ubiquitous. Consequently, the ability to predict toxicities of RT has long been of interest. With better understanding of the pathobiology of radiation injury, using genomics as the basis for toxicity risk prediction has been the focus of active research.<sup>29</sup> In contrast to the toxicity presented in this paper, the primary toxicity phenotypes studied have been tissue-centric injuries such as mucositis, dermatitis, and pneumonitis and fibrosis.<sup>30</sup> And the primary approaches used to try to identify predictive relationships between genes or SNPs and toxicities have primarily relied on candidate gene or genome-wide association analyses. In both cases, the majority of investigations have sought to identify one or two genes or SNPs associated with the phenotype of interest. The resulting lack of consistency of results has been disappointing.<sup>31</sup>

Our approach differed in that we proposed that the risk of a complex disease, such as CTRF, could well be more easily defined by identifying groups of simultaneously expressed, synergistically functioning genes. While this hypothesis is supported by studies in which Bayesian network development was used to identify SNP clusters predictive of chemotherapy-related side effects,<sup>11–13</sup> we sought to accelerate and simplify the analytical process through the use of a novel method in which we used a sequence of supervised and learned (unsupervised) “filters” to identify the most predictive cluster of genes for CTRF. Our finding that the gene cluster so identified was then able to predict CTRF risk with an accuracy of >75% suggests that the approach has validity.

The process of selecting the most predictive cluster of genes revealed informative considerations. For example, the genes with the highest fold change did not coincide with those exhibiting the highest FR because the means of both distributions were different, hence their tails did not overlap. So, in this method we concluded that FR was a better feature selection method than fold change. While, in the case of fold-change analysis, noisy genes are typically penalized by the FR selection method because of an increase of their variance; the noise might be amplified by the fold-change ratio. Genes with



**Figure 9.** High Fatigue (HF)/Low Fatigue (LF) median expression signatures and misclassified samples at validation. It can be observed that sample xrt33 is closer to LF median signature, while xrt14, xrt36 and xrt39 are closer to the HF median signature (values for the expressions are given in tables 2 and 3).

the highest FR and fold change have the biggest discriminatory power and are assumed to be involved in the genesis of fatigue.

Interestingly, the histogram analysis of the first 360 genes that most discriminated between HF and LF subjects was informative in that the shift of the mode of distribution showed lower expressions of these genes among HF subjects. It seems possible that it is this distributional shift that ultimately is responsible for discriminating the fatigue phenotype in this population.

We were unable to correctly predict four samples, based on our phenotypic approach, since the consensus provides the opposite class in all the cases. These classified samples were close to the border of separation between both fatigue classes (Fig. 8). There are three possibilities: (1) these samples are behavioral outliers, (2) the phenotypic approach needs further review and improvement, especially dealing with samples that are bordering the cut-off scores set for fatigue grouping, and (3) possible use of more sophisticated algorithms (black box neural networks) to classify the samples may be needed, which could run the risk of losing the clarity in the interpretation.

We recognize that this study was limited by its small sample size. Nonetheless, the fact that the analysis was successful in predicting LF/HF in an unrelated population with reasonable accuracy suggests that increasing the number of subjects in the training population would likely improve the predictive model's ability. Nevertheless, this analysis confirms that it is possible to separate both classes of the LF/HF phenotype by reducing the dimension to the most discriminatory genes, provided by their FR.

The importance of predicting toxicity or adverse event risk associated with cancer treatment regimens cannot be

understated as the clinical implications in personalizing cancer therapy and prospectively attenuating toxicity risk are significant. Furthermore, this type of information provides patients and their care-givers more specific knowledge upon which to make treatment decisions.

## Conclusion

A novel analytical algorithm introduced in this study that incorporates fold-change differential analysis, linear discriminant analysis, and a k-NN can predict radiation-related fatigue in men with non-metastatic prostate cancer. Applicability of this novel algorithm to detect other treatment-related toxicities in other cancer populations would be worthwhile to pursue.

## Author Contributions

Conceived and designed the experiments: LS, SS. Analyzed the data: LS, JLFM, EdG, SS. Wrote the first draft of the manuscript: LS, JLFM. Contributed to the writing of the manuscript: LS, JLFM, EdG, SS. Agree with manuscript results and conclusions: LS, JLFM, EdG, SS. Jointly developed the structure and arguments for the paper: LS, JLFM, EdG, SS. Made critical revisions and approved final version: LS, JLFM, EdG, SS. All authors reviewed and approved of the final manuscript.

## REFERENCES

1. Carlotto AA, Hoqsett VL, Maiorini EM, Razulis JG, Sonis ST. The economic burden of toxicities associated with cancer treatment: review of the literature and analysis of nausea and vomiting, diarrhea, oral mucositis and fatigue. *Pharmacoeconomics*. 2013;31:753–66.
2. Minton O, Richardson A, Sharpe M, Hotopf M, Stone P. A systematic review and meta-analysis of the pharmacological treatment of cancer-related fatigue. *J Natl Cancer Inst*. 2008;100:1155–66.



3. Mock V. Clinical excellence through evidence-based practice: fatigue management as a model. *Oncol Nurs Forum*. 2003;30:787–96.
4. Fransson P. Fatigue in prostate cancer patients treated with external beam radiotherapy: a prospective 5-year long-term patient-reported evaluation. *J Cancer Res Ther*. 2010;6:516–20.
5. Hofman M, Ryan JL, Figueroa-Moseley CD, Jean-Pierre P, Morrow GR. Cancer-related fatigue: the scale of the problem. *Oncologist*. 2007;12:4–10.
6. Miaskowski C, Paul SM, Cooper BA, et al. Trajectories of fatigue in men with prostate cancer before, during, and after radiation therapy. *J Pain Symptom Manage*. 2008;35(6):632–43.
7. Bower JE, Ganz PA, Tao ML, et al. Inflammatory biomarkers and fatigue during radiation therapy for breast and prostate cancer. *Clin Cancer Res*. 2009;15:5534–40.
8. Ryan JL, Carroll JK, Ryan EP, et al. Mechanisms of cancer-related fatigue. *Oncologist*. 2007;12:22–34.
9. Courtier N, Gambling T, Enright S, Barrett-Lee P, Abraham J, Mason MD. A prognostic tool to predict fatigue in women with early-stage breast cancer undergoing radiotherapy. *Breast*. 2013;22:504–9.
10. Hwang SS, Chang VT, Rue M, Kasimis B. Multidimensional independent predictor of cancer-related fatigue. *J Pain Symptom Manage*. 2003;26:604–14.
11. Schwartzberg LS, Sonis ST, Walter MS, et al. Single nucleotide polymorphism Bayesian networks predict risk of chemotherapy-induced side effects in patients with breast cancer receiving dose dense doxorubicin/cyclophosphamide plus paclitaxel (AC+T). *Cancer Res*. 2012;72(suppl):1–15–12.
12. Sonis ST, Schwartzberg LS, Walker MS, et al. Predicting risk of chemotherapy-induced side effects in patients with colon cancer with single-nucleotide polymorphisms (SMP) Bayesian networks. *J Clin Oncol*. 2013;Suppl 4;abstr 344.
13. Sonis S, Haddad R, Posner M, et al. Gene expression changes in peripheral blood cells provide insight into the biological mechanisms associated with regimen-related toxicities in patients being treated for head and neck cancers. *Oral Oncol*. 2007;43(3):289–300.
14. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *J Pain Symptom Manage*. 1997;13:63–74.
15. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six patient-reported outcomes measurement information system–cancer scales in advanced-stage cancer patients. *J Clin Epidemiol*. 2011;64(5):507–16.
16. Moberg PJ, Lazarus LW, Mesholam RI, et al. Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric inpatients. *Am J Geriatr Psychiatry*. 2001;9(1):35–40.
17. Hsiao CP, Wang D, Kaushal A, Chen MK, Saligan L. Differential expression of genes related to mitochondrial biogenesis and bioenergetics in fatigued prostate cancer men receiving external beam radiation therapy. *J Pain Symptom Manage*. 2014;Epub ahead of print.
18. Mayer C, Popanda O, Greve B, et al. A radiation-induced gene expression signature as a tool to predict acute radiotherapy-induced side effects. *Cancer Lett*. 2011;302:20–8.
19. Qiagen. Ingenuity pathway analysis: calculating and interpreting the p-values for functions, pathways and lists in IPA. Retrieved from <http://www.qiagen.com/ingenuity>
20. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*. 2001;98(18):5116–21.
21. Fisher RA. Has Mendel's work been rediscovered? *Ann Sci*. 1936;1:115–37.
22. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1–3):389–422.
23. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc Fourteenth Int Joint Conf Artif Intell*. 1995;2(12):1137–43.
24. Prestini E. *The Evolution of Applied Harmonic Analysis: Models of the Real World*. New York, NY: Springer; 2004.
25. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inform Theory*. 1967;13(1):21–7.
26. Epstein JI, Allsbrook WC Jr, Amin MB, Egevad LL. ISUP Grading Committee: the 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am J Surg Pathol*. 2005;29:1228–42.
27. Campbell T, Blasko J, Crawford ED, et al. Clinical staging of prostate cancer: reproducibility and clarification of issues. *Int J Cancer*. 2001;96:198–209.
28. Gareth J. *Majority Vote Classifiers: Theory and Applications* [PhD dissertation]. Stanford University; 1998.
29. Andreassen CN, Alsner J. Genetic variants and normal tissue toxicity after radiotherapy: a systematic review. *Radiother Oncol*. 2009;92:299–309.
30. West CM, Barnett GC. Genetics and genomics of radiotherapy toxicity: towards prediction. *Genome Med*. 2011;3:52–67.
31. Andreassen CN. The biological basis for differences in normal tissue response to radiation therapy and strategies to establish predictive assays for individual complication risk. In: Sonis S, Keefe DM, eds. *Pathobiology of Cancer Regimen-Related Toxicities*. New York, NY, USA: Springer; 2013:19–33.
32. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003;4(4):210.

## **A.4 Genomic data integration in Chronic Lymphocytic Leukemia**

Under review in the Journal of Gene Medicine.





## Genomic Data Integration in Chronic Lymphocytic Leukemia

Journal:	<i>Journal of Gene Medicine</i>
Manuscript ID	JGM-16-0013
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	08-Mar-2016
Complete List of Authors:	Fernández-Martínez, Juan; Universidad de Oviedo, Mathematics deAndrés-Galiana, Enrique; Universidad de Oviedo, Mathematics Sonis, Stephen; Biomodels, Chief Scientific Officer
Keywords:	cancer - leukemia/ hematologic, gene - expression, mathematical modeling, oncology

SCHOLARONE™  
Manuscripts

Review

## Genomic Data Integration in Chronic Lymphocytic Leukemia

Juan Luis Fernández-Martínez<sup>1,†</sup>, Enrique J. deAndrés-Galiana<sup>1,2,+</sup>, Stephen T. Sonis<sup>3\*</sup>

<sup>1</sup>Group of Inverse Problems, Optimization and Machine Learning.  
Department of Mathematics, University of Oviedo, Oviedo, Asturias, Spain  
<sup>2</sup>Artificial Intelligence Center, University of Oviedo, Oviedo, Asturias, Spain.  
<sup>3</sup>Biomodels LLC, Watertown, MA, USA.

[†jlfm@uniovi.es](mailto:†jlfm@uniovi.es)  
[+eag@aic.uniovi.es](mailto:+eag@aic.uniovi.es)  
[\\*ssonis@biomodels.com](mailto:*ssonis@biomodels.com)

Corresponding author: Juan Luis Fernández-Martínez, [jlfm@uniovi.es](mailto:jlfm@uniovi.es), 0034 985 103 199.

### Abstract

**Background** B-cell Chronic Lymphocytic Leukemia (CLL) is a heterogeneous disease and the most common adult leukemia in western countries. *IgVH* mutational status distinguishes two major types of CLL, which were associated with different prognosis and survival. Sequencing identified *NOTCH1* and *SF3B1* as the two main recurrent mutations. We described a novel method to elucidate how these mutations affect gene expression by finding small-scale signatures to predict the *IgVH*, *NOTCH1* and *SF3B1* mutations. We subsequently defined the biological pathways and correlation networks that are involved in the disease development with the potential goal of identifying new druggable targets.

**Methods** We modeled a microarray data set consisting of 48807 probes derived from 163 subject samples. Using Fisher's ratio and Fold-change combined with backwards feature elimination allowed us to identify the minimum number of genes with the highest predictive mutation power and subsequently applied network and pathway analyses of these genes to identify their biological roles.

**Results** The mutational status of the patients was accurately predicted (94 to 99%) using small-scale gene signatures: 13 genes for *IgVH*, 60 for *NOTCH1*, and 22 for *SF3B1*. *LPL* plays an important role in the case of the *IgVH* mutation, while *MSI2*, *LTK*, *TFEC* and *CNTAP2* in the *NOTCH1* mutation, and *RPL32* and *PLAGL1* in the *SF3B1* mutation. Four high discriminatory genes (*IGHG1*, *MYBL1*, *NRIP1* and *RGS1*) are common to these three mutations. This analysis suggests an important role of the immune response mechanisms and antigen presentation.

**Keywords** cancer - leukemia/ hematologic, gene – expression, mathematical modeling, oncology.

## 1. Introduction

B-cell chronic lymphocytic leukemia (CLL) is a complex heterogeneous disease characterized by the accumulation of malignant B-cells in blood and lymphoid organs [1]. Clinical diagnosis of CLL is based on the demonstration of an abnormal population of B lymphocytes in the blood, bone marrow, or tissues that display an unusual but characteristic pattern of molecules on the cell surface (*CD5* and *CD23* clusters of differentiation). Rai [2] or the Binet [3] staging systems are currently used to determine the extent of the disease and are primarily based on a low platelet or red cell counts.

DNA analysis distinguishes two major types of CLL with different survival times [4]. This distinction is based on lymphocyte maturity, as discerned by the immunoglobulin variable-region heavy chain (*IgVH*) gene mutation status [5]. Since the determination of the *IgVH* mutation status is very labor-intensive and expensive, alternative markers have been investigated to better prognosticate disease progression. *ZAP-70* became a very popular surrogate marker of the *IgVH* mutational status [6-10], and cell membrane expression of *CD38* has been described as a reliable prognostic value for CLL [11]. Both *ZAP-70* and *CD38* have been established as good predictors of the *IgVH* mutational status [12, 13].

Gene expression profiles were also used to understand the genesis and progression of CLL. Due to the high dimensionality of the data, on first inspection, both subtypes of CLL showed quite homogeneous expression profiles irrespective of their *IgVH* mutational status [4, 14-16]. This fact has suggested that both CLL types derive from a common pathogenic pathway. Nevertheless, different subsets of genes specifically expressed by CLL cells with potential pathogenesis and clinical relevance were identified [17-21].

1  
2  
3 57 Subsequently four major genomic aberrations have identified in CLL cells that are  
4  
5 58 strongly associated with the disease behavior [22]. More recently, *NOTCH1* and *SF3B1*  
6  
7 59 have been described as the most frequently mutated genes that were predictive of CLL  
8  
9 60 prognosis [23]. Additionally, disease progression has been associated with a number of  
10  
11 61 genetic alterations that include cytogenetic abnormalities and specific gene mutations  
12  
13 62 [23-29] and epigenetic alterations, including aberrant methylation CLL [30].

14  
15  
16 63 Given the low incidence of *NOTCH1* (9%) and *SF3B1* (8%) mutations, it seemed  
17  
18 64 unlikely to us that CLL progression could be solely ascribed to the two. We therefore  
19  
20 65 sought to identify shared/synergistic mechanisms among the three most common  
21  
22 66 mutations (*IgVH*, *NOTCH1* and *SF3B1*) which might better predict and explain disease  
23  
24 67 progression and behavior.

25  
26  
27 68 Obviously CLL is a multi-cause disease, and for that reason is so important to integrate  
28  
29 69 the modelling of different mutations. In the current research we describe a novel method  
30  
31 70 in which we developed a methodology to elucidate how these mutations affect gene  
32  
33 71 expression by finding small-scale signatures to predict the *IgVH*, *NOTCH1* and *SF3B1*  
34  
35 72 mutations (genomic data integration). We subsequently applied our method to define  
36  
37 73 and understanding the biological pathways and correlation networks that are involved in  
38  
39 74 the disease development with the potential goal of identifying new druggable targets.

40  
41 75 We can affirm that the integration of microarray data and the main mutational status of  
42  
43 76 patients attained by CLL is a novel approach to understand the main causes of disease  
44  
45 77 progression.

46  
47  
48  
49  
50 78

## 51 79 **2. Material and methods**

### 52 53 54 80 **2.1. Dataset**

1  
2  
3 81 We used a publically accessible dataset (European Bioinformatics Institute  
4  
5 82 EGAS00001000374) in which microarray results consisting of 48807 probes were  
6  
7 83 derived from 163 patients with a diagnosis of CLL [31]. The expression data were  
8  
9 84 originally presented in logarithmic scale ( $\log_2$ ) after the corresponding RMA  
10  
11 85 preprocess. Of the original cohort of 163 patients, 92 had mutated *IgVH*, which was  
12  
13 86 associated with a favorable prognosis, while *IgVH* was not mutated in the remainder  
14  
15 87 ( $n=71$ ) and prognosticated an unfavorable outcome. The exome sequencing data is  
16  
17 88 described by Quesada and colleagues [25], who identified 1246 mutations resulting in  
18  
19 89 protein coding changes. Six genes appeared to be most frequently mutated (>5%):  
20  
21 90 *NOTCH1*, *SF3B1*, *NOP16*, *CHD2*, *ATM* and *LRP1B*. Amongst the 163 samples we  
22  
23 91 evaluated, *NOTCH1* and *SF3B1* mutational status were determined for 117 patients. Of  
24  
25 92 these, 106 were unmutated for *NOTCH1* and 107 were unmutated for *SF3B1*.  
26  
27 93 These two classification problems are naturally unbalanced, and the designed classifier  
28  
29 94 has to take this feature into account. It will consider that the modelling has been correct  
30  
31 95 if the predictive accuracies of the small-scale signatures found are higher than those  
32  
33 96 provided by the corresponding majority class classifiers. In that case we can affirm that  
34  
35 97 we are really learning the set of discriminatory genes for these mutations.  
36  
37 98 These signatures have been validated by cross-validation. For that purpose, the dataset  
38  
39 99 was divided in two folds according to the different mutations: one fold was used for  
40  
41 100 training and the other for validation. This process is repeated until the whole dataset is  
42  
43 101 processed. Besides these results have been confirmed using different holds (75% for  
44  
45 102 learning and 25% for validation), but in the paper we only present the list of most  
46  
47 103 discriminatory genes found by Leave-One-Out-Cross-validation.  
48  
49 104 This methodology has been successfully used to design biomedical robots in phenotype  
50  
51 105 prediction problems [32].  
52  
53  
54  
55  
56  
57  
58  
59  
60

106

## 107 **2.2. Analysis**

108 Factor selection is one of the major challenges of any genotype-based phenotypic  
109 prediction problem is that the analysis is based on a genomic dataset in which the  
110 number of probes far exceeds the number of samples. To directly address this issue we  
111 used a combination of two well-known ranking algorithms: Fisher's ratio (FR) [33] and  
112 Fold change (FC) [34] (see Appendix for further details). We first ranked genes  
113 according to their discriminatory power (FR or absolute FC value) and then applied a  
114 Nearest Neighbor ( $k$ -NN) based algorithm to establish the accuracy of the different  
115 ranked sets of genes using Leave-One-Out-Cross-Validation (LOOCV) modeling. The  
116 combination of this procedure with a backwards feature elimination algorithm produced  
117 the shortest list of high discriminatory gene and served to validate the prognostic value  
118 of these gene signatures over the existing dataset by cross-validation [35 and Figure 1].  
119 Applying the Pearson correlation coefficient [36] and the Normalized Mutual  
120 Information [37] to the set of highly predictive genes, we then developed two different  
121 kinds of correlation networks as both values are measures of mutual dependence  
122 between random variables (see Appendix for further details). In this case they served  
123 to analyze inter-relationships between genes, which impacted both expression, and  
124 function. We then identified the pathways ontology using GeneAnalytics™ [38] to  
125 cover the altered and disease pathways.

## 126 **3. Results and Discussion**

### 127 **3.1. *IgVH* mutational status**

128 We determined the best set of genes that discriminates *IgVH* mutational status based on  
129 microarray expression and the class information defined by the *IgVH* phenotype using  
130 92 mutated and 71 unmutated samples.

1  
2  
3 131 - *Gene ranking*: The shortest list with the highest predictive accuracy (93.3%) was  
4  
5 132 composed by 13 first probes: *LPL* (2 probes), *CRY1*, *LOC100128252* (2 probes),  
6  
7 133 *SPG20* (2 probes), *ZBTB20*, *NRIP1* (2 probes), *ZAP-70*, *LDOC1* and *COBLL1*. Table 1  
8  
9 134 shows the list of these genes, their associated FR and the mean LOOCV accuracy. FR  
10  
11 135 was applied to the log<sub>2</sub> of the expressions. Table 2 shows the first 28 genes with the  
12  
13 136 highest predictive accuracy (93.3%), ordered by decreasing absolute FC (under or over  
14  
15 137 expressed in logarithmic scale,  $fc(log)$ ), the mean ( $\mu_1, \mu_2$ ) and the standard deviation ( $\sigma_1,$   
16  
17 138  $\sigma_2$ ) for each group, and the LOOCV accuracy ( $Acc(\%)$ ) in which over expression  
18  
19 139 implies that the average expression is higher in the mutated group. The gene with the  
20  
21 140 highest under-expression is *LPL*, and the gene with the highest over expression is  
22  
23 141 *RSG13*.

24  
25  
26  
27 142 - *Correlation networks*: Figure 1A shows the Pearson Correlation (PC) network of  
28  
29 143 the most discriminatory genes (defined by FR) of *IgVH* mutational status. The  
30  
31 144 Normalized Mutual Information (NMI) correlation network is shown in figure 1B.  
32  
33

### 34 145

### 35 146 **3.2. Modeling the NOTCH1 mutation status**

36  
37 147 It has been demonstrated that *NOTCH1* mutation influence survival in CLL patients  
38  
39 148 [39, 40]. We recognized the challenge analyzing of those genes for which the *NOTCH1*  
40  
41 149 mutation impacted expression given the highly unbalanced sample mix (106 of 117  
42  
43 150 samples did not show the *NOTCH1* mutation).  
44  
45

46  
47 151 - *Gene ranking*: The shortest list with the highest predictive accuracy (95.7%) was  
48  
49 152 composed by 60 probes with FR's between 4.6 and 1.4 (see Table 3). The first five  
50  
51 153 probes of this list corresponded to *MSI2*. Also using the two first probes of *MSI2*, the  
52  
53 154 *NOTCH1* mutation is predicted with 94.9% of accuracy. All *MSI2* probes had lower  
54  
55 155 expression in *NOTCH1*-mutation negative patients. One probe of the *LPL* gene  
56  
57  
58  
59  
60

1  
2  
3 156 appeared in eighth position in this list. Therefore the incremental accuracy from probe  
4  
5 157 5 to 60 was minimal (0.8%). That means the genes from the 6<sup>th</sup> position to the 60<sup>th</sup> serve  
6  
7 158 to add high frequency details in the discrimination, as it has been pointed in [35]. All  
8  
9 159 these genes show a correlation network that is analyzed later in this section.  
10  
11 160 The FC provided a longer list composed of 126 with the same predictive accuracy than  
12  
13 161 the FR list (see Table 4) with probes for *TFEC* and *CNTNAP2* being most consistently  
14  
15 162 differentially expressed. The high variability of these genes in samples with unmutated  
16  
17 163 *NOTCH1* precluded their inclusion among the leaders in the FR list.

18  
19  
20 164 -*Correlation networks*: Figure 2A shows the Pearson Correlation network of the  
21  
22 165 most discriminatory genes of the *NOTCH1* mutation in which three main networks  
23  
24 166 associated to *MSI2* through *WSB2*, *ACSL5* and *CNTNAP2* are apparent. The Normalized  
25  
26 167 Mutual Information network (Figure 2B) demonstrates a main connection through  
27  
28 168 *NCK2*.

### 170 3.3. *SF3B1* mutation status

171  
172 171 *SF3B1* gene (Splicing Factor 3b, Subunit 1) is located in chromosome 2. Its  
173  
174 172 importance in CLL has been analyzed by [26] and [41]. As with *NOTCH1*, the *SF3B1*  
175  
176 173 classification problem was also highly unbalanced, since 107 CLL samples (out of 117)  
177  
178 174 did not show the mutation.

179  
180 175 - *Gene ranking*: The shortest list with the highest predictive accuracy (99.1%) was  
181  
182 176 composed of 22 probes with FR's between 2.6 and 1.7. The most discriminatory gene  
183  
184 177 was RPL32 (Table 5). The FC provided a predictive accuracy of 96.6% using the list of  
185  
186 178 the first 118 genes ranked by absolute FC (Table 6). This accuracy was lower when  
187  
188 179 compared to the list of 22 most discriminatory genes provided by the FR. Seven  
189  
190 180 different probes of *ANXA4* appeared in this list.



1  
2  
3 181 - *Correlation networks*: Figure 3A shows the Pearson Correlation network of the  
4  
5 182 most discriminatory genes of the *SF3B1* mutation. In general correlations between  
6  
7 183 discriminatory genes are low, implying that these genes are independent predictors for  
8  
9 184 this mutation. Two main networks were noted to be associated to the most  
10  
11 185 discriminatory gene *RPL32*, through *YWHAB* and *KLF8*. Conversely, the correlation  
12  
13 186 network using the Normalized Mutual Information (figure 3B) demonstrated a single  
14  
15 187 network associated with *CNPY2-STK38*.  
16  
17  
18  
19

188

### 189 3.4. Gene intersections for *IgVH*, *NOTCH1* and *SF3B1* mutations

20  
21  
22 190 We analyzed the intersection between the most discriminatory genes for *IgVH*,  
23  
24 191 *NOTCH1*, and *SF3B1* mutations as defined by FR and FC analyses. We consolidated  
25  
26 192 both lists for each mutation, and then performed pairwise intersections to establish  
27  
28 193 shared genes. This would serve to understand which effects might be amplified by these  
29  
30 194 mutations. Figure 4 shows the result for these intersections. The intersection with the  
31  
32 195 greater number of genes is *NOTCH1-SF3B1* (19 genes), followed by *IgVH-NOTCH1*  
33  
34 196 (11 genes) and *IgVH-SF3B1* with only 5 genes. Only four genes were common to all  
35  
36 197 mutations: *IGHG1*, *MYBL1*, *NRIP1* and *RGS13*.  
37  
38  
39  
40

198

## 199 4. Conclusions

44  
45 200 In this paper we show the genomic data integration in CLL patients, by linking together  
46  
47 201 microarray expression data and their *IgVH*, *NOTCH1* and *SF3B1* mutational status. The  
48  
49 202 paper focuses on a novel methodological approach to define hierarchical gene  
50  
51 203 relationships among CLL patients expressing these 3 different mutations and  
52  
53 204 establishing the predictive accuracy of gene clusters relative to each mutation. Due to  
54  
55 205 the high dimensionality of the microarray data with respect to the number of available  
56  
57  
58  
59  
60

1  
2  
3 206 samples, this kind of phenotype prediction problems have a very high underdetermined  
4  
5 207 character. Therefore, simple and efficient methods are needed to rank genes according  
6  
7 208 to their discriminatory power and establishing their predictive accuracy. We have used  
8  
9 209 two well-known filter techniques: Fisher's ratio and Fold Change. For each mutation  
10  
11 210 and ranking method we have determined the shortest list of high discriminatory genes  
12  
13 211 with its corresponding LOOCV predictive accuracy. Using this methodology, we have  
14  
15 212 predicted the *IgVH* mutational status and how the *NOTCH1* and *SF3B1* mutations affect  
16  
17 213 the expression of different genes and their correlation networks via the Pearson's and  
18  
19 214 the Normalized Mutual Information similarity coefficients. In this discussion we also  
20  
21 215 provide the top ontological pathways that are involved in the disease progression, using  
22  
23 216 GeneAnalytics™. Correlation networks and canonical pathways provide effective  
24  
25 217 methodologies to understand the mechanisms that are involved in the disease  
26  
27 218 progression. This methodology served us to depict the gene clusters that are most  
28  
29 219 strongly associated with the expression of each selective mutation (networks of  
30  
31 220 synergistically working genes), and their relationship between mutation expressions  
32  
33 221 with a particular clinical outcome (survival).  
34  
35  
36  
37

38 The main conclusions for each of these mutation are the following:

39  
40  
41 **1. *IgVH***

42  
43 224 The *IgVH* mutational status was predicted with very high accuracy (94%) using a small-  
44  
45 225 scale signature composed of 13 genes. *LPL* (Lipoprotein Lipase) is the most  
46  
47 226 discriminatory gene of the list with 2 probes having a FR of 4.6 and 3.7. The predictive  
48  
49 227 power of the first *LPL* probe is also very high providing a LOOCV predictive accuracy  
50  
51 228 of 87.1%. *LPL* has a lower expression in the patients with mutated *IgVH*. This fact has  
52  
53 229 also been pointed out by [20]. These authors also found that high *LPL* mRNA  
54  
55 230 expression is associated with shorter treatment-free survival. *LPL* is a very specific  
56  
57  
58  
59  
60

1  
2  
3 231 biomarker since its activity is low or absent in other blood cells types. Kaderi et al  
4  
5 232 (2011) [42] noted that *LPL* was the strongest prognostic factor in comparative analysis  
6  
7 233 of RNA-based markers in early CLL. The result shown in this paper confirms that *LPL*  
8  
9 234 has almost twice discriminatory power than *ZAP-70*. Several genes of this list have  
10  
11 235 several probes involved and belong to the list of 24 genes differentially expressed in the  
12  
13 236 study for the identification and validation of biomarkers of *IgVH* mutational status using  
14  
15 237 PCR [43]. Particularly, *CRY1*, *LDOC1*, and *LPL* were overexpressed in *IgVH*-  
16  
17 238 unmutated compared with *IgVH*-mutated cases. Conversely, *COBLL1* and *ZBTB20*  
18  
19 239 were under-expressed. The analysis of differential expression shows that *LPL* is also  
20  
21 240 the second gene with the highest absolute FC (3.24) and it is overexpressed in the group  
22  
23 241 with unmutated *IgVH* of worst prognosis. *RSG13* is the other gene that is highly under  
24  
25 242 expressed in the same group. Other high discriminatory genes of the *IgVH* mutational  
26  
27 243 status are *CRY1*, *SPG20*, *ZBTB20*, *NRIP1* and *ZAP-70*, confirming previous findings by  
28  
29 244 other research groups [18-20].  
30  
31  
32

33  
34 245 The Pearson Correlation network shows a main branch relating *LPL* and *ZBTB20*.  
35  
36 246 *ZBTB20* is a transcription factor that may be involved in hematopoiesis, oncogenesis  
37  
38 247 and innate immune responses [44]. *ZBTB20* and *LPL* have been shown to predict  
39  
40 248 survival in B-cell CLL [45]. Diseases associated with *ZBTB20* include bone lymphoma.  
41  
42 249 *ZBTB20* expression is increased in hepatocellular carcinoma associated with poor  
43  
44 250 prognosis [46]. This network also shows connections between *LPL* and *COBLL1*,  
45  
46 251 *LDOC1*, *LOC100128252*, *KANK2*, *WSB2* and *ANKRD57-SEPT10*. Some of these genes  
47  
48 252 are known to have important roles in CLL and cancer. *COBLL1* (Cordon-Bleu Protein-  
49  
50 253 Like 1) is a gen related to actin binding. Actins participate in important processes such  
51  
52 254 as muscle contraction, cell motility, cell division and cytokinesis, cell signaling, etc.  
53  
54 255 This gene is down-regulated in CLL groups with poor prognostic [47]. *LDOC1*  
55  
56  
57  
58  
59  
60

1  
2  
3 256 (Leucine Zipper, Down-Regulated in Cancer 1) has been proposed as a tumor  
4  
5 257 suppressor gene whose protein product may have an important role in the development  
6  
7 258 and/or progression of some cancers [48]. It is thought to regulate transcriptional  
8  
9 259 responses by NF-kappa B that plays a key role in regulating the immune response to  
10  
11 260 infection. It has been also shown that *LDOC1* is differentially expressed in CLL and it  
12  
13 261 is a good predictor for overall survival in untreated patients [49]. The Normalized  
14  
15 262 Mutual Information shows two main branches with *TBC1D2B* and *SEPT10*. *SEPT10*  
16  
17 263 has been associated to B-cell Chronic Lymphocytic Leukemia and Chronic  
18  
19 264 Lymphocytic Leukemia [50, 51]. GO annotations related to *TBC1D2B* include  
20  
21 265 phospholipid binding.  
22  
23  
24

25 266 Also, the analysis of the networks for the most differentially expressed genes shows the  
26  
27 267 importance of the connection *RGS13-SPG20*. *RGS13* encodes a protein of the *RGS*  
28  
29 268 family, and it has been associated to mantle cell lymphoma. *SPG20* encodes the protein  
30  
31 269 called Spartin that seems to be related to endocytosis. This gene has been associated to  
32  
33 270 *ZAP70* and *LPL* as a good prognostic biomarker in CLL survival [18]. Hyper-  
34  
35 271 methylation of *SPG20* in early stage CLL has been positively associated with  
36  
37 272 progression free survival, supporting the fact that epigenetic changes have clinical  
38  
39 273 impact in CLL [52]. Two different probes of this gene are within the set of 13 most  
40  
41 274 discriminatory genes.  
42  
43  
44

45 275 The Gene-Analytics software has shown that the most important pathways involved are  
46  
47 276 the Inflammatory Response and the PAK pathways. The main GO biological processes  
48  
49 277 involved were related to Aging and Vasoconstriction, and the main GO molecular  
50  
51 278 functions were Cholesterol binding, Antigen Binding, Patched Binding and Nitric-oxide  
52  
53 279 Synthase Binding. Finally the main compounds that were identified target *IGHG1* and  
54  
55 280 *IGKC*.  
56  
57  
58  
59  
60

281 **2. NOTCHI**

282 The *NOTCHI* mutation was also predicted with very high accuracy (96%) using a set of  
283 60 most discriminatory genes. *MSI2* was the most important gene with 5 probes affected  
284 by this mutation. This gene plays an important role in posttranscriptional gene  
285 regulation and tumorigenesis. *MSI1* and *MSI2* are cooperatively involved in the  
286 proliferation and maintenance of CNS stem cell populations [53]. Also, their expression  
287 levels in human myeloid leukemia directly correlate with decreased survival in patients,  
288 defining *MSI2* as a new prognostic marker and new target for therapy [54]. The  
289 Musashi-Numb pathway can control the differentiation of chronic myeloid leukemia  
290 cells. *MSI2* expression is upregulated during human chronic myeloid leukemia  
291 progression and also an early indicator of poorer prognosis [55].

292 The analysis of most differentially expressed genes in this mutation showed the  
293 importance of *TFEC* and *CNTNAP2*. *TFEC* (Transcription Factor EC) gene encodes a  
294 member of the microphthalmia (*MiT*) family. *MiT* transcription factors play important  
295 role in multiple cellular processes including survival, growth and differentiation.  
296 *CNTNAP2* encodes a member of the neurexin family with functions in cell adhesion.  
297 This protein contains epidermal growth factor and laminin G domains. Annotations  
298 related to this gene include enzyme binding and receptor binding. This gene has been  
299 associated to genetic risk prediction for acute myeloid leukemia [56] and is involved in  
300 the genomic abnormalities for this illness [57].

301 The Pearson correlation network of the most discriminatory genes shows two main  
302 connections: *MSI2-ACSL5* and *MSI2-CNTNAP2*, while the Normalized Mutual  
303 Information network shows the connections to *NCK2*, *LPL* and *SPG20*.

304 The FC networks are simpler and show in both cases (PC and NMI) one main  
305 connection between *TFEC* and *NRIP1*. *NRIP1* encodes a nuclear protein that

1  
2  
3 306 specifically interacts with the hormone-dependent activation domain *AF2* of nuclear  
4  
5 307 receptors and modulates transcriptional activity of the estrogen receptor. This gene has  
6  
7 308 been shown to have predictive value in CLL together with LPL and ZAP-70 [18].  
8

9 309 The pathways analysis using Gene-Analytics have shown that the main discriminatory  
10 310 and differentially expressed genes are related to multiple cellular processes, including  
11 311 survival, growth and differentiation, apoptosis and host defense. The mains super-  
12 312 pathways were the creation of C4 and C2 activators, the FCGR dependent Phagocytosis  
13 313 and Adipogenesis. The main GO biological processes involve the regulation of  
14 314 Smoothened Signaling pathway, Immune response, Retina Homeostasis, Positive  
15 315 regulation of cardiac muscle hypertrophy, Fc-gamma receptor signaling pathway  
16 316 involved in phagocytosis, Triglyceride Biosynthetic Processes and Neural Tube  
17 317 formation. The main GO molecular function is Antigen Binding and the compounds  
18 318 found also target the genes *IGHG1* and *IGKC*. A second type of compounds are retinoid  
19 319 (*APOD*, *BMP6*, *NRIP1*, *PRKCA*, *RORA*, *THRB*), and valine (*FCGR3A*, *IGKC*, *LPL*,  
20 320 *PRKCA*, *THRB*).  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35

### 36 321 3. *SF3B1*

37  
38 322 Finally, the *SF3B1* mutation was predicted with almost 100% accuracy using a list of  
39 323 the 22 most discriminatory genes, including *RPL32*, *KLF8*, *PDGFD*, three different  
40 324 probes of *PLAGL1* (also named *ZAC1*) and two different probes of *HBB*. *PLAGL1*  
41 325 encodes a *C2H2* zinc finger protein with transactivation and DNA-binding activities and  
42 326 has been shown to have anti-proliferative properties as a tumor suppressor [58]. *HBB*  
43 327 (hemoglobin beta) expression in CLL samples with *SF3B1* mutation is almost 6 times  
44 328 the *HBB* expression in the unmutated samples. The analysis of most differentially  
45 329 expressed genes in this mutation showed the importance of *ANXA4* with seven different  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 330 probes in this list. *ANXA4* (Annexin IV) belongs to the annexin family of calcium-  
4  
5 331 dependent phospholipid binding proteins.  
6

7 332 The Pearson network of the most discriminatory genes shows two main networks of  
8  
9 333 *RPL32* with *YWHAB* and *KLF8*. *YWHAB* (Tyrosine 3-Monooxygenase/Tryptophan 5-  
10  
11 334 Monooxygenase Activation Protein, Beta) encodes a protein that has been shown to  
12  
13 335 interact with *RAF1* and *CDC25* phosphatases, suggesting that it may play a role in  
14  
15 336 linking mitogenic signaling and the cell cycle machinery. *KLF8* (Kruppel-Like Factor  
16  
17 337 8) encodes a protein, which is a member of the *Sp/KLF* family of transcription factors,  
18  
19 338 and is thought to play an important role in metastasis [59]. Diseases associated with  
20  
21 339 *KLF8* include ovarian epithelial cancer, and mental retardation.  
22  
23

24 340 The NMI correlation shows one main network relating *RPL32* with *CNPY2-STK38*.  
25  
26 341 Related to this last gene are hemoglobin *HBA1* and *HBB*. Also, the analysis of the  
27  
28 342 networks for the most differentially expressed genes shows the importance of the  
29  
30 343 connections of *ANXA4* with *CYBB* and *FCRL3* (PC network) and one main connection  
31  
32 344 with *ADM* (NMI network).  
33  
34

35 345 The most important pathways found using Gene-Analytics were *NFAT* in Immune  
36  
37 346 Response, *ERK* signaling, Immune Response of *DAPI2*, Creation of C4 and C2  
38  
39 347 activators, Inhibitory Action of Lipoxins on Super-Oxide Production in Neutrophils,  
40  
41 348 *HIF-1* alpha transcription factor network, Fc-gamma receptor signaling pathway, *PAK*  
42  
43 349 pathway, immune response *CCR* signaling in eosinophils and *GPCR* pathway. The main  
44  
45 350 GO molecular function involved is Oxygen transporter activity and the main GO  
46  
47 351 biological processes are Oxygen Transport, Innate Immune Response, Fc-gamma  
48  
49 352 receptor signaling pathway involved in phagocytosis and signal transduction. Finally,  
50  
51 353 the main compounds found were Thymidine (*ADM*, *CD69*, *CD72*, *FOS*, *HBB*, *NT5E*,  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 354 *PF4*, *PPBP*, *RNGTT* and *SMAD3*), and also the drugs targeting the genes *HBA1* and  
4  
5 355 *HBB* concerning the hemoglobin.

6  
7 356 Most of the genes affected by the *SF3B1* mutation are not known to play a role in CLL.

8  
9 357 In these three classification problems the FR always provided the shortest list with  
10  
11 358 highest discriminatory power whilst FC gave longer lists of genes with lower predictive  
12  
13 359 accuracy in general terms. Therefore, we can conclude that the most differentially  
14  
15 360 expressed genes are not the most discriminatory due to the high variability of these  
16  
17 361 genes in the groups of patients with unmutated *NOTCH1* and *SF3B1*.

18  
19 362 Regarding commonalities, *NOTCH1* and *SF3B1* mutations share a longer list of high  
20  
21 363 discriminatory genes than with the *IgVH* mutation. Besides the relationship *IgVH*-  
22  
23 364 *NOTCH1* is stronger than for the *SF3B1* mutation. Using an expanded list of high  
24  
25 365 discriminatory and differentially expressed genes we have shown that these three  
26  
27 366 mutations share only four genes (*IGHG1*, *MYBL1*, *NRIP1* and *RGS1*) that are related to  
28  
29 367 immune diseases, blood diseases, rare diseases and cancer.

30  
31 368 Finally, using GeneAnalytics™ we have identified the main pathways, GO molecular  
32  
33 369 and biological functions that are involved in each mutation and also the compounds that  
34  
35 370 are at disposal and the genes that could be targeted. These analyses suggest in the 3  
36  
37 371 cases an important role of the immune response and antigen presentation. This  
38  
39 372 methodology could also be applied to analyze the effect of other mutations in CLL and  
40  
41 373 to understand the genesis of other illnesses with genetic background. The aim of this  
42  
43 374 retrospective analysis was to provide a deeper understanding on the effects of the  
44  
45 375 different mutations in the CLL disease progression, hoping that these findings will be  
46  
47 376 used clinically in the near future with the development of new drugs. A future  
48  
49 377 verification of these findings with other independent cohorts could lead to a better  
50  
51 378 design of the therapeutic targets.  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 3794  
5 **380 Acknowledgments**

6  
7 381 The authors declare that there are no conflicts of interest; they have also read and  
8  
9 382 accepted the journal's authorship agreement. We would like to thank different  
10  
11 383 researchers from the Biochemistry Department (University of Oviedo, Spain) for having  
12  
13 384 introduced us to this interesting problem. Enrique J. de Andrés-Galiana salary has been  
14  
15 385 covered by the Spanish Ministerio de Economía y Competitividad (grant TIN2011-  
16  
17 386 23558). No additional research funds were provided to perform this research.

18  
19  
20 **387 Authors' contributions**

21  
22 388 JLFM and EJAG prepared the data, designed the machine learning methodology,  
23  
24 389 carried out the experiment, analyzed and interpreted the results and drafted the  
25  
26 390 manuscript. SS revised the design of the methodology critically, analyzed and  
27  
28 391 interpreted the results and drafted the manuscript. All authors read and approved the  
29  
30 392 final manuscript.

31  
32  
33 **393 Statement on conflicts of interest**

34  
35 394 There are no any competing interests (political, personal, religious, ideological,  
36  
37 395 academic, intellectual, commercial or any other) to declare in relation to this  
38  
39 396 manuscript.

40  
41  
42 39743  
44  
45 398  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

399 **References**

- 400 [1] Rodríguez-Vicente AE, Díaz MG, Hernández-Rivas JM. Chronic lymphocytic  
401 leukemia: a clinical and molecular heterogeneous disease. *Cancer Genet.*  
402 2013;206(3): 49–62.
- 403 [2] Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, Pasternack BS  
404 Clinical staging of chronic lymphocytic leukemia. *Blood.* 1975;46(2): 219–34.
- 405 [3] Binet JL, Auquier A, Dighiero G, Chastang C, Piguët H, Goasguen J, et al. A  
406 new prognostic classification of chronic lymphocytic leukemia derived from a  
407 multivariate survival analysis. *Cancer.* 1981;48(1): 198–206.
- 408 [4] Rosenwald A, Alizadeh AA, Widhopf G, Simon R, Davis RE, Yu X, et al.  
409 Relation of gene expression phenotype to immunoglobulin mutation genotype in  
410 b cell chronic lymphocytic leukemia. *J Exp Med.* 2001;194(11): 1639–47.
- 411 [5] Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated ig v(h)  
412 genes are associated with a more aggressive form of chronic lymphocytic  
413 leukemia. *Blood.* 1999;94(6): 1848–54.
- 414 [6] Crespo M, Bosch F, Villamor N, Bellosillo B, Colomer D, Rozman M, et al.  
415 Zap-70 expression as a surrogate for immunoglobulin-variable-region mutations  
416 in chronic lymphocytic leukemia. *N Engl J Med.* 2003;348(18): 1764–75.
- 417 [7] Wiestner A, Rosenwald A, Barry TS, Wright G, Davis RE, Henrickson SE, et al.  
418 Zap-70 expression identifies a chronic lymphocytic leukemia subtype with  
419 unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene  
420 expression profile. *Blood.* 2003;101(12): 4944–51.
- 421 [8] Orchard JA, Ibbotson RE, Davis Z, Wiestner A, Rosenwald A, Thomas PW, et  
422 al. Zap-70 expression and prognosis in chronic lymphocytic leukemia. *Lancet.*  
423 2004;363(9403): 105–11.
- 424 [9] Kim SZ, Chow KU, Kukoc-Zivojnov N, Boehrer S, Brieger A, Steimle-Grauer  
425 SA, et al. Expression of zap-70 protein correlates with disease stage in chronic  
426 lymphocytic leukemia and is associated with, but not generally restricted to,  
427 non-mutated ig vh status. *Leuk Lymphoma.* 2004;45(10): 2037–45.
- 428 [10] Smolej L, Saudkova L, Spacek M, Kozak T. Zap-70 in b-cell chronic  
429 lymphocytic leukemia: clinical significance and methods of detection. *Vnitr Lek.*  
430 2006;52(12): 1194–9.
- 431 [11] Mainou-Fowler T, Dignum HM, Proctor SJ, Summerfield GP. The prognostic  
432 value of cd38 expression and its quantification in b cell chronic lymphocytic  
433 leukemia (b-cell). *Leuk Lymphoma.* 2004;45(3): 455–62.
- 434 [12] Cruse JM, Lewis RE, Webb RN, Sanders CM, Suggs JL. Zap-70 and cd38 as  
435 predictors of IgVH mutation in cll. *Exp Mol Pathol.* 2007;83(3): 459–61.

- 1  
2  
3 436 [13] Shanafelt TD, Byrd JC, Call TG, Zent CS, Kay NE. Narrative review: initial  
4 437 management of newly diagnosed, early-stage chronic lymphocytic leukemia. *Ann*  
5 438 *Intern Med.* 2006;145(6): 435–47.
- 6  
7 439 [14] Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, et al. Gene  
8 440 expression profiling of b cell chronic lymphocytic leukemia reveals a  
9 441 homogeneous phenotype related to memory b cells. *J Exp Med.* 2001;194(11):  
10 442 1625–38.
- 11  
12 443 [15] Ferrer A, Ollila J, Tobin G, Nagy B, Thunberg U, Aalto Y, et al. Different gene  
13 444 expression in immunoglobulin-mutated and immunoglobulin- unmutated forms of  
14 445 chronic lymphocytic leukemia. *Cancer Genet Cytogenet.* 2004;153(1): 69–72.
- 15  
16 446 [16] Haslinger C, Schweifer N, Stilgenbauer S, Döhner H, Lichter P, Kraut N, et al.  
17 447 Microarray gene expression profiling of b-cell chronic lymphocytic leukemia  
18 448 subgroups defined by genomic aberrations and vh mutation status. *J Clin Oncol.*  
19 449 2004;22(19): 3937–49.
- 20  
21 450 [17] Vasconcelos Y, De Vos J, Vallat L, Rème T, Lalanne AI, Wanherdrick K, et al.  
22 451 Gene expression profiling of chronic lymphocytic leukemia can discriminate  
23 452 cases with stable disease and mutated ig genes from those with progressive  
24 453 disease and unmutated ig genes. *Leukemia.* 2005;19(11): 2002–5.
- 25  
26 454 [18] Van't Veer MB Brooijmans AM, Langerak AW, Verhaaf B, Goudswaard CS,  
27 455 Graveland WJ, et al. The predictive value of lipoprotein lipase for survival in  
28 456 chronic lymphocytic leukemia. *Haematologica.* 2006;91(1): 56–63.
- 29  
30 457 [19] Nuckel H, Hüttmann A, Klein-Hitpass L, Schroers R, Führer A, Sellmann L, et  
31 458 al. Lipoprotein lipase expression is a novel prognostic factor in b-cell chronic  
32 459 lymphocytic leukemia. *Leuk Lymphoma.* 2006;47(6): 1053–61.
- 33  
34 460 [20] Hartman ML, Kilianska ZM Lipoprotein lipase: a new prognostic factor in  
35 461 chronic lymphocytic leukaemia. *Contemp Oncol (Pozn).* 2012;16(6): 474–9.
- 36  
37 462 [21] Oppezzo P, Vasconcelos Y, Settegrana C, Jeannel D, Vuillier F, Legarff-  
38 463 Tavernier M, et al. The LPL/ADAM29 expression ratio is a novel prognosis  
39 464 indicator in chronic lymphocytic leukemia. *Blood.* 2005;106(2): 650–7.
- 40  
41 465 [22] Döhner H, Stilgenbauer S, Benner A, Leupolt E, Kröber A, Bullinger L, et al.  
42 466 Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J*  
43 467 *Med.* 2000;343(26): 1910–6.
- 44  
45 468 [23] Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, et al.  
46 469 Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic  
47 470 leukemia. *Nature.* 2011;475(7354): 101–5.
- 48  
49 471 [24] Fabbri G, Rasi S, Rossi D, Trifonov V, Khiabani H, Ma J, et al. Analysis of  
50 472 the chronic lymphocytic leukemia coding genome: role of notch1 mutational  
51 473 activation. *J Exp Med.* 2011;208(7): 1389–401.
- 52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 474 [25] Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al.  
4 475 Exome sequencing identifies recurrent mutations of the splicing factor sf3b1 gene  
5 476 in chronic lymphocytic leukemia. *Nat Genet.* 2012;44(1): 47–52.  
6  
7 477 [26] Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al.  
8 478 Sf3b1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J*  
9 479 *Med.* 2011;365(26): 2497–506.  
10  
11 480 [27] Oscier DG, Rose-Zerilli MJ, Winkelmann N, Gonzalez de Castro D, Gomez B,  
12 481 Forster J, et al. The clinical significance of notch1 and sf3b1 mutations in the uk  
13 482 lrf cll4 trial. *Blood.* 2013;121(3): 468–75.  
14  
15 483 [28] Ramsay AJ, Quesada V, Foronda M, Conde L, Martínez-Trillos A, Villamor N,  
16 484 et al. Pot1 mutations cause telomere dysfunction in chronic lymphocytic  
17 485 leukemia. *Nat Genet.* 2013;45(5): 526–30.  
18  
19 486 [29] Ghia P, Guida G, Stella S, Gottardi D, Geuna M, Strola G, et al. The pattern of  
20 487 cd38 expression defines a distinct subset of chronic lymphocytic leukemia (cll)  
21 488 patients at risk of disease progression. *Blood.* 2003;101(4): 1262–9.  
22  
23 489 [30] Rush LJ, Raval A, Funchain P, Johnson AJ, Smith L, Lucas DM, et al.  
24 490 Epigenetic profiling in chronic lymphocytic leukemia reveals novel methylation  
25 491 targets. *Cancer Res.* 2004;64(7): 2424–33.  
26  
27 492 [31] Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N,  
28 493 et al. Transcriptome characterization by RNA sequencing identifies a major  
29 494 molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome*  
30 495 *Res.* 2014;24(2): 212–26.  
31  
32 496 [32] deAndrés-Galiana EJ, Fernández-Martínez JL, Sonis ST. Design of biomedical  
33 497 robots for phenotype prediction problems. *J Comput Biol.* In press 2016.  
34  
35 498 [33] Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of*  
36 499 *Eugenics.* 1936;7(7): 179–88.  
37  
38 500 [34] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to  
39 501 the ionizing radiation response. *Proc Natl Acad Sci USA.* 2001;98(9): 5116–21.  
40  
41 502 [35] Saligan L, Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Supervised  
42 503 classification by filter methods and recursive feature elimination predicts risk of  
43 504 radiotherapy-related fatigue in patients with prostate cancer. *Cancer Informatics.*  
44 505 2014;13:141–152.  
45  
46 506 [36] Witten IH, Eiben F., Hall MA. *Data Mining: Practical Machine Learning Tools*  
47 507 *and Techniques.* 3rd ed. Morgan Kaufmann;2011.  
48  
49 508 [37] Peng HC, Long F., Ding C. Feature selection based on mutual information:  
50 509 criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans*  
51 510 *Pattern Anal Mach Intell.* 2005;27(8): 1226–1238.  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 511 [38] Stelzer G1, Inger A, Olender T, Iny-Stein T, Dalah I, Harel A, et al. GeneDecks:  
4 512 paralog hunting and gene-set distillation with GeneCards annotation. OMICS.  
5 513 2009;13(6): 477-87.  
6  
7 514 [39] Rossi D, Rasi S, Fabbri G, Spina V, Fangazio M, Forconi F, et al. Mutations of  
8 515 notch1 are an independent predictor of survival in chronic lymphocytic leukemia.  
9 516 Blood. 2012;119(2): 521–9.  
10  
11 517 [40] Willander K, Dutta RK, Ungerback J, Gunnarsson R, Juliusson G, Fredrikson  
12 518 M, et al. Notch1 mutations influence survival in chronic lymphocytic leukemia  
13 519 patients. BMC Cancer. 2013;13: 274.  
14  
15 520 [41] Wan Y, Wu CJ. SF3B1 mutations in chronic lymphocytic leukemia. Blood.  
16 521 2013;121(23): 4627–34.  
17  
18 522 [42] Kaderi MA, Kanduri M, Buhl AM, Sevov M, Cahill N, Gunnarsson R et al. Lpl  
19 523 is the strongest prognostic factor in a comparative analysis of rna-based markers  
20 524 in early chronic lymphocytic leukemia. Haematologica. 2011;96(8): 1153–1160.  
21  
22 525 [43] Abruzzo LV, Barron LL, Anderson K, Newman RJ, Wierda WG, O'brien S, et  
23 526 al. Identification and validation of biomarkers of IGV(h) mutation status in  
24 527 chronic lymphocytic leukemia using microfluidics quantitative real-time  
25 528 polymerase chain reaction technology. J Mol Diagn. 2007;9(4): 546–55.  
26  
27 529 [44] Liu X, Zhang P, Bao Y, Han Y, Wang Y, Zhang Q, et al. Zinc finger protein  
28 530 ZBTB20 promotes Toll-like receptor-triggered innate immune responses by  
29 531 repressing IκBα gene transcription. Proc Natl Acad Sci USA. 2013;110(27):  
30 532 11097-102.  
31  
32 533 [45] Nikitin EA, Malakho SG, Biderman BV, Baranova AV, Lorie YY, Shevelev  
33 534 AY, et al. Expression level of lipoprotein lipase and dystrophin genes predict  
34 535 survival in B-cell chronic lymphocytic leukemia. Leuk Lymphoma. 2007;48(5):  
35 536 912-22.  
36  
37 537 [46] Wang Q, Tan YX, Ren YB, Dong LW, Xie ZF, Tang L, et al. Zinc finger protein  
38 538 ZBTB20 expression is increased in hepatocellular carcinoma and associated with  
39 539 poor prognosis. BMC Cancer. 2011;11:271.  
40  
41 540 [47] Ronchetti D, Mosca L, Cutrona G, Tuana G, Gentile M, Fabris S, et al. Small  
42 541 nucleolar RNAs as new biomarkers in chronic lymphocytic leukemia. BMC Med  
43 542 Genomics. 2013;6:27.  
44  
45 543 [48] Nagasaki K, Manabe T, Hanzawa H, Maass N, Tsukada T,, Yamaguchi K.  
46 544 Identification of a novel gene, LDOC1, down-regulated in cancer cell lines.  
47 545 Cancer Lett. 1999;140(1-2): 227-34.  
48  
49 546 [49] Duzkale H, Schweighofer CD, Coombes KR, Barron LL, Ferrajoli A, O'Brien S,  
50 547 et al. LDOC1 mRNA is differentially expressed in chronic lymphocytic leukemia  
51 548 and predicts overall survival in untreated patients. Blood. 2011;117(15): 4076-84.  
52  
53 549 [50] Benedetti D, Bomben R, Dal-Bo M, Marconi D, Zucchetto A, Degan M, et al.  
54 550 Are surrogates of IgVH gene mutational status useful in b-cell chronic  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 551 lymphocytic leukemia? The example of septin-10. *Leukemia* 2008;22(1): 224–  
552 226.
- [51] Travella A, Ripollés L, Aventin A, Rodríguez A, Bezares RF, Caballín MR, et  
553 al. Structural alterations in chronic lymphocytic leukaemia. cytogenetic and fish  
554 analysis. *Hematol Oncol.* 2013;31(2): 79–87  
555
- [52] Ronchetti D, Tuana G, Rinaldi A, Agnelli L, Cutrona G, Mosca L, et al. Distinct  
556 patterns of global promoter methylation in early stage chronic lymphocytic  
557 leukemia. *Genes Chromosomes Cancer.* 2014;53(3): 264-73.  
558
- [53] Sakakibara S, Nakamura Y, Yoshida T, Shibata S, Koike M, Takano H, et al.  
559 RNA-binding protein Musashi family: roles for CNS stem cells and a  
560 subpopulation of ependymal cells revealed by targeted disruption and antisense  
561 ablation. *Proc Natl Acad Sci USA.* 2002;99(23): 15194-9.  
562
- [54] Kharas MG, Lengner CJ, Al-Shahrour F, Bullinger L, Ball B, Zaidi S, et al.  
563 Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid  
564 leukemia. *Nat Med.* 2010;16(8): 903-8.  
565
- [55] Ito T, Kwon HY, Zimdahl B, Congdon KL, Blum J, Lento WE, et al. Regulation  
566 of myeloid leukaemia by the cell-fate determinant Musashi. *Nature.*  
567 2010;466(7307): 765-8.  
568
- [56] Heo SG, Hong EP, Park JW. Genetic risk prediction for normal-karyotype acute  
569 myeloid leukemia using whole-exome sequencing. *Genomics Inform.* 2013;11(1):  
570 46-51.  
571
- [57] De Weer A, Poppe B, Vergult S, Van Vlierberghe P, Petrick M, De Bock R, et  
572 al. Identification of two critically deleted regions within chromosome segment  
573 7q35-q36 in EVI1 deregulated myeloid leukemia cell lines. *PLoS One.*  
574 2010;5(1):e8676.  
575
- [58] Varrault A, Ciani E, Apiou F, Bilanges B, Hoffmann A, Pantaloni C, et al. hZac  
576 encodes a zinc finger protein with antiproliferative properties and maps to a  
577 chromosomal region frequently lost in cancer. *Proc Natl Acad Sci USA.*  
578 1998;95(15): 8835–40.  
579
- [59] Wan W, Zhu J, Sun X, Tang W. Small interfering RNA targeting Krüppel-like  
580 factor 8 inhibits U251 glioblastoma cell growth by inducing apoptosis. *Mol Med*  
581 *Rep.* 2012;5(2): 347-50.  
582  
583

584 **Tables**585 Table 1: *IgVH* mutational status prediction using Fisher's ratio.

586 List of the 13 most discriminatory genes list with the highest predictive accuracy  
 587 (93.3%), ordered by decreasing Fisher's ratio.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean  
 588 expression and standard deviation in class 1, (mutated *IgVH*), and  $\mu_2$  and  $\sigma_2$  for the  
 589 unmutated group. FR (log) stands for the logarithmic Fisher's ratio.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
LPL	40	70	380	272	4.6	87.1
LPL	26	33	146	102	3.7	86.5
CRY1	62	125	352	298	3.1	90.2
LOC100128252	29	43	224	194	3.0	90.2
LOC100128252	30	42	220	172	3.0	89.6
SPG20	24	35	111	85	2.9	91.4
ZBTB20	1943	505	982	417	2.8	91.4
NRIP1	275	183	63	81	2.7	91.4
SPG20	30	53	148	126	2.6	91.4
ZAP70	103	151	273	140	2.4	92.6
LDOC1	20	19	50	27	2.3	92.6
COBLL1	186	107	85	100	2.3	92.6
NRIP1	85	60	24	24	2.1	93.3

590

591

592 Table 2: *IgVH* mutational status prediction using Fold Change.

593 List of the 28 most discriminatory genes with the highest predictive accuracy (93.3%),  
 594 ordered by decreasing absolute Fold Change.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean  
 595 expression and standard deviation in class 1 (mutated *IgVH*), while  $\mu_2$  and  $\sigma_2$  do for the  
 596 unmutated group. *fc* stands for the Fold Change.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	<i>fc</i> (log)	Acc(%)
RGS13	261	568	22	25	3.6	57.7
LPL	40	70	380	272	-3.2	86.5
PXDNL	269	838	29	52	3.2	87.1
SEPT10	22	50	177	232	-3.0	88.3
SEPT10	21	46	172	232	-3.0	86.5
LOC100128252	29	43	224	194	-3.0	89.0
SEPT10	22	48	158	206	-2.9	85.3
LOC100128252	30	42	220	172	-2.9	86.5
KANK2	40	75	258	335	-2.7	87.1
NPTX1	28	32	168	363	-2.6	85.9
LMNA	30	54	178	446	-2.6	85.9
PTCH1	140	163	24	42	2.6	89.0
IGHG1	162	687	942	1893	-2.5	88.3
MYBL1	424	526	73	88	2.5	89.6
PPP1R9A	45	90	260	278	-2.5	90.2
CRY1	62	125	352	298	-2.5	90.8
LPL	26	33	146	102	-2.5	92.0
TFEC	412	580	77	230	2.4	92.6
PPP1R9A	33	63	171	175	-2.4	92.6
SPG20	30	53	148	126	-2.3	92.6
STK32B	32	43	153	301	-2.3	92.6
ANKRD57	20	28	98	97	-2.3	91.4
SPG20	24	35	111	85	-2.2	92.0
ADAM29	205	203	45	97	2.2	92.0
RBMS3	19	27	84	122	-2.2	91.4
PPP1R9A	27	36	120	126	-2.2	91.4
NRIP1	275	183	63	81	2.1	92.6
PTCH1	115	123	27	33	2.1	93.3

597

598



599 Table 3: *NOTCH1* mutational status prediction using Fisher's ratio.

600 List of the 60 most discriminatory genes to predict the *NOTCH1* mutation list with the  
 601 highest predictive accuracy (95.7%), ordered by decreasing Fisher's ratio. Class 1  
 602 corresponds to samples with mutated *NOTCH1* and class 2 corresponds to those with  
 603 unmutated *NOTCH1*.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard  
 604 deviation in class 1 (mutated *NOTCH1*), and  $\mu_2$  and  $\sigma_2$  for the unmutated group. FR  
 605 (log) stands for the logarithmic Fisher's ratio.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
MSI2	157	74	43	26	4.6	93.2
MSI2	238	123	62	49	4.1	94.9
MSI2	73	25	31	16	3.0	91.5
MSI2	283	149	92	61	2.8	90.6
MSI2	58	19	32	15	2.7	92.3
C10orf137	193	86	392	135	2.4	90.6
LAG3	236	155	77	103	2.4	90.6
LPL	357	250	170	254	2.3	92.3
NCK2	838	219	1560	529	2.2	93.2
CNTNAP2	66	96	667	799	2.1	92.3
ST3GAL1	38	11	85	36	2.1	90.6
CCDC24	109	73	48	44	2.0	92.3
LTK	216	96	103	132	2.0	90.6
FLNB	59	30	33	17	1.9	94.0
ZNF333	38	5	57	16	1.9	92.3
PREPL	190	62	329	108	1.9	93.2
C19orf28	120	37	217	80	1.9	93.2
C1orf38	365	148	189	109	1.8	91.5
LTK	107	52	52	64	1.8	91.5
SPG20	182	150	71	106	1.8	92.3
SAP30L	74	38	111	32	1.8	94.0
MYST1	248	37	322	60	1.7	93.2
C10orf137	99	41	187	66	1.7	94.9
ATP6V0B	831	198	596	183	1.7	91.5
LPL	130	89	75	99	1.7	92.3
SLC4A7	47	39	150	120	1.7	90.6
LOC100128252	161	126	112	156	1.7	89.7
HNRNPR	57	22	110	48	1.7	89.7
REEP5	41	18	80	39	1.6	90.6
SRSF1	110	60	175	52	1.6	94.0
GNPNAT1	37	8	64	24	1.6	94.0
SHPRH	270	64	383	83	1.6	94.0
CNTNAP2	101	140	804	1105	1.6	94.9
PHF2	119	44	175	60	1.6	92.3
FCRL1	234	180	525	308	1.6	93.2
WSB2	804	329	489	258	1.6	93.2
ATP6V0B	624	145	448	134	1.6	94.9
LYL1	87	31	140	47	1.5	94.9
ACSL5	230	85	332	106	1.5	94.9
STX17	50	21	75	25	1.5	94.0

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
SPG20	125	98	55	74	1.5	94.0
NHEJ1	29	7	37	8	1.5	94.0
ZNF248	48	25	89	45	1.5	93.2
MPST	55	20	35	10	1.5	93.2
CDK13	69	42	132	75	1.5	93.2
TRMT1	58	17	86	30	1.5	92.3
PI4K2A	224	101	115	84	1.5	93.2
ELOVL5	254	97	504	188	1.5	93.2
FAM30A	588	900	1535	1495	1.5	93.2
PTDSS1	129	21	190	44	1.5	94.0
PLGLB1	74	47	152	103	1.5	94.0
C5orf53	51	22	125	74	1.5	94.0
PSMD7	608	175	414	141	1.5	94.9
NASP	117	26	176	52	1.5	94.0
ATP6V0B	768	170	566	172	1.5	94.9
WDR36	108	36	164	43	1.4	94.9
LTN1	511	52	645	99	1.4	94.9
GAL3ST3	22	2	19	2	1.4	94.9
PDE7A	102	67	214	120	1.4	94.9
CAPRIN2	1098	345	1511	368	1.4	<b>95.7</b>

606

For Peer Review

607 Table 4: *NOTCH1* mutational status prediction using Fold Change.

608 List of the most discriminatory genes (126) to predict the *NOTCH1* mutation ordered by  
 609 decreasing absolute fold change with an accuracy of 95.7%. Class 1 corresponds to  
 610 samples with mutated *NOTCH1*, and class 2 corresponds to those with unmutated  
 611 *NOTCH1*.  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in  
 612 class 1 (mutated *NOTCH1*), whilst  $\mu_2$  and  $\sigma_2$  do for the unmutated group. *fc* stands for  
 613 the fold change.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	<i>fc</i> (log)	Acc(%)
TFEC	19	4	289	488	-3.9	85.5
TFEC	15	2	180	331	-3.6	85.5
CNTNAP2	66	96	667	799	-3.3	82.9
RASSF6	15	1	144	352	-3.2	84.6
PXDNL	18	3	168	598	-3.2	82.1
CNTNAP2	101	140	804	1105	-3.0	79.5
DEFA1	272	816	1866	3452	-2.8	77.8
ADAM29	24	20	156	199	-2.7	76.9
NRIP1	31	10	200	187	-2.7	84.6
IGF2BP3	33	47	186	281	-2.5	88.0
MYBL1	47	65	262	407	-2.5	86.3
ZNF208	18	6	100	143	-2.5	89.7
FGL2	67	84	359	422	-2.4	91.5
CLC	29	17	141	184	-2.3	90.6
ZNF208	18	4	81	108	-2.2	91.5
APOD	32	34	142	204	-2.1	90.6
APOD	115	139	459	637	-2.0	89.7
MSI2	238	123	62	49	1.9	90.6
SORL1	58	72	221	415	-1.9	90.6
NRIP1	17	2	64	60	-1.9	90.6
IGKV3D-11	2509	5892	657	2843	1.9	88.9
TCTN1	46	31	173	440	-1.9	88.0
MSI2	157	74	43	26	1.9	94.0
IGJ	299	429	1079	1859	-1.9	92.3
HOMER3	68	39	239	337	-1.8	91.5
RGS13	420	808	121	382	1.8	92.3
TFEC	15	1	51	71	-1.8	92.3
PTCH1	25	29	86	135	-1.8	93.2
FCER1A	27	15	93	145	-1.8	93.2
FCRL2	81	85	276	214	-1.8	92.3
CD9	94	118	316	466	-1.7	91.5
LAIR1	31	27	105	114	-1.7	92.3
IGJ	82	106	273	463	-1.7	91.5
IGKC	1561	4218	471	2074	1.7	90.6
LOC728175	135	131	41	52	1.7	91.5
CD9	78	97	258	379	-1.7	92.3
PRKCA	119	132	36	54	1.7	92.3
TSHZ2	28	39	92	143	-1.7	92.3
FCRL2	650	530	2102	1440	-1.7	93.2
RGS1	1049	1956	327	670	1.7	91.5

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	$f_c$ (log)	Acc(%)
SLC4A7	47	39	150	120	-1.7	91.5
CD9	89	86	281	433	-1.7	92.3
SORL1	49	34	153	273	-1.6	92.3
THRB	78	150	25	30	1.6	91.5
MSI2	283	149	92	61	1.6	91.5
LAG3	236	155	77	103	1.6	91.5
FCRL3	133	125	400	362	-1.6	93.2
FGL2	64	57	191	193	-1.6	92.3
CNTNAP2	21	4	62	79	-1.6	92.3
FCRL2	342	261	1018	724	-1.6	93.2
ATF5	146	152	49	45	1.6	92.3
IgVH5-78	56	46	166	232	-1.6	91.5
IGHG1	1369	2081	468	1313	1.5	91.5
FCRL2	460	334	1334	872	-1.5	92.3
FCRL5	226	185	656	590	-1.5	93.2
FCGR3A	65	76	189	269	-1.5	93.2
FCRL2	420	303	1215	788	-1.5	91.5
CD9	74	70	215	302	-1.5	92.3
FCRL5	564	492	1629	1333	-1.5	93.2
FCRL2	469	323	1351	877	-1.5	91.5
NBPF3	27	19	78	88	-1.5	91.5
FCRL2	220	261	629	541	-1.5	90.6
CD9	37	27	106	152	-1.5	90.6
FCRL5	266	210	749	628	-1.5	91.5
CD9	151	180	425	595	-1.5	91.5
FCRL2	725	482	2017	1253	-1.5	92.3
FGL2	17	7	46	50	-1.5	92.3
C21orf7	66	100	178	225	-1.4	92.3
FCRL3	1692	1627	4567	3202	-1.4	92.3
FCRL2	739	560	1992	1191	-1.4	91.5
MAP4K4	33	16	90	65	-1.4	90.6
TRIB2	51	55	135	214	-1.4	89.7
EDNRB	326	368	123	280	1.4	91.5
FCRL3	1653	1573	4376	3115	-1.4	91.5
TUBB6	409	540	156	210	1.4	92.3
ATF5	186	192	71	68	1.4	91.5
LOC728175	69	67	26	26	1.4	92.3
FAM30A	588	900	1535	1495	-1.4	92.3
ACSM3	483	397	185	208	1.4	93.2
PYHIN1	262	264	683	447	-1.4	92.3
C1orf38	593	309	229	195	1.4	88.9
MEF2C	286	166	739	393	-1.4	88.0
SPG20	182	150	71	106	1.4	90.6
FCRL5	128	98	325	300	-1.3	89.7
LAIR1	29	16	75	81	-1.3	89.7
PLGLB2	78	52	198	187	-1.3	89.7
TCF7	316	337	802	780	-1.3	89.7
H3F3C	47	24	119	118	-1.3	89.7

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	$f_c$ (log)	Acc(%)
SLC15A2	53	33	133	107	-1.3	91.5
PLGLB2	88	66	219	136	-1.3	92.3
MEF2C	193	107	481	252	-1.3	91.5
CD24	1070	624	2656	1495	-1.3	93.2
BMP6	26	19	65	60	-1.3	93.2
C5orf53	51	22	125	74	-1.3	92.3
C1orf38	1176	588	477	377	1.3	90.6
SERPINB6	245	176	100	119	1.3	93.2
HIST1H2BD	413	362	1014	1104	-1.3	90.6
RORA	347	546	142	256	1.3	92.3
TUBB6	410	630	168	344	1.3	91.5
C21orf7	207	281	504	587	-1.3	92.3
SESN3	29	7	70	61	-1.3	92.3
SLAMF1	76	75	185	197	-1.3	92.3
ATF5	247	251	102	93	1.3	90.6
MSI2	73	25	31	16	1.3	92.3
FCRL2	2055	1421	4932	2543	-1.3	92.3
FCRL5	1062	723	2521	1444	-1.2	92.3
TCTN1	469	238	1103	1509	-1.2	93.2
PPAPDC1B	474	394	1113	832	-1.2	93.2
ACSM3	839	726	358	424	1.2	94.0
ZADH2	50	26	118	71	-1.2	93.2
MNDA	168	139	391	235	-1.2	94.0
PER1	110	102	48	39	1.2	93.2
CRIP3	55	30	127	104	-1.2	92.3
SERPINB6	224	149	97	108	1.2	91.5
PYHIN1	161	150	370	223	-1.2	92.3
FCRL1	850	631	1955	1048	-1.2	92.3
KCTD7	62	37	143	96	-1.2	93.2
SLC30A1	65	60	150	266	-1.2	93.2
IL15	167	176	73	113	1.2	93.2
SPG20	125	98	55	74	1.2	92.3
HIST1H2AC	132	156	301	367	-1.2	91.5
DNASE1L3	63	98	28	60	1.2	92.3
SERPINB6	341	219	151	164	1.2	91.5
B4GALT2	135	95	60	59	1.2	94.0
KLF7	29	12	64	46	-1.2	94.0
ABCA9	387	282	172	124	1.2	95.7

614

615

616 Table 5: *SF3B1* mutational status prediction using Fisher's ratio.

617 List of most discriminatory genes (22) to predict the *SF3B1* mutation, ordered by  
 618 decreasing Fisher's ratio with an accuracy of 99.1%. Class 1 corresponds to samples  
 619 with mutated *SF3B1*, and class 2 corresponds to those with unmutated *SF3B1*.  $\mu_1$  and  $\sigma_1$   
 620 refer respectively to the mean expression and standard deviation in class 1 (mutated  
 621 *SF3B1*), while  $\mu_2$  and  $\sigma_2$  do for the unmutated group. FR (log) stands for the logarithmic  
 622 Fisher's ratio.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	FR(log)	Acc(%)
RPL32	859	228	513	115	2.6	94.0
KLF8	131	45	59	30	2.4	94.0
PDGFD	85	34	42	20	2.2	95.7
PLAGL1	171	87	336	118	2.2	94.0
KLF3	40	29	239	221	2.2	94.0
UQCC	27	7	41	7	2.1	94.9
HBA1	3650	2978	755	2218	2.1	96.6
CNPY2	206	73	317	70	2.1	97.4
TMC6	322	74	546	155	2.0	97.4
CSNK2B	71	37	141	38	2.0	97.4
PLAGL1	282	135	507	174	2.0	97.4
PIP5K1B	55	32	212	200	1.9	98.3
DGKG	44	16	115	70	1.9	97.4
HBB	12044	6627	2783	5082	1.9	98.3
PLAGL1	138	83	252	92	1.9	98.3
ZNF76	34	8	61	20	1.8	98.3
AMT	48	8	97	41	1.8	97.4
STK38	206	108	368	156	1.8	97.4
HBB	8359	5278	1777	3669	1.8	97.4
ACTR2	3113	266	3789	506	1.8	97.4
GLIPR1	115	107	359	261	1.7	97.4
MAST4	136	89	59	60	1.7	99.1

623

624

625 Table 6: *SF3B1* mutational status prediction using Fold Change.

626 List of the most discriminatory genes (118) to predict the *SF3B1* mutation ordered by  
 627 decreasing absolute fold change with an accuracy of 96.6%. Class 1 corresponds to  
 628 samples with mutated *SF3B1*, and class 2 corresponds to those with unmutated *SF3B1*.  
 629  $\mu_1$  and  $\sigma_1$  refer respectively to the mean expression and standard deviation in class 1  
 630 (mutated *SF3B1*), whilst  $\mu_2$  and  $\sigma_2$  do for the unmutated group. *fc* stands for the fold  
 631 change.

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	<i>fc</i> (log)	Acc(%)
ANXA4	42	19	430	524	-3.3	87.2
IGHG1	63	37	599	1472	-3.3	88.0
ANXA4	41	17	333	408	-3.0	85.5
ANXA4	55	26	419	498	-2.9	87.2
ANXA4	47	17	335	408	-2.8	88.9
ANKRD36BP2	32	31	227	683	-2.8	90.6
TSPAN13	29	17	204	338	-2.8	86.3
ANXA4	52	24	364	460	-2.8	87.2
ANXA4	44	17	279	343	-2.7	88.9
MYBL1	41	36	261	406	-2.7	93.2
ANXA4	28	10	173	218	-2.6	91.5
KLF3	40	29	239	221	-2.6	92.3
NT5E	17	2	96	210	-2.5	92.3
GNB4	24	7	130	253	-2.4	90.6
TRPV3	198	525	38	67	2.4	92.3
ZNF608	16	2	82	166	-2.4	91.5
RBM20	26	25	132	212	-2.3	93.2
KLRK1	53	63	266	610	-2.3	93.2
CNTNAP2	135	188	655	801	-2.3	91.5
HBA1	3650	2978	755	2218	2.3	90.6
PPP1R9A	22	18	103	145	-2.2	90.6
HBB	8359	5278	1777	3669	2.2	87.2
HTRA3	202	409	45	55	2.2	88.9
KLRK1	51	48	227	512	-2.2	88.9
HBB	12044	6627	2783	5082	2.1	88.9
TUBB6	48	42	204	398	-2.1	88.9
CNTNAP2	189	266	789	1105	-2.1	88.0
TCTN1	42	32	172	438	-2.0	88.0
PTPRK	14	1	58	117	-2.0	88.9
HOMER3	59	31	239	336	-2.0	88.0
PIP5K1B	55	32	212	200	-2.0	89.7
PPP1R9A	39	43	148	227	-1.9	90.6
S100A9	1224	2650	339	591	1.9	90.6
HBB	15100	6062	4201	5929	1.8	90.6
VASH1	39	44	136	161	-1.8	88.9
CD69	166	259	48	56	1.8	88.9
PPP1R9A	21	9	71	99	-1.8	89.7
CD69	362	484	106	153	1.8	88.9
FCRL3	118	154	399	359	-1.8	89.7
MAP3K8	43	39	145	212	-1.7	91.5

Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	$f_c$ (log)	Acc(%)
FOSB	659	1112	199	418	1.7	91.5
RASSF6	362	720	110	274	1.7	92.3
SDPR	51	67	168	308	-1.7	93.2
CYBB	91	112	295	433	-1.7	93.2
PSD3	235	211	74	175	1.7	93.2
ADM	153	206	48	84	1.7	93.2
CD72	148	94	471	297	-1.7	91.5
FOS	928	1556	292	513	1.7	91.5
RAB20	25	8	80	123	-1.7	91.5
IGKC	365	1057	1151	2084	-1.7	90.6
BCL7A	61	56	193	248	-1.7	92.3
GLIPR1	115	107	359	261	-1.6	93.2
SDPR	96	160	296	533	-1.6	94.9
FOS	1105	1870	360	631	1.6	94.0
FCER1A	30	16	92	144	-1.6	93.2
CX3CR1	32	14	97	141	-1.6	92.3
GNG11	60	85	179	322	-1.6	94.0
GLIPR1	127	79	373	246	-1.6	94.0
FCRL3	1498	1626	4365	3100	-1.5	90.6
SLAMF1	64	42	185	196	-1.5	90.6
FCRL3	1565	1755	4552	3186	-1.5	90.6
ADM	163	221	56	102	1.5	89.7
RPL31	803	501	279	140	1.5	95.7
TUBB1	176	297	505	870	-1.5	95.7
BCL7A	20	3	57	77	-1.5	94.9
IGKV1-5	367	1058	1047	2114	-1.5	94.0
HOMER3	17	2	49	80	-1.5	94.0
PLAC8	559	386	1588	1047	-1.5	94.0
BTNL9	201	384	71	300	1.5	94.0
FCRL1	186	105	527	307	-1.5	94.0
CNR1	301	362	106	236	1.5	94.0
CLEC4C	24	9	69	95	-1.5	94.9
FCGBP	51	26	145	209	-1.5	95.7
GLIPR1	189	117	532	361	-1.5	94.9
CYBB	28	19	80	108	-1.5	94.9
CLLU1	449	597	161	355	1.5	94.9
GEN1	77	31	215	374	-1.5	94.0
GAPT	64	38	177	139	-1.5	94.0
PPBP	624	1065	1730	2686	-1.5	94.9
RASSF6	124	258	45	100	1.5	94.9
NRIP1	23	17	63	60	-1.5	94.9
CNTNAP2	22	9	62	78	-1.5	94.0
RGS13	355	777	130	399	1.5	93.2
SIGLEC6	37	47	100	164	-1.4	92.3
GLIPR1	207	120	554	354	-1.4	94.0
SMAD3	51	33	136	198	-1.4	94.0
PRDM1	128	201	48	51	1.4	94.9
NRIP1	73	119	195	187	-1.4	94.0



Gene	$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$	$f_c$ (log)	Acc(%)
RNGTT	981	825	370	627	1.4	94.0
PF4	85	114	226	319	-1.4	93.2
CYB5A	49	17	130	104	-1.4	93.2
FHDC1	52	43	137	148	-1.4	94.0
GLIPR1	168	135	444	292	-1.4	94.0
FCGR3A	71	82	187	268	-1.4	94.9
GLIPR1	346	186	907	595	-1.4	95.7
RNGTT	673	558	257	419	1.4	94.9
CYB5A	62	34	163	135	-1.4	94.9
TCF7	306	247	798	780	-1.4	95.7
GEN1	38	8	99	213	-1.4	95.7
RNGTT	972	755	374	622	1.4	95.7
ITGAX	196	217	510	356	-1.4	95.7
DGKG	44	16	115	70	-1.4	94.0
TCF7	100	54	259	267	-1.4	94.9
RASGRP1	177	137	455	315	-1.4	94.9
SSBP2	44	29	112	87	-1.4	94.0
PDK4	31	14	78	74	-1.3	92.3
KLF3	22	4	56	73	-1.3	92.3
KLF3	27	6	68	56	-1.3	92.3
PF4	107	163	272	399	-1.3	93.2
FGL2	138	152	349	424	-1.3	92.3
IPCEF1	101	100	40	34	1.3	95.7
FCRL1	775	460	1952	1051	-1.3	95.7
FCRL1	642	350	1610	883	-1.3	95.7
FRMD5	58	82	23	14	1.3	95.7
NSUN7	171	168	69	95	1.3	95.7
FCRL2	109	75	271	216	-1.3	95.7
HBM	41	31	103	708	-1.3	95.7
RGS1	871	1001	351	864	1.3	96.6

632  
633

## LIST OF CAPTIONS

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

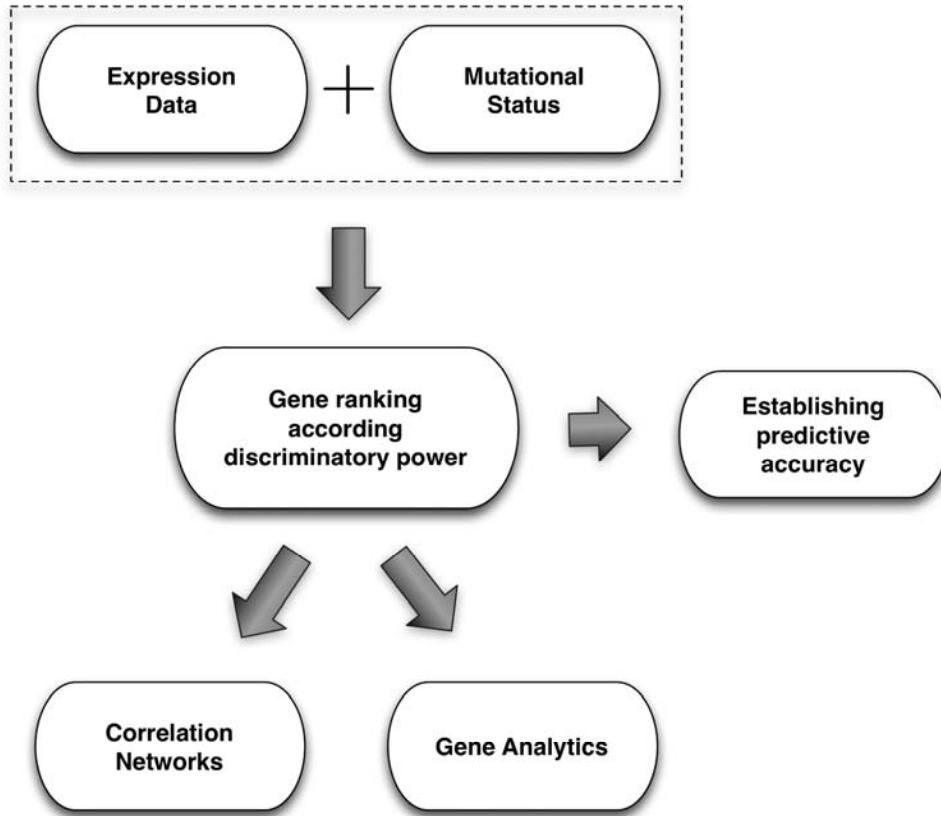
**Figure 1:** Flow diagram of the analytical procedure.

**Figure 2:** Correlation network of the most discriminatory genes for the *IgVH* mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

**Figure 3:** Correlation network of the most discriminatory genes for the *NOTCH1* mutational status prediction. A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

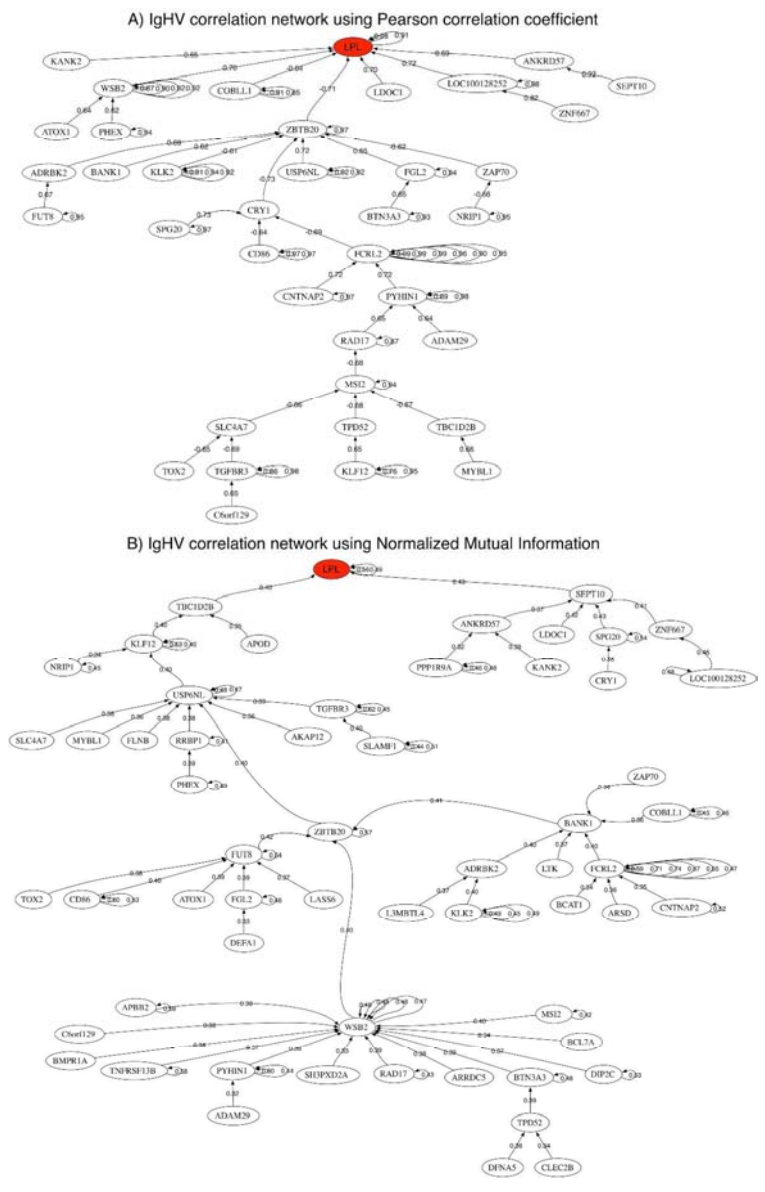
**Figure 4:** Correlation network of the most discriminatory genes for the *SF3B1* mutational status prediction. A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

**Figure 5:** Intersection among the most discriminatory genes of the *IgVH*, *NOTCH1* and *SF3B1* mutations. The three main mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with three edges (surrounded by a dot rectangle) are common to these three main mutations.



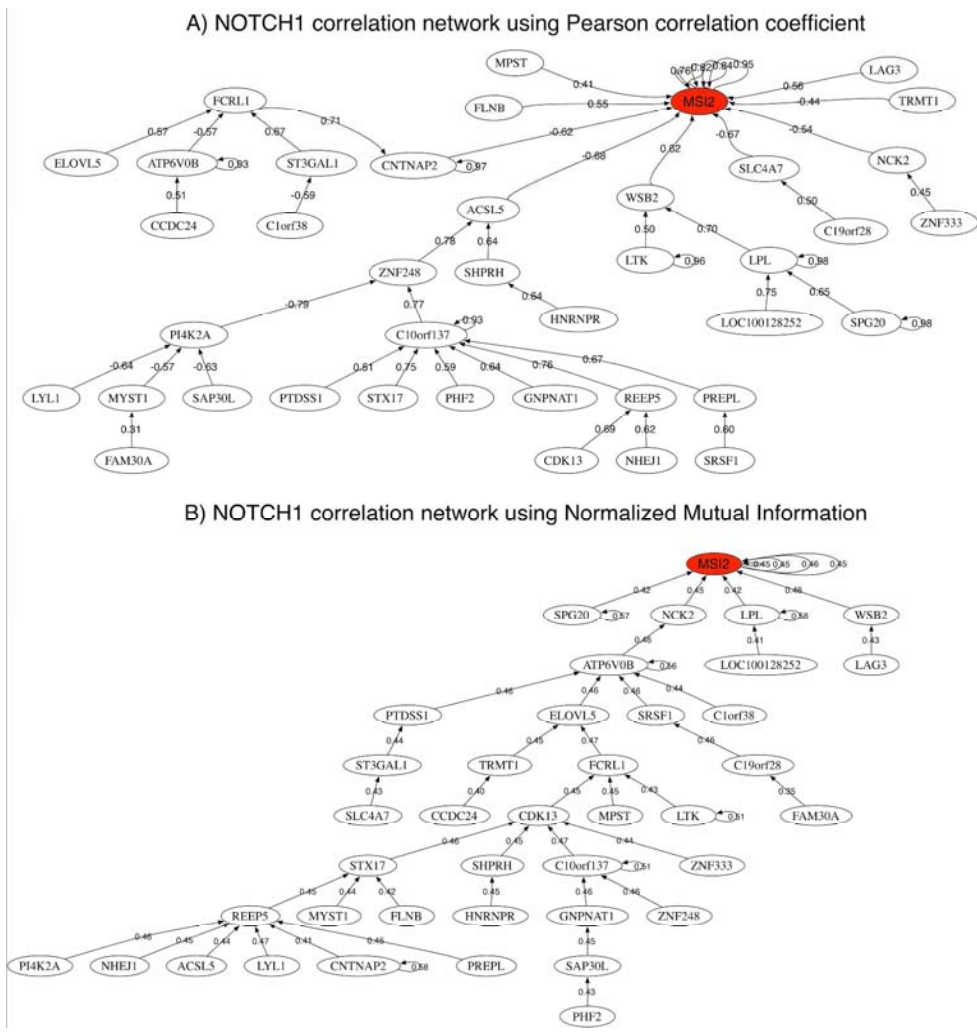
Flow diagram of the analytical procedure  
214x184mm (300 x 300 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Correlation network of the most discriminatory genes for the IgVH mutational status prediction: A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

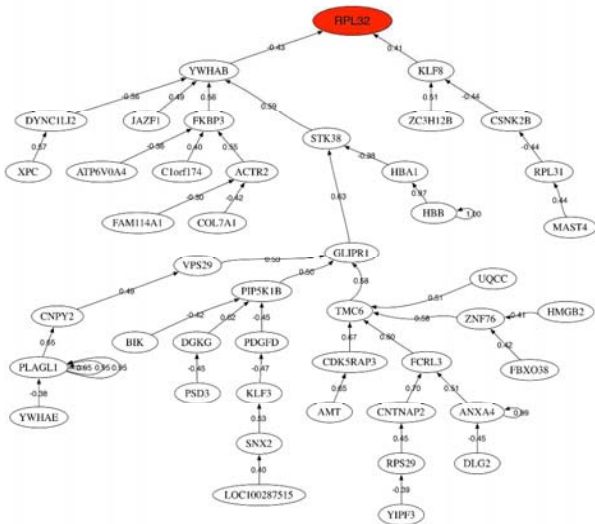
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



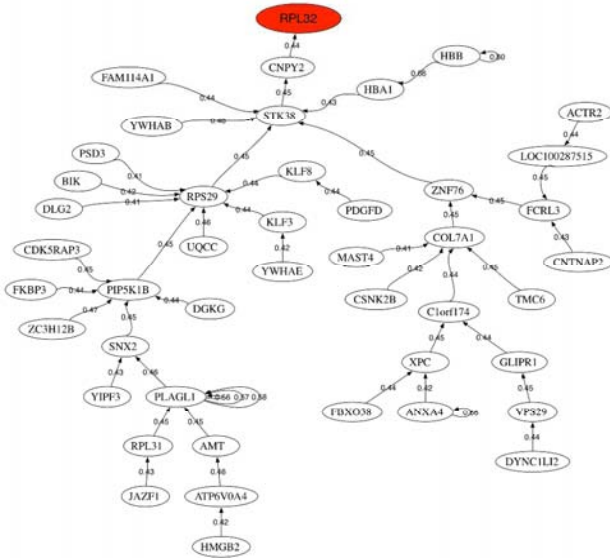
Correlation network of the most discriminatory genes for the NOTCH1 mutational status prediction. A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

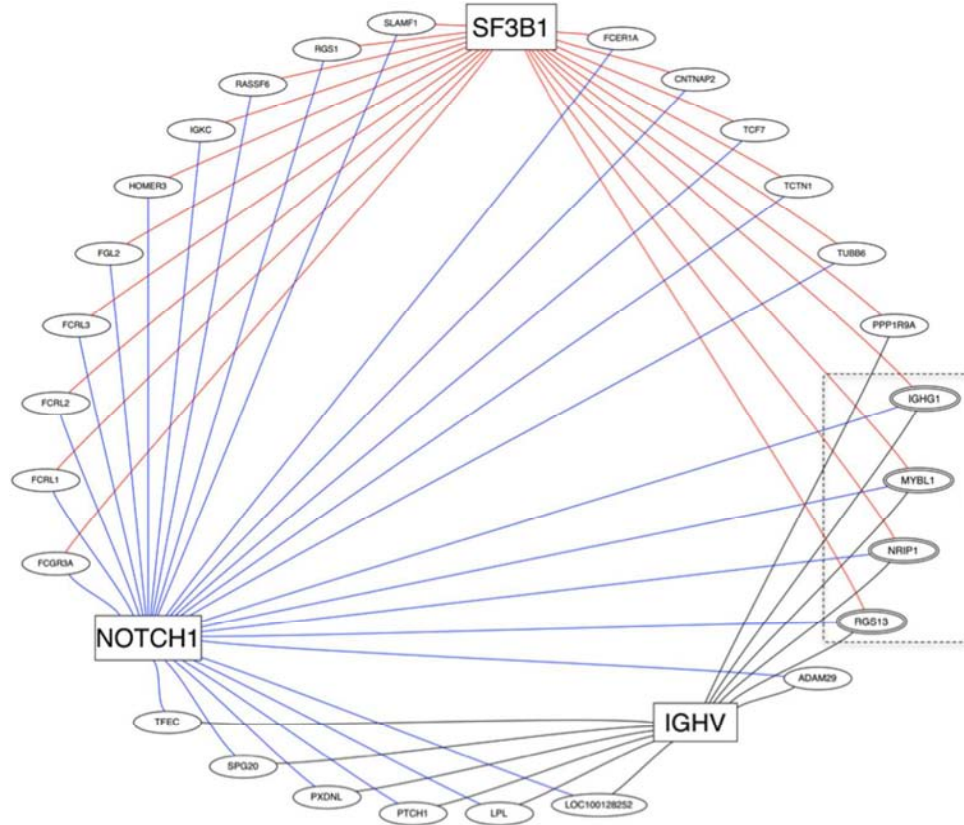
A) SF3B1 correlation network using Pearson correlation coefficient



B) SF3B1 correlation network using Normalized Mutual Information



Correlation network of the most discriminatory genes for the SF3B1 mutational status prediction. A) Using the Pearson correlation coefficient. B) Using the Normalized Mutual information.



Intersection among the most discriminatory genes of the IgVH, NOTCH1 and SF3B1 mutations. The three main mutations are represented with a rectangle and the most discriminatory genes are surrounded by ellipses. An edge represents that the gene appears as most discriminatory for a specific mutation. Genes with three edges (surrounded by a dot rectangle) are common to these three main mutations.

## Appendix

### 1. Gene discriminatory power

**Fisher's Ratio** (FR) uses the class information from the samples to find the set of genes that separate the centers of the distribution of both classes, keeping the dispersion in each class quite low (homogenous), that is, the method looks for very stable biomarkers. For that purpose, the Fisher's ratio of a gene  $j$ , in two-class problem,  $c_1, c_2$ , is defined as follows:

$$FR_j(c_1, c_2) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2},$$

where,  $\mu_{j1}, \mu_{j2}$  are measures of the center of the distribution (means or medians) of gene  $j$  in classes 1 and 2, and  $\sigma_{j1}^2, \sigma_{j2}^2$  are measures of the dispersion (variance) within these classes. Genes with the highest FR have the biggest discriminatory power of the phenotype and are expected to be involved in the genesis of the illness that is analyzed. Using the medians instead of the means makes the FR definition more robust to the presence of noise. Also FR penalizes noisy genes since their total variance  $\sigma_{j1}^2 + \sigma_{j2}^2$  increases.

The **Fold Change** (FC) of a gene  $j$  is defined as follows:

$$fc_j(c_1, c_2) = \log_2 \frac{\mu_{j1}}{\mu_{j2}}.$$

This method selects differentially expressed genes according to their absolute FC value,  $|fc_j(c_1, c_2)|$ . The fact that genes are differentially expressed does not imply that their discriminatory power in the phenotype prediction is high since the tails of both



distributions might still overlap introducing ambiguity in the discrimination. This happens in our case in the unmutated class where the differentially expressed genes typically exhibit higher expression variability, with respect to the mutated class, which is more homogeneous in expression.

## 2. Correlation networks

The **Pearson correlation coefficient**  $\rho_{ij}$  measures the linear correlation of two random variables. The formula for  $\rho_{ij}$  is:

$$\rho_{ij} = \frac{C(E_i, E_j)}{\sqrt{\text{var}(E_i)\text{var}(E_j)}},$$

where  $C(E_i, E_j)$  is the covariance between the expressions of two genes  $E_i, E_j$  considered as random variables and  $\text{var}(E_i)$  is the variance of the expression in gene  $i$ .

$\rho_{ij}$  is zero when the variables are uncorrelated, that is, linearly independent, and varies between -1 (negative correlation between expressions) and 1 (positive correlation).

The **Mutual Information**  $NM_{ij}$  of two random variables is a measure of mutual dependence of both variables. In our case we have used the normalized mutual information, which is similar to a correlation coefficient:

$$NM_{ij} = \frac{I(E_i, E_j)}{\sqrt{H(E_i)H(E_j)}},$$

where  $I(E_i, E_j)$  is the mutual information content and  $H(E_i)$  the entropy of gene  $i$  calculated based in the ordering of its expression with respect to the class assignment.

The mutual information  $I(E_i, E_j)$  content is calculated as follows:

$$I(E_i, E_j) = H(E_i) + H(E_j) - H(E_i \cup E_j),$$

being  $H(E_i \cup E_j)$  the joint entropy. This metric serves can be interpreted how much knowing one of these variables reduces uncertainty about the other.

The normalized mutual information can be interpreted as a correlation coefficient based exclusively in the ordering (entropy) in  $E_i$  and  $E_j$ . Nevertheless it varies between 0 (totally independent) and 1 (totally dependent):

$$NM_{ij} = 0 \Leftrightarrow H(E_i \cup E_j) = H(E_i) + H(E_j).$$

Therefore, the normalized mutual information is null when one variable does not reduce the uncertainty about the other, that is, they are independent descriptors.

## **A.5 Sensitivity analysis of gene ranking methods in phenotype prediction**

Under review in the Journal of Biomedical Informatics.

# Sensitivity analysis of gene ranking methods in phenotype prediction

Enrique J. deAndrés-Galiana<sup>1,2</sup>, Juan L. Fernández-Martínez<sup>1,§</sup>, and Stephen T. Sonis<sup>3</sup>

<sup>1</sup>Department of Mathematics, University of Oviedo, Spain.

<sup>2</sup>Artificial Intelligence Center, University of Oviedo, Spain.

<sup>3</sup>Biomodels, LLC, Watertown, MA, USA.

<sup>§</sup>Corresponding author

Corresponding author: JLFM, [jlfm@uniovi.es](mailto:jlfm@uniovi.es), 00-34 -985 103 199.

## Abstract

**Introduction:** It has become clear that noise generated during the assay and analytical processes has the ability to disrupt accurate interpretation of genomic studies. Not only does such noise impact the scientific validity and costs of studies, but when assessed in the context of clinically translatable indications such as phenotyping prediction, it can lead to inaccurate conclusions that could ultimately impact patients. We applied a sequence of ranking methods to dampen noise associated with microarray outputs, and then tested the utility of the approach in three disease indications using publically available datasets.

**Materials and Methods:** This study was performed in three phases. We first theoretically analyzed the effect of noise in phenotype prediction problems showing that it can be expressed as a modeling error that partially falsifies the pathways. Secondly, via synthetic modeling, we performed the sensitivity analysis for the main gene ranking methods to different types of noise. Finally, we studied the predictive accuracy of the gene lists provided by these ranking methods in synthetic data and in three different datasets related to cancer, rare and neurodegenerative diseases to better understand the translational aspects of our findings. **Results and Discussion:** In the case of synthetic modelling, we showed that Fisher's Ratio (FR) was the most robust gene ranking method in terms of precision for all the types of noise at different levels. Significance Analysis of Microarrays (SAM) provided slightly lower performance and the rest of the methods (fold change, entropy and maximum percentile distance) were much less precise and accurate. The predictive accuracy of the smallest set of high discriminatory probes was similar for all the methods in the case of Gaussian and Log-Gaussian noise. In the case of class assignment noise, the predictive accuracy of SAM and FR is higher. Finally, for real datasets (Chronic Lymphocytic Leukemia, Inclusion Body Myositis and Amyotrophic Lateral Sclerosis) we found that FR and SAM provided the highest predictive accuracies with the smallest number of genes. Biological pathways were found with an expanded list of genes whose discriminatory power has been established via FR. **Conclusions:** We have shown that noise in expression data and class assignment partially falsifies the sets of discriminatory probes in phenotype prediction problems. FR and SAM better exploit the principle of parsimony and are able to find subsets with less number of high discriminatory genes. The predictive accuracy and the precision are two different metrics to select the important genes, since in the presence of noise the most predictive genes do not completely coincide with those that are related to the phenotype. Based on the synthetic results, FR and SAM are recommended to unravel the biological pathways that are involved in the disease development.

**Keywords** Noise analysis, Machine learning, Gene expression, Cancer genomics.

## 1. INTRODUCTION

The revolution in molecular biology and the development of high-throughput technologies for sequencing in genetic and genomic analyses has generated an explosion in the amount of genetic data. These technologies, which have been firstly

52 applied in research, are now increasingly applied in translational medicine. Particularly,  
53 gene expression analysis through hybridization microarrays or RNA sequencing is now  
54 a conventional component in many areas of biomedical research. This kind of  
55 experiments has a very high under-determined character since the number of samples  
56 (patients) is much lower than the number of monitored probes (genes). Therefore, gene-  
57 ranking methods are needed to establish the discriminatory power of the genes in the  
58 phenotype prediction.

59 In this paper we first theoretically analyzed the effect of noise in phenotype prediction  
60 problems by casting them into abstract optimization problems. To accomplish this, we  
61 first show that noise in data can be expressed as a modeling error that partially falsifies  
62 the set of discriminatory probes that are phenotype-related, and therefore the biological  
63 pathways that are involved. Secondly, the sensitivity to different kind of noise (in  
64 expression and class assignment) for the main gene ranking methods (Fold Change,  
65 Fisher's Ratio, Percentile Distance and Entropy) compared to well-established  
66 Significance Analysis of Microarrays (SAM) [1] is performed via synthetic microarray  
67 modeling. This analysis has shown that in general terms Fisher's ratio is the most robust  
68 method in terms of precision closely followed by SAM. Besides, both methods provided  
69 the smallest sets with the highest discriminatory power. The effect of noise increases the  
70 number of genetic probes that are needed to slightly improve the predictive accuracy.  
71 Therefore, an optimum method to find the biological pathways in translational problems  
72 will consist of ranking the differential expressed genes decreasingly by their  
73 corresponding Fisher's ratio. The results of these analyses are confirmed using three  
74 different datasets concerning the study of cancer (Chronic Lymphocytic Leukemia), rare  
75 diseases (Inclusion Body Myositis) and neurodegenerative diseases (Amyotrophic  
76 Lateral Sclerosis). We found that FR and SAM provide the highest predictive accuracies

77 with the smallest number of genes, exploiting the principle of parsimony. Besides, we  
78 show their corresponding biological found with an expanded list of genes whose  
79 discriminatory power has been established via FR. In these three cases, the effect of  
80 viral infections in the corresponding pathways is clear. We expect that the results of this  
81 analysis will help optimize the use of these methods in translational medicine,  
82 particularly in the biological understanding of different diseases and in drug  
83 optimization problems.

84

## 85 **2. MATERIAL AND METHODS**

### 86 **2.1. The effect of noise in phenotype prediction**

87 One of the main obstacles in the analysis of genomic data is the absence of a conceptual  
88 model that relates the different genes/probes to the class prediction (phenotype).  
89 Machine-learning algorithms are therefore needed to model these complex  
90 relationships. For this reason, a classifier  $L^*(\mathbf{g})$  has to be constructed and it is defined as  
91 an application between the set of genetic signatures  $\mathbf{g}$  and the set of classes  
92  $C = \{c_1, c_2, \dots, c_n\}$  in which the phenotype is divided:

$$93 \quad L^*(\mathbf{g}): \mathbf{g} \in \mathcal{G} \rightarrow C = \{c_1, c_2, \dots, c_n\} \quad (1)$$

94 In most cases, the classification problem involved is binary.

95 The machine learning procedure is composed of two stages:

- 96 1. The learning process, that consists in giving a subset of samples  $\mathbf{T}$  (training data  
97 set) whose class vector is known,  $\mathbf{c}^{obs}$ , finding the subset of genetic signatures  $\mathbf{g}$   
98 that maximizes the learning accuracy, that is, the number of samples whose class is  
99 correctly predicted. This can be written as the result of the following optimization  
100 problem:

$$\begin{aligned}
O(\tilde{\mathbf{g}}) &= \min_{\mathbf{g} \in \mathbb{R}^S} O(\mathbf{g}), \\
O(\mathbf{g}) &= \left\| \mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs} \right\|_p, \\
\mathbf{L}^*(\mathbf{g}) &= (L^*(\mathbf{g}_1), \dots, L^*(\mathbf{g}_m)),
\end{aligned} \tag{2}$$

102 where  $\mathbf{L}^*(\mathbf{g})$  is the set of predicted classes,  $\mathbf{g}_i$  is the genetic signature  
103 corresponding to the sample  $i$  in the training dataset  $\mathbf{T}$ , and  $\left\| \mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs} \right\|_p$  stands for  
104 the percentage of corrected predicted classes with respect to the total number of  
105 samples in  $\mathbf{T}$ .

106 2. The generalization, that consists in predicting the class of a new sample ( $\mathbf{g}_{new}$ )  
107 whose class is unknown using the genetic signatures that have been found during  
108 the learning process.

109 One of the main numerical difficulties in learning is the high dimensionality of the  
110 genomic data since the number of monitored probes (or genes) is much greater than the  
111 number of samples (or patients). This fact provokes that the phenotype prediction in the  
112 learning stage will have a very high underdetermined character. Therefore, several gene  
113 lists with similar predictive accuracy might exist. This fact can be easily understood  
114 considering the classification as a parameter identification or inverse problem [2]: the  
115 topography of the cost function  $O(\mathbf{g})$  in the region of lower misfits (or higher  
116 predictive accuracies) corresponds to flat elongated valleys with null gradients where  
117 the high predictive genetic signatures are located. Obviously, the topography changes if  
118 the space where the optimization is performed ( $\mathbb{R}^S$ ) changes. All these high predictive  
119 lists are expected to be involved in the genetic pathways that explain the phenotype. The  
120 smallest-scale signature is the one that has the least number of discriminatory genes. In  
121 practice, the predictive accuracy of a genetic signature,  $O(\mathbf{g})$ , is performed via cross-

122 validation. This knowledge could be very important for early diagnosis and treatment  
123 optimization.

124 The presence of noise in the genomic data will impact the classification and obviously  
125 the pathway analysis resulting from this procedure. There are at least two main sources  
126 of noise in phenotype prediction problems:

127 • **Noise in the gene expression** induced by the process of measurement. In this  
128 case, the observed genetic expression of a sample,  $\mathbf{g}^{obs}$ , can be expressed as the  
129 sum of the true genetic expression array,  $\mathbf{g}^{true}$ , and the measurement noise,  $\delta\mathbf{g}$ :  
130  $\mathbf{g}^{obs} = \mathbf{g}^{true} + \delta\mathbf{g}$ . Therefore, using a simple Taylor expansion we get:

$$131 \quad L^*(\mathbf{g}^{obs}) = L^*(\mathbf{g}^{true}) + \delta L^*(\mathbf{g}^{true}) = L^*(\mathbf{g}^{true}) + \sum_{k=1}^s \frac{\partial L^*}{\partial g_k}(\mathbf{g}^{true}) \delta g_k + o(\delta\mathbf{g}), \quad (3)$$

132 where  $o(\delta\mathbf{g})$  vanishes when the noise term  $\delta\mathbf{g} \rightarrow \mathbf{0}$ . Obviously, this analysis is  
133 theoretical because  $\mathbf{g}^{true}$  and  $\delta\mathbf{g}$  are unknown.

134 • **Noise in the class assignment** since some samples could be wrongly annotated or  
135 might belong to a different class, not yet discovered. Naming  $\mathbf{c}^{true}$  the true class  
136 assignment array and  $\delta\mathbf{c}$  the noise in the class assignment, then the observed  
137 class array will be  $\mathbf{c}^{obs} = \mathbf{c}^{true} + \delta\mathbf{c}$ .

138 In presence of these types of noise, the genetic signature with the highest predictive  
139 score will never perfectly coincide with the genetic signature that explains the disease.

140 Both types of noises ( $\delta\mathbf{g}$ ,  $\delta\mathbf{c}$ ) induce a modeling error, ( $\delta\mathbf{L}^*(\mathbf{g})$ ), in the classifier  
141 related to the phenotype prediction.

142 In the case of class assignment noise the cost function writes:



143 
$$\begin{aligned} O^p(\mathbf{g}) &= \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p = \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true} - \delta\mathbf{c}\|_p = \\ &= \|\mathbf{L}^*(\mathbf{g})\|_p + \delta\mathbf{L}^*(\mathbf{g}) = O^t(\mathbf{g}) + \delta\mathbf{L}^*(\mathbf{g}), \end{aligned} \quad (4)$$

144 where  $O^p(\mathbf{g})$ ,  $O^t(\mathbf{g})$  stand respectively for the perturbed and noise-free cost functions,  
145 and  $\delta\mathbf{L}^*(\mathbf{g})$  for the modeling error term induced by the noise in the class assignment.

146 For instance, if the squared Euclidean norm is used to define the cost function, we have:

147 
$$\begin{aligned} O^p(\mathbf{g}) &= \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_2^2 = \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true} - \delta\mathbf{c}\|_2^2 = \\ &= (\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true} - \delta\mathbf{c})^T (\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true} - \delta\mathbf{c}) = \\ &= \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true}\|_2^2 - 2(\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true})^T \delta\mathbf{c} + \delta\mathbf{c}^T \delta\mathbf{c}. \end{aligned} \quad (5)$$

148 Therefore the modeling error is:

149 
$$\delta\mathbf{L}^*(\mathbf{g}) = \delta\mathbf{c}^T \delta\mathbf{c} - 2(\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{true})^T \delta\mathbf{c}, \quad (6)$$

150 and  $\delta\mathbf{L}^*(\mathbf{g}) \rightarrow \mathbf{0}$  when  $\delta\mathbf{c} \rightarrow \mathbf{0}$ .

151 What's more, the classifier, which is constructed ad-hoc and is a priori unknown,  
152 induces modeling errors into the classification. Therefore, the automatic conclusion is  
153 that  $O^p(\mathbf{g})$  and  $O^t(\mathbf{g})$  will never achieve their corresponding minima for the same  
154 genetic signatures ( $\mathbf{g}$ ). For that reason, it is also desirable to inspect the genetic  
155 signatures having a lower predictive accuracy than the optimum.

156 To alleviate the high underdetermined character of genomic-phenotype prediction  
157 problems, feature selection methods are used to reduce the dimensionality of the  
158 genomic data. The problem of determining the genes that separate two (or more) classes  
159 corresponding to given phenotypes has been traditionally been addressed by filter,  
160 wrapper and embedded methods [3]. In the case of filter methods, the gene selection  
161 and the classifier for phenotype prediction are independent (uncoupled). Wrapper and  
162 embedded techniques are most sophisticated approaches where the gene selection is the  
163 solution of an optimization problem; therefore selection and classification are coupled.

164 Wrapper and embedded methods usually involve the use of neural network, support  
165 vector machines, decision trees and global optimization algorithms. Filter methods rank  
166 different genes according to different measures of their discriminatory power in  
167 phenotype prediction problems.

168

## 169 **2.2. Gene selection ranking methods and noise**

170 To determine the stability and robustness of supervised ranking algorithms in mitigating  
171 microarray-generated noise, we compared the Fisher's ratio (FR), Fold Change (FC),  
172 Percentile Distance (PD), Entropy (EN) and SAM (Significance Analysis of  
173 Microarrays) using a synthetic dataset and publically available datasets associated with  
174 B-chronic lymphocytic leukemia, inclusion body myositis, and amyotrophic lateral  
175 sclerosis. At a translational level, the aim of this analysis is to establish an optimum  
176 way to find the most discriminatory genes in a phenotype prediction and the biological  
177 pathways that are involved.

178 **Fisher's ratio** uses class information from the samples to find the set of genes that  
179 separate the centers of the distribution of both classes, keeping the dispersion in each  
180 class quite low (homogenous), that is, the FR method looks for very stable biomarkers  
181 along classes. The FR of a gene  $j$  in two-class problem,  $c_1, c_2$ , is defined as follows [4]:

$$182 \quad FR_j(c_1, c_2) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \quad (7)$$

183 where,  $\mu_{j1}, \mu_{j2}$ , are measures of the center of the distribution (means or medians) of the  
184 gene  $j$  in classes 1 and 2, and  $\sigma_{j1}^2, \sigma_{j2}^2$  are measures of the dispersion (variance) within  
185 these classes. As a result, high FR values correspond to high discriminatory genes of the  
186 phenotype and are expected to be involved in the genesis of the illness that is analyzed.  
187 Using the medians instead of means makes the FR definition more robust against noise.

188 Noisy genes are penalized in gene selection by FR since uncorrelated white noise in the  
189 expression increases the standard deviation, as shown:

$$190 \quad E_j^p = E_j^t + n_j, \quad (8)$$

191 where  $E_j^p$  is the expression of gene  $j$  affected by the noise  $n_j$  which is independent of  
192 the true gene expression  $E_j^t$ . Then:

$$193 \quad \begin{aligned} \mu_j^p &= \mu_j^t + \mu_{n_j}, \\ \sigma_j^p &= \sigma_j^{true} + \sigma_{n_j}. \end{aligned} \quad (9)$$

194 The presence of uncorrelated white noise increases the variance of the expression,  
195 decreasing the FR of the noisiest genes. The genes are therefore penalized in the  
196 selection by FR.

197 The **Fold Change** of a gene  $j$  is defined as follows:

$$198 \quad fc_j(c_1, c_2) = \log_2 \frac{\mu_{j1}}{\mu_{j2}}. \quad (10)$$

199 This method selects differentially expressed genes according to their absolute FC value  
200  $|fc_j(c_1, c_2)|$ . FC analysis is sensible to noise that obviously affects the mean of the  
201 expression in different ways, generating numerical artifacts of under-expression and  
202 over-expression. Furthermore, the fact that genes are differentially expressed does not  
203 imply that their discriminatory power is high, since the tails of both distributions might  
204 still overlap introducing ambiguity in the discrimination. Due to this fact, the volcano  
205 plot [1, 5, 6] has been introduced, combining a statistical test ( $p$ -value) with FC and  
206 enabling for quick visual identification of statistically significant and large-magnitude  
207 changes. Nevertheless, the additional  $p$ -value discrimination does not solve the problem  
208 of distribution tails overlapping.

209 **Percentile Distance** [7] selects genes with higher distance between the corresponding  
 210 cumulative probability functions (percentile array) within each class. The percentile  
 211 distance for attribute  $j$  is defined as follows:

$$212 \quad d_j(c_1, c_2) = \frac{\|\mathbf{p}_{j1} - \mathbf{p}_{j2}\|_2}{\max(\|\mathbf{p}_{j1}\|_2, \|\mathbf{p}_{j2}\|_2)}; j = 1, \dots, N_{att}. \quad (11)$$

213 where  $\mathbf{p}_{j1}, \mathbf{p}_{j2}$  stand for the percentile vector of the gene  $j$  in classes 1 and 2. Percentiles  
 214 vary from 5 to 95 (or 10 to 90) to avoid the possible effect of outliers. Variables with  
 215 higher percentile distances between the corresponding discrete cumulative probability  
 216 functions (percentile array) in each class will be selected. The percentile array takes into  
 217 account both, the distance between the centers of the distributions (medians or  
 218 percentile 50%), the interquartile range (percentiles 25% and 75%), and also the  
 219 distance between the tails of the distributions (percentiles 5-10% and 90-95%).

220 The **Entropy** selection method takes into account the ordering mismatch in the  
 221 expression of a gene with respect to the class vector for a given phenotype [8, 9].  
 222 Therefore, lower entropies imply in principle bigger discriminatory power. Entropy is  
 223 affected by noise to the extent that this affects the ordering of the expression signature  
 224 with respect to the one provided by the phenotype class.

225 Finally, **Significance Microarray Analysis** (SAM) [1] uses as score the absolute  
 226 difference between the means in both classes divided by the sum of the total standard  
 227 deviation ( $\sigma_j^T$ ) and a tunable exchangeability factor ( $\sigma_{j0}$ ) used to damp the effect of  
 228 outliers, that is, genes with very small  $\sigma_j^T$  that will bring an anomalous score:

$$229 \quad SAM_j(c_1, c_2) = \frac{|\mu_{j1} - \mu_{j2}|}{\sigma_j^T + \sigma_{j0}}. \quad (12)$$

230

231 A variety of analyses have been performed to study the sensitivity of some of these

232 methods to noise in the expression data [10-13]. However, so far the robustness against  
233 different kind of noises for all these ranking methods has not been addressed.

234 For that purpose, we used a synthetic dataset where three different types of noise were  
235 introduced: additive Gaussian noise, lognormal noise and noise in the class assignment.

236 The Gaussian noise has been introduced through a random number generator following  
237 a normal distribution  $n_j \rightarrow N(0, r_k E_j^t)$  for each gene, being  $r_k$  the noise level, and  $E_j^t$   
238 is the noise-free expression of the gene  $j$ . Therefore, the noisy expression corresponding  
239 to the gene  $j$  would be:

$$240 \quad E_j^p = E_j^t + n_j. \quad (13)$$

241 The lognormal noise has been obtained by adding Gaussian noise to the logarithms of  
242 the expression:

$$243 \quad \ln_j = \log_2 s_j \rightarrow N(0, r_k \log_2 E_j^t). \quad (14)$$

244 Therefore, the lognormal noise has a scaling effect, since:

$$245 \quad \begin{aligned} \log_2 E_j^p &= \log_2 E_j^t + \ln_j, \\ E_j^p &= s_j E_j^t. \end{aligned} \quad (15)$$

246 In the case of class assignment noise, a given number of samples are misclassified. The  
247 class assignment and lognormal noises belong to the category of non-Gaussian noise.

248 The synthetic dataset was built with a predefined number of differentially expressed  
249 genes. We subsequently introduced different levels of noise: 1 to 6% for Gaussian and  
250 log-Gaussian noises and 10 to 40% for the class assignment noise.

251 To check the performance of the different ranking methods we used the *Precision*  
252 metric:

$$253 \quad \text{Precision} = \frac{|\{DE\_genes\} \cap \{Selected\_genes\}|}{|\{Selected\_genes\}|}, \quad (16)$$

254 where  $\{DE\_genes\}$  is the set of the differentially expressed genes and  
255  $\{Selected\_genes\}$  the set of genes selected by the ranking algorithm.

256

### 257 2.3 The synthetic and disease datasets

258 A flow diagram for the methodology used in this paper is shown in figure 1.

259 The synthetic datasets was created to compare the various filtering methods against a  
260 known dataset and then, based on these findings, create a hierarchy which defines the  
261 effectiveness of the ranking methods against different kind of noise and to understand  
262 how to find optimally the biological pathways in disease datasets.

263 The synthetic dataset was built simulating a real dataset related to Chronic Lymphocytic  
264 Leukemia (163 samples and 48807 probes) [14] using the OC-plus package available  
265 for The Comprehensive R Archive Network [15]. The original data was compound of  
266 163 samples and 48807 probes. We have chosen this dataset for building the synthetic  
267 dataset because it has a good sample size and the class is well balanced. The experiment  
268 was set up as follows:

- 269 1. The class of the synthetic dataset was the same as the one observed for the *IgVH*  
270 status [14]: 92 vs 71.
- 271 2. The noise-free synthetic data set (expression) was generated using as main  
272 parameters  $D = 2$  and  $P_0 = 0.47$ , where  $D$  is the effect size for differentially  
273 expressed genes expressed in units of the gene-specific standard deviation and  $P_0$   
274 is the proportion of differentially expressed genes. This simulation made 229  
275 genes be differentially expressed which we will try to recover via the different  
276 gene-ranking methods. These genes are supposed in the synthetic dataset to  
277 optimally differentiate the known *IgVH* status.

278

279 Furthermore, we have modeled different real microarray datasets to confirm these  
280 findings:

281 1. B-cell Chronic Lymphocytic Leukemia (CLL) dataset composed by 163 samples  
282 and 48807 probes [14]. CLL is a complex and molecular heterogeneous disease  
283 which is the most common adult Leukemia in western countries. DNA analyses  
284 served to distinguish two major types of CLL with different survival times based  
285 on the maturity of the lymphocytes, as discerned by the Immunoglobulin Heavy  
286 chain Variable-region (*IgVH*) gene mutation status. 92 samples had the *IgVH*  
287 gene mutated versus 71 samples with worse prognosis. The aim of this analysis  
288 is to find the pathways that are associated with bad prognosis in CLL patients.

289 2. Inclusion Body Myositis (IBM): microarray studies (with 22283 probes) were  
290 performed on muscle biopsy specimens from 34 patients with inclusion body  
291 myositis and 11 samples without neuromuscular disease [16]. IBM is a muscle  
292 disease characterized by chronic, progressive muscle inflammation accompanied  
293 by muscle weakness. The aim of this analysis is to find the pathways that are  
294 associated to the development of IBM with respect to healthy controls.

295 3. Amyotrophic Lateral Sclerosis (ALS) dataset composed by 85 samples (57  
296 samples are ALS cases and 28 healthy controls) and 54675 probes [17]. ALS is a  
297 fatal neurodegenerative disease characterized by progressive loss of motor  
298 neurons. These authors have shown that the co-stimulatory pathway is  
299 upregulated in the blood of a high percentage of human patients with ALS (56%).  
300 The aim of this analysis is to define the genes that are associated with a diagnosis  
301 of ALS, the possible causes and the biological pathways that are involved.

302

303 These datasets are representative of 3 different types of diseases: cancer, rare and

304 neurodegenerative diseases. Besides, they have a reasonable sample size and a good  
305 balance between both classes in each case. Although all the microarray datasets treated  
306 herein are post processed via the RMA algorithm that performs an estimation and  
307 correction of the noise [18], noise is still present due to the complexity of the data  
308 acquisition. Because the genes which are differentially expressed in real datasets are  
309 unknown, we have applied the methodology explained in [19] to select the smallest  
310 subset of high discriminatory probes. In summary, the method consists in ranking the  
311 genes according to their discriminatory power, selecting different lists of genes through  
312 Backwards Feature Elimination (BFE) and establishing their predictive accuracy via  
313 Leave-One-Out-Cross-Validation (LOOCV). These datasets have also been modeled  
314 through biomedical robots [20] that exploit the uncertainty space of the classifiers in  
315 phenotype prediction problems. Therefore, this paper is also useful for analyzing the  
316 sensitivity towards noise of the main ranking methods used in the design of biomedical  
317 robots.

318

### 319 **3. RESULTS AND DISCUSSION**

#### 320 **a. Synthetic data**

321 In order to compare the performance of each method we calculated the precision for  
322 each method, considering the set of 229 genes that were differentially expressed in the  
323 synthetic dataset. Table 1 provides the precision for all the ranking methods mentioned  
324 above for different noise types and levels. Table 2 shows the LOOCV mean accuracy  
325 and the number of selected genes in each method. What's more, we have also calculated  
326 the empirical Cumulative Distribution Functions (CDF) of the positions of the  
327 differentially expressed genes captured by each method. For the sake of clearness we  
328 only used the first 1000 gene positions. A perfect CDF would be a straight line reaching



329 the value of 1 at position 229. Figure 2, 3 and 4 shows these CDF curves for every type  
330 of noise and noise level.

331 It can be observed the following:

332 1. As we expected, the precision decreases for all the methods as the noise level  
333 increases (refer to Table 1). The FR provides the best precision score for all the  
334 noise types and levels. These differences decrease very fast with the noise level  
335 in the case of lognormal noise. The precision figures for SAM, in some cases,  
336 are very close to FR. In the case of class assignment noise FR keeps precision  
337 levels up to 90% for 10 to 25% of noise, showing a very good robustness against  
338 this type of noise (Table 1). This result has an important translational impact in  
339 real datasets to find the biological pathways that are involved in the disease  
340 development.

341 2. The differences in the LOOCV mean accuracy (table 2) is not so clear and all  
342 methods provide similar results for the three types of noise at the different levels  
343 in the expenses of increasing the number of probes needed to improve the  
344 LOOCV predictive accuracy. In the case of Gaussian noise, SAM and FR show  
345 very similar results obtaining 100% of predictive accuracy with a much more  
346 reduced set of selected probes. Regarding lognormal noise, entropy seems to be  
347 the best for lower level of noises, while SAM and FR behave better when the  
348 noise level increases. FR and SAM are the best methods with a very little  
349 difference between them in the case of class assignment noise. These  
350 conclusions can also be clearly observed in the CDF curves (figures 2 to 4).

351 We have also combined the Gaussian noise and the Log-Gaussian noise with the  
352 noise in the class assignment obtaining similar results. Adding the class  
353 assignment noise to a noisy dataset (for both Gaussian and Log-Gaussian noises)

354 affects much more in finding the differentially expressed genes since the  
355 Precision decreases drastically (see supplementary material Table 1). What is  
356 interesting is that the fold change seems to work better in terms of precision for  
357 a combination of class assignment and log-Gaussian noise. In terms of  
358 predictive accuracy more genes are needed to have a high predictive accuracy  
359 when class assignment noise is present (see supplementary material, Table 2). In  
360 this case, FR and SAM provide the best results. Furthermore, it is possible to  
361 observe that for high levels of noise we can achieve high predictive accuracy  
362 with null precision at the expenses of adding a lot of genes to the predictive  
363 genetic signature. In this case, the biological pathways are clearly falsified.

364

365 In conclusion, noise in class assignment affects the selection of the important  
366 discriminatory genes in phenotype prediction problems more than noise in the  
367 expression data. This result emphasizes the importance in translational medicine of  
368 having at disposal a correct class assignment of the samples, provided by the doctors.

369

#### 370 **b. Disease datasets**

371 Concerning the real datasets we selected the smallest subset of high discriminatory  
372 probes using the methodology described in [19]. Table 3 shows the mean accuracy and  
373 number of selected probes for each ranking method and dataset. For these three datasets  
374 we achieved accuracies higher than 90% with a very small subset of probes. The  
375 selection was performed via backwards feature elimination and a nearest-neighbor  
376 based algorithm through the LOOCV accuracy [19].

377 In the case of CLL, the difference between all the methods is very small. The entropy  
378 method achieved 94% of accuracy with 99 probes. However SAM got almost 94% of

379 accuracy with 26 probes and FR 93% with only 6 probes. High discriminatory genes of  
380 the *IgVH* phenotype include: LPL, CRY1, LOC100128252, SPG20, ZBTB20, NRIP1,  
381 ZAP-70, LDOC1, COBLL1 and NRIP1.

382 The pathway analysis has revealed the importance of the Inflammatory Response, the  
383 PAK pathway and the ERK signaling super pathway that includes ERK signaling, ILK  
384 signaling, MAPK signaling, Molecular Mechanisms of cancer and Rho Family GTPases  
385 pathway. These pathways control Proliferation, Differentiation, Survival and Apoptosis.  
386 Also, other important pathways found were Allograft Rejection, the Inflammatory  
387 Response Pathway, CD28 Co-stimulation, TNF-alpha/NF-kB Signaling Pathway, Akt  
388 Signaling, PAK Pathway and TNF Signaling. The presence of some of these pathways  
389 opens the hypothesis of viral infection as a cause of CLL. Figure 5 shows the  
390 correlation tree between the most discriminatory genes found by FR that provided the  
391 highest precision in synthetic modeling. Note that the most important branch is  
392 associated to the connection LPL/ZBTB 20.

393 Regarding the IBM dataset, we found that SAM and FR were able to correctly predict  
394 97% of the samples just with 2 and 1 probes respectively. Differences between SAM  
395 and FR and other methods are remarkable. The list of most discriminatory genes of the  
396 IBM phenotype include: HLA-C, HLA-B, TMSB10, S100A6, HLA-G, STAT1, TIMP1,  
397 HLA-F, IRF9, BID, MLLT11 and PSME2. Note the presence of different HLA-x genes  
398 of major histocompatibility. Particularly, the function of the gene HLA-B would explain  
399 alone the genesis of IBM: "HLA-B (major histocompatibility complex, class I, B) is a  
400 human gene that provides instructions for making a protein that plays a critical role in  
401 the immune system. HLA-B is part of a family of genes called the human leukocyte  
402 antigen (HLA) complex. The HLA complex helps the immune system to distinguish the  
403 body's own proteins from proteins made by foreign invaders such as viruses and

404 bacteria". The analysis of biological pathways has revealed the importance of viral  
405 infections, mainly in IBM patients: Allograft Rejection, Influenza A, Class I MHC  
406 Mediated Antigen Processing and Presentation, Staphylococcus Aureus Infection,  
407 Interferon Signaling, Immune Response IFN Alpha/beta Signaling Pathway,  
408 Phagosome, Tuberculosis, Cell Adhesion Molecules (CAMs), Epstein-Barr Virus  
409 Infection, and TNF Signaling. We can see several viral infections in this list. It is  
410 interesting to remark that 75% of the cases of viral myositis are due to Staphylococcus  
411 Aureus infection [21]. Figure 6 shows the correlation tree for the IBM phenotype,  
412 indicating the importance of the major histocompatibility gene family.

413 Finally, in the case of ALS dataset, SAM reached an accuracy of 95% with 42 probes,  
414 whilst FR and PD got a 94% with 12 and 17 probes respectively. High discriminatory  
415 genes of the ALS phenotype include: CASP1, ZNF787 and SETD7. The pathway  
416 analysis has revealed the importance of the GPCR Pathway, RhoA Signaling Pathway,  
417 EPHB Forward Signaling, EphrinA-EphR Signaling, EBV LMP1 Signaling, and  
418 Regulation of Microtubule Cytoskeleton. These pathways have different important  
419 signaling roles and suggest a possible link to the Epstein-Barr virus (EBV). Finally  
420 Figure 7 shows the correlation tree between the most discriminatory genes found by FR,  
421 highlighting the link between caspases, zinc finger proteins and the gene NKAP that  
422 encodes a protein that is involved in the activation of the ubiquitous transcription factor  
423 NF-kappaB. The activation of caspases plays a central role in cell apoptosis and  
424 activates interleukin-1, a cytokine involved in the processes such as inflammation.  
425 Caspases have been also associated to the pathogenesis of Huntington disease.

426

#### 427 **4. CONCLUSION**

428 We have theoretically showed that noise in expression data and class assignment

429 partially falsifies the sets of discriminatory probes in phenotype prediction problems.  
430 Via synthetic modeling we have shown that FR and SAM are the most robust gene  
431 selection methods for different kind of noises. Besides, FR and SAM seem to exploit  
432 the parsimony principle and are able to find the smallest-scale high discriminatory gene  
433 signature. Nevertheless, SAM is much more computationally expensive than FR while  
434 the achieved results are similar. We have also found that noise in class assignment  
435 affect the predictive accuracy and the precision much more than noise in the expression  
436 data. Nevertheless, the No-Free-Lunch Theorem in search and optimization [22] states  
437 that all these algorithms are needed to understand the complex relationships hidden in  
438 the genomic datasets. Therefore, the prior knowledge provided by the doctors is of  
439 paramount importance in the search for solutions of the different diseases. From the  
440 translational point of view this analysis shows the importance of establishing the  
441 discriminatory power of the genes in phenotype prediction problems to correctly find  
442 the biological pathways that are involved. To accomplish this task in the most efficient  
443 way possible, suggested in this paper, we suggest ranking the most differentially  
444 expressed genes according to their Fisher' ratio (or SAM ratio). Examples to cancer  
445 (CLL), rare (IBM) and neurodegenerative diseases (ALS) are also outlined in this paper  
446 obtaining very interesting conclusions that might imply an important role of several  
447 viral infections.

448

#### 449 **Acknowledgements**

450 Enrique J. de Andrés salary was supported by the Spanish Ministerio de Economía y  
451 Competitividad (grant TIN2011-23558). No other financial support has been received to  
452 perform this retrospective analysis. We acknowledge Celia Fernández-Brillet for style  
453 corrections.

454

455 **Author Disclosure Statement**

456 No competing financial interests exist.

457

458

459 **REFERENCES**

- 460 [1] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the  
461 ionizing radiation response. *Proc Natl Acad Sci USA* 2001, 98(9): 5116-21.
- 462 [2] Fernández-Martínez JL, Fernández-Muñiz Z, Tompkins MJ. On the topography of  
463 the cost functional in linear and nonlinear inverse problems. *Geophysics* 2012; 77(1):  
464 1-15.
- 465 [3] Saeys Y, Inza In, Larrañaga P. A review of feature selection techniques in  
466 bioinformatics. *Bioinformatics* 2007; 23(19): 2507–17.
- 467 [4] Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of*  
468 *Eugenics* 1936, 7(7): 179–88.
- 469 [5] Cui X, Churchill GA. Statistical tests for differential expression in cDNA  
470 microarray experiments. *Genome Biol.* 2003; 4(4): 210.
- 471 [6] Li W1. Volcano plots in analyzing differential expressions with mRNA microarrays.  
472 *J Bioinform Comput Biol.* 2012; 10(6): 1231003.
- 473 [7] deAndrés-Galiana EJ, Fernández-Martínez JL, Luaces O, et al. On the prediction of  
474 Hodgkin lymphoma treatment response. *Clin Transl Oncol* 2015, 17(8):612-9.
- 475 [8] Shannon C. A mathematical theory of communication. *Bell System Technical*  
476 *Journal* 1948; 27: 379–423, 623.
- 477 [9] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA:  
478 Morgan Kaufmann Publishers Inc; 1993.
- 479 [10] Kooperberg C, Fazio T, Delrow J, Tsukiyama T. Improved background correction  
480 for spotted DNA microarrays. *J Comput Biol* 2002, 9:55-66.
- 481 [11] Larsson O1, Wahlestedt C, Timmons JA. Considerations when using the  
482 significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics* 2005, 6:  
483 129.
- 484 [12] Jeffery I, Higgins D, Culhane A: Comparison and evaluation of methods for  
485 generating differentially expressed gene lists from microarray data. *BMC*  
486 *Bioinformatics* 2006, 7(1): 359.
- 487 [13] Dinu I, Potter JD, MuellerT, et al. Improving gene set analysis of microarray data  
488 by SAM-GS. *BMC Bioinformatics* 2007, 8: 242.
- 489 [14] Ferreira PG, Jares P, Rico D, et al. Transcriptome characterization by RNA  
490 sequencing identifies a major molecular and clinical subdivision in chronic  
491 lymphocytic leukemia. *Genome Res* 2014, 24(2): 212–26.
- 492 [15] Pawitan Y and Ploner A. OCplus: Operating characteristics plus sample size and  
493 local fdr for microarray experiments. R package version 1.40.0.
- 494 [16] Greenberg SA, Bradshaw EM, Pinkus JL, et al. Plasma cells in muscle in inclusion  
495 body myositis and polymyositis. *Neurology.* 2005 Dec 13;65(11):1782-7.
- 496 [17] Lincecum JM, Vieira FG, Wang MZ, et al. From transcriptome analysis to  
497 therapeutic anti-cd40l treatment in the sod1 model of amyotrophic lateral sclerosis.  
498 *Nat Genet* 2010, 42(5):392–399.
- 499 [18] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries  
500 of high density oligonucleotide array probe level data. *Biostatistics.* 2003, 4(2):249-  
501 64.

- 502 [19] Saligan LN, Fernández-Martínez JL, deAndrés-Galiana EJ, Sonis S. Supervised  
503 classification by filter methods and recursive feature elimination predicts risk of  
504 radiotherapy-related fatigue in patients with prostate cancer. *Cancer Inform* 2014,  
505 13:141–152.
- 506 [20] deAndrés-Galiana EJ, Fernández-Martínez JL, Sonis S. Design of biomedical  
507 robots for phenotype prediction problems. *Journal of Computational Biology*, 2016.  
508 Accepted for publication.
- 509 [21] Fayad LM, Carrino JA, and Fishman EK (2007). Musculoskeletal infection: role of  
510 CT in the emergency department. *Radiographics*, 27(6):1723–1736.
- 511 [22] Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans*  
512 *Evol Comput* 1997, 1:67–82.
- 513



514 **TABLE CAPTIONS**

		1%	2%	3%	4%	5%	6%
GAUSSIAN	FR	<b>1.00</b>	<b>0.97</b>	<b>0.86</b>	<b>0.72</b>	<b>0.65</b>	<b>0.55</b>
	FC	0.64	0.64	0.61	0.56	0.53	0.46
	EN	0.85	0.75	0.68	0.55	0.5	0.43
	PD	0.28	0.31	0.32	0.34	0.34	0.34
	SAM	0.94	0.91	0.80	0.67	0.60	0.51
LOG-GAUSSIAN	FR	<b>0.84</b>	<b>0.62</b>	<b>0.41</b>	<b>0.26</b>	<b>0.21</b>	<b>0.16</b>
	FC	0.60	0.54	0.38	<u>0.27</u>	0.23	<b>0.16</b>
	EN	0.67	0.45	0.31	0.18	0.12	0.10
	PD	0.32	0.36	0.28	0.24	0.19	0.14
	SAM	0.79	0.57	0.38	0.25	0.20	0.15
		10%	15%	20%	25%	30%	35%
CLASS	FR	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>0.94</b>	<b>0.68</b>	<b>0.40</b>
	FC	0.53	0.52	0.41	0.29	0.25	0.16
	EN	0.87	0.88	0.82	0.77	0.50	0.32
	PD	0.27	0.26	0.22	0.19	0.18	0.12
	SAM	0.94	0.94	0.93	0.88	0.64	0.37

515 **Table 1:** Synthetic modeling. Precision for each of the noise types at different noise  
516 levels.

517

		1%	2%	3%	4%	5%	6%
GAUSSIAN	FR	100.00 / 8	<b>100.00 / 5</b>	<b>100.00 / 6</b>	100.00 / 5	100.00 / 13	<b>100.00 / 8</b>
	FC	100.00 / 9	100.00 / 12	100.00 / 12	100.00 / 9	<b>100.00 / 9</b>	100.00 / 12
	EN	100.00 / 17	100.00 / 11	100.00 / 8	100.00 / 12	100.00 / 19	100.00 / 28
	PD	100.00 / 21	100.00 / 19	100.00 / 23	100.00 / 17	100.00 / 17	100.00 / 22
	SAM	<b>100.00 / 6</b>	<b>100.00 / 5</b>	<b>100.00 / 6</b>	<b>100.00 / 4</b>	100.00 / 14	<b>100.00 / 8</b>
LOG-GAUSSIAN	FR	100.00 / 6	100.00 / 22	100.00 / 47	<b>100.00 / 29</b>	100.00 / 37	<b>100.00 / 88</b>
	FC	100.00 / 9	100.00 / 16	100.00 / 48	<b>100.00 / 29</b>	100.00 / 37	100.00 / 119
	EN	<b>100.00 / 4</b>	<b>100.00 / 14</b>	<b>100.00 / 24</b>	100.00 / 38	100.00 / 45	100.00 / 132
	PD	100.00 / 22	100.00 / 23	100.00 / 111	100.00 / 37	100.00 / 46	100.00 / 128
	SAM	100.00 / 8	100.00 / 18	100.00 / 47	<b>100.00 / 29</b>	<b>100.00 / 28</b>	100.00 / 90
		10%	15%	20%	25%	30%	35%
CLASS	FR	90.18 / 3	85.28 / 14	<b>83.44 / 2</b>	<b>76.07 / 4</b>	<b>73.62 / 2</b>	69.94 / 213
	FC	90.18 / 10	84.66 / 8	80.98 / 188	76.07 / 52	72.39 / 183	<b>69.94 / 85</b>
	EN	90.18 / 25	85.28 / 28	81.60 / 18	75.46 / 2	73.62 / 3	71.17 / 4
	PD	90.80 / 121	85.89 / 180	80.98 / 29	75.46 / 23	71.17 / 33	66.87 / 46
	SAM	<b>90.80 / 5</b>	<b>85.28 / 4</b>	<b>83.44 / 2</b>	<b>76.07 / 4</b>	<b>73.62 / 2</b>	69.33 / 5

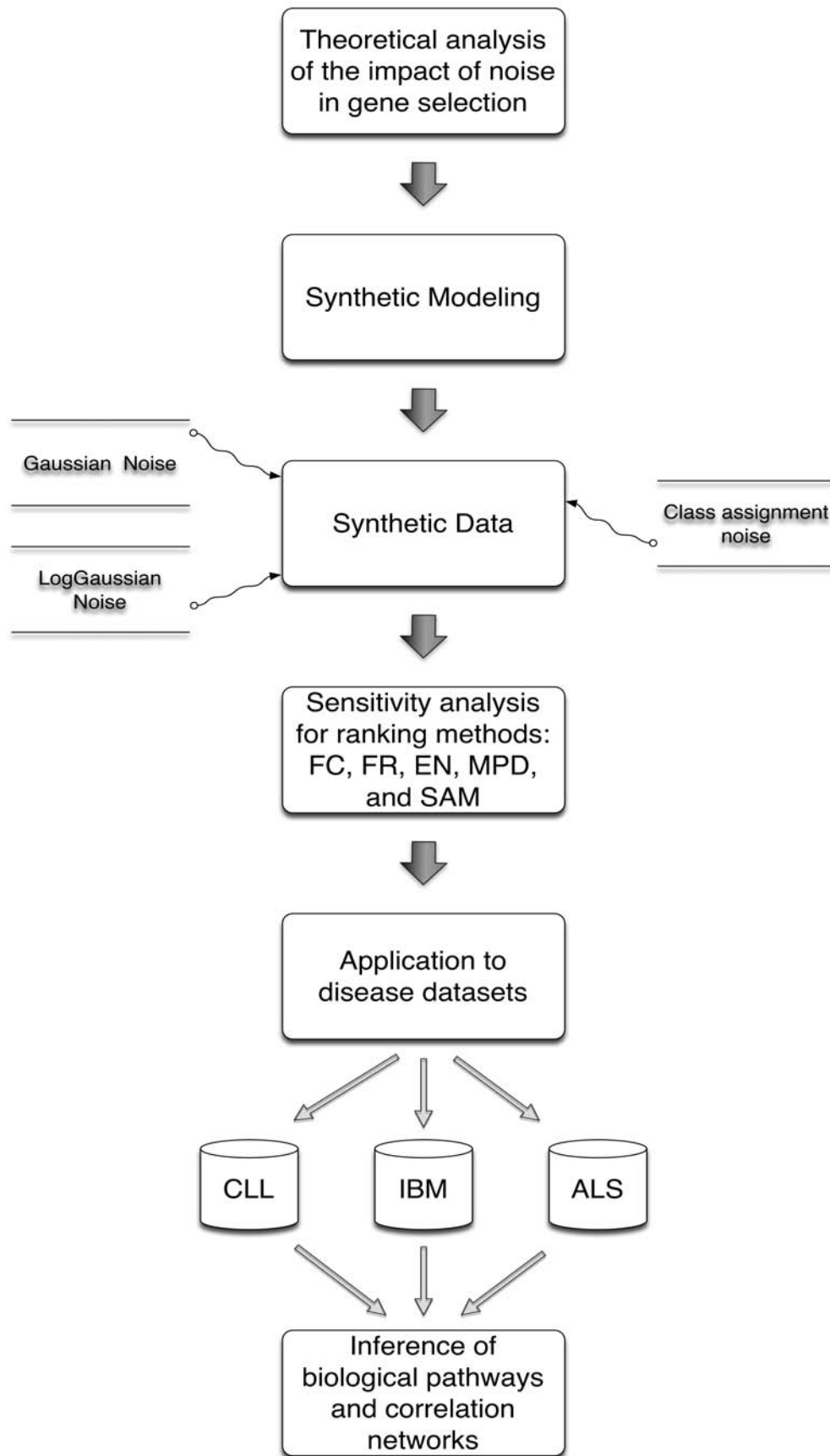
519 **Table 2:** Synthetic modeling. Mean LOOCV predictive accuracy for each of the noise  
520 types at different noise levels.

	CLL	IBM	ALS
FR	93.25 / 6	97.06 / 2	94.12 / 12
FC	93.87 / 35	79.41 / 2	87.06 / 254
PD	93.25 / 7	91.18 / 32	94.12 / 17
EN	94.48 / 99	79.41 / 6	88.24 / 114
SAM	<b>93.87 / 26</b>	<b>97.06 / 1</b>	<b>95.29 / 42</b>

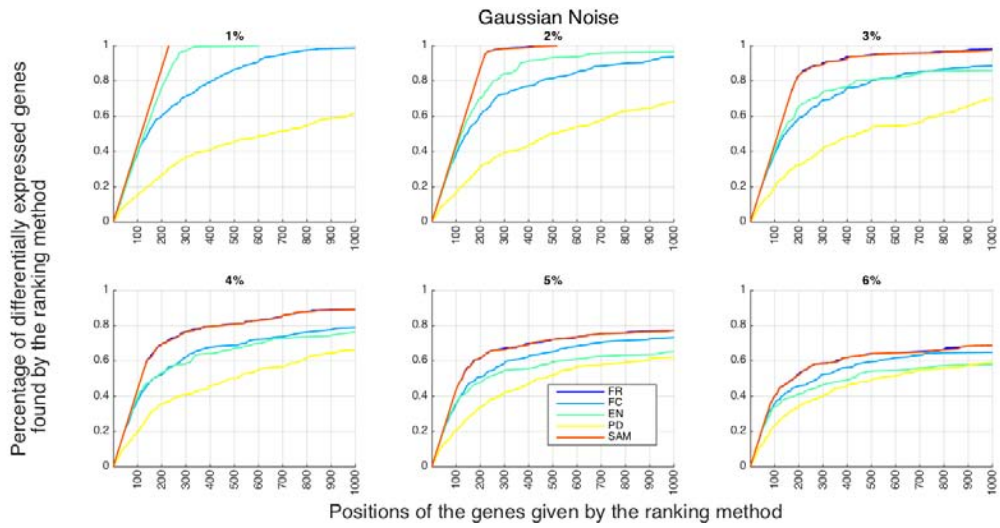
523 **Table 3.** Mean LOOCV accuracy / Number of selected probes for CLL/IBM/ALS  
524 datasets.

525

526

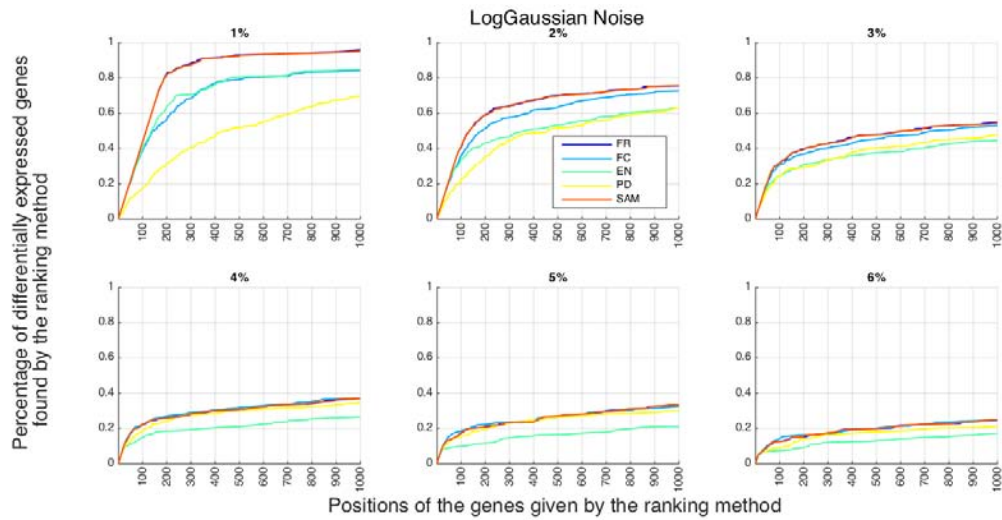


528  
529 **Fig 1** Flow diagram of the methodology shown in this paper.



530  
 531  
 532  
 533  
 534  
 535  
 536

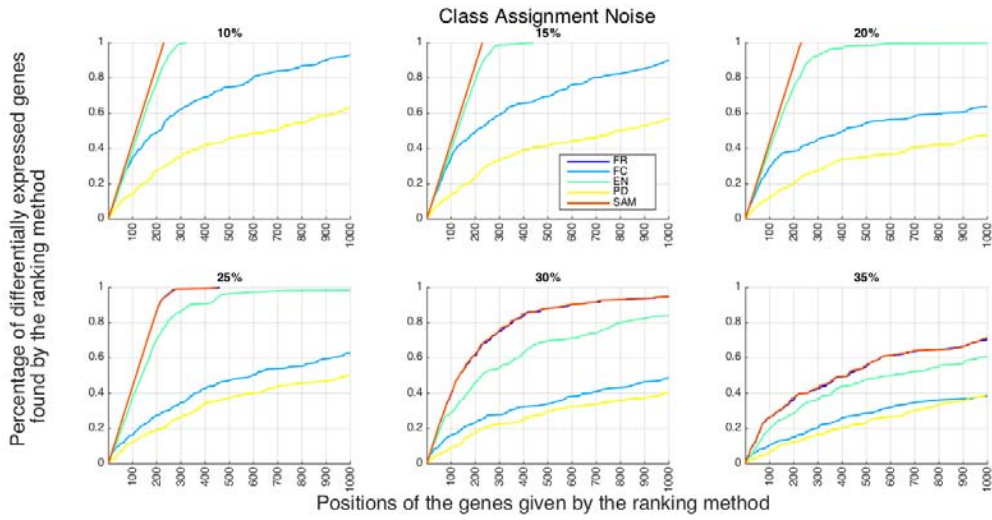
**Fig 2** Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for Gaussian noise.



538  
 539  
 540  
 541  
 542  
 543  
 544

**Fig 3** Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for log-Gaussian noise.

545

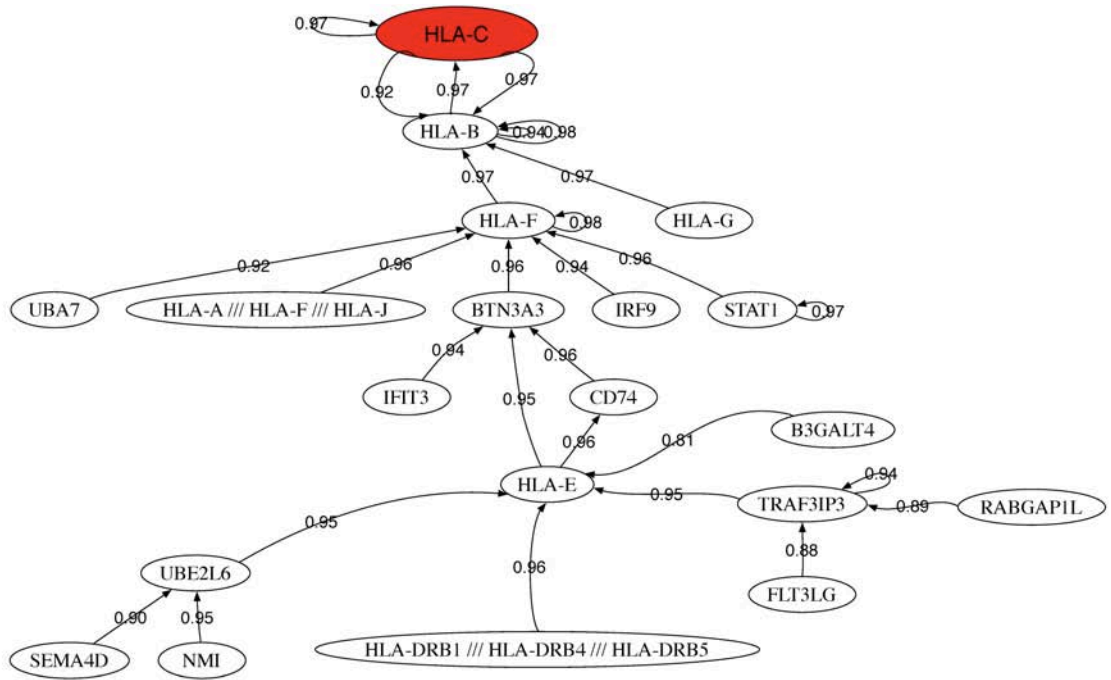


546  
547  
548  
549  
550  
551  
552

**Fig 4** Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes in the set of the first 1000 selected genes for class assignment noise.



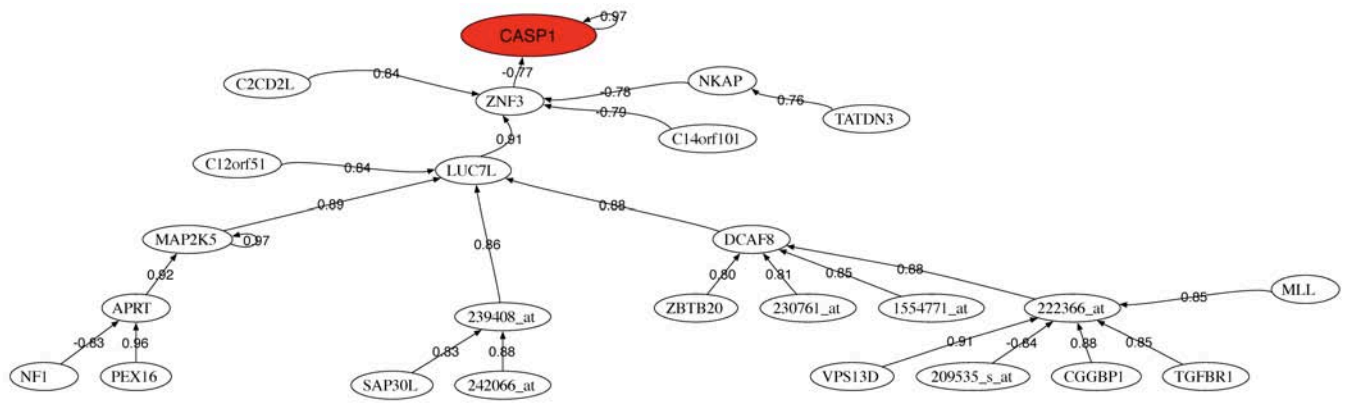




562  
563  
564  
565  
566  
567

**Fig 6 Correlation tree for IBM using the first 30 most discriminatory genes.**

568  
569  
570  
571



572  
573  
574  
575  
576  
577  
578

**Fig 7 Correlation tree for ALS using the first 30 most discriminatory genes.**

## **A.6 Impact of microarray preprocessing techniques in unraveling biological pathways**

Accepted for publication in the Journal of Computational Biology.

# Journal of Computational Biology

**Decision Letter (JCB-2016-0042)**

**From:** sorin@cs.brown.edu

**To:** jlfr@uniovi.es

**CC:**

**Subject:** Journal of Computational Biology - Decision on Manuscript ID JCB-2016-0042

**Body:** 14-Mar-2016

Dear Prof. Fernández-Martínez:

It is a pleasure to accept your manuscript entitled "Impact of microarray preprocessing techniques in unraveling biological pathways" in its current form for publication in Journal of Computational Biology.

Please be sure to cite this article to ensure maximum exposure of your work.

All authors will get a follow-up email with instructions on how to complete our online Copyright Agreement form. The corresponding author is responsible for communicating with coauthors to make sure they have completed the online copyright form. Authors not permitted to release copyright must still return the forms acknowledging the statement of the reason for not releasing the copyright.

FAILURE BY ALL AUTHORS TO SUBMIT THIS FORM MAY RESULT IN A DELAY IN PUBLICATION.

Consider Liebert Open Option to have your paper made free online immediately upon publication for a one-time fee. Benefits of Liebert Open Option include: accelerated e-pub ahead of print publication; email message highlighting the article; increased readers, citations and downloads; an identifying icon in the table of contents showing that the paper is permanently available for free to all readers; and immediate deposition into PubMed Central®. Please contact [OpenAccess@liebertpub.com](mailto:OpenAccess@liebertpub.com) or call (914) 740-2194 for more information.

If your institution is not currently subscribing to this journal, please ensure that your colleagues have access to your work by recommending this title ([http://www.liebertpub.com/mcontent/files/lib\\_rec\\_form.pdf](http://www.liebertpub.com/mcontent/files/lib_rec_form.pdf)) to your Librarian.

Thank you for your fine contribution. On behalf of the Editors of Journal of Computational Biology, we look forward to your continued contributions to the Journal.

Sincerely,  
Dr. Sorin Istrail  
Editor-in-Chief, Journal of Computational Biology  
[sorin@cs.brown.edu](mailto:sorin@cs.brown.edu)

**Date Sent:** 14-Mar-2016

 Close Window

## Impact of microarray preprocessing techniques in unraveling biological pathways

Enrique J. deAndrés-Galiana · Juan Luis  
Fernández-Martínez · Leorey N. Saligan ·  
Stephen T. Sonis

Received: date / Accepted: date

**Abstract** To better understand the impact of microarray preprocessing normalization techniques on the analysis of biological pathways in the prediction of chronic fatigue (CF) following radiation therapy (RT), this study have compared the list of predictive genes found using the Robust Microarray Averaging (RMA) and the Affymetrix's MAS5 method, with the list that is obtained working with raw data (without any preprocessing). First we modeled the spiked-in dataset where differentially expressed genes were known and spiked-in at different known concentrations, showing that the precisions established by different gene ranking methods were higher than working with raw data. The results obtained from the spiked-in experiment were extrapolated to the chronic fatigue dataset to run learning and blind validation. RMA and MAS5 provided different sets of discriminatory genes that have a higher predictive accuracy in the learning phase, but lower predictive accuracy during the blind validation phase, suggesting that the genetic signatures generated using both preprocessing techniques cannot be generalizable. The pathways found using the raw dataset described better what is a priori known for the CF disease. Besides, RMA produced more reliable pathways than MAS5. Understanding the strengths of these two preprocessing techniques in phenotype prediction is critical for precision medicine. Particularly, this

---

E.J. deAndrés-Galiana  
Mathematics department, Universidad de Oviedo, Asturias, Spain.  
E-mail: eag@aic.uniovi.es

J.L. Fernández-Martínez  
Mathematics department, Universidad de Oviedo, Asturias, Spain.  
Tel.: +34-985103199  
Fax: +34-985103354  
E-mail: jlfm@uniovi.es

Leo Saligan  
NINR/NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA.  
E-mail: saliganl@mail.nih.gov

S.T. Sonis  
Biomodels, LLC, Watertown, MA, USA.  
E-mail: ssonis@biomodels.com

manuscript concludes that biological pathways might be better unraveled working with raw expression data. Moreover, the interpretation of the predictive gene profiles generated by RMA and MAS5 should be done with caution. This is an important conclusion with a high translational impact that should be confirmed in other disease datasets.

**Keywords** CANCER GENOMICS, DNA arrays, GENE EXPRESSION, GENE NETWORKS

## 1 Introduction

Microarray data analysis is used to identify important genes to predict at-risk phenotypes, understand biologic underpinning of health conditions, and identifying therapeutics targets. However, microarray data are notorious for containing noise which historically contributed to issues around reproducibility, especially as related to gene / clinical phenotype relationships (Dinu et al., 2007; Jeffery et al., 2006; Kooperberg et al., 2002; Larsson et al., 2005). Further, genomic noise also impedes accurate mechanistic conclusions by partially falsifying the biological pathways that are involved in the disease development (deAndrés-Galiana et al., 2016). To address this concern, it is common practice to apply different kinds of preprocessing techniques to the microarray data in order to amplify the gene signal and limit the noise caused by experimental factors (Irizarry et al., 2003). Noise might impact the results provided by the bioinformatics techniques used to identify the most discriminatory genes in phenotype prediction problems. Due to the high dimension and complexity of microarray datasets, filtering/ranking methods are often applied as a first step in order to preselect the set of most discriminatory genes.

In this manuscript we compared the precision of identifying biologically relevant genes obtained from a raw dataset and preprocessed datasets using Robust Multi-array Average (RMA) and Affymetrix Microarray Suit 5.0 algorithm (MAS5). For that purpose, we used the most common ranking methods, Fisher's Ratio (FR) and Fold Change (FC), to measure their predictive accuracy using a Leave-One-Out-Cross-validation approach. We first modeled the Affymetrix Latin Square Data for Expression Algorithm Assessment (Human Genome U133 Data Set Affymetrix (2015)), where 42 different control genes are spiked-in at known concentrations. This is commonly known as the Spiked-In experiment. We observed that working with raw data provided better results than using the RMA and MAS5 preprocessed datasets to locate the spiked-in genes. To our knowledge this a novel observation that warrants confirmations in other diseases datasets. In this study we also present the results obtained for a radiotherapy-related fatigue dataset in patients with prostate cancer (Saligan et al., 2014), obtaining some interesting and unexpected conclusions.

## 2 Microarrays preprocessing techniques

Microarrays are manufactured using photo-lithographic techniques to attach hundreds of thousands of different oligonucleotide sequences on the surface of a glass

slide. These oligonucleotides correspond to known DNA or RNA sequences that are arranged in different probe sets. Quantification of the levels of transcripts in a sample is performed via hybridization to the specific probes and measurement of the expression through fluorescence-based methods. Generally, raw data contains about 20 pairs of oligonucleotides for each DNA or RNA target (gene) known as probe sets. The first component of these pairs is referred to as the Perfect Match (PM) probe. Each PM probe is paired with a Mismatch (MM) probe that is artificially created by changing the middle base with the intention of measuring non-specific binding. Typically, to define a measure of gene expression, probe intensities are summarized for each probe set into a single value. Different studies have been performed to analyze the accuracy of these measurements and to correct the effect of noise in microarrays (Benito et al., 2004; Chen et al., 2011; Scherer, 2009). Two techniques of particular importance are RMA (Irizarry et al., 2003) and MAS5 (Affymetrix, 2001), and are analyzed in this paper.

## 2.1 MAS5

The Affymetrix Microarray Suite 5.0 (MAS5) algorithm uses both PM and MM probes to summarize gene expression. The MAS5 signal of a probe set  $i$  is defined as the anti-log of the Tuckey's biweight robust mean (Huber and Ronchetti, 2009) of the following values:

$$u_{ij} = \log(PM_{ij} - CT_{ij}), \quad j = 1, \dots, N, \quad (1)$$

where

$$CT_{ij} = \begin{cases} MM_{ij} & \text{if } MM_{ij} < PM_{ij}, \\ PM_{ij} - \varepsilon^2 & \text{if } MM_{ij} > PM_{ij}, \end{cases} \quad (2)$$

being  $N$  de number of probes in the probe set (or gene)  $i$  and  $\varepsilon^2$  a given positive amount that has to be individually adjusted for each probe set. Therefore, the robust Tuckey's mean of a probe set  $i$  is

$$\bar{u}_i = \frac{\sum_{j=1}^N \psi(u_{ij}; c) u_{ij}}{\sum_{j=1}^N \psi(u_{ij}; c)}, \quad (3)$$

where

$$\psi(x; c) = \begin{cases} x \left(1 - \frac{x^2}{c^2}\right)^2 & \text{for } |x| < c, \\ 0 & \text{for } |x| > c. \end{cases} \quad (4)$$

## 2.2 RMA

Robust Multiarray Average (RMA), basically consists in three steps:

1. Background correction using the following additive probabilistic model:

$$PM_{ij} = s_{ij} + bg_{ij}, \quad (5)$$

where  $PM_{ij}$  is the Perfect Match of the probe  $j$  in gene  $i$ ,  $s_{ij}$  is the gene signal and it is supposed to follow an exponential distribution  $s_{ij} \sim Exp(\lambda_i)$ , and  $bg_{ij}$  is the background correction caused by the optical noise and non-specific binding and it is supposed to follow a normal distribution  $bg_{ij} \sim N(\mu_i, \sigma_i^2)$ . This identification problem has three unknown parameters  $(\lambda_i, \mu_i, \sigma_i)$  and  $N$  different realizations for  $PM_{ij}$ . This problem can be typically solved by least squares and the maximum likelihood estimation.

2. Normalization across all arrays to make all distributions the same. This task is performed by quantile normalization, and consists in normalizing the background corrected array to a common set of quantiles. This process is aimed at correcting for array biases and avoiding the effect of outliers. This process provided a set of normalized probe values  $sn_{ij}$ .
3. Probe set summarizing, where the final expression is calculated separately for each gene  $i$  using the following linear model in  $\log_2$  scale:

$$Y_{ij} = \mu_i + \alpha_{ij} + \varepsilon_{ij}, \quad (6)$$

where  $Y_{ij}$  are the background corrected, normalized, log transformed probe intensities ( $Y_{ij} = \log_2(sn_{ij})$ ),  $\mu_i$  is the log-expression level for gene  $i$ ,  $\alpha_{ij}$  is the probe affinity effect of probe  $j$  in the gene  $i$ , and  $\varepsilon_{ij}$  is the independent identically distributed error term with zero mean. The probe affinities  $\alpha_{ij}$  should verify  $\sum_{j=1}^N \alpha_j = 0$ . This linear model is solved using the median polish algorithm and provides the final summarized gene intensity value  $\mu_i$ , that is commonly used in phenotype prediction problems.

### 3 Material and Methods

The methodology shown herein has two main parts: A) Analysis of the precision of the ranking methods using a synthetic data set for both raw and preprocessed datasets. B) Analysis of the accuracy of predictive genes by inspecting the biological pathways for the cancer-related fatigue raw and preprocessed datasets.

In part A, we used the Affymetrix Latin Square Data for Expression Algorithm Assessment. Knowing the genes that are differentially expressed, we first ranked the genes according to a combination of Fold Change and Fisher's ratio and then analyzed the precision of the generated gene ranking using raw and preprocessed data. Subsequently, we performed gene selection to study the discrimination power of the selected genes in both cases (raw and preprocessed). In part B, we used a cancer-related fatigue data set. In this case, we did not know the differentially expressed



genes, therefore, we performed gene selection based on the same ranking methods used in the synthetic dataset, identified the predictive genes, and conducted correlation networks and pathway analysis to understand the biological pathways that are associated with these selected genes. Then, we compared the biological pathways and correlation networks associated with the selected genes from the raw and preprocessed data. A flow chart of this methodology is shown in figure 1.

### 3.1 Ranking methods and gene selection

To alleviate the high under-determined character of the genomic-phenotype prediction problem, filter methods are applied to reduce the dimensionality of the genomic data to select the most discriminatory genes. Filter methods rank the different genes according to different measures of their discriminatory power in the phenotype prediction problem. In this study, we analyzed the precision on the selection of the differential expressed genes using a combination between the most common and well-known ranking methods: Fold Change (FC) and Fisher’s ratio (FR). The Precision  $P$  was defined as follows:

$$P = \frac{|\{DE\_genes\} \cap \{Selected\_genes\}|}{|\{Selected\_genes\}|}, \quad (7)$$

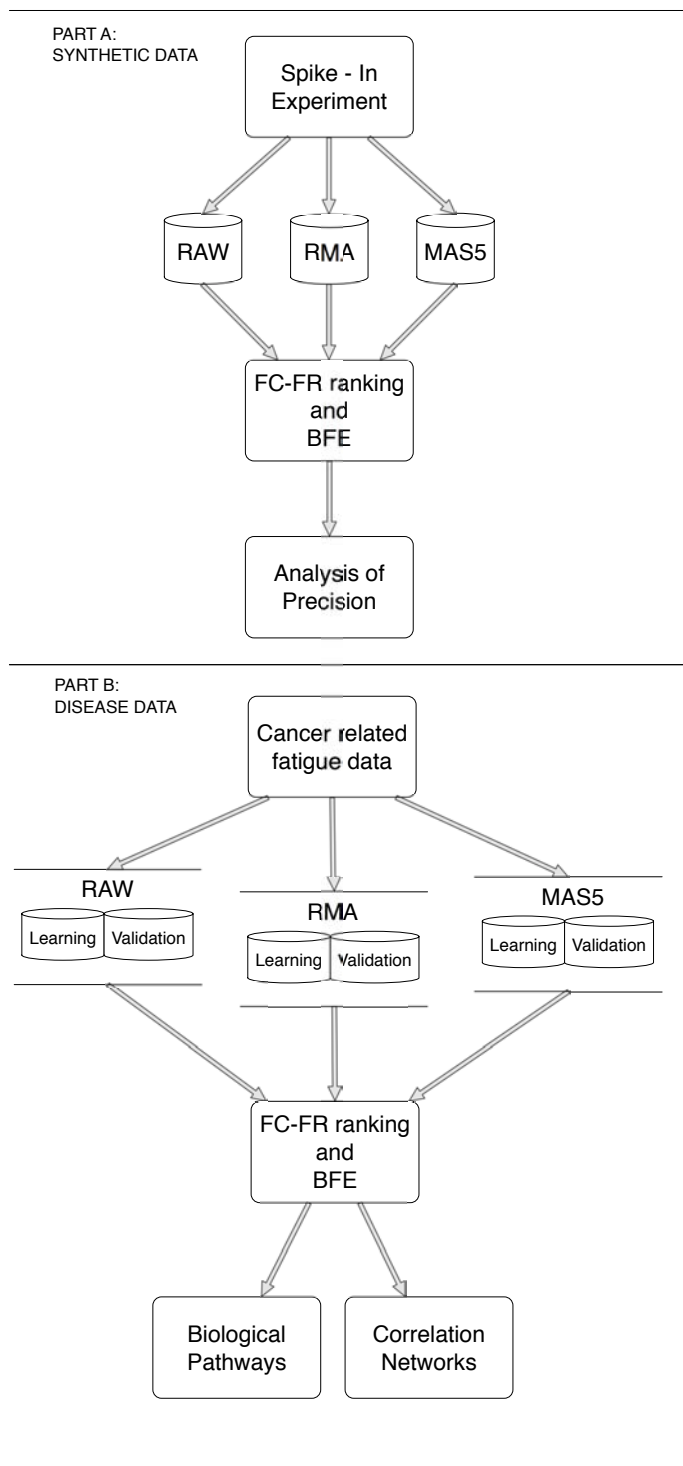
where  $DE\_genes$  is the set of differentially expressed genes and  $Selected\_genes$  is the set of selected genes/probes.

We work with binary classification problems, first knowing the gene expression in the different samples of each class. Our ranking algorithm is a combination of Fold Change (Schena et al., 1996) and Fisher’s ratio (Fisher, 1936). The algorithm first preselected the most differentially expressed genes above a certain absolute FC value and then the preselected genes are ranked according to their FR. The reason to first preselect with FC is to avoid low dispersions in both classes which could provide high FR values, when in fact the centers of both distributions in expressions are very close.

Once we ranked the preselected genes, we identify the most discriminatory genes. The selection of the most predictive genes, followed the same procedure that was described in Saligan et al. (2014): the shortest list of genes with the highest predictive accuracy was selected via Backwards Feature Elimination (BFE) and a distance-based nearest-neighbor classifier. To measure the discriminatory power of the different embedded lists we used the Leave-One-Out Cross-Validation (LOOCV) predictive accuracy. For comparison purposes, the same procedure is used for raw and preprocessed data through MAS5 and RMA preprocessing techniques.

### 3.2 The Spike-in experiment

In order to check the precision of the above described ranking method using both raw and preprocessed data, we needed a dataset where we know the genes that are differentially expressed. In such case we used the Affymetrix Latin Square Data for

**Fig. 1** Flow chart of the methodology

Expression Algorithm Assessment (Human Genome U133 Data Set) that consists of 3 technical replicates of 14 separate hybridization of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125pM to 512pM. The concentrations in the first experiment, composed by three replicas, are 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512pM (see supplementary material). Each subsequent experiment and its three replicas rotated the spiked-in concentrations by one group; i.e. experiment 2 and its three replicas began with 0.125pM and ended at 0pM, up to experiment 14 and its three replicas, which began with 512pM and ended with 256pM. Further details can be consulted in Affymetrix (2015).

### 3.3 The cancer related fatigue dataset

The cancer-related fatigue microarray dataset was obtained from men who were 18 years or older, diagnosed with non-metastatic prostate cancer with or without a history of prostatectomy, and scheduled to receive External Beam Radiation Treatment (EBRT) with or without concurrent androgen deprivation therapy (ADT). A total of 44 men with non-metastatic prostate cancer were enrolled, in a NIH IRB-approved study. Data from 27 subjects were used in the training set and data from 17 subjects were included in the validation blind set Saligan et al. (2014). The training set was from the array outputs of 27 subjects; 18 High Fatigue (HF) and 9 Low Fatigue subjects, phenotyped using a 3-point decline in fatigue score measured by the Functional Assessment of Cancer Therapy -Fatigue (Cella et al., 2002). We managed a raw microarray dataset with 604,258 probes and the preprocessed dataset with 54,675 different probes in both cases, using RMA and MAS5 preprocessing techniques.

Once the most discriminatory genes from raw and preprocessed data were selected, pathways analysis was performed using Gene-Analytics software (Stelzer et al., 2009). Further, we built correlation networks to understand how the expressions of the most discriminatory genes are interrelated. Correlation networks were generated using Pearson correlation coefficient (Pearson, 1895), and Kruskal's algorithm (Kruskal, 1956) to find the minimum-spanning-tree.

## 4 Results and Discussion

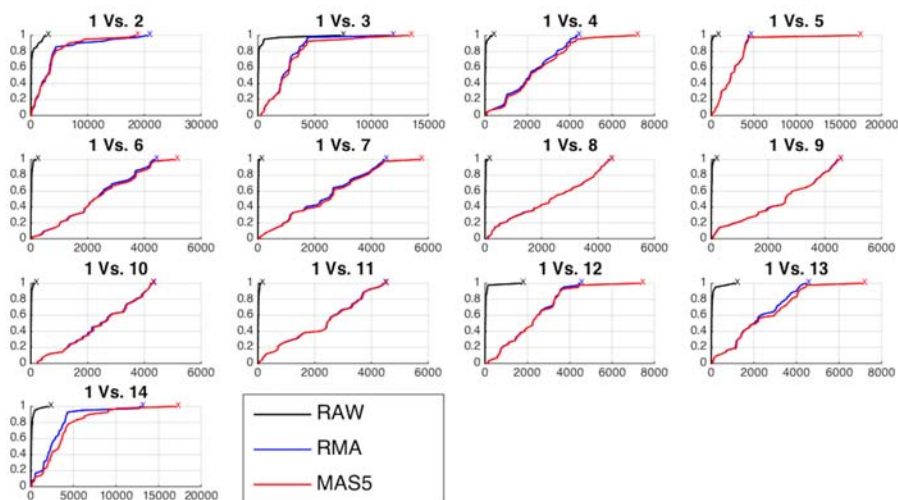
### 4.1 The Spike-In experiment

Using the Affymetrix Latin Square Data for Expression Algorithm Assessment (Human Genome U133 Data Set) we have checked the Precision of the FC/FR ranking algorithm described in section 3. There are 42 differentially expressed probes and we selected the first 42 probes in the ranking. We compared the first group with the rest of groups to cover all the possible concentration comparisons. In the first comparison (group 1 Vs. group 2) the difference of concentration between all the differentially expressed probes was 0.125pM, in the second comparison (group 1 Vs. group 3) the difference was 0.5pM, on up to the 12 comparisons (group 1 Vs. group 13), which was 256pM. Due to the rotation of the concentrations, the last comparison (group 1

Vs. 14) had again a difference of 0.125pM in concentration among all the differentially expressed probes.

Table 2 shows the Precision for each comparison using RAW, RMA and MAS5 datasets, showing the mean Precision along the different comparisons. In almost all the comparisons we got better results in terms of precision working with RAW data than with RMA and MAS5 datasets. Also, the higher Mean Precision was obtained with RAW data.

We have also calculated the empirical Cumulative Distribution Functions (CDF) of the positions of the differentially expressed genes. A perfect CDF would be a straight line reaching the value of 1 at position 42, corresponding to the total number of differentially expressed genes. These curves served to visualize how many genes we have to select in order to locate all the differentially expressed genes. Figure 2 shows these CDF curves for each comparison and type of data. As the raw data obviously have more genes/probes (248,152 for raw data and 22,300 for preprocessed data, see Affymetrix (2015)), the positions given by the ranking method are divided by a correction factor:  $C = nR/nP$  where  $nR$  is the number of raw probes/genes equal to 248,152 and  $nP$  is the number of preprocessed probes/genes equal to 22,300. Therefore,  $C = 11.13$ , for the spiked-in experiment.



**Fig. 2** Empirical Cumulative Distribution Function (CDF) of the positions of the differentially expressed genes ranked by the FC/FR methods for each comparison and different types of data.

In this figure the X-axis represents the positions of the genes/probes given by the ranking method and the Y-axis represents the percentage of differentially expressed genes that were located. Therefore, in the first comparison we were able to find all the differentially expressed genes (42) selecting less than 5000 (0.5E4 in the X-axis of the graphic) while working with preprocessed data, we needed almost all the probes/genes (2.23E4). In all the comparisons, we were able to find all the dif-

ferentially expressed genes selecting rather less number of genes with raw data than with preprocessed data.

#### 4.2 The Chronic fatigue dataset

The aim of this study is to find the list of most discriminatory genes that serve to differentiate between high and low chronic fatigue induced by the radiotherapy in prostate cancer patients (Saligan et al., 2014). The differences on the selection of most discriminatory genes using raw and preprocessed data are shown. For the sake of clarity we are showing the first 50 most discriminatory genes in each case.

Table 3 shows the LOOCV accuracy of first 50 most discriminatory probes/genes in each case. The highest predictive accuracy we obtained a 92.59% of accuracy with only the first 3 probes/genes. However, using RMA and MAS5 we achieved a 100% with 6 and 44 probes/genes respectively. Obviously the dimensionality of the raw data set is 11.05 times higher than the preprocessed datasets, that is, using the raw data, the probe sets have not been summarized in one gene like in the preprocessed data. For that reason the repetition of a probe in the raw data indicates the importance of the corresponding gene. This is the case of *TUBB2A*, *HLA-DQA1*, *TUBB3*, *HLA-DQB1*, and *BTNL3*. It can be observed that RMA also found these genes within the most discriminatory set, but not using MAS5.

Additionally, a blind validation of these results has been performed using the set of 17 subjects, independent of the training set, originally used in Saligan et al. (2014) to assess the validity of the learned predictive model. The result of this blind validation using raw data was 76.47% accurate, while using MAS5 and RMA, the accuracy dropped to 58.82% and 64.7%, respectively. This result is very important and shows that RMA and MAS5 increase the accuracy in the learning process at the prize of decreasing the accuracy in blind validation. Therefore, this implies that the biological pathways associated with the predictive genes found using raw data are more meaningful, and both preprocessing techniques (RMA and MAS5) highly impact the biological pathway analysis and the corresponding phenotype prediction problem.

#### 4.3 Pathway analysis and Correlation networks

In this section we provide the main pathways associated with the discriminatory genes that can predict the Chronic Fatigue phenotype using raw, RMA and MAS5 datasets. These genes are shown in table 3.

The raw data generated predictive genes associated with pathways mainly related to pathogenic infections (*HLA-DQX* genes), as well as pathways associated with oligomerization of connexins into connexons (*TUBB2A* and *TUBB3*) involved in intercellular signals and metabolic communication (Koval, 2006). These are crucial mechanisms in the development of many human diseases (Kelsell et al., 2001).

The main pathways associated with predictive genes generated by RMA are related to mitotic prometaphase (*BIRC5*, *CLIP1*, *STAG2*, *TUBB3*) that controls the nu-

clear membrane breaking apart into numerous membrane vesicles, cytoskeleton remodeling neurofilaments (*EEPK1*, *KRT6A*, *TUBB2A* and *TUBB3*) and mitotic metaphase and ana-phase (*BIRC5*, *CLIP1*, *TUBB2A* and *TUBB3*). The beta-tubulin gene family controls the tubulin protein super-family of globular proteins. Beta-tubulins polymerize into micro-tubules which is a major component of the cytoskeleton formation. Micro-tubules function in many essential cellular processes, including mitosis. For instance, tubulin-binding drugs serve to kill cancerous cells by inhibiting micro-tubule dynamics that are required for DNA segregation and cell division. The main pathways associated with predictive genes generated by MAS5 are *GADD45* pathway, *EGFR1* signaling pathway, and interferon type I related to the *MAP3KX* genes (Jordan and Wilson, 2004; McKean et al., 2001).

We also provide the correlation graphs for the 50 most discriminatory genes for each dataset. Figure 3, 4 and 5 shows the correlation graphs for raw, RMA and MAS5 respectively. In the case of raw data we can observe one main tree connecting the tubulin genes to the major histocompatibility complex gene and other genes that serve to expand the tree. RMA privileges the connection between the beta-tubuline genes and two probes (241238\_at and 1566585\_at) whose gene name is unknown. MAS5 privileges the role of *SOCS3*. This gene encodes a member of the STAT-induced STAT inhibitor (SSI), also known as suppressor of cytokine signaling (SOCS), family. SSI family members are cytokine-inducible negative regulators of cytokine signaling. The expression of *SOCS3* gene is induced by various cytokines, including IL6, IL10, and interferon (IFN)-gamma (Masuhara et al., 1997; Minamoto et al., 1997).

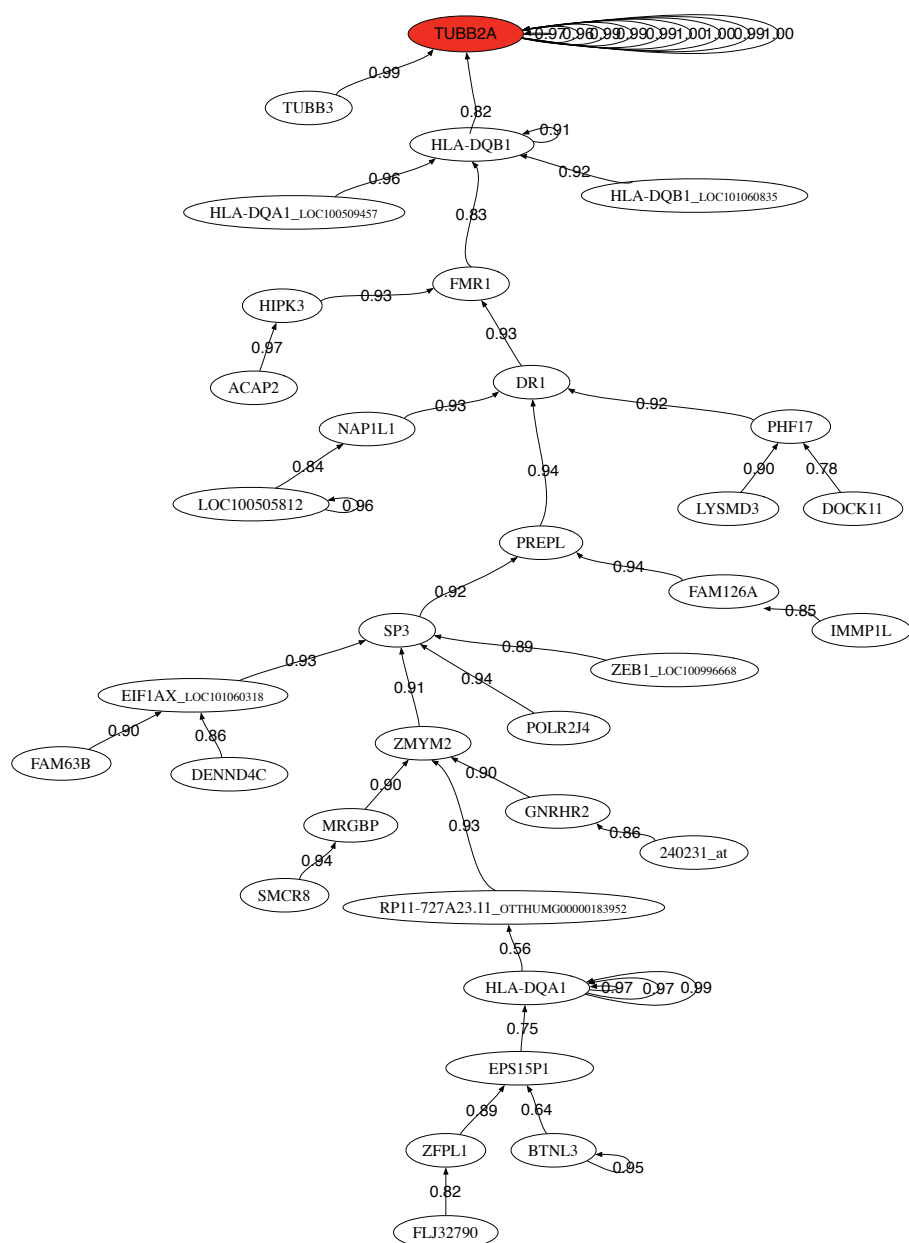
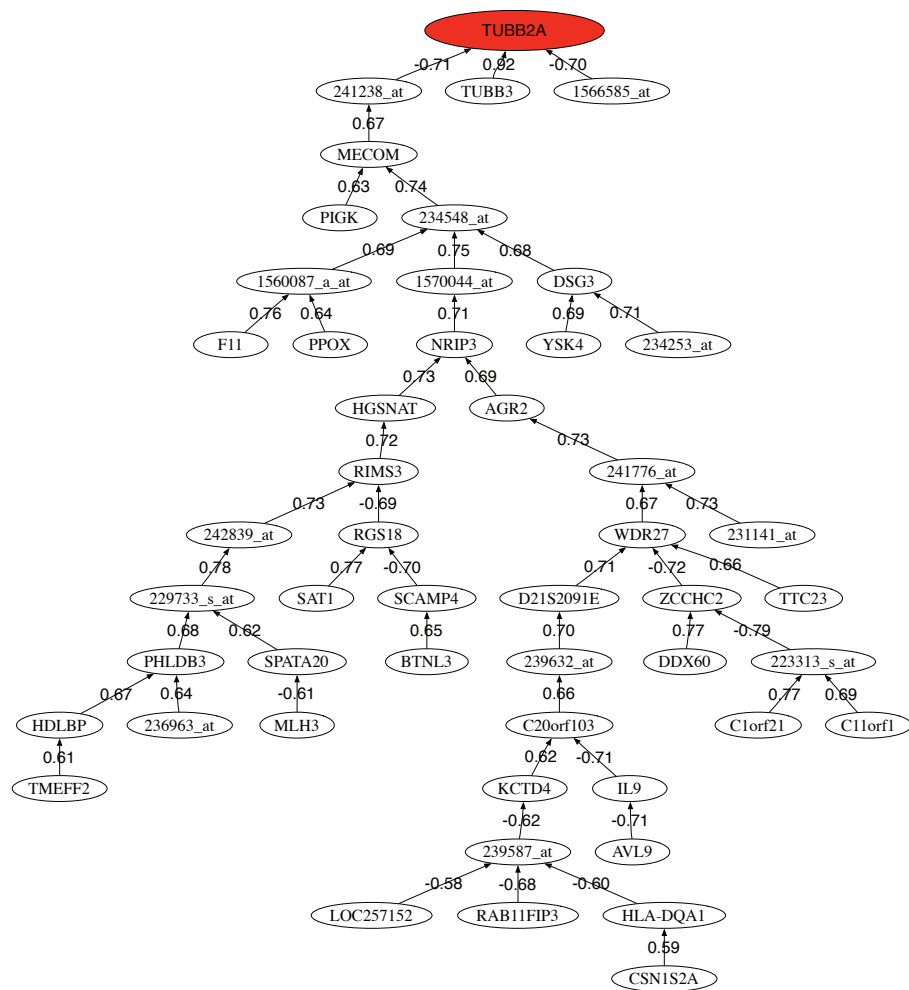


Fig. 3 Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using raw data.



**Fig. 4** Pearson correlation coefficient minimum-spanning-tree of the 50 first selected probes using pre-processed data with RMA.





is an important conclusion with a high translational impact that should be confirmed in other disease datasets.

## References

- Affymetrix (2001). Microarray suite user guide, version 5. <http://www.affymetrix.com/support/technical/manuals.affx>.
- Affymetrix (2015). Latin square data for expression algorithm assessment. [http://www.affymetrix.com/support/technical/sample\\_data/datasets.affx](http://www.affymetrix.com/support/technical/sample_data/datasets.affx).
- Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114.
- Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE*, 6(2):e17238.
- deAndrés-Galiana, E. J., Fernández-Martínez, J. L., and Sonis, S. (2016). Design of biomedical robots for phenotype prediction problems. *Journal of Computational Biology*, page to appear.
- Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by sam-gs. *BMC Bioinformatics*, 8:242.
- Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–88.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics, Second Edition*. New York: Wiley.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.
- Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:359.
- Jordan, M. A. and Wilson, L. (2004). Microtubules as a target for anticancer drugs. *Nat Rev Cancer*, 4(4):253–265.
- Kelsell, D. P., Dunlop, J., and Hodgins, M. B. (2001). Human diseases: clues to cracking the connexin code? *Trends Cell Biol*, 11(1):2–6.
- Kooperberg, C., Fazio, T. G., Delrow, J. J., and Tsukiyama, T. (2002). Improved background correction for spotted dna microarrays. *J Comput Biol*, 9(1):55–66.
- Koval, M. (2006). Pathways and control of connexin oligomerization. *Trends Cell Biol*, 16(3):159–166.
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50.
- Larsson, O., Wahlestedt, C., and Timmons, J. A. (2005). Considerations when using the significance analysis of microarrays (sam) algorithm. *BMC Bioinformatics*, 6:129.

- Masuhara, M., Sakamoto, H., Matsumoto, A., Suzuki, R., Yasukawa, H., Mitsui, K., Wakioka, T., Tanimura, S., Sasaki, A., Misawa, H., Yokouchi, M., Ohtsubo, M., and Yoshimura, A. (1997). Cloning and characterization of novel cis family genes. *Biochem Biophys Res Commun*, 239(2):439–446.
- McKean, P. G., Vaughan, S., and Gull, K. (2001). The extended tubulin superfamily. *J Cell Sci*, 114(Pt 15):2723–2733.
- Minamoto, S., Ikegame, K., Ueno, K., Narazaki, M., Naka, T., Yamamoto, H., Matsumoto, T., Saito, H., Hosoe, S., and Kishimoto, T. (1997). Cloning and functional analysis of new members of stat induced stat inhibitor (ssi) family: Ssi-2 and ssi-3. *Biochem Biophys Res Commun*, 237(1):79–83.
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.
- Saligan, L. N., Fernandez-Martinez, J. L., deAndres Galiana, E. J., and Sonis, S. (2014). Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Inform*, 13:141–152.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P., and Davis, R. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *PNAS*, 20(93):10614–10619.
- Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. John Wiley and Sons.
- Stelzer, G., Inger, A., Olender, T., Iny-Stein, T., Dalah, I., Harel, A., Safran, M., and D, L. (2009). Genedecks: paralog hunting and gene-set distillation with genecards annotation. *OMICS*, 13(6):477–87.

## TABLES

**Table 1** Precision on the selection of the differential expressed genes using raw data or preprocessed data with RMA and MAS5. The data is the Affymetrix Latin Square Data for Expression Algorithm Assessment. The selection is performed between the first group and the rest to include all the differences between the spike-in concentrations.

Group comparison	RAW	RMA	MAS5
1 vs 2	7.14	<b>9.52</b>	4.76
1 vs 3	<b>26.19</b>	16.67	16.67
1 vs 4	<b>38.10</b>	11.90	14.29
1 vs 5	<b>28.57</b>	<b>28.57</b>	16.67
1 vs 6	26.19	<b>28.57</b>	<b>28.57</b>
1 vs 7	<b>40.48</b>	26.19	23.81
1 vs 8	<b>35.71</b>	21.43	30.95
1 vs 9	<b>40.48</b>	23.81	23.81
1 vs 10	<b>35.71</b>	19.05	21.43
1 vs 11	<b>38.10</b>	14.29	21.43
1 vs 12	<b>23.81</b>	16.67	9.52
1 vs 13	<b>23.81</b>	<b>23.81</b>	14.29
1 vs 14	7.14	4.76	<b>9.52</b>
Mean Precision	<b>28.57</b>	18.86	18.13

**Table 2** Probe/Gene name and Accuracy (Acc %) of the selected genes/probes for raw data and preprocessed data with RMA and MAS5

RAW		RMA		MAS5	
Probe/Gene	Acc(%)	Probe/Gene	Acc(%)	Probe/Gene	Acc(%)
<b>TUBB2A</b>	85.19	<b>TUBB2A</b>	88.89	<b>SOCS3</b>	85.19
<b>HLA-DQA1</b>	<b>96.3</b>	<b>C11orf1</b>	88.89	<b>TMEM194A</b>	92.59
TUBB2A	92.59	<b>PPOX</b>	96.3	<b>1561478.at</b>	92.59
TUBB2A	92.59	<b>TTC23</b>	92.59	<b>CIB3</b>	96.3
TUBB2A	88.89	<b>NRIP3</b>	96.3	<b>ESYT2</b>	92.59
TUBB2A	85.19	<b>SCAMP4</b>	<b>100</b>	<b>ABHD1</b>	92.59
TUBB2A	85.19	HLA-DQA1	100	<b>JTB</b>	92.59
HLA-DQA1	88.89	234253.at	100	<b>1556412.at</b>	92.59
TUBB2A	88.89	223313.s.at	96.3	<b>207371.at</b>	96.3
TUBB2A	88.89	BTNL3	100	<b>LOC100131756</b>	92.59
BTNL3	88.89	YSK4	96.3	<b>CDK6</b>	92.59
TUBB2A	88.89	236963.at	100	<b>ALS2CR8</b>	96.3
HLA-DQA1	92.59	ZCCHC2	100	<b>SEL1L2</b>	96.3
TUBB2A	88.89	DSG3	100	<b>FLJ35220</b>	96.3
TUBB3	88.89	TMEFF2	100	<b>215626.at</b>	96.3
HLA-DQB1	85.19	1566585.at	100	<b>SPAM1</b>	96.3
HLA-DQB1_LOC101060835	85.19	231141.at	100	<b>FTCD</b>	96.3
HLA-DQA1	88.89	SPATA20	100	<b>1570285.at</b>	96.3
IMMP1L	85.19	CSN1S2A	100	<b>216795.at</b>	96.3
BTNL3	85.19	RAB11FIP3	100	<b>MAP3K2</b>	96.3
240231.at	85.19	239587.at	100	<b>MTSS1L</b>	96.3
ZFPL1	85.19	RIMS3	100	<b>GMEB1</b>	96.3
GNRHR2	85.19	234548.at	100	<b>SOCS7</b>	96.3
DR1	88.89	C20orf103	100	<b>GNA12</b>	96.3
DOCK11	88.89	AGR2	100	<b>244274.at</b>	96.3
HLA-DQB1	88.89	SAT1	100	<b>PLP2</b>	96.3
FMR1	88.89	RGS18	100	<b>ATG9B</b>	96.3
ACAP2	85.19	1570044.at	100	<b>1564056.at</b>	96.3
HLA-DQB1	85.19	TUBB3	100	<b>PCCB</b>	96.3
ZEB1_LOC100996668	85.19	HDLBP	100	<b>239370.at</b>	96.3
FLJ32790	85.19	1560087.a.at	100	<b>ANK1</b>	96.3
LOC100505812	88.89	AVL9	100	<b>SCAND2</b>	96.3
DENND4C	88.89	241238.at	100	<b>1564872.at</b>	96.3
PREPL	88.89	PHLDB3	100	<b>SMAD2</b>	96.3
LOC100505812	85.19	PIGK	100	<b>CMTM3</b>	96.3
FAM63B	88.89	F11	100	<b>INSR</b>	96.3
LYSMD3	85.19	C1orf21	100	<b>PSG1</b>	96.3
RP11-727A23.11_OTTHUMG00000183952	85.19	IL9	100	<b>1560169.at</b>	96.3
HIPK3	85.19	229733.s.at	100	<b>MAP3K1</b>	96.3
POLR2J4	85.19	241776.at	100	<b>KCNRG</b>	96.3
PHF17	85.19	WDR27	100	<b>DOCK7</b>	96.3
SP3	85.19	D21S2091E	100	<b>1560995.s.at</b>	96.3
MRGBP	85.19	239632.at	100	<b>WNT5A</b>	96.3
NAP1L1	85.19	HGSNAT	100	<b>1562673.at</b>	<b>100</b>
FAM126A	85.19	242839.at	100	GSK3B	100
EPS15P1	85.19	KCTD4	100	NCKIPSD	100
SMCR8	85.19	MECOM	100	215439.x.at	100
HLA-DQA1_LOC100509457	85.19	LOC257152	100	CDHR3	96.3
ZMYM2	85.19	MLH3	100	PCGEM1	96.3
EIF1AX_LOC101060318	85.19	DDX60	100	GNG13	96.3

## **A.7 Design and application of biomedical robots to phenotype prediction problems**

Accepted for publication in the Journal of Computational Biology.

# Journal of Computational Biology

**Decision Letter (JCB-2016-0008)**

**From:** sorin@cs.brown.edu

**To:** jlfm@uniovi.es

**CC:**

**Subject:** Journal of Computational Biology - Decision on Manuscript ID JCB-2016-0008

**Body:** 09-Feb-2016

Dear Prof. Fernández-Martínez:

It is a pleasure to accept your manuscript entitled "Design of biomedical robots for phenotype prediction problems" in its current form for publication in Journal of Computational Biology.

Please be sure to cite this article to ensure maximum exposure of your work.

All authors will get a follow-up email with instructions on how to complete our online Copyright Agreement form. The corresponding author is responsible for communicating with coauthors to make sure they have completed the online copyright form. Authors not permitted to release copyright must still return the forms acknowledging the statement of the reason for not releasing the copyright.

FAILURE BY ALL AUTHORS TO SUBMIT THIS FORM MAY RESULT IN A DELAY IN PUBLICATION.

Consider Liebert Open Option to have your paper made free online immediately upon publication for a one-time fee. Benefits of Liebert Open Option include: accelerated e-pub ahead of print publication; email message highlighting the article; increased readers, citations and downloads; an identifying icon in the table of contents showing that the paper is permanently available for free to all readers; and immediate deposition into PubMed Central®. Please contact [OpenAccess@liebertpub.com](mailto:OpenAccess@liebertpub.com) or call (914) 740-2194 for more information.

If your institution is not currently subscribing to this journal, please ensure that your colleagues have access to your work by recommending this title ([http://www.liebertpub.com/mcontent/files/lib\\_rec\\_form.pdf](http://www.liebertpub.com/mcontent/files/lib_rec_form.pdf)) to your Librarian.


Thank you for your fine contribution. On behalf of the Editors of Journal of Computational Biology, we look forward to your continued contributions to the Journal.

Sincerely,  
Dr. Sorin Istrail  
Editor-in-Chief, Journal of Computational Biology  
[sorin@cs.brown.edu](mailto:sorin@cs.brown.edu)

Referee report

An interesting paper that would make a lovely contribution to JCB. It will have a positive follow up. It could be accepted as is.

**Date Sent:** 09-Feb-2016

 Close Window

## **Design of biomedical robots for phenotype prediction problems**

**Enrique J. deAndrés-Galiana · Juan Luis  
Fernández-Martínez · Stephen T. Sonis**

Received: date / Accepted: date

**Abstract** Genomics has been used with varying degrees of success in the context of drug discovery and in defining mechanisms of action for diseases like cancer, neurodegenerative and rare diseases in the quest for orphan drugs. To improve its utility, accuracy and cost-effectiveness optimization of analytical methods, especially those that translate to clinically relevant outcomes is critical. Here we define a novel tool for genomic analysis termed a biomedical robot in order to improve phenotype prediction, identifying disease pathogenesis and significantly defining therapeutic targets. Biomedical robot analytics differ from historical methods in that they are based on melding feature selection methods and ensemble learning techniques. The biomedical robot mathematically exploits the structure of the uncertainty space of any classification problem conceived as an ill-posed optimization problem. Given a classifier, there exist different equivalent small-scale genetic signatures that provide similar predictive accuracies. We perform the sensitivity analysis to noise of the biomedical robot concept using synthetic microarrays perturbed by different kind of noises in expression and class assignment. Finally, we show the application of this concept to the analysis of different diseases, inferring the pathways and the correlation networks. The final aim of a biomedical robot is to improve knowledge discovery and provide decision systems to optimize diagnosis, treatment and prognosis. This analysis shows that the biomedical robots are robust against different kind of noises and particularly to a wrong class assignment of the samples. Assessing the uncertainty that is inher-

---

E.J. deAndrés-Galiana  
Artificial Intelligence Center, Universidad de Oviedo, Asturias, Spain.  
E-mail: eag@aic.uniovi.es

J.L. Fernández-Martínez  
Mathematics department, Universidad de Oviedo, Asturias, Spain.  
Tel.: +34-985103199  
Fax: +34-985103354  
E-mail: jlfr@uniovi.es

S.T. Sonis  
Biomodels, LLC, Watertown, MA, USA.  
E-mail: ssonis@biomodels.com



ent to any phenotype prediction problem is the right way of addressing this kind of problems.

**Keywords** Translational genomics · Biomedical robots · Phenotype prediction · Uncertainty assessment

## 1 Introduction

Despite all of its promises, clinical translation of genomics findings has been tempered by analytical limitations, the requirement for extensive numbers of subjects, and cost. To help address these issues, we have developed a coordinated set of bioinformatics algorithms derived from a combination of applied mathematics, statistics and computer science that are capable of analyzing dynamically (as a function of time) high dimensional data. Aside from specifically addressing the interpretation of genomic data, strength of the method is its ability to synchronously include non-genomic inputs (epigenetics, demographic variables, etc.) as a component of a comprehensive analysis. To best describe the concept and potential applications of the Biomedical Robot, we first present the generic and broadly applicable problem of phenotype prediction. For a clinical perspective, this problem applies to linking a set of genes to a specific disease or condition. Second, we describe the design and construction of the biomedical robot, and finally, we provide specific applications of the methodology to different disease datasets: Chronic Lymphocytic Leukemia (CLL), Inclusion Body Myositis (IBM)-Polymyositis (PM) and Amyotrophic Lateral Sclerosis (ALS).

## 2 The phenotype prediction problem

The primary objective of phenotype discrimination is to define sets of genes/probes that optimally differentiate between populations expressing or not expressing a particular phenotype such as disease risk, drug responsiveness or medication toxicity. This concept can be useful in identifying biological and molecular pathways differences between normal and cancer cells, or investigating drug mechanisms of action for a certain type of disease.

We built a conceptual model that related different genes/probes to class prediction (phenotype) as a nonlinear classification problem, since the classifier and the genetic features that serve to achieve an optimum prediction of the phenotype are unknown. Therefore, as a first step a given type of classifier (nearest-neighbor, neural networks, support vector machines, etc) should be built ad-hoc to relate the genetic features to the observed phenotype classes. The classification problem of phenotype discrimination does not need necessarily to be binary, it could be multi-class. This can be considered as the first source of uncertainty in the phenotype prediction problem, since the perfect classifier is a priori unknown.

First we start with a set of expressions of  $n$  genes/probes for a set of  $m$  samples whose phenotype classes are defined, usually by medical expert annotation. This information is typically organized in the expression matrix  $E \in M_{m \times n}(\mathbb{R})$  with  $m \ll n$

and in the class phenotype vector  $\mathbf{c}^{obs} \in \mathbb{R}^m$ . The classifier  $L^*(\mathbf{g})$  can be formally defined as an application between the set of genetic features  $\mathbf{g} \in M \subset \mathbb{R}^s$  and the set of classes  $C = \{c_1, c_2, \dots, c_n\}$ :

$$L^*(\mathbf{g}) : \mathbf{g} \in \mathbb{R}^s \rightarrow C = \{c_1, c_2, \dots, c_n\}. \quad (1)$$

Importantly, not all the genes/probes provide useful information to the phenotype prediction inverse problem. These extraneous genes are noisy and can be analytically disruptive. Fortunately, it is possible to discard irrelevant features, that is, those genes that do not provide any useful information for the phenotype discrimination, since these features introduce ambiguity in the classification. The relevant genes would be defined as the ones that minimize a given target function  $O(\mathbf{g})$  related to the class prediction array:

$$\mathbf{g} : O(\bar{\mathbf{g}}) = \min_{\mathbf{g} \in \mathbb{R}^s} O(\mathbf{g}), \quad (2)$$

$$O(\mathbf{g}) = \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p \quad (3)$$

$$\mathbf{L}^*(\mathbf{g}) = (L^*(\mathbf{g}_1), \dots, L^*(\mathbf{g}_i), \dots, L^*(\mathbf{g}_m)), \quad (4)$$

where  $\mathbf{c}^{obs}$  is the set of observed classes,  $p$  is the norm applied in the distance criterion,  $\mathbf{L}^*(\mathbf{g})$  is the set of predicted classes and  $\mathbf{g}_i \in \mathbb{R}^{N^s}$  is the genetic signature corresponding to sample  $i$ . Otherwise said, the relevant genes would be the ones that allow us to predict the phenotype of new incoming samples. Three considerations are particularly relevant:

- First, several equivalent genetic signatures exist which explain the phenotype class equally well or having a similar predictive accuracy. This is known as the ill-posed character of the phenotype classification problem. Thus, we can apply the parsimony principle to identify small scale signatures by introducing the concept of redundancy. Given a genetic signature  $\mathbf{g} \in \mathbb{R}^s$  characterized by its class predictive accuracy and length  $s$ , redundant features (or genes) are those that provide no additional information than the currently selected features, that is, the prediction accuracy does not increase by adding these genetic features to  $\mathbf{g}$  in the classifier. Interestingly, the fact that the parsimony principle is applied does not avoid the existence of other equivalent signatures that form the equivalence space of the phenotype prediction problem.
- Second, the ill-posed character of the classification is due to the high underdetermined character of the inverse problem involved, since the number of samples  $m$  is much lower than the total number of genetic probes  $n$ . Fernández-Martínez et al. (2012, 2013) analyzed the uncertainty space of linear and nonlinear inverse and classification problems showing that the topography of the cost function  $O(\mathbf{g})$  in the region of lower misfits (or higher predictive accuracies) correspond to one or several flat elongated valleys with null gradients, where the high predictive genetic signatures reside. This valley is unique and rectilinear if the classification/inverse problem is linear, and bends and might be composed of several disconnected basins if the inverse problem is nonlinear and the classification problem becomes nonlinear separable. Also, if we are somehow able to define the

discriminatory power of the different genes, a classification problem could be interpreted as the Fourier expansion of a signal, that is, there will be genes that provide high accuracy for the classification problem alone (head genes), while others will assist in expanding the high frequency details (helper genes) in order to improve the predictive accuracy. Nevertheless, there is a time when adding more details to the classifier do not increase its predictive accuracy. The smallest scale signature is the one that has the least number of highest discriminatory genes. This knowledge could be important for diagnosis and treatment optimization since it allows a fast and cheap genetic data gathering.

- Third, genomic data is notorious for containing noise which has historically contributed to issues around reproducibility, especially as related to gene/clinical phenotype relationships. Similarly, genomic noise also impedes accurate mechanistic conclusions by-partially falsifying biological pathways. There are two main sources of noise:
  - First, noise in the genes expressions that is introduced by the process of data filtering and measurement. The observed genetic expression of a sample,  $\mathbf{g}^{obs}$ , can be expressed as then sum of the noiseless expression  $\mathbf{g}^{true}$  and the measurement noise  $\delta\mathbf{g}$ :  $\mathbf{g}^{obs} = \mathbf{g}^{true} + \delta\mathbf{g}$ .
  - Second, noise in the class assignment  $\delta\mathbf{c}$ , that is due to an incorrect labeling of the samples by the experts. Therefore the observed class vector can be expressed as the sum of the true class vector  $\mathbf{c}^{true}$  and the class assignment noise  $\delta\mathbf{c}$ . For instance, sometimes the classification problem is parameterized as binary when in fact there are more than two classes. Therefore, assigning two different classes to the samples will input noise in the classification. In this case, finding a predictive accuracy lower than 100% would be the expected result, otherwise the algorithm will find a wrong genetic signature in order to fit (or explaining) the wrong class assignment. Obviously this situation is always difficult to detect, since the strategy that is normally followed, consists in achieving a perfect classification. This is not the point of view that is proposed in this paper.

It is straight forward to show that both kinds of noise induce a modeling error in the classifier. Therefore, in presence of these types of noise ( $\delta\mathbf{g}$  and  $\delta\mathbf{c}$ ) the genetic signature with the highest predictive accuracy (and therefore the lowest misfit error) will never perfectly coincide with the genetic signature(s) that explains the disease (noise-free phenotype classification problem). For that reason it is desirable to look also for genetic signatures with lower predictive accuracy than the optimum. Besides, the classifier  $\mathbf{L}^*$  is built ad-hoc and it is just a mathematical abstraction used to discover the genes that are involved in the phenotype discrimination, but it is not the reality itself.

### 3 Biomedical robots

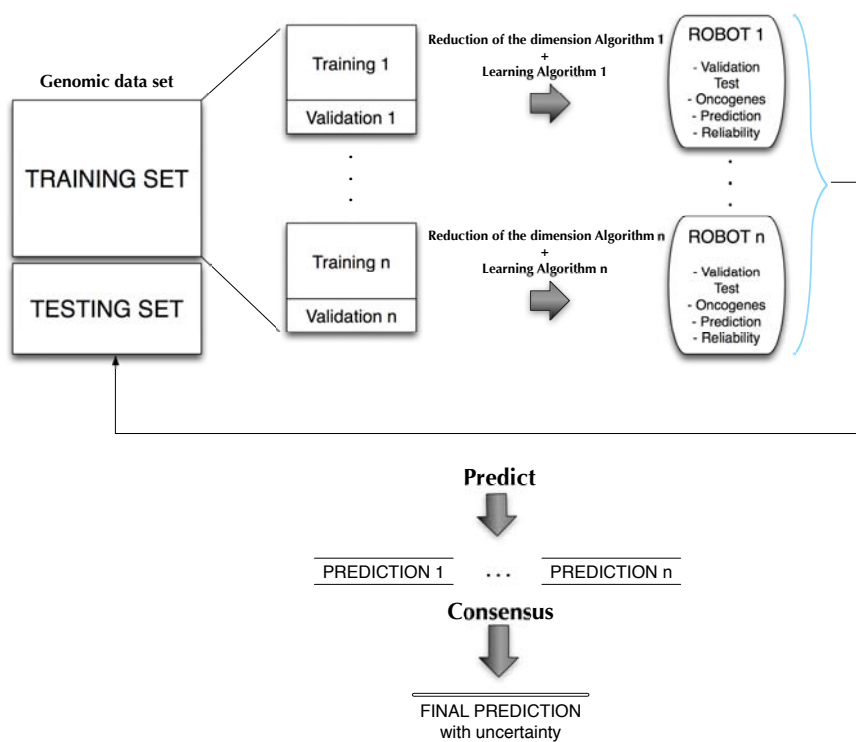
A biomedical robot is a set of algorithms derived from applied mathematics, statistics and computer science that are capable of dynamically analyzing high dimensional data, discovering knowledge, generating new biomedical working hypothesis, and

supporting medical decision making with its corresponding uncertainty assessment. In this definition the data does not need necessarily to be of the same type, that is, several types of data could be used for decision-making purposes. In the present case the data come from microarray differential expression analysis, between individuals that develop the illness and others that do not.

Generating new working hypothesis in the present case includes the analysis of biomarkers and mechanisms of action involved in the illness development, and finding existing drugs that could target the main actionable genes. Also, a benefit of this approach could be the design of intelligent systems to support medical doctors/researchers in the decision making process of new incoming uncatalogued samples to decide questions relative to their diagnosis, treatment and prognosis before any decision is taken. These techniques can help for instance in segmenting patients with respect to response to treatment (deAndrés-Galiana et al., 2015) and also to drug response, to predict the development of induced toxicities (Saligan et al., 2014), to infer the possible surgical risk, etc, among many different applications that we can imagine.

Figure 1 shows a conceptual scheme of the biomedical robot concept. From a training data set we built  $N_r$  robots. The robot is in this case a set of classifiers characterized by their small scale genetic signatures  $\mathbf{g}$  and their corresponding set of parameters needed to perform the classification of the samples. These robots will be deduced from the dataset by applying different supervised filter feature selection methods and dimensional reduction algorithms. Each robot will be also characterized by its predictive accuracy according to the classification cost function  $O(\mathbf{g})$  in a testing dataset. The design of the cost function is important because the set of genetic signatures found might depend on that design. In this paper we have used a Leave-One-Out-Cross-Validation (LOOCV) average error because it makes use of most of the samples that are at disposal, and also mimic the process that we will find in real practice: predicting the class of a new sample using a set of samples that were previously observed and annotated by medical experts (training data set).

It is important to remark that we are not interested in building a black-box methodology, but also being able of inferring the mechanisms of action and the genetic and biological pathways that are involved. The final decision approach is as follows: given a new incoming sample, each of the equivalent robots will perform a prediction. A final prediction with its uncertainty assessment will be given using all these predictions via a consensus strategy such as majority voting. This approach has been used by Fernández-Martínez and Cernea (2014) in a face recognition problem obtaining very high stable accuracies. Ensemble classification and majority vote decisions are based on Condorcet's jury theorem, which is a political science theorem about the probability of a given group of individuals arriving at a correct decision. In the context of biomedical robots and ensemble learning, it implies that the probability of being correct for a majority of independent voters is higher than the probability of any of the individual voters, and tends to 1 when the number of voters (or weak classifiers) tends to infinite. In this case the weak classifiers are any of the biomedical robots of the ensemble that have a high predictive accuracy. These classifiers are guaranteed to be independent since they use different high discriminatory genetic signatures, measured by their corresponding discriminatory power.



**Fig. 1** Conceptual scheme for the design of biomedical robots.

Several methods exist to assign the discriminatory power of the genes: Fold-Change (Schena et al., 1996), Fisher's ratio (Fisher, 1936), Entropy (Shannon, 1948), Mutual Information (Quinlan, 1993), Significance Analysis of Microarrays (SAM) (Tusher et al., 2001), percentile distance between statistical distributions (deAndrés-Galiana et al., 2015), etc. Generally speaking high discriminatory genes are those that have very different distributions in both classes (in a binary problem) and whose expression remains quite stable or homogeneous within each class.

The algorithm used in this paper is similar to the one that was introduced in Saligan et al. (2014) and deAndrés-Galiana et al. (2015) and consists in several steps (see figure 1):

1. Applying several filter feature selection methods to find different lists of high discriminatory genes.
2. Establishing the predictive accuracy of these lists of genes using a Leave-One-Out-Cross-Validation (LOOCV) cost function via a k-Nearest-Neighbor (k-NN) classifier. Others classifiers could be also used. This sampling procedure of the phenotype prediction uncertainty space aims at obtaining from these lists different biomedical robots with their corresponding predictive accuracy. For that purpose we can use backwards feature elimination and/or random sampling methodologies.

3. Selecting robots above a certain predictive accuracy (or below a given error tolerance) and performing the consensus prediction through majority voting.

According to the definitions stated in (1), (2), (3), and (4) we can formally define a biomedical robot as the set of classifiers:

$$L_{tol} = \{L^*(\mathbf{g}_k) : k = 1, \dots, m\}, \quad (5)$$

whose predictive error (the number of misclassified samples) is lower than a given bound  $tol$ . The phenotype prediction problem with uncertainty estimation consists in, giving an incoming sample  $\mathbf{s}_{new}$ , applying the set of Biomedical robots  $L_{tol}$  (with predictive accuracy higher than  $(100 - tol)\%$ ) and performing the consensus classification. Following the rules of importance sampling, and supposing that the uncertainty analysis was correctly performed, then the probability of  $\mathbf{s}_{new}$  to belong to class  $c_i$  is calculated as the number of robots that predicted the sample to belong to class  $c_i$  divided by the total number of selected robots in the set  $L_{tol}$ .

In this paper we apply this concept to the analysis of 3 kinds of diseases: Cancer (CLL), Neurodegenerative (ALS) and Rare diseases (IBM-PM). Although the concept is theoretically correct before applying it to these datasets, we have analyzed its robustness against different type of noises using synthetic microarrays. This analysis helped us to extract interesting conclusions regarding the interpretation of the results obtained for real datasets.

### 3.1 Noise Sensitivity Analysis

We have generated different synthetic data sets using three types of noise: additive Gaussian noise, lognormal noise, and noise in class assignment. These last two types belong to the category of non-Gaussian noise, since they are multiplicative and systematic random noises. The method consists in building a synthetic dataset with a predefined number of differentially expressed discriminatory genes, and subsequently introducing different types of noise, and determining the predictive accuracy ( $Acc$ ) as a function of the number of applied robots ( $\#R$ ). The synthetic dataset was built using the OCplus package available for The Comprehensive R Archive Network (Pawitan and Ploner, 2015).

Table 1 shows the results obtained for the sensitivity analysis.  $\delta$  represents the level of noise imputed for each type of noise,  $Acc$  the mean LOOCV accuracy,  $P$  the Precision established using the set of genes constructed with the union of all the genes found by the robots, and  $\#R$  the number of robots used in the consensus strategy. The precision is defined as follows:

$$precision = \frac{|\{DE\_genes\} \cap \{Selected\_genes\}|}{|\{Selected\_genes\}|} \quad (6)$$

where  $\{DE\_genes\}$  stands for the set of the differentially expressed genes that we introduced in the synthetic dataset and  $\{Selected\_genes\}$  is the union set of the high discriminatory genes selected by the different robots. This analysis is very important since the correlation networks and biological pathways will be established this way. The results can be summarized as follows:

**Table 1** Noise results. The following information is given:  $\delta$  the percentage of noise introduced, *Acc* the mean LOOCV predictive accuracy, *P* the precision of the selection using the union of all the genes found by the robots and *#R* the number of robots applied in the consensus strategy.

$\delta$ (%)	Class Assignment			Gaussian			Log Gaussian		
	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>	<i>Acc</i> (%)	<i>P</i>	<i>#R</i>
1	98.77	1.00	98	100.00	1.00	98	100.00	1.00	98
3	96.93	1.00	98	100.00	1.00	98	100.00	0.74	98
5	94.48	1.00	98	100.00	1.00	98	100.00	0.35	98
10	90.18	1.00	98	100.00	0.60	98	100.00	0.14	10
15	87.12	1.00	3	100.00	0.33	98	99.39	0.05	37
20	80.98	1.00	1	100.00	0.22	11	100.00	0.03	43
25	77.30	1.00	10	99.39	0.13	81	98.77	0.04	98
30	73.62	0.92	43	99.39	0.14	7	100.00	0.03	14

- The Precision *P* keeps quite stable when noise in class assignment is increased. This result is very interesting since the biomedical robots are able to find the differentially expressed genes when the noise in class assignment is introduced. In the case of Gaussian noise the precision is very high for noise levels less than 5%. The worst result was obtained when multiplicative noise is added to the expressions. The fact that the precision gradually decreases when noise in the expression increases, implies that some of the biological pathways that are inferred might be partially falsified. Therefore, any filtering step that it is usually performed in the microarray data will have important consequences with respect to the pathway analysis. Future research will be devoted to this important subject.
- The mean predictive accuracy (*Acc*) systematically decreases when a higher level of the noise is added to the class assignment vector, and is very stable when Gaussian and non-Gaussian noises are added to the expression data, meaning that the biomedical robots are robust in terms of accuracy with respect to the presence of noise in the expressions. This result also suggests that noise acts as regularization with respect to the accuracy in the prediction as it has been theoretically proved by Fernández-Martínez et al. (2014a,b) in inverse problems. It can be also concluded that if the biomedical robots are unable to improve the accuracy of the best prediction, the dataset must have some wrong class assignment that prevents achieving a perfect classification. Other possibility is that parameterization of the samples is incorrect, that in the present case would mean that none of the genes that have been measured bring enough information to achieve a good phenotype discrimination.

### 3.2 Chronic Lymphocytic Leukemia

B-cell Chronic Lymphocytic Leukemia (CLL) is a complex and molecular heterogeneous disease, being the most common adult Leukemia in western countries. In our cohort DNA analyses served to distinguish two major types of CLL with different survival times based on the maturity of the lymphocytes, as discerned by the Immunoglobulin Heavy chain Variable-region (IgVH) gene mutation status (Ferreira

et al., 2014). In this first example we had at disposal a microarray data set consisting of 163 samples and 48807 probes.

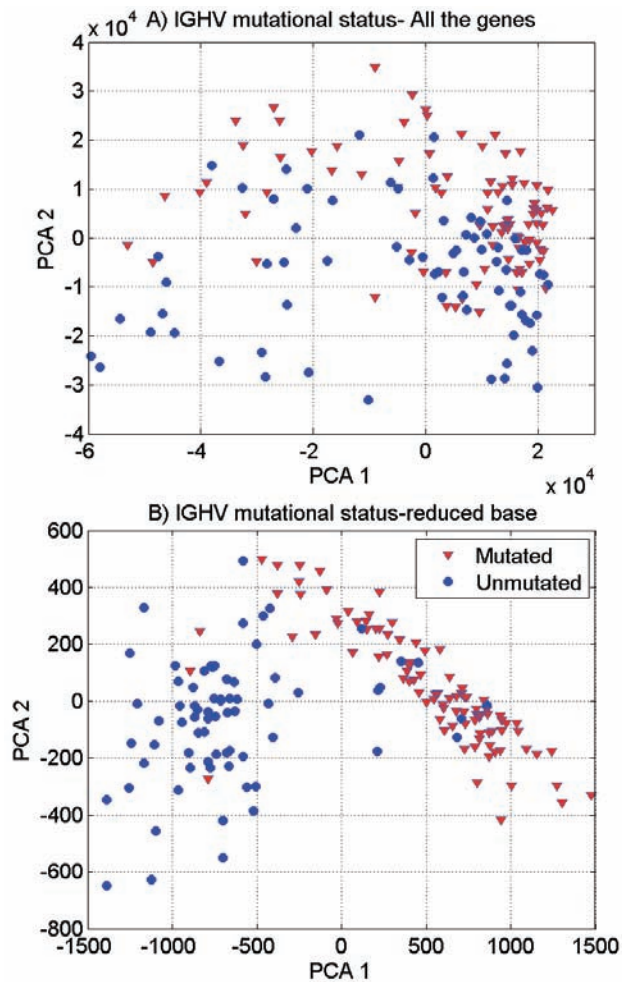
The best robot predicted the IgVH mutational status with 93.25% accuracy using small-scale signature composed by 13 genes: LPL (2 probes), CRY1, LOC100128252 (2 probes), SPG20 (2 probes), ZBTB20, NRIP1, ZAP-70, LDOC1, COBLL1 and NRIP1.

Table 2 shows the results of applying the methodology of biomedical robots to this problem. In this case the highest prediction accuracy obtained by the set of biomedical robots equal the accuracy provided by the best robot (93.25%). This result implies that some samples are behavioral outliers or might be misclassified. This happened with 11 samples that are identified in the PCA graphic in two dimensions (figure 2) using the genetic signature composed of these 13 genes. It can be observed how the classification in this reduced set of genes becomes almost linearly separable while using all the genetic information that we have at disposal the classification is nonlinear. Therefore, as an important conclusion we can affirm that reducing the dimension to the set of discriminatory genes helps to linearize the phenotype classification problem.

**Table 2** CLL, IBM & PM and ALS results

CLL			IBM & PM			ALS		
<i>Acc(%)</i>	<i>tol</i>	<i>#R</i>	<i>Acc(%)</i>	<i>tol</i>	<i>#R</i>	<i>Acc(%)</i>	<i>tol</i>	<i>#R</i>
92.64	85.89	488	87.50	82.50	223	84.71	83.53	547
92.64	86.50	487	87.50	85.00	159	85.88	84.71	441
92.64	89.57	486	90.00	87.50	138	87.06	85.88	241
92.64	90.18	479	90.00	90.00	71	88.24	87.06	197
92.64	90.80	446	92.50	92.50	32	90.59	88.24	134
92.64	91.41	373	100.00	95.00	2	91.76	89.41	96
93.25	92.02	255	97.50	97.50	1	90.59	90.59	54
93.25	92.64	120				92.94	91.76	32
93.25	93.25	22				95.29	92.94	20
						94.12	94.12	10
						95.29	95.29	6
						96.47	96.47	1





**Fig. 2** IgVH classification in CLL: A) Considering all the genes of the microarray, the classification problem is nonlinear. B) Using the most discriminatory genes (13 probes) the classification problem becomes linearly separable.

Figure 3 also shows the correlation network of the most discriminatory genes of the CLL-IgVH mutational status found in this analysis. This is an interesting tool to understand how the most discriminatory genes regulate the expression of other genes involved in different biological pathways. The head of graph is the gene that has the highest discriminatory power LPL. It can be observed one main network associated to ZBTB20.

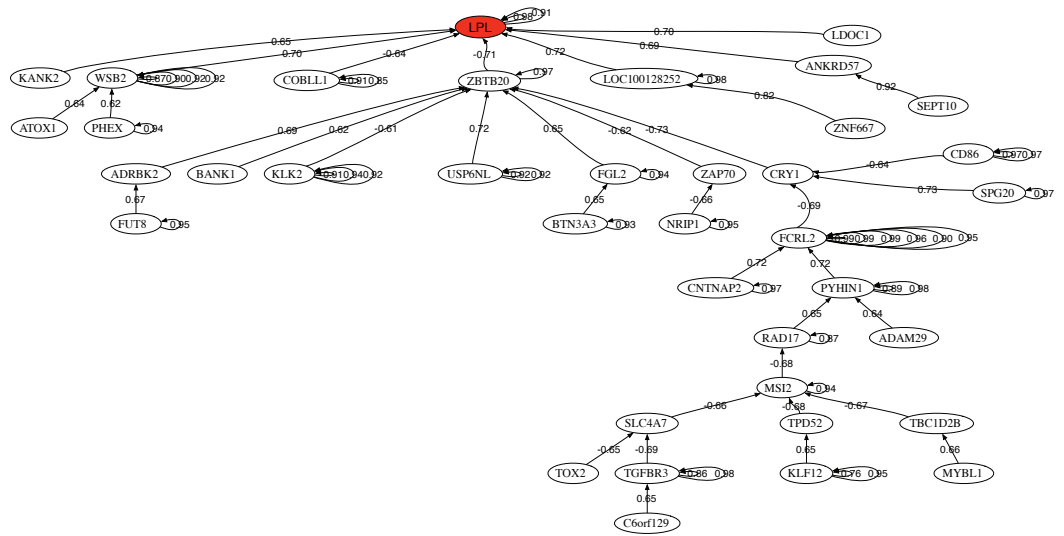


Fig. 3 Correlation network for IgVH mutational status in Chronic Lymphocytic Leukemia.

Finally the pathway analysis deduced from the biomedical robots has revealed the importance of the ERK signaling super pathway that includes ERK signaling, ILK signaling, MAPK signaling, Molecular Mechanisms of cancer and Rho Family GTPases pathway. These pathways control Proliferation, Differentiation, Survival and Apoptosis. Also, other important pathways found were Allograft Rejection, the Inflammatory Response Pathway, CD28 Co-stimulation, TNF-alpha/NF-kB Signaling Pathway, Akt Signaling, PAK Pathway and TNF Signaling. The presence of some of these pathways suggests viral infection as a possible cause for CLL.

### 3.3 Inclusion Body Myositis and Polymyositis

Myositis means muscle inflammation, and can be caused by infection, injury, certain medicines, exercise, and chronic disease. Some of the chronic, or persistent, forms are idiopathic inflammatory myopathies whose cause is unknown. We have modeled the Inclusion Body Myositis /Polymyositis (IBM/PM) dataset published by Greenberg et al. (2005). The data consisted in the microarray analysis of 23 patients with IBM, 6 with PM and 11 samples corresponding to healthy controls. The best robot performed the classification of the IBM+PM vs control obtaining a predictive accuracy of 97.5% using a reduced base with only 17 probes. The genes belonging to the highest predictive small-scale genetic signature are HLA-C (3 probes), HLA-B (4 probes), TMSB10, S100A6, HLA-G, STAT1, TIMP1, HLA-F, IRF9, BID, MLLT11 and PSME2. It can be observed the presence of different HLA-x genes of the major histocompatibility. Particularly the function of the gene HLA-B would explain alone the genesis of IBM: "HLA-B (major histocompatibility complex, class I, B) is a human gene that provides instructions for making a protein that plays a critical role in the immune system. HLA-B is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria".

Table 2 shows the results using the biomedical robots methodology. In this case we are able to hit the 100% of the samples with two robots, improving the results of the best robot. The analysis of biological pathways has revealed the importance of viral infections, mainly in IBM patients: Allograft Rejection, Influenza A, Class I MHC Mediated Antigen Processing and Presentation, Staphylococcus Aureus Infection, Interferon Signaling, Immune Response IFN Alpha/beta Signaling Pathway, Phagosome, Tuberculosis, Cell Adhesion Molecules (CAMs), Epstein-Barr Virus Infection, and TNF Signaling. Several viral infections appeared in this list. Interesting, it has been found that 75% of the cases of viral myositis are due to Staphylococcus Aureus infection (Fayad et al., 2007).

Figure 4 shows the correlation network of the most discriminatory genes found in this analysis. It can be observed the presence of one main dense network involving different HLA-X genes. Among its related pathways are ERK Signaling and Apoptosis Pathway. GO annotations related to this gene include calcium ion binding and cysteine-type peptidase activity.

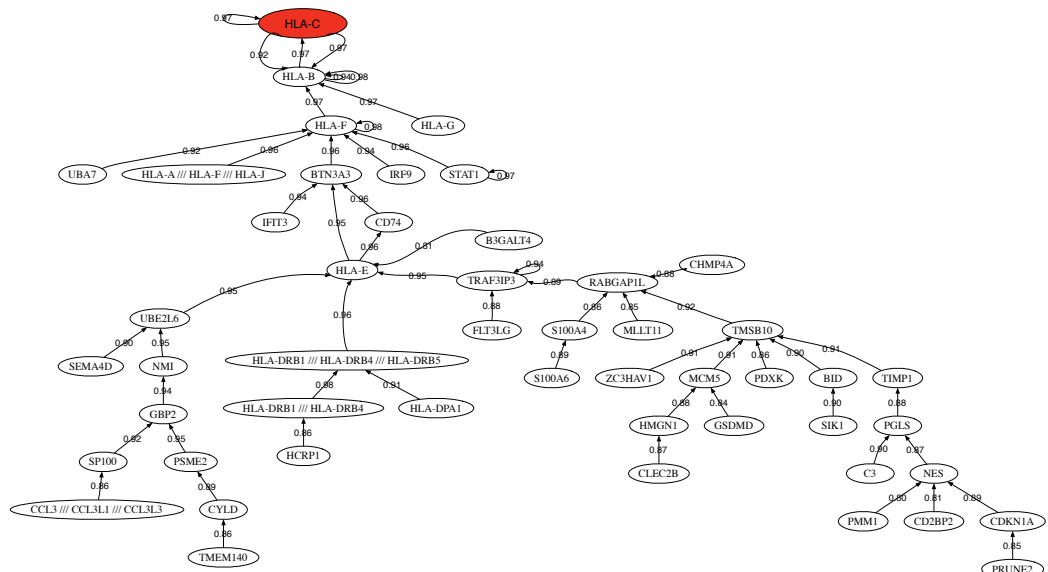
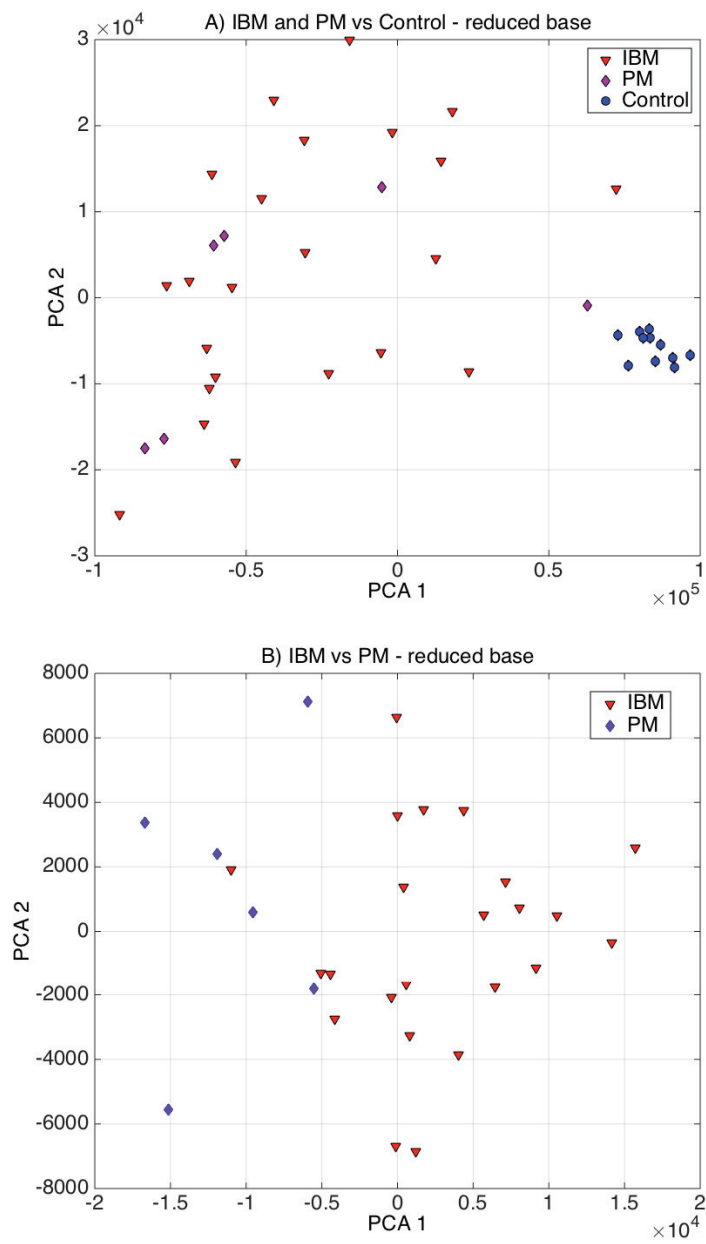


Fig. 4 Correlation network for Inclusion Body Myositis/Polymyositis.

Figure 5 A) shows the PCA projection for the IBM+PM versus control samples using the optimum reduced base. It can be observed that the separability is almost perfect and only one PM sample that is close to the control samples might be misclassified. This graphic also explains that this basis set is not optimum to perform the classification of IBM vs PM. This separability can be achieved with 100% accuracy using a reduced base composed by the following genes: RHOBTB2, MT1P2, FBXL8, HIF3A, C17orf101, RPL12, RBM19, MT1G, WT1-AS, HEXIM1, NQO2, ENOSF1, ADRM1, EIF5A, CSF2RA, CPLX3 /// LMAN1L, C10orf95, NFIC, POLR2J2. The main pathways involved in the IBM vs PM phenotype differentiation is: FOXA1 Transcription Factor Network, O<sub>2</sub>/CO<sub>2</sub> Exchange in Erythrocytes, Methotrexate Pathway, Drug Induction of Bile Acid Pathway, Bile Secretion and Statin Pathway. Figure 5 B) shows the PCA graphic of the IBM vs PM classification, and how this separability can be achieved.



**Fig. 5** Classification of IBM, PM and Control: A) PCA graphic for IBM+PM versus control samples. B) PCA graphic for IBM versus PM.

### 3.4 Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is a motor neuron disease that characterized by stiff muscles, muscle twitching, and gradually worsening weakness. Between 5 and 10% of the cases are inherited from a relative, and for the rest of cases, the cause is still unknown (NINDS, 2013). It is a progressive disease that the average survival from onset to death is three to four years, in which most of them die from a respiratory failure. There is no cure yet.

We reinterpreted the dataset published by Lincecum et al. (2010) consisting of 57 ALS cases and 28 healthy controls. The best result yields an accuracy of 96.5% with small scale signature involving the following genes: CASP1, ZNF787 and SETD7. Table 2 shows the results of applying this methodology to this problem. The biomedical robots in this case did not improve this prediction. The pathway analysis has revealed the importance of the GPCR Pathway, RhoA Signaling Pathway, EPHB Forward Signaling, EphrinA-EphR Signaling, EBV LMP1 Signaling, and Regulation of Microtubule Cytoskeleton. These pathways have different important signaling roles and suggest a possible link to the Epstein-Barr virus (EBV).

Figure 6 shows the correlation network of the most discriminatory genes found in this analysis. The head of the network is the CASP1 that is connected to MAP2K5, through ZNF3 and LUC7. MAP2K5 acts as a scaffold for the formation of a pathway that seems to play a critical role in protecting cells from stress-induced apoptosis, neuronal survival, cardiac development and angiogenesis. Also DCAF8 has been associated to neuropathies.

Figure 7 shows the PCA graphic for the ALS vs control samples. It can be observed that the accuracy of the classification could be easily improved by discarding 5-6 control samples that lie very close to the border defined by the ALS samples. Also it can be observed that one ALS sample is clearly a behavioral outlier.



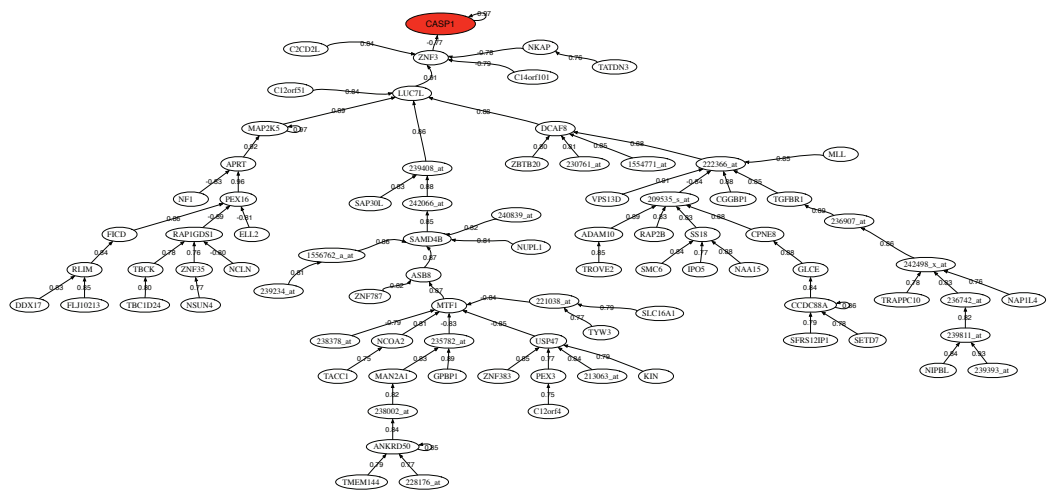


Fig. 6 Correlation network for Amyotrophic Lateral Sclerosis. Probe names are used when gene names are unknown.

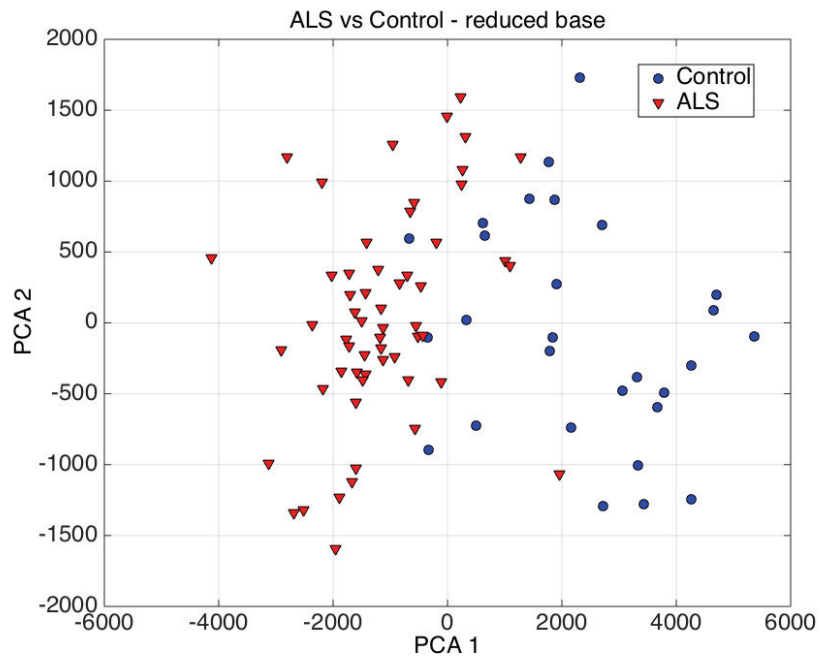


Fig. 7 PCA graphic for ALS versus control samples

## 4 Conclusion

In this paper we have introduced the concept of biomedical robots, performing its sensitivity analysis to different kind of noises, and showing its application to the analysis of cancer, rare and neurodegenerative diseases. The concept of biomedical robot is based in exploring the uncertainty space of the phenotype classification problem involved, and using the structure of the uncertainty space to adopt decisions and inferring knowledge. The synthetic dataset modeling has shown the robustness and stability of this methodology, particularly to class assignment noise. The presence of high noise levels in expressions might falsify the biological pathways that are inferred. Nevertheless, the predictive accuracy remains very high. Finally, we have shown the application of this novel concept to 3 different illnesses: CLL, IBM-PM and ALS, proving that it is possible to infer at the same time, both, high discriminatory small-scale signatures and the description of the biological pathways involved. We have shown that referring to the set of most discriminatory genes these classification problems becomes linearly separable. Generally speaking in the 3 cases no high class assignment errors have been detected, being CLL the case where more samples (11) have been found to be behavioral outliers. The pathway analyses revealed in the three cases a possible link to viral infections and served to identify actionable genes and drug targets. The methodology shown in this paper is not computationally very expensive, since all the simulations shown in this paper were done with a personal computer in real time (several minutes).

## References

- deAndrés-Galiana E. J., Fernández-Martínez J. L., Luaces O., Del Coz J. J., Fernández R., Solano J., Nogues E. A., Zanabilli Y., Alonso J. M., Payer A. R., Vicente J. M., Medina J., Taboada F., Vargas M., Alarcon C., Moran M., Gonzalez-Ordóñez A., Palicio M. A., Ortiz S., Chamorro C., Gonzalez S., and Gonzalez-Rodriguez A. P. (2015). On the prediction of hodgkin lymphoma treatment response. *Clin Transl Oncol*, 17(8):612–619.
- Fayad L. M., Carrino J. A., and Fishman E. K. (2007). Musculoskeletal infection: role of ct in the emergency department. *Radiographics*, 27(6):1723–1736.
- Fernández-Martínez J. L. and Cernea A. (2014). Exploring the uncertainty space of ensemble classifiers in face recognition. *Int J Pattern Recognit Artif Intell*, 108:186–193.
- Fernández-Martínez J. L., Fernández-Muñiz M., and Tompkins M. (2012). On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics*, 77(1):1–15.
- Fernández-Martínez J. L., Fernández-Muñiz Z., Pallero J., and Pedruelo-González L. (2013). From bayes to tarantola: New insights to understand uncertainty in inverse problems. *Journal of Applied Geophysics*, 98:62–72.
- Fernández-Martínez J. L., Pallero J., and Fernández Muñiz Z. (2014a). The effect of noise and tikhonov’s regularization in inverse problems. part I: The linear case. *Journal of Applied Geophysics*, 108:176 – 185.

- Fernández-Martínez J. L., Pallero J., and Fernández Muñiz Z. (2014b). The effect of noise and tikhonov's regularization in inverse problems. part II: The nonlinear case. *Journal of Applied Geophysics*, 108:186 – 193.
- Ferreira P. G., Jares P., Rico D., Gomez-Lopez G., Martinez-Trillos A., Villamor N., Ecker S., Gonzalez-Perez A., Knowles D. G., Monlong J., Johnson R., Quesada V., Djebali S., Papasaikas P., Lopez-Guerra M., Colomer D., Royo C., Cazorla M., Pinyol M., Clot G., Aymerich M., Rozman M., Kulis M., Tamborero D., Gouin A., Blanc J., Gut M., Gut I., Puente X. S., Pisano D. G., Martin-Subero J. I., Lopez-Bigas N., Lopez-Guillermo A., Valencia A., Lopez-Otin C., Campo E., and Guigo R. (2014). Transcriptome characterization by rna sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res*, 24(2):212–226.
- Fisher R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–88.
- Greenberg S. A., Bradshaw E. M., Pinkus J. L., Pinkus G. S., Burleson T., Due B., Bregoli L., O'Connor K. C., and Amato A. A. (2005). Plasma cells in muscle in inclusion body myositis and polymyositis. *Neurology*, 65(11):1782–1787.
- Lincecum J. M., Vieira F. G., Wang M. Z., Thompson K., De Zutter G. S., Kidd, J., Moreno A., Sanchez, R., Carrion I. J., Levine B. A., Al-Nakhala B. M., Sullivan S. M., Gill A., and Perrin S. (2010). From transcriptome analysis to therapeutic anti-cd40l treatment in the sod1 model of amyotrophic lateral sclerosis. *Nat Genet*, 42(5):392–399.
- NINDS (2013). *Motor Neuron Diseases Fact Sheet*. National Institute of Neurological Disorders and Stroke.
- Pawitan Y. and Ploner A. (2015). Ocplus: Operating characteristics plus sample size and local fdr for microarray experiments. *R package*.
- Quinlan J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.
- Saligan L. N., deAndrés-Galiana E. J., Fernández-Martínez J. L., and Sonis S. (2014). Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Inform*, 13:141–152.
- Schena M., Shalon D., Heller R., Chai A., Brown P., and Davis R. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *PNAS*, 93(20):10614–10619.
- Shannon C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(379–423, 623).
- Tusher V., Tibshirani R., and Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*, 98(9):5116–21.

## **Appendix B**

### **Concentrations in the Spike-In experiment**

Sample ID	ID Group	203508_at	204563_at
12_13_02_U133A_Mer_Latin_Square_Expt1_R1	1	0	0
12_13_02_U133A_Mer_Latin_Square_Expt1_R2	1	0	0
12_13_02_U133A_Mer_Latin_Square_Expt1_R3	1	0	0
12_13_02_U133A_Mer_Latin_Square_Expt2_R1	2	0.125	0.125
12_13_02_U133A_Mer_Latin_Square_Expt2_R2	2	0.125	0.125
12_13_02_U133A_Mer_Latin_Square_Expt2_R3	2	0.125	0.125
12_13_02_U133A_Mer_Latin_Square_Expt3_R1	3	0.25	0.25
12_13_02_U133A_Mer_Latin_Square_Expt3_R2	3	0.25	0.25
12_13_02_U133A_Mer_Latin_Square_Expt3_R3	3	0.25	0.25
12_13_02_U133A_Mer_Latin_Square_Expt4_R1	4	0.5	0.5
12_13_02_U133A_Mer_Latin_Square_Expt4_R2	4	0.5	0.5
12_13_02_U133A_Mer_Latin_Square_Expt4_R3	4	0.5	0.5
12_13_02_U133A_Mer_Latin_Square_Expt5_R1	5	1	1
12_13_02_U133A_Mer_Latin_Square_Expt5_R2	5	1	1
12_13_02_U133A_Mer_Latin_Square_Expt5_R3	5	1	1
12_13_02_U133A_Mer_Latin_Square_Expt6_R1	6	2	2
12_13_02_U133A_Mer_Latin_Square_Expt6_R2	6	2	2
12_13_02_U133A_Mer_Latin_Square_Expt6_R3	6	2	2
12_13_02_U133A_Mer_Latin_Square_Expt7_R1	7	4	4
12_13_02_U133A_Mer_Latin_Square_Expt7_R2	7	4	4
12_13_02_U133A_Mer_Latin_Square_Expt7_R3	7	4	4
12_13_02_U133A_Mer_Latin_Square_Expt8_R1	8	8	8
12_13_02_U133A_Mer_Latin_Square_Expt8_R2	8	8	8
12_13_02_U133A_Mer_Latin_Square_Expt8_R3	8	8	8
12_13_02_U133A_Mer_Latin_Square_Expt9_R1	9	16	16
12_13_02_U133A_Mer_Latin_Square_Expt9_R2	9	16	16
12_13_02_U133A_Mer_Latin_Square_Expt9_R3	9	16	16
12_13_02_U133A_Mer_Latin_Square_Expt10_R1	10	32	32
12_13_02_U133A_Mer_Latin_Square_Expt10_R2	10	32	32
12_13_02_U133A_Mer_Latin_Square_Expt10_R3	10	32	32
12_13_02_U133A_Mer_Latin_Square_Expt11_R1	11	64	64
12_13_02_U133A_Mer_Latin_Square_Expt11_R2	11	64	64
12_13_02_U133A_Mer_Latin_Square_Expt11_R3	11	64	64
12_13_02_U133A_Mer_Latin_Square_Expt12_R1	12	128	128
12_13_02_U133A_Mer_Latin_Square_Expt12_R2	12	128	128
12_13_02_U133A_Mer_Latin_Square_Expt12_R3	12	128	128
12_13_02_U133A_Mer_Latin_Square_Expt13_R1	13	256	256
12_13_02_U133A_Mer_Latin_Square_Expt13_R2	13	256	256
12_13_02_U133A_Mer_Latin_Square_Expt13_R3	13	256	256
12_13_02_U133A_Mer_Latin_Square_Expt14_R1	14	512	512
12_13_02_U133A_Mer_Latin_Square_Expt14_R2	14	512	512
12_13_02_U133A_Mer_Latin_Square_Expt14_R3	14	512	512

204513_s_at	204205_at	204959_at	207655_s_at	204836_at	205291_at	209795_at
0	0.125	0.125	0.125	0.25	0.25	0.25
0	0.125	0.125	0.125	0.25	0.25	0.25
0	0.125	0.125	0.125	0.25	0.25	0.25
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.25	0.5	0.5	0.5	1	1	1
0.25	0.5	0.5	0.5	1	1	1
0.25	0.5	0.5	0.5	1	1	1
0.5	1	1	1	2	2	2
0.5	1	1	1	2	2	2
0.5	1	1	1	2	2	2
1	2	2	2	4	4	4
1	2	2	2	4	4	4
1	2	2	2	4	4	4
2	4	4	4	8	8	8
2	4	4	4	8	8	8
2	4	4	4	8	8	8
4	8	8	8	16	16	16
4	8	8	8	16	16	16
4	8	8	8	16	16	16
8	16	16	16	32	32	32
8	16	16	16	32	32	32
8	16	16	16	32	32	32
16	32	32	32	64	64	64
16	32	32	32	64	64	64
16	32	32	32	64	64	64
32	64	64	64	128	128	128
32	64	64	64	128	128	128
32	64	64	64	128	128	128
64	128	128	128	256	256	256
64	128	128	128	256	256	256
64	128	128	128	256	256	256
128	256	256	256	512	512	512
128	256	256	256	512	512	512
128	256	256	256	512	512	512
256	512	512	512	0	0	0
256	512	512	512	0	0	0
256	512	512	512	0	0	0
512	0	0	0	0.125	0.125	0.125
512	0	0	0	0.125	0.125	0.125
512	0	0	0	0.125	0.125	0.125

207777_s_at	204912_at	205569_at	207160_at	205692_s_at	212827_at	209606_at
0.5	0.5	0.5	1	1	1	2
0.5	0.5	0.5	1	1	1	2
0.5	0.5	0.5	1	1	1	2
1	1	1	2	2	2	4
1	1	1	2	2	2	4
1	1	1	2	2	2	4
2	2	2	4	4	4	8
2	2	2	4	4	4	8
2	2	2	4	4	4	8
4	4	4	8	8	8	16
4	4	4	8	8	8	16
4	4	4	8	8	8	16
8	8	8	16	16	16	32
8	8	8	16	16	16	32
8	8	8	16	16	16	32
16	16	16	32	32	32	64
16	16	16	32	32	32	64
16	16	16	32	32	32	64
32	32	32	64	64	64	128
32	32	32	64	64	64	128
32	32	32	64	64	64	128
64	64	64	128	128	128	256
64	64	64	128	128	128	256
64	64	64	128	128	128	256
128	128	128	256	256	256	512
128	128	128	256	256	256	512
128	128	128	256	256	256	512
256	256	256	512	512	512	0
256	256	256	512	512	512	0
256	256	256	512	512	512	0
512	512	512	0	0	0	0.125
512	512	512	0	0	0	0.125
512	512	512	0	0	0	0.125
0	0	0	0.125	0.125	0.125	0.25
0	0	0	0.125	0.125	0.125	0.25
0	0	0	0.125	0.125	0.125	0.25
0.125	0.125	0.125	0.25	0.25	0.25	0.5
0.125	0.125	0.125	0.25	0.25	0.25	0.5
0.125	0.125	0.125	0.25	0.25	0.25	0.5
0.25	0.25	0.25	0.5	0.5	0.5	1
0.25	0.25	0.25	0.5	0.5	0.5	1
0.25	0.25	0.25	0.5	0.5	0.5	1



205267_at	204417_at	205398_s_at	209734_at	209354_at	206060_s_at	205790_at
2	2	4	4	4	8	8
2	2	4	4	4	8	8
2	2	4	4	4	8	8
4	4	8	8	8	16	16
4	4	8	8	8	16	16
4	4	8	8	8	16	16
8	8	16	16	16	32	32
8	8	16	16	16	32	32
8	8	16	16	16	32	32
16	16	32	32	32	64	64
16	16	32	32	32	64	64
16	16	32	32	32	64	64
32	32	64	64	64	128	128
32	32	64	64	64	128	128
32	32	64	64	64	128	128
64	64	128	128	128	256	256
64	64	128	128	128	256	256
64	64	128	128	128	256	256
128	128	256	256	256	512	512
128	128	256	256	256	512	512
128	128	256	256	256	512	512
256	256	512	512	512	0	0
256	256	512	512	512	0	0
256	256	512	512	512	0	0
512	512	0	0	0	0.125	0.125
512	512	0	0	0	0.125	0.125
512	512	0	0	0	0.125	0.125
0	0	0.125	0.125	0.125	0.25	0.25
0	0	0.125	0.125	0.125	0.25	0.25
0	0	0.125	0.125	0.125	0.25	0.25
0.125	0.125	0.25	0.25	0.25	0.5	0.5
0.125	0.125	0.25	0.25	0.25	0.5	0.5
0.125	0.125	0.25	0.25	0.25	0.5	0.5
0.25	0.25	0.5	0.5	0.5	1	1
0.25	0.25	0.5	0.5	0.5	1	1
0.25	0.25	0.5	0.5	0.5	1	1
0.5	0.5	1	1	1	2	2
0.5	0.5	1	1	1	2	2
0.5	0.5	1	1	1	2	2
1	1	2	2	2	4	4
1	1	2	2	2	4	4
1	1	2	2	2	4	4

200665_s_at	207641_at	207540_s_at	204430_s_at	203471_s_at	204951_at	207968_s_at
8	16	16	16	32	32	32
8	16	16	16	32	32	32
8	16	16	16	32	32	32
16	32	32	32	64	64	64
16	32	32	32	64	64	64
16	32	32	32	64	64	64
32	64	64	64	128	128	128
32	64	64	64	128	128	128
32	64	64	64	128	128	128
64	128	128	128	256	256	256
64	128	128	128	256	256	256
64	128	128	128	256	256	256
128	256	256	256	512	512	512
128	256	256	256	512	512	512
128	256	256	256	512	512	512
256	512	512	512	0	0	0
256	512	512	512	0	0	0
256	512	512	512	0	0	0
512	0	0	0	0.125	0.125	0.125
512	0	0	0	0.125	0.125	0.125
512	0	0	0	0.125	0.125	0.125
0	0.125	0.125	0.125	0.25	0.25	0.25
0	0.125	0.125	0.125	0.25	0.25	0.25
0	0.125	0.125	0.125	0.25	0.25	0.25
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.125	0.25	0.25	0.25	0.5	0.5	0.5
0.25	0.5	0.5	0.5	1	1	1
0.25	0.5	0.5	0.5	1	1	1
0.25	0.5	0.5	0.5	1	1	1
0.5	1	1	1	2	2	2
0.5	1	1	1	2	2	2
0.5	1	1	1	2	2	2
1	2	2	2	4	4	4
1	2	2	2	4	4	4
1	2	2	2	4	4	4
2	4	4	4	8	8	8
2	4	4	4	8	8	8
2	4	4	4	8	8	8
4	8	8	8	16	16	16
4	8	8	8	16	16	16
4	8	8	8	16	16	16

AFFX-r2-TagA_at	AFFX-r2-TagB_at	AFFX-r2-TagC_at	AFFX-r2-TagD_at	AFFX-r2-TagE_at
64	64	64	128	128
64	64	64	128	128
64	64	64	128	128
128	128	128	256	256
128	128	128	256	256
128	128	128	256	256
256	256	256	512	512
256	256	256	512	512
256	256	256	512	512
512	512	512	0	0
512	512	512	0	0
512	512	512	0	0
0	0	0	0.125	0.125
0	0	0	0.125	0.125
0	0	0	0.125	0.125
0.125	0.125	0.125	0.25	0.25
0.125	0.125	0.125	0.25	0.25
0.125	0.125	0.125	0.25	0.25
0.25	0.25	0.25	0.5	0.5
0.25	0.25	0.25	0.5	0.5
0.25	0.25	0.25	0.5	0.5
0.5	0.5	0.5	1	1
0.5	0.5	0.5	1	1
0.5	0.5	0.5	1	1
1	1	1	2	2
1	1	1	2	2
1	1	1	2	2
2	2	2	4	4
2	2	2	4	4
2	2	2	4	4
4	4	4	8	8
4	4	4	8	8
4	4	4	8	8
8	8	8	16	16
8	8	8	16	16
8	8	8	16	16
16	16	16	32	32
16	16	16	32	32
16	16	16	32	32
32	32	32	64	64
32	32	32	64	64
32	32	32	64	64

AFFX-r2-TagF_at	AFFX-r2-TagG_at	AFFX-r2-TagH_at	AFFX-DapX-3_at	AFFX-LysX-3_at
128	256	256	256	512
128	256	256	256	512
128	256	256	256	512
256	512	512	512	0
256	512	512	512	0
256	512	512	512	0
512	0	0	0	0.125
512	0	0	0	0.125
512	0	0	0	0.125
0	0.125	0.125	0.125	0.25
0	0.125	0.125	0.125	0.25
0	0.125	0.125	0.125	0.25
0.125	0.25	0.25	0.25	0.5
0.125	0.25	0.25	0.25	0.5
0.125	0.25	0.25	0.25	0.5
0.25	0.5	0.5	0.5	1
0.25	0.5	0.5	0.5	1
0.25	0.5	0.5	0.5	1
0.5	1	1	1	2
0.5	1	1	1	2
0.5	1	1	1	2
1	2	2	2	4
1	2	2	2	4
1	2	2	2	4
2	4	4	4	8
2	4	4	4	8
2	4	4	4	8
4	8	8	8	16
4	8	8	8	16
4	8	8	8	16
8	16	16	16	32
8	16	16	16	32
8	16	16	16	32
16	32	32	32	64
16	32	32	32	64
16	32	32	32	64
32	64	64	64	128
32	64	64	64	128
32	64	64	64	128
64	128	128	128	256
64	128	128	128	256
64	128	128	128	256

AFFX-PheX-3_at	AFFX-ThrX-3_at
512	512
512	512
512	512
0	0
0	0
0	0
0.125	0.125
0.125	0.125
0.125	0.125
0.25	0.25
0.25	0.25
0.25	0.25
0.5	0.5
0.5	0.5
0.5	0.5
1	1
1	1
1	1
2	2
2	2
2	2
4	4
4	4
4	4
8	8
8	8
8	8
16	16
16	16
16	16
32	32
32	32
32	32
64	64
64	64
64	64
128	128
128	128
128	128
256	256
256	256
256	256

