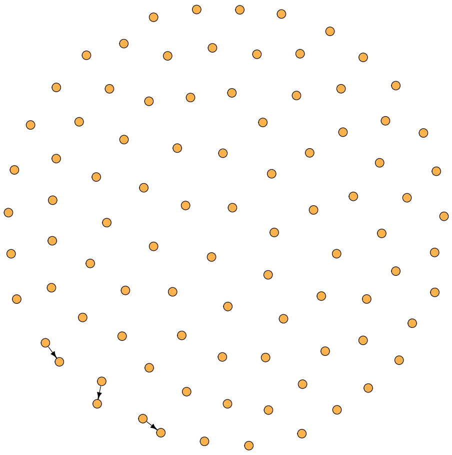


Machine Learning methods for finding predictors of Rheumatoid Arthritis' treatment response

Joana Alter Palhinha

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Alexandra Sofia Martins de Carvalho
Prof. Susana de Almeida Mendes Vinga Martins

Examination Committee

Chairperson: Prof. Maria Margarida Campos da Silveira
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Dr. Sara Guilherme Oliveira da Silva

December 2020

Ao avô Tó e ao avô Palhinha.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at INESC-ID and IT, in collaboration with the Epi-DoC Unit, CEDOC, during the period of October 2019-December 2020, under the supervision of Prof. Alexandra Carvalho and Prof. Susana Vinga. This work is a result of the Project PREDICT (PTDC/CCI-CIF/29877/2017), funded by Fundo Europeu de Desenvolvimento Regional (FEDER), through Programa Operacional Regional LISBOA (LISBOA2020), and by national funds, through Fundação para a Ciência e Tecnologia (FCT).

Acknowledgments

I would first like to thank Professors Susana and Alexandra without whom this thesis could not have been completed. Thank you for all the understanding and for showing me I had a story to tell. I am also grateful to Cláudia, Professor Gökmen, Carolina, Rodrigo, Cristiano and Ricardo, whose wisdom lead to valuable input throughout this project. To my family, thank you for your unwavering support and fantastic laughs.

Finally, I can not go without mentioning my friends, whom with I have shared great melodies and unforgettable stories. *Irish and proud, baby, naturally.*

Resumo

A artrite reumatóide (AR) é uma doença autoimune caracterizada por uma resposta inflamatória crónica provocando inicialmente lesões nas articulações e podendo levar à perda da sua função. A incerteza da sua causa e a heterogeneidade dos pacientes dificultam o processo terapêutico. A combinação de medicamentos biológicos modificadores de AR, nomeadamente inibidores do fator de necrose tumoral (anti-TNF) com metotrexato constitui uma abordagem terapêutica comum. Porém, a dificuldade em prever o tipo de resposta do paciente à medicação constitui um grande obstáculo, dado que no eventual caso desta funcionar, os seus efeitos são apenas sentidos meses após o início da administração, o que leva a uma evolução dos sintomas e acarreta custos financeiros. A identificação de biomarcadores tem sido um tema incansável que, com base em métodos de aprendizagem automática, visa compreender os mecanismos desencadeadores da AR e alcançar a melhor terapêutica, que no limite constitui a "medicina de precisão". Esta tese envolveu a análise de dados transcriptómicos de pacientes com AR em instâncias diferentes de tratamento com anti-TNF. Regressão logística esparsa permitiu a seleção das características relevantes. Redes Bayesianas identificaram duas interações entre proteínas (*MPO–CTSG* e *CTSG–AZU1*) indicadoras da eficácia do tratamento e sabidas relevantes na comunidade científica. e os seus desempenhos comparados a nível de esparsidade, influência das funções de normalização/transformação e tipo de algoritmo. A regressão logística esparsa aliada à análise Bayesiana permitiu identificar biomarcadores com potencial clínico.

Palavras-chave: Artrite reumatóide, Biomarcador, Modelos esparsos, Previsão de resposta, Rede Bayesiana.

Abstract

Rheumatoid arthritis (RA) is an autoimmune disease characterized by chronic inflammatory response causing joint damage and ultimately severe disability. There is no cure for RA and it is regarded as therapeutically challenging due to patient heterogeneity and variability. Biologic agents such as anti-TNF (tumor necrosis factor) in combination with metothrexate are a common first approach. However, the problem remains to be the prediction of the patient's response, since the eventual positive results are only observable months after initiating treatment. This is unsettling because, regardless of the patient's response, the anticipation period may cause irreversible repercussions and have a socio-economic impact. Many researches have focused the use of machine learning algorithms on finding biomarkers (*e.g.* at transcriptomic or protein level) which can aid the understanding of RA pathogenesis and consequently find the appropriate treatment. The development of improved analysis strategies is leading towards precision medicine. This thesis applied a sparse logistic regression framework to transcriptomic data of RA patients collected at day 0 and day 90 of anti-TNF treatment in order to select the significant features. Bayesian network learning allowed for the identification of known protein-protein interactions, such as *MPO–CTSG* and *CTSG–AZU1* for patients regarded, respectively, as good-responders and non-responders, according to the EULAR criteria. Finally, different classification algorithms were tested in order to evaluate parameters such as sparsity, influence of normalization methods and performance based on their continuous/discrete/voom-based nature. Structured sparse regression conjugated with Bayesian learning identified RA biomarkers which potentially can support the clinical domain.

Keywords: Bayesian network, Biomarker, Response prediction, Rheumatoid arthritis, Sparse models.

Contents

List of Tables

List of Figures

Nomenclature

Acc	Accuracy
ACPA	Anticitrillunated protein antibodies
ACR	American College of Rheumatology
AUC	Area Under the Curve
bDMARD	Biologic Disease-modifying anti rheumatic drug
BL	Baseline dataset
BN	Bayesian Network
CDAI	Clinical Disease Activity Score
csDMARD	Conventional synthetic Disease-modifying anti rheumatic drug
CV	Cross-Validation
DAG	Directed Acyclic Graph
DAS28	Disease Activity Score at 28 Joints
DAS	Disease Activity Score
DESeq	Differential Expression Analysis for Sequence Count
DLDA	Diagonal Linear Discriminant Analysis
DMARD	Disease-modifying anti rheumatic drug
DNA	Deoxyribonucleic acid
EULAR	European League Against Rheumatism
GLM	Generalized Linear Model
HAQ-DI	Health Assessment Questionnaire Disability Index
HAQ	Health Assessment Questionnaire
log-cpm	logarithm of counts per million reads

LOOCV	Leave-one-out Cross-Validation
M03	Post-three month dataset
MHC	Histocompatibility complex
NBLDA	Negative BinomialLinear Discriminant Analysis
NR	Non-Responder
NSC	Nearest Shrunken Centroid
OLS	Ordinary Least Squares
PLDA2	Power-transformed Poisson Linear Discriminant Analysis
PLDA	Poisson Linear Discriminant Analysis
PrGA	Provider/Physician Global assessment of Disease Activity
PRO	Patient-reported outcome
PtGA	Patient Global assessment of Disease Activity
RA	Rheumatoid Arthritis
RF	Rheumatoid Factor
rf	Random Forest
rlog	regularized logarithmic
RNA-Seq	Ribonucleic acid Sequencing
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
R	Responder
SDAI	Simplified Disease Activity Score
SVM	Support Vector Machine
TMM	Trimmed Mean of the M-values
TNF	Tumor Necrosis Factor
tsDMARD	Targated Synthetic Disease-modifying anti rheumatic drug
VAS	Visual Analog Scale
voom	variance modeling at the observational level
vst	variance stabilizing transformation

Chapter 1

Introduction

What does it implicate to be diagnosed with rheumatoid arthritis in the current decade? How does artificial intelligence aid the medical physician, specifically in the rheumatic diseases field? Is transcriptomic data collection the direction to the ultimate goal of precision medicine? These are some of the questions that instigated the present study. This chapter intends to broaden their answers and introduce this thesis subject.

1.1 Motivation

Rheumatoid arthritis (RA) is an auto-immune inflammatory progressive disorder affecting primarily the joint system. In the absence of appropriate treatment, it causes joint destruction, leading to reduced life quality, decreased life expectancy and increased risk of cardiovascular diseases. Being a chronic disease, there is no cure for RA. Its prevalence amongst the Portuguese population is estimated to be between 0.8% and 1.5%, being women more likely to be affected than men. An early recognition is fundamental because if RA is diagnosed between the first three to six months of disease activity and treated correctly, there are great chances of preventing functional disabilities and thus contributing to a better life quality [1, 2].

Biologic agents have revolutionized the treatment of RA over the last decade. Their efficacy and safety has been clearly demonstrated in the setting of a multitude of randomized controlled trials [3]. However, assuming that once the diagnose is complete the treatment choice is pretty straight forward would be a misconception. In fact, regardless of the major improvements researchers have made in the RA field in the last decades, there are some obstacles: (1) considerably long list of options and possible combinations regarding medical treatments; (2) uncertainty about their effect on a specific patient's health and (3) waiting time, sometimes it can reach several months, until the drug exerts its effect. The long document produced by the European League Against Rheumatism (EULAR, [4]) suggests the unlikely chance of hitting the right course of action at first. Unfortunately this means patients will undergo a trial-and-error approach facing a long period of drug experiment until the right therapy is found. Successive changes in drug administration can worsen patient disability and may have a big financial impact.

The possibility of predicting the patient's response to a specific treatment is very encouraging considering it would on the one hand, prevent the disease to cause irreversible damage in the joints and bone erosion, and on the other hand reduce duration of time until clinical remission or low activity disease. For these reasons, this study is devoted to identifying gene factors influencing response to anti-TNF therapy in RA (TNF, tumour necrosis factor, is an important host defence molecule long defined as a good target in RA [5]).

For the last four decades, modern society has witnessed tremendous advances in data storage and computing processing power which will dramatically mark the era in which we live. Machine learning, a field that falls under the artificial intelligence concept, has prospered with an astounding acceleration rate, changing every industry and having a substantial impact on our day-to-day lives. Medical approaches backed by machine learning provide a powerful tool by bringing together data, extracting insights and presenting it to physicians for their evaluation [6].

The significant increase in data storage has brought a new dilemma. Nowadays, the focus is no longer on how to collect data but rather how to analyse it and extract relevant information from it. This area of expertise is called data mining. The goal is to develop new intelligent analytics and workflow technologies that enable finding interesting patterns amongst enormous databases without any *a priori* hypothesis and usually from observational data.

High-dimensional data has also become increasingly available in all fields of research and thus it is of great interest to effectively analyse it *i.e.*, to be able to, from a big dataset with an abounding number of features, find the essential ones that influence or are related to a certain event under study. This idea can be translated as sparsity. Thus, sparse models, which are obtained with regularization methods, perform feature selection. They are of special interest when dealing with situations in which the number of variables (for example, in a genetic study, they would correspond to the several thousand genes) is much greater than the number of observations (corresponding to the subjects under study, in the same example).

Graphical models are a common framework used in a broad group of fields such as genetics, oncology, computational biology, medicine and healthcare, finance, among others. This probabilistic tool has aided in the discovery of novel biological mechanisms [7]. Not disregarding the regularization methods, which are essential for feature selection, they do not take into account the possible interrelations among the variables. Bayesian networks (BNs), the most widely known directed graphical model, emerge as an intuitive approach that allows understanding the dependency structure of the underlying data distribution, *e.g.*, whether two variables are in direct interaction [8].

Transcriptomics is the study of the transciptome, a term widely understood as the complete set of all the ribonucleic acid (RNA) molecules expressed in some given entity, for example, a cell, tissue or organism. Complex diseases are characterized by a variety of molecular aberrations including gene expression changes. Therefore, a multidimensional understanding of the molecular features underlying a complex disease phenotype is required for the development of effective intervention strategies. Transcriptome sequencing or gene expression profiling can be achieved by *microarray* (a high-throughput method involving hybridization of micro-RNA to an array of complementary deoxyribonucleic acid [DNA] probes corresponding to genes of interest), allowing to monitor thousands of gene expression levels simultaneously to study how they are affected by certain treatments or diseases. This type of experiments produce massive amounts of data which requires suitable computational tools to process it [9]. Alternatively to microarray, *RNA sequencing* (RNA-Seq) is an ever increasingly popular tool for transcriptome profiling, but also other aspects of RNA biology, with clear advantages over the former method: it can identify transcripts in species without genomic known sequences; it has a much lower noise level than microarray, which is caused by cross-hybridization; it is much more sensitive than microarray-based methods in detecting low and high expressed genes. The quantity and sequences of RNA in a sample are examined using next-generation sequencing and then quantified by counting the number of reads [10, 11]. RNA-Seq data is usually represented by a matrix of counts showing the expression levels of micro-RNAs for a set of samples. For each sample, millions of reads can be measured by the RNA-Seq technique.

Conventional bioinformatics approaches have largely been designed for making population-level inferences about "average" disease processes but they do not adequately capture and describe individual variability i.e., they only allow physicians to stratify treatments according to some patient characteristics. Unlike them, transcriptomics allied with the computational development is a promising tool that can bring novel insights in disease mechanisms specific of a patient and unveil potential patient-specific treatments. This is a growing field essential for *precision medicine* which aims at devising a different treatment for each individual patient [12].

A common and relevant concept in this type of clinical research is biomarker. This portmanteau of "biological marker" includes the medical signs which can be measured accurately and reproducibly. They contrast with medical symptoms in the way that the former are objective indications of the medical state observed from outside the patient while the latter are subject to the patient's perception of health or illness [13].

1.2 State-of-the-art

The beginning of the search for RA response predictors coincided with exploring readily available clinical parameters in routine care. One of the few successes at start was the finding that absence in the blood of the autoantibodies rheumatoid factor (RF) and anticitrillunated protein antibodies (ACPA) can, very marginally, predict a decreased chance of response to rituximab, a type of biological treatment drug. However, no recommendations for a tailored treatment approach according to serologic autoantibody status are embedded in the official guidelines, most likely due to the small differences in response according to serological status. Besides, rituximab is still considered an effective biological in all RA patients and a good first choice after methotrexate (a commonly prescribed drug) failure [14].

Due to the limited additive success achieved by clinical parameters to discriminate patients' responses to biologicals, the focus shifted to biochemical markers more closely related to RA pathogenesis, arising from the cascade DNA, RNA, epigenetics, proteins, and metabolites. However, the still unresolved heterogeneity in RA disease processes and in clinical response to therapy makes prediction of therapeutic response one of the major challenges in RA. The exploration of these associations has been extensively reviewed [14–20].

For example, the use of clinical and molecular biomarkers was proposed for informing the choice of biologic treatment on a group level in the context of stratified medicine by Wijbrandts and Tak [16]. Farutin et al., using the same dataset in this present study, has shown there are associations between differences in innate/adaptive imumune cell-type-specific at the beginning of anti-TNF therapy and the patient's response within three months [17]. Just very recently, Yoosuf et al. obtained very similar conclusions, by observing clear differences in certain gene expression levels between patients who responded to therapy and patients who did not [18]. In 2019 Kim et al. performed a meta-analysis of RA synovial transcriptomic data showing that differences in the activation of genes involved in several key and targetable signalling pathways could predict the response to infliximab with high accuracy [19]. In the same year, Guan et al. demonstrated that genetic heterogeneity, along with robust clinical assessment, can together be used for improving treatment strategies for patients with RA [20]. Nonetheless, the mentioned research works present some limitations, such as not including genetic information or the association of RA with clinical factors including age, sex and disease duration, or a limited number of samples/observations having been treated with other drugs. Consequently, individualisation of biological therapy in RA based on baseline predictors remains an unsolved problem since the identification of biomarkers has yet to reproducibly manifest relevant predictive value [21, 22].

The clinical utility of machine learning will likely further increase in the coming years. Thus obtaining robust and reproducible results will take the scientific community a step further to reach regular patient care in the form of precision medicine.

1.3 Objectives

This work's cornerstone is to find gene signatures across the human genome with a potential role on predicting the response treatment to RA when biologic agents are administrated. This systemic autoimmune disease's etiology and pathogenesis, complex and multifaceted, are not fully understood although it is known that genetic factors have a great degree of influence [23–25]. Thus, improved understanding of the root pathogenesis of the disease holds the promise of improved diagnostic and prognostic tools based upon this information.

From a dataset containing gene expression levels from RA patients undergoing treatment with biologic agents, the main goal was to obtain sparse models which enabled the identification of predictor genes regarding the response to anti-TNF treatment and to apply BN learning to uncover protein-protein interactions underlying that same response. The methodology further included a comparison of different machine learning algorithms when predicting the treatment response.

1.4 Contributions

The main contributions of this thesis are:

- A characterization of RA including a detailed explanation of T cell role in its pathogenesis, the disease activity scores and the possible treatment options;

- Benchmarking sparse model proposed for finding biomarkers from transcriptomic data;
- Identification of novel potential biomarkers for anti-TNF treatment response;
- Brief comparison of alternative classification tools in the context of anti-TNF response modelling using transcriptomic data.

1.5 Thesis Outline

Rheumatoid arthritis is first described in Chapter 2. Following, Chapter 3 includes a review of the methods applied in this thesis and, at the end, the data description. The work methodology is explained in Chapter 4 and the experimental results and consequent discussion are analysed in Chapter 5. Finally, in Chapter 6 the conclusions are presented, which incorporate the work achievements and possible future research directions.

Chapter 2

Rheumatoid Arthritis

From antiquity to the Renaissance, infrequent and fragmentary descriptions of joint diseases have been preserved. An early reference to arthritis can be found in the age of Pericles, when Hippocrates (460-370 B.C.) described: “a disease with fever, severe joint pain, fixing itself in one joint now, then in another, of short duration, acute, not leading to death, more apt to attack the young than the old” [3]. *La Familia de Jordaens en un Jardín* by Jacob Jordaens (c. 1630) constitutes a more contemporary RA-like findings, one of several examples portrayed by the Dutch Masters, in which swelling of the metacarpal-phalangeal and proximal interphalangeal joint can be observed [26].

2.1 Brief description

Rheumatoid arthritis (RA) is the most common inflammatory arthritis and it is a major cause of disability [26]. The word “rheumatoid” derives from the Greek *rheuma*, meaning that which flows, and the suffix *oid*, meaning like or in the form of, which denotes “any defluxion of thin humor”. The term “arthritis” stems from *arthros*, meaning joint, and it suggests inflammation [27].

RA is a systemic inflammatory disease characterized by a chronic inflammatory response which causes joint swelling, joint tenderness, and destruction of synovial joints, leading to severe disability and premature mortality [1]. The small joints of the hands and feet are involved most often, although larger joints (such as hips, shoulders or knees) may become involved later on. Joints are typically affected in a symmetric pattern; hence if the left foot is affected, for example, then both feet will be affected. Patients with RA report that their joint pain and stiffness is worse in the morning after they get out of bed. Although the joint system is the main focus of the disease, it can also damage a wide variety of non-articular tissues, including skin, eyes, lungs, heart or blood vessels, leading to different manifestations [23].

RA affects from 0.5% to 1.5% of the world’s population, and in Portugal its occurrence is estimated to be between 0.8% and 1.5% [2]. It affects more women than man, having a proportion of 4:1 in Portugal, according to the Portuguese Directorate-General of Health. Although it can appear at any age, its symptoms are most likely to manifest during adulthood [28].

During recent years, it has become clear that RA is composed of several phenotypes with defined and different genetic and environmental risk factors. Two major phenotyping criteria are the presence of serologic autoantibodies such as rheumatoid factor (RF) and anticitrullinated protein antibody (ACPA) [15]. Being an autoimmune disease, the case is these autoantibodies lead to abnormal immune reactions, attacking the patient's own tissues and organs. The presence in RA patients of these molecules is 70% to 90% and 75%. Research which has focused on the presence of these circulating autoantibodies has made major discoveries, one of them being that the autoantibodies precede the clinical manifestations of the disease by many years [1, 29]. Certain risk factors such as cigarette smoking or obesity are expected to worsen symptoms rather than triggering the systemic dysregulation *per se* [2].

Although RA prevalence is somewhat constant across the globe, there are some interesting exceptions. In the Chinese population RA occurrence is somewhat lower (0.3%) whereas it is substantially higher in native populations from North America, like the Prima Indians (5%) [23]. This is evidence that genetic predispositions play an important role. The class II major histocompatibility complex (MHC) genes, specifically the HLA (human leukocyte antigen) DRB1 alleles, which are associated to antigen presentation, have been proven to be a risk factor of RA. Tumour necrosis factor (TNF) alleles have also been linked with RA. However, it is estimated that these genes can explain only 50% of the genetic effect [24].

The exact cause that leads to the deregulation of the immune system and the synovium inflammation is unknown. The human immune response is characterized by two distinctive but interconnected stages. The first line of defense is called the innate response, which is non-specific and actuated mainly by neutrophils and eosinophils. The second line of defense is constituted by the adaptive response, a specific and period-long mechanism. The latter involves a humoral reaction (T cells lead to activation of B cells which produce antigen specific antibodies) and a cellular reaction (T helper cells, or Th cells, release cytokines which synergize T cells binding to infected cell's MHC-antigen complex, causing the latter's lysis).

T cells have long been implicated in mediating many aspects of joint inflammation [25, 30]:

- T cells may regulate osteoclast activation and thus joint destruction;

- $CD4^+$ T cells (which are Th cells), being CD4 a coreceptor, have been identified to inhibit immune reaction and suppress established immunity;

- synovial T cells produce CD40 ligand, a member of TNF receptor superfamily which is responsible for promoting B cell proliferation and immunoglobulin production;

- $CD4+$ T cells expressing CD25, called Treg cells, have a big repertoire of antigen specificity and their generation is at least in part developmentally and genetically controlled. Thus genetic defects may affect their development/function and be a primary cause of autoimmune and other inflammatory diseases;

- Th17 cells (Interleukin-17-producing $CD4+$ Th cells) induce the release of proinflamatory mediators;

- Costimulation molecules (Immunoglublins, TNF receptor and cytokine receptors) may be present at elevated levels in rheumatoid tissue, inducing T cell activation even in the absence of antigens.

T cells undergo differentiation and maturation in the thymus. After positive selection, they leave for the tissue's peripheral region. However, some of them when released are still function devoid, which is the only way there could be an adaptive immune response. Thus, following thymic selection, mechanisms are needed to maintain tolerance toward self-structures. The failure of immunologic self-tolerance leads to development of autoimmune diseases, such as RA [30].

2.2 Disease activity

The presentation and progression over time of RA are highly variable between individuals. In the best case scenario the patient has a very mild case of RA in which he or she remains undiagnosed. At the other end, there is the group of severe cases: the disease progresses fast and leads to a debilitating state. However, the majority of patients presents with an intermediate form of the illness during which the symptoms worsen for a finite time period.

An early initiation of treatment and a frequent assessment and monitoring of the disease activity enables a timely adoption of appropriate therapies. In the long run this control prevents radiographic disease progression and improves the patient's physical function and quality of life.

There is no single test defining RA. The standard and accepted means of defining RA is by use of classification criteria. Thus several scores have been created over the years to monitor RA. They include different variables and quantitative evidence about the patient state. Some of those scores are the disease activity score at 28 joints (DAS28), the Clinical Disease Activity Index (CDAI) and the Simplified Disease Activity Index (SDAI). The Visual Analog Scale (VAS) and the Health Assessment Questionnaire Disability Index (HAQ-DI) are important scales which aid and complement, respectively, those scores.

The *VAS* is an unidimensional measure of pain intensity commonly used in several rheumatic disorders. The VAS continuum pain scale ranges from “no pai” to “worst pain”, and patients mark a line to indicate how they are feeling [31]. Patient global assessment of disease activity (PtGA) and provider/physician global assessment of disease activity (PrGA/PhGA) are simple patient-completed or provider completed, respectively, on a 0–10cm or 0–100mm VAS, measuring the overall way RA affects the patient at a point in time [32].

The *health assessment questionnaire* (HAQ) is an example of patient-reported outcome (PRO). Introduced in 1980, it is one of the most cited and employed PRO instruments, particularly but not exclusively in the rheumatic disease literature. The HAQ measures the extent of the functional damage caused by a disease. It is based on five patient centered dimensions: disability, pain, medication effects, costs of care and mortality.

There are two versions of the HAQ: the full HAQ, which assesses all five dimensions and the short or 2-page which contains only the HAQ disability index (HAQ-DI) and the HAQ's patient global and pain VAS. The HAQ-DI includes items that assess fine movements of the upper extremity, locomotor activities of the lower extremity and activities that involve both the upper and lower extremities. The items, organized in 8 categories, represent functional activities, for example dressing, eating or walking. To calculate the HAQ-DI there has to be a response to at least six of the eight categories [33, 34].

Regarding the disease activity scores, they are described as follows:

The *DAS* is a clinical index of RA disease activity that combines information from swollen joints, tender joints, the acute phase response and general health. Due to its complexity and difficult computation requirements, a simplified version was developed: the DAS at 28 joints. DAS28 is a joint index that includes a maximum of 28 joints which are evaluated for swelling and tenderness. It also comprises erythrocyte sedimentation rate. If the score is higher than 5.1 it is considered that the disease is active and if it is lower than 2.6 it is indicative of a remission state. The DAS28 is the most common score used [35].

SDAI combines single measures into an overall continuous measure of RA disease activity, including a 28-swollen joint count, 28-tender joint count, the patient global assessment of disease activity, provider global assessment of disease activity and C-reactive protein level [36].

Finally, an alternative and simplified tool for assessment of disease activity, which does not include C-reactive protein levels, is called *CDAI*. Its greater advantage is the potential to be employed in evaluation of patients with RA consistently with close frequency and independently of any calculating device, therefore, it can essentially be used everywhere and anytime for disease activity assessment in RA patients [37].

The explicit formulas for the different score models and their respective range is indicated in Table 2.1. Their interpretation is found in Table 2.2.

Table 2.1: Formulas for calculation of RA disease activity scores: DAS, DAS28, SDAI and CDAI [32, 38].

Score Model	**Formula**	**Range**
DAS	$0.53938\sqrt{RAI} + 0.06465\ SJC44 + 0.33\ln ESR + 0.00722\ GH$	0 - 10
DAS28	$0.56\sqrt{TJC28} + 0.28\sqrt{SJC28} + 0.7\ln ESR + 0.014\ GH$	0 - 9.4
SDAI	TJC28 + SJC28 + PtGA + PhGA + CRP	0 - 86
CDAI	TJC28 + SJC28 + PtGA + PhGA	0 - 76

Disease activity score (DAS and DAS28); simplified disease activity index (SDAI); clinical disease activity index (CDAI). Ritchie articular index (RAI); Tender joint count (TJC); swollen joint count (SJC). SJC can be determined using 44 or 28 joints. C-reactive protein (CRP) in mg/dL; erythrocyte sedimentation rate (ESR) in mm/h. DAS and DAS28 use the general health (GH) or patient global assessment (PtGA) on a 0 to 100mm Visual Analog Scale.

2.3 Remission criteria

Remission represents an absence of disease activity. However, practically speaking, to require a complete absence of disease activity in RA would define a state that almost no patient with disease could meet. In fact, it has been shown that subclinical inflammation may be present on imaging despite the absence of clinical findings of disease activity. Thus, a realistic treatment expectation in clinical practice may be

Table 2.2: Interpretation of DAS, DAS28, SDAI and CDAI in the context of disease activity state [32].

Score Interpretation			
Remission	Low disease activity	Moderate disease activity	High Disease activity
DAS < 1.6	$1.6 \leq$ DAS < 2.4	$2.4 \leq$ DAS ≤ 3.7	DAS > 3.7
DAS28 < 2.6	$2.6 \leq$ DAS28 < 3.2	$3.2 \leq$ DAS28 ≤ 5.1	DAS28 > 5.1
SDAI ≤ 3.3	$3.3 <$ SDAI ≤ 11	$11 <$ SDAI ≤ 26	SDAI > 26
CDAI $\leq$2.8	$2.8 <$ CDAI ≤ 10	$10 <$ CDAI ≤ 22	CDAI > 22

Disease activity score (DAS and DAS28); simplified disease activity index (SDAI); clinical disease activity index (CDAI).

to achieve a level of disease activity so low that it is not troublesome to the patient and portends a later good prognosis [39, 40].

Several remission criteria have been defined. However, varying definitions of what constitutes remission make it difficult to compare results of current drug trials and to know how to apply those results to clinical practice. Therefore it is fundamental to have a standardize remission measure. With this in mind, in 2010 the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) met to define a uniform remission criterion for RA in trials and practice. The resulting work produced two definitions for evaluating remission: one is Boolean-based and the other is based on the composite index SDAI [41]. According to ACR/EULAR, in order to consider a patient in a remission state, at any point in time one of the conditions should be verified (see Table 2.3).

Table 2.3: ACR/EULAR boolean and index based definition of remission for clinical trials and clinical practice [41].

	Boolean-based	**Index-based**
Clinical trials	SJC, TJS, PtGA, CRP all ≤ 1	SDAI ≤ 3.3
Clinical practice	SJC, TJS, PtGA all ≤ 1	CDAI ≤ 2.8

Swollen joint count (SJC) using 28 joints, tender joint count (TJS) using 28 joints, patient global assessment (PtGA) on a 0 to 10 scale, C-reactive protein (CRP) in mg/dL, simplified disease activity index (SDAI), clinical disease activity index (CDAI)

The committee recommends that, even though the inclusion of ankles and forefeet in the assessment of remission is not mandatory, it should be taken into account in the examination. The new ACR/EULAR definitions are stringent and achievable, and they should be a major outcome for trials. However, they were developed using trial data and should be validated for use in practice settings [1, 41].

2.4 Response criteria

In RA clinical trials, treatment response is often assessed via the EULAR criteria, developed by the European League Against Rheumatism. This evaluation is based on change from baseline (which is a statistical approach) and the individual change in DAS reached during followup (low moderate or high, based on treatment decisions, which is a judgmental approach). This criteria is used to classify the participants as good, moderate or non-responders in relation to the efficacy of treatment according to Table 2.4.

Table 2.4: EULAR response criteria using DAS28 [38].

DAS28 at endpoint	DAS28 improvement from baseline (ΔDAS28)		
	> 1.2	0.6 - 1.2	≤ 0.6
≤ 3.2	GR	MR	NR
3.2 – 5.1	MR	MR	NR
> 5.1	MR	NR	NR

GR: good responder; MR: moderate responder; NR: non-responder

A change of 1.2 (*i.e.*, 2 times the measurement error, 95% confidence, since the variables used to calculate the DAS28 are transformed to have a Gaussian distribution) in a patient's DAS28 is considered indicative of a statistically significant change. For example, a patient must show a significant change (ΔDAS28 $> -$ 1.2), but must also reach low disease activity (DAS28 $\leq$ 3.2) to be classified as a good responder. The EULAR response criteria can also be applied using the DAS [38].

2.5 Disease treatment

Thirty decades ago the diagnose of RA was synonymous of a devastating quality of life, with progressive joint destruction, reduced life expectancy, early unemployment and considerable disability. The therapeutic agents that existed then were scarce and not efficacious. Today research has revolutionized the way RA patients live with this diagnose.

The most recent update of the EULAR recommendations for RA treatment happened in 2019. Since 2010, when they were initially developed, there has been progresses in the classification criteria, novel information on optimal clinical targets, evolution of treatment algorithms and introduction of new drugs. These contributions lead to the necessity of creating new or updating the RA management principles and recommendations. It is widely accepted nowadays that clinical remission is the main therapeutic target for patients with RA, with low disease activity as a best possible alternative, and that the best treatment approach is a treat-to-target strategy.

The so-called overarching principles, based more on the common sense nature rather than on specific scientific evidence, constitute the foundation on which the actual recommendations are based. They include the patient right to be a part of the treatment decision and to require access to multiple drugs with different modes of action [4].

Regarding the recommendations, as soon as the diagnose is concluded, it is highly indicated to initiate therapy with disease-modifying antirheumatic drugs (DMARDs), which are immunosuppressive and immunomodulatory agents that slow the progression of joint damage. Methotrexate, a conventional synthetic DMARD (csDMARD), is the most common prescribed one and it is efficacious used as monotherapy. Additionally, it is also the basis for combination therapies with other DMARDs or glucocorticoids. Glucocorticoids are recommended as a short bridging therapy when initiating or changing conventional synthetic DMARD therapies; once the treatment exhibits efficacy, their use should be rapidly tapered (within 3 months) [4].

The patient should be monitored frequently in active disease (every 1–3 months). When there is no improvement by at most three months after initiating the treatment or the target has not been reached by six months, therapy should be adjusted.

On the one hand, in case the treatment target is not achieved with the first csDMARD strategy, and in the absence of poor prognostic factors (such as high disease activity, presence of erosions and autoantibody positivity at high titres), other csDMARDs should be considered (leflunomide or sulfasalazine are common alternatives). On the other hand, if poor prognostic factors are present, then it is recommended adding a biologic DMARD (bDMARD) or a targeted synthetic DMARD (tsDMARD). EULAR recommends that bDMARDs and tsDMARDs should be combined with a csDMARD since they are less efficacious in monotherapy. A series of different combinations should be applied until the right therapy is found [4].

When the patient is in persistent remission, after tapering the glucocorticoids, the physician may consider tapering the comedication, if that is the case, and later on if improving, the first-line treatment may also be reduced [4].

Produced with biotechnology, bDMARDs are highly specific and engineered to act like a natural human protein and interrupt immune system signals. Some biological therapies are called TNF inhibitor or anti-TNF. They block the TNF alpha (TNF-α), a cytokine that induces local inflammation and pannus formation (abnormal tissue which invades the space in between a joint's bones, covering the bones and their protective layer of articular cartilage), leading to erosion of cartilage and bone destruction. So, by taking anti-TNF, the TNF is prevented from acting and thus triggering inflammation. The clinical use of anti-TNF includes the drugs infliximab, adalimumab, etanercept, golimumab and certolizumab pepol [42].

The bDMARDS that do not target the TNF-α are called non-TNFi. They may act by targeting CD20 (cluster of differentiation 20) proteins and destroying B cells (drug called rituximab) or CD80/86 proteins preventing T-cell activation (drug called abatacept), both of which cells play an important role in RA, as mentioned before. Moreover, non-TNFi therapies may block the interleukin-1 receptor, preventing the action of proinflammatory cytokine IL1 (anakinra) or the interleukin-6 receptor, which ultimate effect is bone erosion (tocilizumab) [43].

Being a type of drug that suppresses the immune system, the administration of any DMARD carries risks and thus after initiating the treatment, patients need to be monitored for potential side effects of the medications. Concerning the bDMARDs, the most concerning adverse effect is increased risk of common and serious infections including bacterial, fungal and viral infections. Besides, their production method makes the price much larger than other csDMARDs and the patient preference must be taken into account [44].

Finally, one needs to bear in mind that RA is regarded a usually incurable disease, hence a drug that proves to be efficacious and is tolerated by the patient should not be stopped, otherwise the possibility of relapse becomes a new reality.

2.6 Socioeconomic impact

RA can become a overwhelming disease with its different possible manifestations which are not necessarily restricted to the joint system. Repercussions may arise from different directions, for example the disease related direct costs, the inability to attend work and thus the risk of income lost or the impact on the quality of life and psychological well-being. Thus, ideally patients should be guided from an early stage and have access to a multitude of health and non-health professionals: family doctor, rheumatologist, physiotherapist, nutritionist, social worker, among others.

Regarding the financial cost, in Portugal, each early diagnose can represent to the state an annual average saving of 30% per each new case, according to Sociedade Portuguesa de Reumatologia [45]. A study conducted in Portugal explored the RA financial impact. It was estimated that the annual mean cost of treating one RA patient is about *3.415€*. The total cost of the disease increases with its stage going from *2.205€* per patient per year in case of remission to *5.634€* in case of high activity [46]. In fact, the emergence of biological therapies for RA in the last decade increased the cost of treatment of rheumatic diseases. It has also been reported the existence of an association between rheumatic diseases and other major chronic diseases with early exit from paid employment in Portugal [47].

Chapter 3

Materials and Methodology

This chapter introduces the supporting tools employed in the discovery of the gene signatures predictable of anti-TNF treatment response. They include sparse logistic regression, Bayesian network learning and examples of other machine learning methods. The type of data and its origin is further described.

3.1 Logistic regression

Regression techniques are versatile in their application to medical research because they enable to predict outcomes and measure associations, and to control for confounding variable effects. Generalized Linear Models (GLMs), proposed by McCullagh and Nelder in 1989, provide a flexible framework to study the association between a family of continuous or categorical outcomes and a set of independent variables [48]. A fundamental aspect of the generalization is the presence in all the models of a linear predictor based on a linear combination of explanatory or stimulus variables.

Consider $\mathbf{y} = \{y_1, \ldots, y_n\}^T$ to be the vector of observations and assumed to be a realization of a random variable $\mathbf{Y}$ whose components are independently distributed with means $\boldsymbol{\mu}$. $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the $n \times p$ matrix containing the set of covariates or explanatory variables. $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ is the vector of unknown regression coefficients associated with each covariate which is intended to be estimated. One can consider the following three-part specification as a generalization of GLMs [48]:

1. Random component: the components of $\mathbf{Y}$ reflect a certain probability distribution;

2. Systematic component: covariates $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ produce a linear predictor $\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta};$$

3. The link function relates the random and systematic components and is defined as

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}.$$

This allows a generalization of GLMs, where the link function says how the expected value of the response $\mathbf{E(Y)}$ relates to the linear predictor of explanatory variables, as follows:

$$\mathbf{E(Y)} = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{3.1}$$

Depending on the three components just mentioned, there are different examples of GLMs. For example, *linear regression* is used when the response variable $\mathbf{y}$ is continuous and $\mathbf{Y}$ is considered to follow a normal distribution. Hence, the link function is the identity function (the simplest of all link functions). Simple linear regression will model how mean expected value of the continuous response variable depends on the explanatory variables set:

$$Y_i = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{3.2}$$

where i indexes the observations; $\boldsymbol{\beta}$ contains the unknown regression coefficients to be estimated and e_i is the error associated with the discrepancy between the estimated and the observed value for the i-th observation. Errors e_i are assumed to have a normal distribution, with mean zero and constant variance. Ordinary least squares (OLS) method estimates the unknown parameters by minimizing the sum of the error parcels e_i,

$$\hat{\boldsymbol{\beta}} = \arg\min_{x} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{3.3}$$

Another example is the *poisson regression* which typically uses the log link function and the Poisson distribution.

However, in case the response variable y is binary, $\mathbf{Y}$ is assumed to follow a Binomial distribution. Considering p_i to be the probability of success *i.e.*, the probability of $Y_i = 1$, given the associated features vector $\mathbf{x}_i$, p_i is defined as:

$$p_i = Prob(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})}. \tag{3.4}$$

Then, using the logit link function (natural logarithm of the odds), the binary *logistic regression* models how the response variable depends on the set of features:

$$E(Y_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i\boldsymbol{\beta}; \qquad i = 1, ..., n, \tag{3.5}$$

where $\boldsymbol{\beta}$ is the vector of unknown regression coefficients to be assessed. The framework used is the maximum likelihood estimation. For a n sized sample, the likelihood function for a binary regression is given by Equation 3.6 and its simplification, after applying the log transformation, is called log-likelihood (Equation 3.7).

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{3.6}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \text{log}(p_i) + (1 - y_i)\text{log}(1 - p_i)]. \tag{3.7}$$

Logistic regression is one of the most commonly used tools for applied statistics and discrete data analysis. It has even shown equally performing results when compared to alternative machine learning techniques in clinical and biological research. This regression technique is an efficient and powerful way to assess independent variable contributions to a binary outcome. Nevertheless, there are important considerations when conducting logistic regression including a careful variable selection. [49, 50]

This will be the subject of the following section.

3.2 Regularization Methods

Situations of high dimensionality, where the number of variables p is much higher than the number of observations n, have become recurrent in genetics research, medical studies, risk management and other fields. The estimates $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, ..., \hat{\beta}_p\}^T$ obtained with logistic regression or other models are usually all nonzero and if $p > n$, they are not unique. Besides, when model selection involves "sparse modeling," the estimation approach that zeroes out all but the most relevant of variables from hundreds or thousands of possible candidates, only a relatively small number of predictors is different from zero. In order to encourage sparsity, it is necessary to constrain the regression problems, which can be achieved through regularization methods.

Ridge regression was proposed by Hoerl and Kennard in 1970 [51]. The basic idea was to constrain the estimates of $\boldsymbol{\beta}$ coefficients which otherwise could "explode", being susceptible to very high variance and affecting the model prediction accuracy. *Ridge* regression imposes a ℓ_2 constraint (sum of the squared error of the coefficients) as follows

$$\sum_{j=1}^{p} \beta_j^2 \leq t; \qquad t \geq 0. \tag{3.8}$$

This constraint technique reduces the model complexity by coefficient shrinkage but sparsity is not encouraged.

Lasso regularization, which means Least Absolute Shrinkage and Selection Operator, was proposed by Tibshirani in 1996 [52]. It allows not only coefficient shrinkage, as in the ridge regression, but also subset selection, by imposing an ℓ_1 constraint (sum of the absolute values of the coefficients):

$$\sum_{j=1}^{p} |\beta_j| \leq t; \qquad t \geq 0. \tag{3.9}$$

Due to the nature of this penalty, *lasso* tends to produce some coefficients that are exactly zero and hence it gives interpretable models. Figure 3.1 illustrates intuitively the reason why this happens by representing a case where $p = 2$. The regularized solution in both methods corresponds to finding the first point where the elliptical contours hit the constraint region. Unlike the disk shape area (ℓ_2 constraint), the diamond (ℓ_1 constraint) has corners. Thus, if the solution occurs at a corner, it will have one parameter β_j equal to zero.

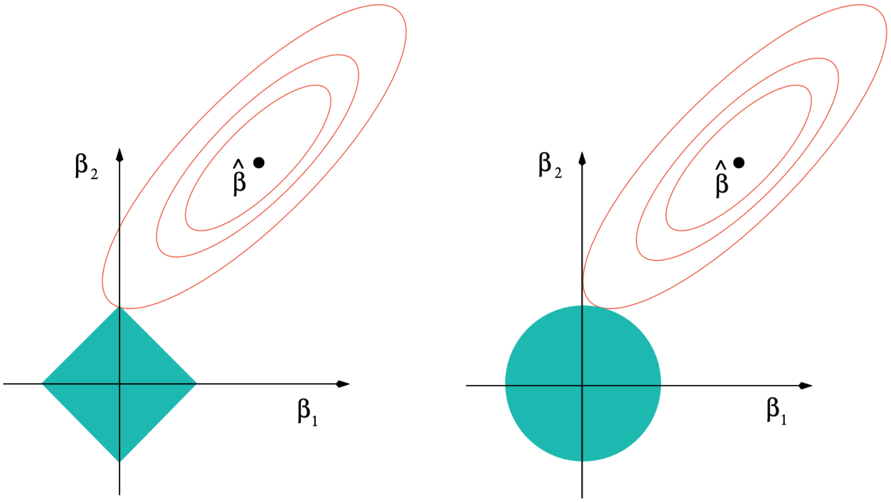


Figure 3.1: Representation of the *lasso* (left) and the *ridge* (right) regressions. The red ellipses show the level curves of the cost function. Point $\hat{\beta}$ shows the (usual) unconstrained OLS estimation. The solid blue areas are the imposed boundaries $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t$, respectively. Figure from [53].

Although the *lasso* has shown success in many situations, it has some limitations. For example when $p > n$, it selects at most n before it saturates, which is a limiting feature for a variable selection method [54]. Another problem is, if in the presence of highly correlated features, it selects only one rather than taking the whole group (there is no clustering). This motivated the development of a new method that could work as well as *lasso* whenever the lasso does its best. The *elastic net* regularization is a weighted combination of both *lasso* and *ridge* regression:

$$\lambda \sum_{j=1}^{p} \{(1-\alpha)\beta_j^2 + \alpha|\beta_j|\}, \tag{3.10}$$

where $\lambda \geq 0$ controls the magnitude of the parameters and $\alpha \in [0, 1]$ controls the relative weight of each penalty. The ℓ_1 norm contributes to a sparse model (thus, increasing α leads to more sparsity) and the ℓ_2 norm removes the limitation on the number of selected variables and encourages the grouping effect. At the limits, if $\alpha = 0$, the *ridge* penalty will be applied otherwise if $\alpha = 1$, the regularization method will be the *lasso*.

3.3 Model Validation Approaches

When learning the parameters of a prediction function from a certain dataset, one should not use the same data to test it, since the model would just repeat the labels of the samples from where it was built, having a perfect score, but failing to predict any new data. This would be a case of over-fitting. One way to combat it is to increase the amount of data from where the algorithm learns, thus making it less likely the model will overlearn the data. However, in the absence of a very large amount of observation data, one common practice is to apply a resampling procedure: *k-fold cross-validation* or simply CV.

This approach involves randomly dividing the set of observations of size m into k groups, or folds, with size m/k. For each k fold, the model is trained using $k-1$ folds and validated on the held-out fold. CV allows, for example, tuning a certain model parameter, by comparing the accuracy of the supervised learning algorithm across all k iterations [55].

Leave-one-out cross-validations (LOOCV) is the particular case of CV when the k number of folds corresponds to the total number of observations m *i.e.*, the model is estimated in each iteration by considering all but one observation $(m-1)$ which is then used to validate the predictive power of the model [55].

In order to evaluate the estimated model a heuristic and straightforward method exists: the Receiver Operating Characteristic (ROC) curve. This graphical plot displays the trade-off between two measurement rates of a binary prediction as the classification probability threshold varies: false positive rate $(1 - specificity)$ and true positive rate (*sensitivity*). Specificity measures how well a test can identify true positives and sensitivity measures how well a test can identify the true negatives. For a certain threshold, the accuracy of the test is given by the sum of the true classifications over the total observations. The overall performance of a classifier, summarized over all possible thresholds, is given by the Area under the ROC curve (AUC). An ideal ROC curve will hug the top left corner which means that the larger the AUC, the better the classifier will be [55].

ROC analysis is used in clinical epidemiology in order to quantify how accurately medical diagnostic tests (or systems) can distinguish between two patient states, typically referred to as "diseased" and "nondiseased" [56].

3.4 Bayesian Networks

Bayes's rule is a fundamental mathematical theorem that describes the conditional probability $P(A|B)$ of a certain event A based on prior knowledge B related to that event. Mathematically speaking, the belief of an event can be updated according to

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.11}$$

Bayesian Networks (BNs) are a type of probabilistic graphical models that aim to model conditional dependence, between variables, allowing the computation of the joint probability distribution (JPD). The graphical representation provides an intuitive and natural way for considering an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor.

BNs are directed acyclic graphs (DAG), commonly defined G = *(V,E)*, in which the vertices or nodes V represent the random variables of interest and the edges or links E represent the informational or causal dependencies amongst those variables. A directed graph is acyclic if there is no directed path $X_1 \to \cdots \to X_n$ so that $X_1 = X_n$.

For instance, an edge between nodes X_i and X_j indicates that the value taken by the variable X_j is conditionally dependent on the value taken by the variable X_i. Node X_i is the parent node of X_j and node X_j is the child node of X_i. Perpetuating this logic, the "descendants" set of a node include all the nodes that can be reached on a direct path from the node and the "ascendants" set of a node contains all the nodes from which that node can be reached on a direct path. Hence, being acyclic means that no node is its own ancestor or descendent. Any node in a BN is always conditionally independent of its all nondescendants given that node's parents.

The understanding of factorization is fundamental to define a BN. Let $\mathbf{X} = \{X_1, ..., X_p\}^T$ be a p-dimensional vector of random variables X_j, where $X_j \subset \mathbb{R}$, that coincides with the nodes V from $G = (V, E)$. G represents a joint probability distribution $P(\mathbf{X})$ over the same space. It can be stated that P factorizes according to G if it can be expressed as a product, called the chain rule, as follows

$$P(X_1, ..., X_p) = \prod_{j=1}^{p} P(X_j | Pa_{X_j}^{G}) = \theta, \tag{3.12}$$

where $Pa_{X_j}^{G} = \{Z_i : Z_i \to Z_j \in E\}$ denotes the set of parents of X_j in G and the factors θ_j are called conditional probability distributions (CPDs) or local probabilistic models. In summary, a BN is a joint probability distribution over the set of random variables, defined as a pair $B = (G, \theta)$ where θ factorises over G, and where θ is specified as a set of CPDs associated with G's nodes [8].

If the random variables in $\mathbf{X}$ are discrete, the CPDs are usually represented in a table that lists the local probabilities that a node takes on each of the possible values for each combination of the values taken by its parents. The joint probability distribution of a collection of variables can be determined uniquely by these local conditional probability tables.

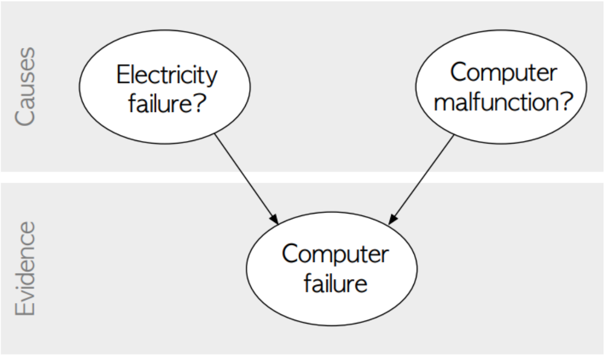
An illustrative BN example is shown in Figure 3.2 describing a trivial example of a computer failure (denoted by C) [57]. If the computer does not initiate when there is an attempt, one would like to know the possible causes for that misfortune. In this example there are only two alternatives: either electricity failure (denoted by E) or computer malfunction (denoted by M). The variables take binary values *i.e.*, either True (denoted as T) or False (denoted as F). The example is a rather simplified case, hence it is assumed some independence between the causal nodes. As a result of the dependencies encoded by the graph, the JPD of the network can be factored as

$$P(C, E, M) = P(E)P(M)P(C|E, M). \tag{3.13}$$

The joint probability defined by the Bayesian Network is given by the product of all conditional probability tables specified in the BN.

In reality, the DAG structure is usually not a given parameter and the BN learning problem is characterized by finding the best network representing the data and the corresponding joint probability distribution parameters. While the latter is rather straightforward once the DAG structure is available, the former can be a difficult challenge.

Computational BN learning from high dimensional data can be very demanding since directed graphical models do not scale well with the number of variables. In order to solve this problem, an R package was developed, called `sparsebn`, which learns the structure of large, sparse graphical models with a focus on BNs [7].



E	***M***	$P(C = T \mid E, M)$
T	T	1
T	F	1
F	T	0.5
F	F	0

Figure 3.2: Directed acyclic graph representing two independent possible causes of a computer failure [57]. E = Electricity failure; M = Computer malfunction; C = Computer failure; T = True; F = False.

To learn a BN from data, the authors have used a score-based approach that relies on regularized maximum likelihood estimation. The criterion considered was

$$\min_{\mathbf{B} \in \mathbb{D}} \quad l(B; \mathbf{X}) + \rho_\lambda(B) \tag{3.14}$$

where $\mathbf{X}$ is a matrix of observations assumed to not have any missing values, l denotes the negative log-likelihood, ρ_λ is some regularizer, matrix B is the weighted adjacency matrix of a DAG and $\mathbb{D}$ the set of weighted adjacency matrices that represent directed graphs without cycles. The BNs are learnt from data using the method *estimate.dag*.

The output of this algorithm is a solution path with multiple graph estimates rather than a single one. It is so because the program depends on the unknown parameter λ, that must be passed to the algorithm. Hence the solution path consists of a sequence of estimates $\{\hat{\boldsymbol{\beta}}(\lambda_{\text{max}}), \hat{\boldsymbol{\beta}}(\lambda_1), ..., \hat{\boldsymbol{\beta}}(\lambda_{\text{min}})\}$ for a predetermined set of lambdas $\lambda_{\text{max}} > \lambda_1 > ... > \lambda_{\text{min}}$. When λ increases, there is less regularization, hence the resulting estimates $\hat{\boldsymbol{\beta}}(\lambda_m)$ become more dense (meaning contain more edges). Since the focus is on sparse graphs, the algorithm is terminated early if the number of edges exceeds some user-defined threshold (which is controlled by the parameter *edge.threshold*).

From the solution path, the user can choose the solution with the preferred number of nodes. Alternatively, a method called *select.parameter* is available with which the algorithm returns the optimal solution, based on a trade-off between the increase in log-likelihood and the increase in complexity between solutions.

3.5 Classification algorithms

The recent advances of next-generation sequencing technologies have allowed measuring the expression levels of tens to thousands of transcripts simultaneously, promising to revolutionize the methodologies used for investigating potential disease markers. However, microarray based classifiers cannot be directly applied due to the discrete nature of RNA-Seq. Consequently, one available option is to develop count-based (or discrete) classifiers. Alternatively, one may wish to bring RNA-Seq samples hierarchically closer to microarrays and apply known algorithms for classification applications of continuous data.

Before describing the used classifiers, several normalization and transformation methods are explained, the latter being crucial when using continuous approaches. All the following classifiers, normalization and transformation methods are available on the `MLSeq` R package [58].

3.5.1 Normalization Methods

During RNA-Seq analysis, normalization is a crucial step, being standardly applied with the intent of reducing the non-biologically derived variability inherent in transcriptomic measurements. These variations may be originated from both between-sample variations including library size (sequencing depth) and the presence of majority fragments, and within-sample variations including gene length and sequence composition (guanine-cytosine content) [59].

There is a multitude of approaches and methods developed to address this problem. One idea is to find the ratio of each read count to the geometric mean of all read counts for that gene across all samples. The median of these ratios for a sample, called the size factor, is used to scale that sample. Anders and Huber designed the method for differential expression analysis for sequence count entitling it *DESeq* [60]. Another example is the trimmed mean of the M-values (*TMM*) normalization. TMM first trims the data in both lower and upper side by log-fold changes (default 30%) to minimize the log-fold changes between the samples and by absolute intensity (default 5%). After trimming, TMM calculates a normalization factor using the weighted mean of data. These weights are calculated based on the inverse approximate asymptotic variances using the delta method [61].

3.5.2 Transformation Methods

RNA-Seq data is usually represented by a matrix of counts showing the expression levels of micro-RNAs (rows) for a set of samples (columns). For each sample, millions of reads can be measured by the RNA-Seq technique. According to the gene annotation and genome build, numbers of features might be different. Different pipelines can result in different properties of the count matrix. Besides, gene expression levels are heavily skewed in linear scale: lower expressed genes have read counts between 0 and 1 while the higher expressed genes between 1 and positive infinity. Thus an appropriate transformation on raw counts is needed.

One simple approach is the logarithm of counts per million reads (*log-cpm*) method [61], which transforms the data from the logarithm of the division of the counts x_{ij} by the library sizes X_j and multiplication by one million, given by:

$$z_{ij} = \log_2\left(\frac{x_{ij}+0.5}{X_j+1} \times 10^6\right) \tag{3.15}$$

Although *log-cpm* transformation provides less-skewed distribution, the gene-wise variances are still unequal and possibly related to the distribution mean. Hence, alternative methods which aim to remove the dependence of the variance on the mean are particularly useful, specially when genes with low expression level and therefore low read counts are present. These tend to have high variance, which is not removed efficiently by the ordinary logarithmic transformation. Two examples of such methods are the variance stabilizing transformation (*vst*), presented by Anders and Huber [60], and regularized logarithmic (*rlog*) transformation, presented by Love et al. [62]. They produce very similar effects, although *rlog* is more robust in the case when the size factors vary widely.

Another major challenge in a differential expression analysis is the frequent possibility of encountering variations in sample quality in small RNA-Seq experiments. In fact, to remove the high variation samples would reduce noise, but at a cost of reducing power, thus limiting the ability to detect biologically meaningful changes. Contrarily, retaining those samples in the analysis may not reveal any statistically significant changes due to the higher noise level. Thus, Law et al. presented a method which reflects the compromise of using all available data, but to down-weight the observations from more variable samples called variance modeling at the observational level or *voom*. Essentially, it applies the *log-cpm* transformation and estimates the mean-variance relationship, and uses this to compute appropriate observational-level weights [63].

3.5.3 Continuous-based Classifiers

Support Vector Machine (SVM) learning is a powerful machine learning tool that creates a decision boundary between two classes, enabling the prediction of labels from one or more feature vectors. This decision boundary, known as the hyperplane, is orientated in such a way that it is as far as possible from the closest data points (called support vectors) from each of the classes [64].

In some occasions it might be necessary to apply a kernel method, which enables the modelling of higher dimensional, non-linear models. In such cases, a kernel function is used to add additional dimensions to the raw data and thus make it a linear problem in the resulting higher dimensional space. Choosing a kernel function is a matter of great importance since it can affect the performance of the SVM algorithm. The characteristic to look for is to be able to separate the data without introducing too many irrelevant dimensions. Some examples of kernel functions include linear, polynomial, radial basis function and sigmoidal kernels. Two major advantages of using kernel functions are the possibility to handle nonvector data and the fact that they provide a mathematical formalism for combining different types of data [65].

A **decision tree** is an intuitive multipurpose support tool like tree structure, in which each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf (terminal) node holds a class label. The decision tree is the main building block of a **random forest**, a classification algorithm that outperforms any of its individual constituent models (the trees). Given an input vector, each decision tree gives a classification and it “votes” for that class. The forest then chooses the class which gathered the majority of votes over all the trees. Each decision tree in the forest considers a random subset of features when classifying and only has access to a random set of the training data points. These decorrelated trees encourage low variance for the ensemble, increasing diversity in the forest. At last, it leads to more robust overall predictions.

This algorithm runs efficiently on large datasets, it can handle thousands of input variables without variable deletion, it generates an internal unbiased estimate of the generalization error as the forest building progresses, to name a few of the many remarkable features [66].

The standard nearest centroid classification consists on computing a standardized centroid for each class, which in gene profiling studies corresponds to the average gene expression for each gene in each class divided by the within-class standard deviation for that gene. Then, for a new sample, the classifier takes its gene expression profile and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample.

Nearest centroid classifier assigns to an observations the label of the class of training samples whose mean (centroid) is closest to the observation. The idea behind **nearest shrunken centroid** (NSC) models, an extension of the former, is to shrink each class centroid towards the overall centroid by an amount defined a *priori* - the threshold. This modification to the standard classification has two advantages. On the one hand it can make the classifier more accurate by reducing the effect of noisy genes. On the other hand it does automatic gene selection when there are more than two classes. In particular, if a gene is shrunk to zero for all classes, then it is eliminated from the prediction rule. Alternatively, it may be set to zero for all classes except one, which is an indicator of how that gene expression characterizes that class [67].

Before using any of the three classifiers, given the discrete nature of RNA-Seq, it is necessary that the data undergoes a transformation, hence the methods described previously in Section 3.5.2.

3.5.4 Discrete-based Classifiers

As RNA-Seq consists of nonnegative data, in matters of expression-based classification it is more appropriate to model it with discrete-count distributions, such as the poisson and the negative binomial. Discriminant functions consist in linear combination of independent variables that are able to discriminate between the categories of the dependent variable. The purpose of discriminant analysis is to assign an unknown subject to one of several classes on the basis of a multivariate observation [68].

The sparse **poisson linear discriminant analysis** (PLDA) is a count-based classifier that extends from nearest shrunken centroids, popularly used for microarray (it takes on continuous values, on the contrary to RNA-Seq) classification, developed by Witten. The authors also suggested applying a power transformation, since poisson distribution underestimates the variation observed from the data, which henceforth will be referred as PLDA2 [69].

Dong et al. proposed a similar method, **negative binomial linear discriminant analysis**, as an alternative to PLDA, which is more appropriate when biological replicates are available and in the presence of overdispersion (*i.e.*, when the variance is larger than or equal to the mean) [70].

3.5.5 Voom-based Classifiers

Novel classification methods integrating voom transformation have been developed to open access microarray based methods for RNA-Seq analysis. One such method is called **voomDLDA**, an extension of Diagonal Linear Discriminant Analysis (DLDA) for RNA-Seq with weighted parameter estimates. DLDA belongs to the family of Naive Bayes classifiers, where the distributions of each class is assumed to be multivariate normal and to share a common covariance matrix. The voomDLDA in a nonsparse method which assumes that the gene specific weighted variances are equal across groups and it uses the weighted pooled covariance matrix in modeling class-conditional densities. The **voomNSC** is a sparse classifier, which accepts either a normalized or non-normalized count data as input, applies *voom* method to data, provides precision weights for each observation and ultimately, fits an adapted NSC classifier by taking these weights into account [71].

3.6 Data Description

The CORRONA (Consortium of Rheumatology Researchers of North America) independent registry, founded in 2001, collects longitudinal data, "real world" data from patients with rheumatologist-diagnosed inflammatory arthritis (which includes RA, osteoarthritis, psoriatic arthritis and/or osteoporosis) and their treating physicians [3]. At the time of this writing, data on 51,649 patients and 769 rheumatologists have been collected. One decade ago, with the objective of expanding the scope of clinical data and focusing the scientific yield on comparative effectiveness, the CERTAIN (Comparative Effectiveness Registry to study Therapies for Arthritis and Inflammatory conditions) registry was launched. This prospective, non-randomized cohort study includes patients with RA fulfilling the ACR criteria (having at least moderate disease activity with a CDAI score higher than 10; see Table 2.2) who are starting or switching biologic agents [3].

The data used in this thesis consists of RNA-Seq of whole blood samples from biologic naïve patients from the CORRONA CERTAIN registry immediately prior to initiation of anti-TNF treatment (at baseline, which will be referred as *BL*) and following three months of therapy (*M03*). Being biologic naïve means that the patients had no previous biologic agent treatment. The patients initiated treatment with adalimumab or infliximab in conjunction with methrotrexate. Data containing 25370 variables (gene expressions) measured from 63 patients at BL and 65 patients at M03 were selected for RNA-Seq, proteomics, and targeted glycopeptide analysis, as explained by Farutin et al. [17]. The public files are deposited in the National Center for Biotechnology Information - Gene Expression Omnibus (NCBI-GEO) database (GSE:129705). The clinical evaluations were performed based on EULAR criteria for clinical response to therapy three months into the treatment and each patient was classified as good responder or non-responder [38]. Patients classified as moderate responders were not selected for this study [17].

STRING is a database of known and predicted protein-protein interactions. These include direct (physical) and indirect (functional) associations which stem from computational prediction, from knowledge transfer between organisms and from interactions aggregated from other (primary) databases. Its aim is to collect and integrate such information for a large number of organisms (it covers 24,584,628 proteins from 5,090 organisms) in order to develop the knowledge of all functional interactions between the expressed proteins and in the late run to widen the understanding of cellular function [72].

Each protein–protein association stored in STRING is given a score. These scores represent confidence levels, and are scaled between zero and one (interactions may be given: (a) highest confidence, score ≥ 0.9; (b) high confidence or better, score ≥ 0.7; (c) medium confidence or better, score ≥ 0.4; (d) low confidence or better, score ≥ 0.15). They indicate the estimated likelihood that a given interaction is biologically meaningful, specific and reproducible, given the supporting evidence. This supporting evidence is provided by *evidence channels*, which depend on the origin and type of the evidence. Examples of these *channels* include: genomic context predictions, high-throughput laboratory experiments, automated text-mining of the scientific literature and data import from curated, among others.

The data used in this work corresponded to the interactions at highest confidence interval for *Homo sapiens.* The file (9606.protein.links.full.v11.0.txt.gz) was acquired directly from the STRING database (downloaded from http://https://string-db.org/cgi/download, assessed on July 13th, 2020).

Chapter 4

Work Methodology

The approach channeled to uncover the gene signatures, the core of this study, is explained and tested. In addition, the research regarding alternative machine learning algorithms and their performance in predicting the patient's treatment response is schematized.

4.1 Finding Biomarkers

The following methodology regarding a sparse approach to unravel gene interactions through Bayesian Network (BN) learning was based on the work developed by Brito [73], where common gene signatures of breast and prostate cancers were investigated, and Constantino et al. [74], whose work was the foundation for this thesis.

Prior to any analysis, and bearing in mind the medical potential of distinguishing patients before and after therapy, two datasets were created regarding the moment of data collection: one at baseline (BL) and another after three months of the beginning of the therapy (M03), both of which maintained the initial 25370 variables.

This work's cornerstone was the possibility that gene expression profiling could give a clue about the efficacy of anti-TNF treatment in RA patients. To that end, prior to performing feature selection, a pre-processing step was carried out. The variables with zero standard deviation were disregarded in both datasets. Following, the variables were log-transformed and normalized to unit variance. An auxiliary vector with the binary response of each patient was created for each dataset: "1" for good responders, R (for simplification reasons, henceforward "R" will be used instead of "GR") and "0" for non-responders, NR, in accordance to the clinical evaluations performed three months into the treatment, as described in Section 3.6.

Following, the dimensionality reduction step was conducted. Sparse logistic regression with elastic net regularization was performed by means of the `glmnet` R package [75]. The procedure, which was applied independently in both datasets, went as follows: in a total of 5,000 times, the data was split in 70% for training the model and the remaining 30% for testing it. In each run, the model was estimated from the training data with logistic regression using method *cv.glmnet*, where the parameter α (Equation 3.10) varied between 0 and 1 with 0.1 intervals. The penalty λ (Equation 3.10) was optimized by 10-fold CV: the chosen λ was the largest one with which the error was within one standard error of the minimum [75]. Lastly, the fitted model was used to predict the treatment response of the test set. For each model, the ROC curve was estimated and the AUC calculated.

Two α values were selected for each dataset, which resulted in obtaining two predictive models for BL and two other predictive models for M03. Afterwards, LOOCV (Leave-one-out cross-validation) approach was used to explore which variables were strongly associated with the treatment response. The premise was that the variables repeatedly selected across all iterations of that procedure could indicate which genes are strongly associated with the treatment response. To evaluate each estimated model, the classifier's specificity and sensitivity trade-off in the validation set was visualized through ROC curves.

In order to uncover the gene networks regulating the anti-TNF treatment in each dataset, BN learning was performed using the `sparsebn` R package [7]. For that purpose, each model had to be split into other two, according to each patient's treatment response. Finally, the BN were obtained. For each BN, the adjacency weights of each edge were inspected. At last, the protein-protein interactions found from learning the BNs were validated by comparing them with the STRING database [72]. Only the highly scored combinations ($score > 0.7$) were taken into account. The flowchart in Figure 4.1 summarizes the overall methodology until this point.

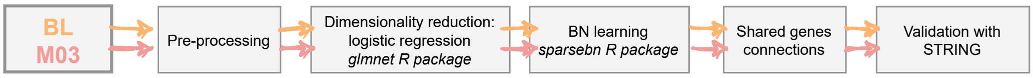


Figure 4.1: Flowchart of procedure used to obtain Bayesian Networks and gene candidates for prediction of treatment response to anti-TNF. The procedure was conducted in parallel for BL and M03 datasets.

Regarding the BN analysis, different network architectures were experimented, as schematized in Figure 4.2. The algorithm's controller *edge.threshold* forces the number of edges in the solution networks to be equal or less than the specified number. Therefore, firstly this parameter matched the number of variables (*i.e.*, it matched the model's number of genes given by the chosen α) and the solution with that number of nodes was chosen. In other words, the BN algorithm received a model with n_var variables and learned a network in which the number of edges was $n\_edges = n\_var$. This corresponds to the labeled S ("*single*") boxes in the scheme. Secondarily, the same parameter was set to be the double of variables number *i.e.*, $n\_edges = 2 \times n\_var$ (D, "*double*"). Lastly, the method *select.parameter*, which automatically returns the optimal solution, was used both when $n\_edges = n\_var$ and $n\_edges = 2 \times n\_var$ (A, "*algorithm*"). These four steps were applied to both models of each dataset and to each group of patients.

4.2 Classification Analysis

The obtained models using the sparse logistic regression inspired the second part of this thesis which consisted in evaluating different classifiers using the transcriptomic data. The goal was not only to

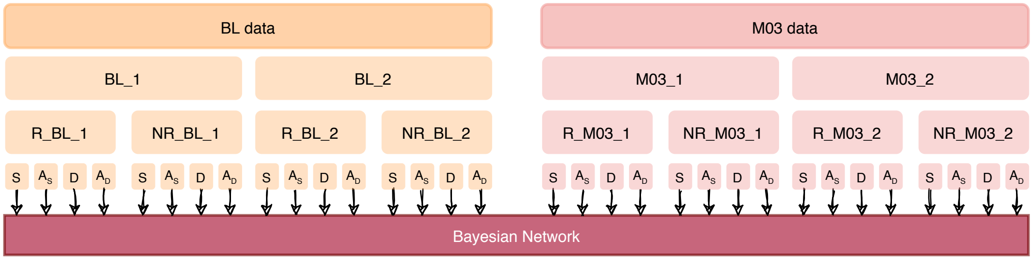


Figure 4.2: Complete set of Bayesian Networks obtained when adjusting the number of edges allowed in the solution. BL_1 and BL_2 indicate the two models obtained with each α from the sparse logistic regression (likewise for $M03\_1$ and $M03\_2$). R and NR indicate the good-responder and non-responder patients cohorts, respectively. S, D and A refer to the number of edges in the solution (S: $n\_edges = n\_var$; D: $n\_edges = 2 \times n\_var$; A_S and A_D: trade-off solution chosen by the algorithm when given maximum number of edges $n\_edges = n\_var$ and $n\_edges = 2 \times n\_var$, respectively).

compare different machine learning tools, but also to inspect how each performed when given different portions of the same data. All the different stages were conducted with the aid of `MLSeq` R package [58]. The first assessment was made using the complete BL and M03 datasets after eliminating the variables (genes) with zero standard deviation. Later, for each dataset, six new sub-datasets were created using a maximum variance filtering: they contained the top 5, 10, 15, 20, 25 and 30 variables with the highest variance. Lastly, the two models of each dataset previously obtained with the sparse logistic regression were used. Figure 4.3 illustrates the final ensemble of the input data used for each dataset (BL and M03).

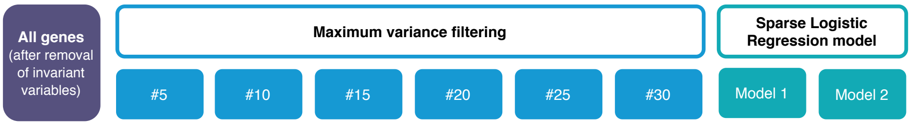


Figure 4.3: Sub-datasets used as starting point for classification analysis for each data group (BL and M03). "*#5*" indicates the sub-dataset with the top 5 features in terms of variable variance, and so on until "*#30*". Model 1 and 2 refer to the models obtained in the previous pipeline (Figure 4.1) which resulted in selecting two α values for each data group.

From the huge variety of classification algorithms, eight were selected to fit the data and predict the patient's response to the anti-TNF treatment. The classifiers used are the ones described in Section 3.5:

- Continuous-based: *svmRadial* (svm algorithm with radial kernel function), *rf* (random forest) and *NSC* (nearest shrunken centroid);
- Discrete-based: *plda* (poisson linear discriminant analysis), *plda2* (plda with power transformation) and *nblda* (negative binomial linear discriminant analysis);
- Voom-based: *voomDLDA* and *voomNSC*.

To all sub-datasets one normalization method (*deseq* and *TMM*, which were exposed in Section 3.5.1) was applied. Regarding the continuous-based classifiers, the sub-datasets were also transformed *a priori* with *vst*, *rlog* and *logcpm* (in detail in Section 3.5.2). Thus, the combinations tested were:

- *deseq-vst*;

- *deseq-rlog*;
- *deseq-logcpm*;
- *tmm-logcpm*.

One should bear in mind that the transformation methods are not applied when using discrete classifiers or voom-based classifiers (the latter perform the voom transformation within itself).

The splitting ratio for training and testing was 70% and 30%, respectively. All the models were trained using 5-fold CV repeated 10 times to assess performance variability across simulations. The test set underwent the same normalization and transformation (in the cases where the classifier was continuous) before the algorithm predicted its class labels. Each model was further evaluated over 16 repeats in order to give robustness to the results. The flowchart in Figure 4.4 describes how the overall approach for the investigation of the different classifiers and their performance across the different datasets created was conducted. For comparison purposes, the accuracy, sensitivity and specificity was assessed and stored. Furthermore, the sparse models' (*NSC*, *plda*, *plda2* and *voomNSC*) sparsity, a measure of proportion of features used in the trained model, was calculated.

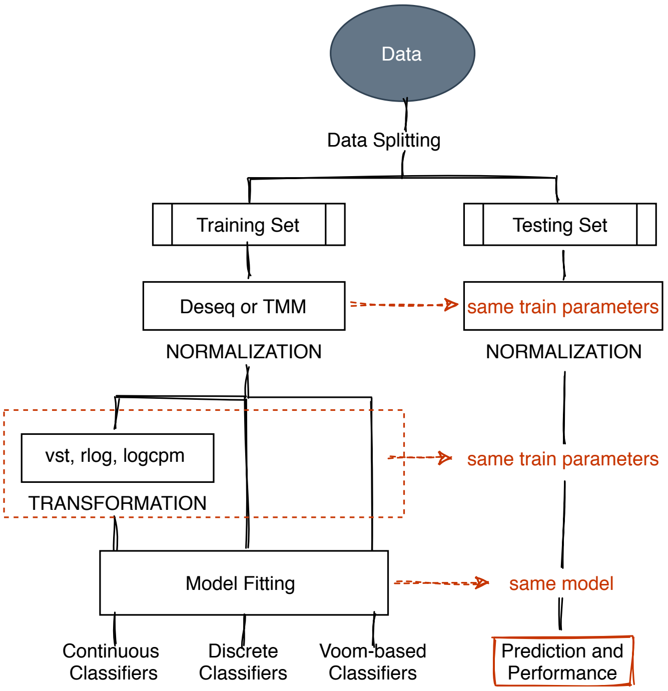


Figure 4.4: Flowchart of procedure used to fit data into classifiers and to compare model's performance after prediction of class labels.

Chapter 5

Experimental Results

Did the sparse logistic regression pipeline originate accurate models? Can Bayesian network learning expose new gene associations not yet known to be associated with anti-TNF treatment response? Did the alternative classifiers contribute with supplementary information? The results produced by the implemented approaches are revealed and examined alongside with the available literature.

5.1 Sparse Logistic Regression

The final datasets used after applying the pre-processing step are summarized in Table 5.1.

Table 5.1: Final datasets after pre-processing.

Dataset	Observations	Variables
	(patients)	*(genes)*
BL	63	21,911
M03	65	22,142

BL: Baseline dataset; M03: Third month dataset

From applying sparse logistic regression, tables 5.2 and 5.3 show the values of the median, maximum, minimum and interquartile (IQR) amplitude of the Area under the curve (AUC) obtained over the 5,000 runs, using each value of α, in each dataset. The corresponding box plots can be found in Figure 5.1.

Through the boxplots from Figure 5.1 two clear traits are observed: across the α parameter range,

Table 5.2: AUC median, maximum, minimum and interquartile amplitude values for different α parameters in the BL dataset.

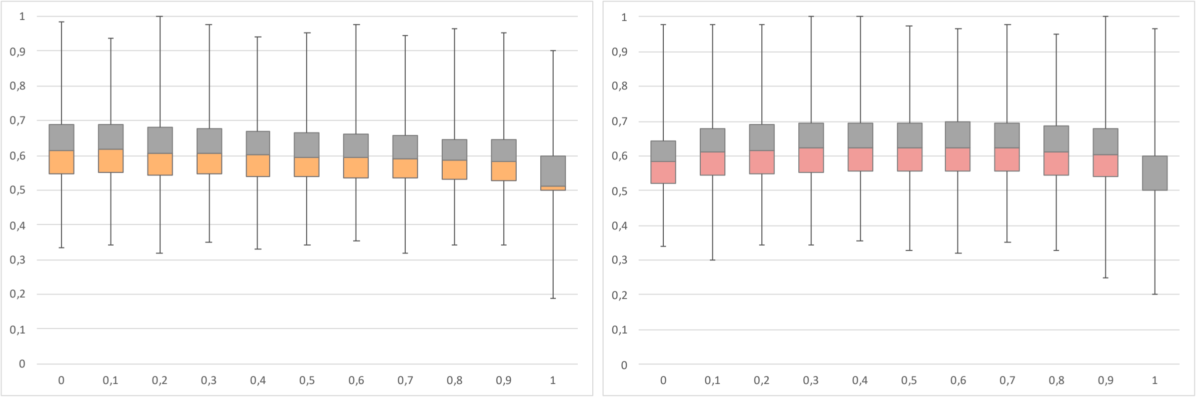
	α										
AUC	0	0.1	**0.2**	**0.3**	0.4	0.5	0.6	0.7	0.8	0.9	1
Median	0.615	0.619	0.607	0.606	0.604	0.596	0.594	0.59	0.586	0.583	0.512
Max	0.984	0.938	1	0.976	0.939	0.952	0.978	0.947	0.964	0.952	0.901
Min	0.333	0.343	0.319	0.352	0.33	0.343	0.354	0.32	0.341	0.343	0.188
IQR	0.141	0.135	0.138	0.131	0.13	0.128	0.126	0.121	0.116	0.118	0.1

AUC: Area under the curve; Max: Maximum value; Min: Minimum value; IQR: Interquartile range

Table 5.3: AUC median, maximum, minimum and interquartile amplitude values for different α parameters in the M03 dataset.

	α										
AUC	0	0.1	0.2	**0.3**	**0.4**	0.5	0.6	0.7	0.8	0.9	1
Median	0.583	0.611	0.614	0.622	0.622	0.625	0.622	0.622	0.611	0.603	0.5
Max	0.978	0.976	0.978	1	1	0.974	0.967	0.977	0.949	1	0.966
Min	0.341	0.3	0.344	0.344	0.356	0.33	0.321	0.352	0.33	0.25	0.2
IQR	0.122	0.135	0.143	0.142	0.138	0.138	0.144	0.138	0.141	0.141	0.1

AUC: Area under the curve; Max: Maximum value; Min: Minimum value; IQR: Interquartile range



(a) BL dataset

(b) M03 dataset

Figure 5.1: Calculated Area under the curve values for each α using the BL (a) and M03 (b) datasets. Box plots show median values (border between grey and orange boxes), the distance between the first and third quartiles (interquartile range; joint grey and orange boxes) and the lower and upper extremes (vertical line that extends from lowest or highest value, respectively).

the obtained median value for the AUC is somewhat constant and there is a great variability of values, all ocurring for each dataset. Considering the few obervations (63 and 65), values of around 0.6 for the AUC medians are satisfactory. Due to the absence of an obvious choice about which α value to choose for each dataset, two values were chosen for each one, and considered to represent the best predictive models. Hereupon these are the four models to be used, and thus they should be clearly defined:

Table 5.4: Designation of the 4 models obtained with the elastic net regularization.

BL		**M03**	
$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.3$
Model 1 BL_1	Model 2 BL_2	Model 1 M03_1	Model 2 M03_2

Regarding the biomarkers possibly in strong association with the treatment response, the Leave-one-Out Cross-Validation (LOOCV) approach was applied to each model. The Receiver Operating Characteristic (ROC) curves can be compared in Figure 5.2. At first look, the BL models seem to perform worst than the M03 models. Setting the threshold for prediction at 0.5, the model which best performed for BL data had an accuracy value of 0.651 and for M03 data an accuracy of 0.4. Due to the apparent contradiction between AUC and accuracy results, and the fact this classification procedure is in fact imbalanced

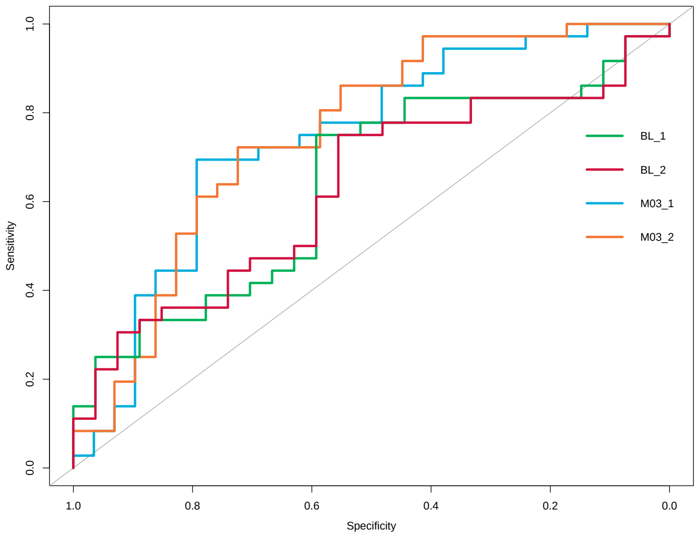


Figure 5.2: ROC curves based on LOOCV for the four models.

(in both datasets the positive class occurrence is higher than the negative one), it was inspected which threshold value represented the best overall prediction. The highest accuracy was achieved with models 1 of each dataset: BL_1 performed with an accuracy of 0.683 and M0_3 of 0.738. All the metrics used to evaluate the LOOCV performances are explicit in table 5.5. This result points at the conclusion that the best treatment response prediction is obtained from the transcriptomic data retrieved after the third month of treatment.

Table 5.5: Leave-one-Out Cross-Validation results of each model when letting the default threshold set at 0.5 and for the best accuracy across the threshold range.

Model	Default (cut off = 0.5)		Optimal cut off value			
	Accuracy	AUC	Threshold	Specificity	Sensitivity	Accuracy
BL_1	0.635	0.637	0.541	0.593	0.750	**0.683**
BL_2	0.651	0.629	0.529	0.556	0.750	0.667
M03_1	0.400	0.739	0.563	0.793	0.694	**0.738**
M03_2	0.369	0.751	0.573	0.724	0.722	0.723

The intersection of genes appearing in all the predictive models calculated with LOOCV may correspond to those that give a better prognostic about the RA treatment. From model 1 (highest α value) to model 2 (smallest α value), the number of genes repeatedly selected by LOOCV for model prediction increased, as it was expected given the influence of that parameter. In fact, the number of genes was 24 (BL_1), 35 (BL_2), 12 (M03_1) and 22 (M03_2). In both datasets, the list of genes in model 2 includes all the genes in model 1 (the Venn diagram in Figure 5.3 illustrates these results). Besides, Constantino et al. presented the intersection between models with $\alpha = 0.1$ and $\alpha = 0.4$, thus obtaining an even smaller list of genes, all of which were found in the present study. The list of genes obtained is presented in Table 5.6 and their corresponding complete name can be found in Table A.1 of Appendix A.

Narrowing the analysis to the genes with a minimum reading count of 20, the boxplots of Figure 5.4

Table 5.6: List of predictive genes in RA treatment after applying Leave-one-Out Cross-Validation in each model for each dataset.

Dataset	**Genes**
BL_1	*ALOX12B*, *CAPNS2*, *CCDC108*, *CTSG*, *EPHX4*, *ERICH6*, *EVPLL*, *FAM133CP*, *FOXD4L3*, *HIST1H3J*, *IGF2BP1*, *LOC339975*, *LRGUK*, *MPO*, *NUAK1*, *ODF3L2*, *PRKG1*, *PRSS30P*, *RAD21L1*, *RCAN3AS*, *ROPN1L-AS1*, *SLC6A19*, *SYT1* and *TGFB2*
BL_2	*ALOX12B*, *CAPN11*, *CAPNS2*, *CCDC108*, *CTSG*, *EPHX4*, *ERICH6*, *EVPLL*, *FAM133CP*, *FOXD4L3*, *HIST1H3J*, *IGF2BP1*, *KCNH4*, *LINC00696*, *LMOD3*, *LOC339975*, *LRGUK*, *MAG*, *MAGEC2*, *MIR941-4*, *MPO*, *NUAK1*, *ODF3L2*, *PMS2L2*, *PRKG1*, *PRSS30P*, *RAD21L1*, *RCAN3AS*, *RNU6-28P*, *ROPN1L-AS1*, *SKA3*, *SLC6A19*, *SYT1*, *TBX2* and *TGFB2*
M03_1	*ADAM33*, *CCDC110*, *ELANE*, *KCNJ8*, *LOC101928222*, *LRRN4CL*, *MTRNR2L3*, *TMEM105*, *TRIM7*, *UBE2QL1*, *VSTM2L* and *ZNF843*
M03_2	*ADAM33*, *CCDC110*, *ELANE*, *FBLIM1*, *GFAP*, *HYAL4*, *KCNJ8*, *KCNK4*, *KNCN*, *LOC100128076*, *LOC100268168*, *LOC101928222*, *LRRN4CL*, *MTRNR2L3*, *SERTM1*, *TMEM105*, *TPBG*, *TRIM7*, *TTC25*, *UBE2QL1*, *VSTM2L* and *ZNF843*

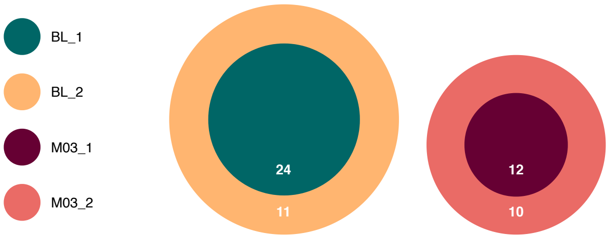


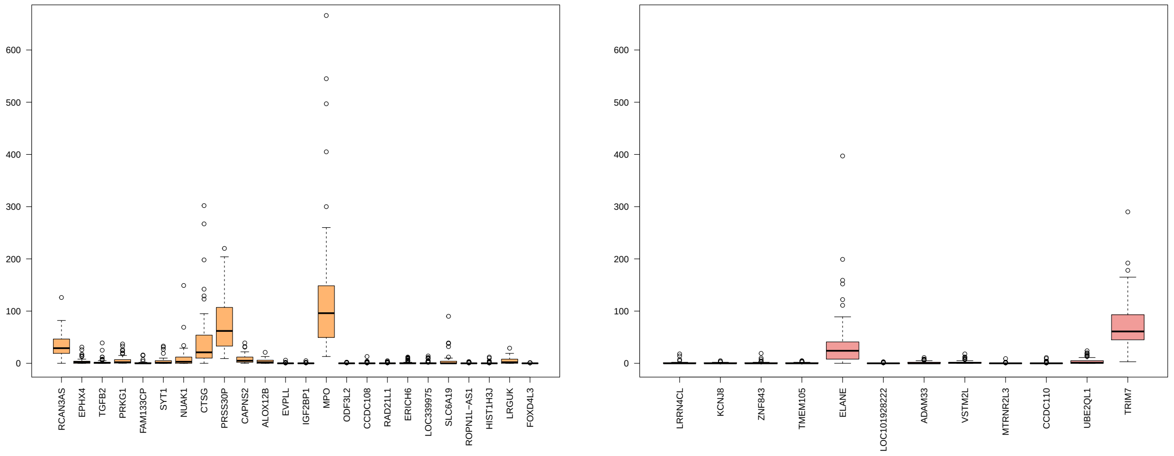
Figure 5.3: Schematic representation of common genes disclosed by LOOCV between the 4 models.

a) reveal that at baseline, the expression *MPO*, *PRSS30P*, *RCAN3AS* and *CTSG* stands out compared to the remaining. Note that until the end of this section all the results refer to models 1 of BL and M03 since in models 2 all the additional genes have a reading count lower than 20.

High serum levels of **myeloperoxidase** (encoded by *MPO* gene), the most frequent protein in mature neutrophils, are known to be associated with RA and other autoimmune complications. When released by this abundant circulating white blood cells, the neutrophils, *MPO* binds to macrophages, a distinctive type of white blood cells, initiating a molecular cascade resulting in secretion of interleukin-1, interleukin-8 or TNF-α [76].

CTSG encodes **Cathepsin G** (*CatG*), which belongs to the neutrophil serine proteases family. Among its many functions, there is a clear role of *CTSG* in immune and inflammation reactions, participating in the pathogenesis of some autoimmune diseases by promoting the migration of neutrophils, monocytes and antigen presenting cells. *CTSG* constitutes a biomarker for inflammatory arthritis and its activity is increased in the synovial fluids of RA patients [77, 78].

Being expressed by RA neutrophils, *MPO* and *CTSG* are directly related to neutrophil granule proteins, which synergize to modulate inflammation and even tumor development. It has been demonstrated that expression of *MPO* and *CTSG* in peripheral blood neutrophils from patients with RA, before therapy with an anti-TNF, can predict a subsequent response to anti-TNF as a first biologic, with specificities and sensitivities of up to 100%. Specifically, they were identified as being significantly different expressed



(a) Reading counts of repeatedly selected genes in BL data

(b) Reading counts of repeatedly selected genes in M03 data

Figure 5.4: Reading counts of common repeatedly selected genes obtained with LOOCV for BL models and M03 models, respectively.

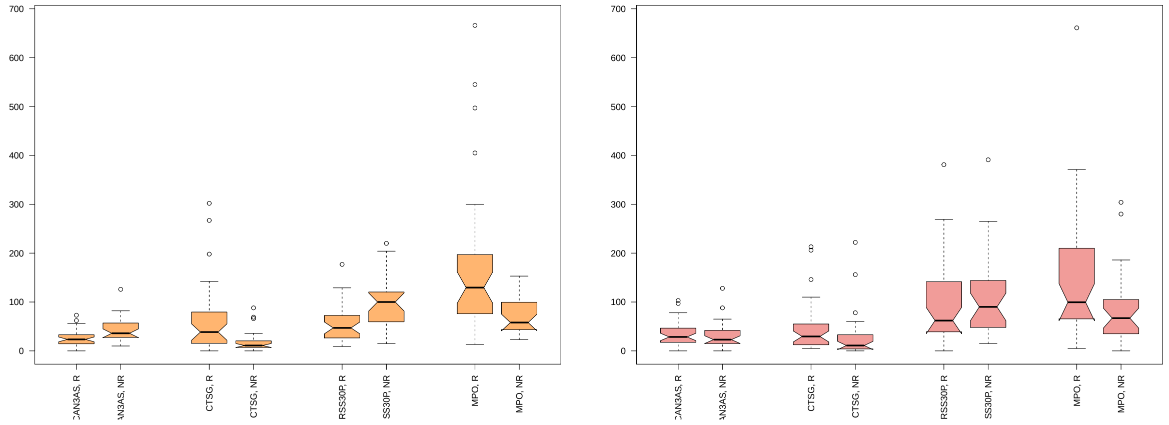
in nonresponder patients [79].

PRSS30P is a pseudogene related to a serine protease but of unknown function. No reference of this gene being related to any pathogenesis or inflammation process was found.

The **regulators of calcineurin** (*RCANs*) are a group of proteins which form a functional subfamily with three members: *RCAN1*, *RCAN2*, and *RCAN3*. They are reported to either facilitate or inhibit calcineurin, depending on *RCAN* protein amount and calcineurin affinity. *RCAN3* specifically has shown to modulate T cell development by increasing positive selection and suppressing pro-inflammatory T cell differentiation in cell culture and in arthritis development induced by collagen injection in murine models, and thus suggesting that it may be an effective treatment for RA [80]. However future research is expected to explore and expand on these functions. Nevertheless, *RCAN3AS* actually refers to the *RCAN3* antisense. Antisense RNAs are unique transcripts that complement mRNA and thus block its translation into a protein [81]. Consequently, the gene *RCAN3* will be under-expressed. One could hypothesize that if in non-responders this protein is prevented from being translated, then in responders its expression levels will be higher and thus considered a biomarker.

To better understand how the previous disclosed four genes relate to the treatment response, their levels of expression in responder and non-responder patient groups were compared, not only at the beginning of the anti-TNF treatment but also at the third month-post treatment initiation in order to understand their time evolution. This is illustrated in Figure 5.5. At baseline, where the four genes were identified as most relevant, the clear differences in expression between responders and non-responders suggest their predictive power: *MPO* and *CTSG* are more expressed in the responder group, whereas *RNAC3AS* and *PRSS30P* are more expressed in the non-responder group. Three months into the treatment, the differences between read counts in responders and non-responders was less noticeable except for *MPO* gene.

Since no literature exists regarding the matter of *RCAN3* and its antisense in RA context, the reading counts of gene *RCAN3* were assessed in hope that the expression would be higher in the responder group



(a) Expression at BL

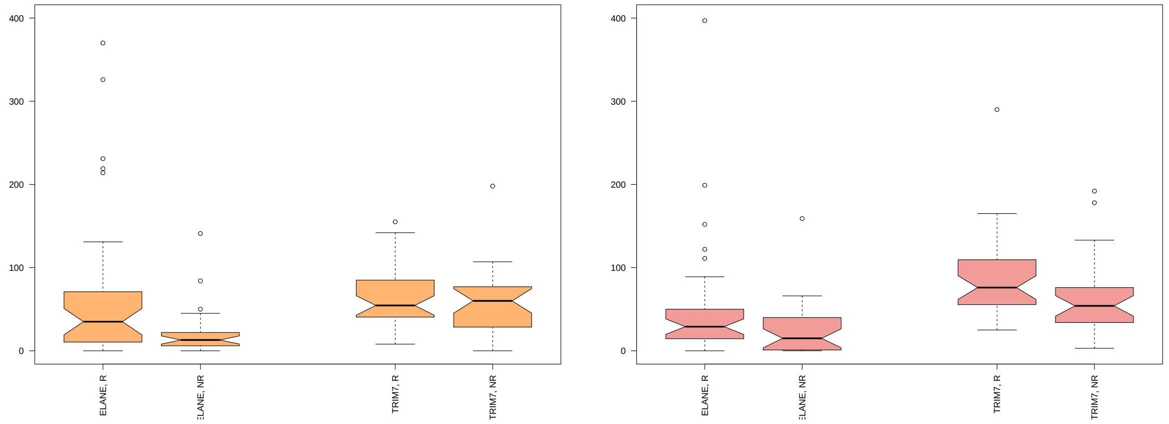
(b) Expression at M03

Figure 5.5: Comparison between responders and non-responders regarding reading counts of *RCAN3AS*, *CTSG*, *PRSS30P* and *MPO* at BL (left, orange color) and M03 (right, red color). Each pair corresponds to one gene, with left and right boxes corresponding to responder and non-responder.

and thus validate the conjecture made before. However, the results showed exactly the opposite *i.e.*, the non-responder group had higher median counts. Nevertheless, this is a study involving a small number of patients and so a more complex investigation focusing gene *RCAN3* should not be disregarded. At last, even though no allusion of *PRSS30P* gene being associated to RA was uncover, given the evidence found relating the remaining genes to the disease, it gives confidence that understanding gene *PRSS30P* might enlighten the complicated process of RA.

Regarding the M03 dataset, the boxplots in Figure 5.4b disclosed a substantial expression of the genes *ELANE* and *TRIM7*. Similarly to *CTSG*, **elastase, neutrophil expressed** (*ELANE*) codes for human neutrophil elastase (*HNE*), a neutrophil serine protease and thus it is involved in the same mechanisms as the former: it is considered a multifunctional enzyme involved in the killing of pathogens, regulation of inflammation and tissue homeostasis. Regarding RA, it can directly degrade the matrix, destroying cartilage components [78]. Being involved in many inflammatory diseases, *HNE* is a therapeutic target of considerable interest as it is demonstrated by the number of *HNE* inhibitors patents developed in the last decade by pharmaceutical companies [82]. In the same research study, Wright et al. also notice the significantly different expression of ELANE, although the focus was only in the patients' transcriptomic data prior to the treatment initiation.

Gene *TRIM7*, **Tripartite Motif Containing 7**, encodes a member protein of the tripartite motif (*TRIM*) family which have been implicated in a broad range of biological processes including cell differentiation, development, oncogenesis and antiviral immunity. The *TRIM7* protein has gained attention in cancer studies, having been described in one of them as a negative regulation in lung tumors [83]. On the contrary, little references were found associating RA and *TRIM7*, except when Stangenberg et al. observed a decrease of its expression in the ankle nerve of patients who had previously suffered from unilateral transection of the sciatic and femoral nerves [84]. Nevertheless, this study was focus on the intriguing observation that these patients, if they evolve to RA later on, its typical inflammatory sym-



(a) Expression at BL

(b) Expression at M03

Figure 5.6: Comparison between responders and non-responders regarding reading counts of *ELANE* and *TRIM7* at BL (left, orange color) and M03 (right, red color). Each pair corresponds to one gene, with left and right boxes corresponding to responder and non-responder.

metry is not observed. Furthermore, Kim et al. findings also suggest that the TRIM family is part of one of the RA subgroups representing a distinct mode of inflammation which is deflected toward a certain combination of signaling pathways [19]. Nonetheless, no clear evidence has been found of a possible connection between TRIM7 and the response to anti-TNF.

Once more, the expression of these last two genes was analysed individually for responder and non-responder patients at both instants in order to understand if the expression patterns were already there at the treatment beginning. Figure 5.6b shows that while being both considered predictors by the LOOCV approach, their expression differences is not so obvious. At baseline, the gene *ELANE* expression was more distinct between responders and non-responders, and later it appeared to have evolved to a similar level. Nonetheless, the methods applied in this work revealed *ELANE* to be a predictor for RA treatment at M03.

The datasets used in this thesis were the basis of previous work [17]. Even though the biomarkers found with the proposed methodology were confirmed with other literature, they do not correspond to the ones obtained by Farutin et al.. However it should be noticed the cited research team used other methods in which not only transcriptomic data but also plasma proteomics was available.

5.2 Bayesian Network Models

Selecting the best α parameters which fitted the data (0.3, 0.2 for BL data and 0.4, 0.3 for M03 data) and applying the elastic net regularization resulted in 4 new sparse datasets with a smaller variable number. The dimension of those 4 new models is indicated in table 5.7. At this point, each model was split into two according to the RA treatment response of each patient contained in it (R *versus* NR).

Accordingly to the scheme presented in Figure 4.2, a total of 32 BN were to be obtained. However, in some of the cases where the algorithm was given the command to choose the network corresponding to the optimal solution, it chose the one with the given number of edges (recalling, that could either lead

Table 5.7: Number of variables selected by elastic net when applied to BL and M03 models.

BL		M03	
BL_1	BL_2	M03_1	M03_2
71	111	61	91

to a network with number of edges as the specified limit or with a smaller number). In other words, it happened that there were no difference between forcing the number of edges and letting the algorithm chose the BN based on the trade-off. To illustrate the cases in which this happened, the scheme 4.2 previously presented in the methodology was rearranged to enlighten which networks were effectively obtained (see Figure 5.7).

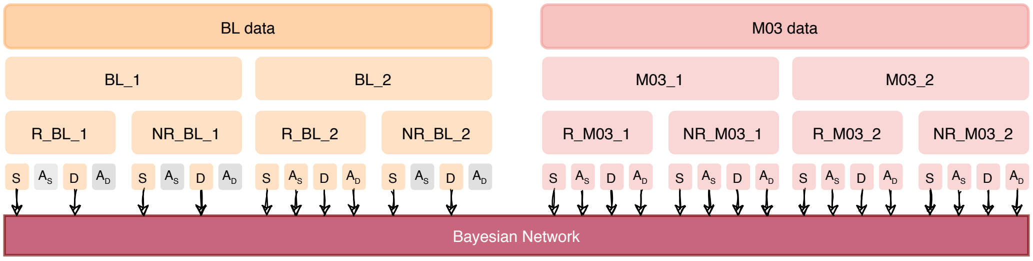


Figure 5.7: Final set of Bayesian Networks obtained. The grey squares indicate the networks which were in repetition and thus not included in the results.

In respect to BL data, letting the algorithm choose the best solution only produced a different network in the responder group of model 2. Contrarily, all the eight "hand-picked" networks of M03 data did not correspond to the optimal solution automatically chosen. The complete set of BN obtained can be found in figures 5.9, 5.10, 5.11 and 5.12. It is important to mention that "forcing" the network to have a certain edge number does not necessarily mean that the algorithm will return a solution with such edge number (*i.e.*, in a few cases the network had slightly fewer edges than what was pretended). However, it was not considered that this affected the analysis.

In order to disclose which gene networks may regulate the response to anti-TNF treatment, the 3 interactions with the highest weight value for every network were assessed. The results are presented in Tables 5.8 and 5.9 (Table A.2 contains the explicit genes' names). In general terms, there was consistency in the edges identified across the different models and the different sizes networks, for both BL and M03 data (the cell colors in the tables intends to highlight examples of that consistency). Increasing the number of allowed edges ("S" cases *vs* "D" cases) revealed a few changes in the interactions. Regarding the comparison between "hand-picked" networks ("S" and "D" cases) and algorithm-chosen ("A_S" and "A_D""cases, accordingly), there was no difference when using the BL models (in the only two occasions the algorithm optimal solution was distinctive, the three most weighted edges remained the same).

Intriguingly, for the M03 models it produced a massive change: the number of edges in each network varied only between 1 and 3. On this account, the few genes connecting those edges were further investigated: *RSPH10B2* and *RSPH10B* correspond to genes encoding for the head components of radial spoke structures (a multi-unit protein structure found in axonemes of eukaryotic cilia and flagella);

Table 5.8: BN interactions obtained (showing only 3) with highest edge weight for **BL** data. Colors highlight four examples of equivalent edges across the different networks obtained. Note that the symbol "-" simply indicates the cases where the obtained network were repeated, as illustrated in Figure 5.7.

		S	D	A_S	A_D
Model 1	R	EPHX4 - LRGUK	TBX2 - CYGB	-	-
		MIR941-4 - MIR941-2	EPHX4 - LRGUK	-	-
		BATF2 - EVPLL	LOC100507156 - LINC00696	-	-
	NR	EVPLL - IGF2BP1	MAG - MAGEC2	-	-
		RCAN3AS - KCNH4	EVPLL - IGF2BP1	-	-
		ERICH6 - SULF1	LOC339975 - LILRB4	-	-
Model 2	R	CDC42EP4 - TCN2	LOC100507156 - LINC00696	CDC42EP4 - TCN2	LOC100507156 - LINC00696
		EPHX4 - LRGUK	TBX2 - CYGB	EPHX4 - LRGUK	TBX2 - CYGB
		LOC100507156 - LINC00696	DRD2 - CAPN11	LOC100507156 - LINC00696	DRD2 - CAPN11
	NR	MIR941-4 - FGD5P1	MIR941-4 - FGD5P1	-	-
		C1orf95 - MAGEC2	SLC25A52 - ADAMTS9	-	-
		SLC25A52 - ADAMTS9	EVPLL - IGF2BP1	-	-

Table 5.9: BN interactions obtained (showing only 3) with highest edge weight for **M03** data. Colors highlight four examples of equivalent edges across the different networks obtained. *none* indicates that the BN learning did not reveal further interactions.

		S	D	A_S	A_D
Model 1	R	KCNK4 - MIR718	KCNK4 - MIR718	RSPH10B2 - RSPH10B	RSPH10B2 - RSPH10B
		RSPH10B2 - RSPH10B	RSPH10B2 - RSPH10B	*none*	*none*
		CTSG - ELANE	MTRNR2L3 - ZNF843	*none*	*none*
	NR	F3 - LOC101927468	C8B - LOC102467224	KNCN - CCDC110	KNCN - CCDC110
		C8B - LOC102467224	F3 - LOC101927468	RSPH10B2 - RSPH10B	RSPH10B2 - RSPH10B
		KNCN - CCDC110	FBLIM1 - UBE2QL1	*none*	*none*
Model 2	R	MTRNR2L3 - MIR4271	MTRNR2L3 - MIR4271	RSPH10B2 - RSPH10B	RSPH10B2 - RSPH10B
		TRIM7 - TMEM51-AS1	KCNK4 - MIR718	*none*	*none*
		KCNK4 - MIR718	FSD2 - RS1	*none*	*none*
	NR	VWA1 - LINC01361	VWA1 - LINC01361	KNCN - CCDC110	KNCN - CCDC110
		FBLIM1 - UBE2QL1	LOC100506071 - HIST1H2AJ	MIR3918 - VWA1	MIR3918 - VWA1
		VWA1 - CES1P1	FBLIM1 - UBE2QL1	RSPH10B2 - RSPH10B	RSPH10B2 - RSPH10B

kinocilin, *KNCN*, has a role in stabilizing dense microtubular networks or in vesicular trafficking [85]; *CCDC110* has been identified as novel cancer/testis antigen recognized by cellular and humoral immune responses [86]; *MIR3918* are short non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multicellular organisms by affecting both the stability and translation of mRNA (messenger RNA) [87] and finally *VWA1* belongs to a superfamily of extracellular matrix proteins and appears to play a role in cartilage structure and function [87]. The possible relation of these protein-protein interactions to RA is not evident in the literature.

Not disregarding the top gene-gene interactions found through the BN learning, it was essential to compare them to what has been published; hence the use of STRING database [72].

Regarding the BN learnt from BL data, Table 5.10 indicates the overlapping interaction found. On the one hand, **CTSG** and **MPO** genes share an interaction in the responders group, which is given a total score of 0.989 in STRING database. Given that they were found to be anti-TNF response predictor in the conducted LOOCV approach, there is strong evidence that their expression levels might be determinant for a future anti-TNF good responder patient. On the other hand, **CTSG** − **AZU1**, which scores 0.964, was an interaction found in the non-responders group. *AZU1* encodes for azurocidin 1 granules, a known important multifunctional inflammatory mediator for recruitment of monocytes in the second wave of inflammation. The protein encoded by the gene SERPINB10 is a protease inhibitor which helps in the regulation of protease activities. It has been reported its influence in inhibiting TNF-α-induced cell death [88].

The Venn diagrams regarding model 2 (see figure 5.8) highlight one interaction common to both responders and non-responders: **MPO – AZU1** (score: 0.985), suggesting that it might be be relevant for both types of patients. An important aspect from Table 5.10 stands out: either increasing the number of variables with the elastic net penalisation (which corresponds to defining a smaller α) or increasing the number of allowed edges in the BN learned did not influence the results in regard to the interactions found in the STRING database.

In relation to the overlaps obtained from the M03 data, only one interaction was found to be in common with the STRING database: **CTSG – ELANE** (score of 0.982). Similarly as previously stated, this protein-protein interaction being found in responders and non-responders advocates for its importance in the mechanisms of anti-TNF treatment. Moreover, the *ELANE* gene was not selected by the elastic net in the regularization applied to the BL data, an hypothesis for that being that its expression became relevant somewhere between the day 1 and day 90 of the anti-TNF treatment. This observation might be worth of exploring in further research. In addition, none of these findings exempt a rigorous analysis by a team of rheumatologists.

Table 5.10: Overlapping protein-protein interactions between learnt Bayesian Networks from **BL data** and STRING Database. Note that the symbol "-" simply indicates the cases where the obtained network were repeated, as illustrated in Figure 5.7.

		S	D	A_S	A_D
Model 1	R-BL ∩ STRING	MPO - CTSG CTSG - SERPINB10 AZU1 - MPO	MPO - CTSG AZU1 - MPO	-	-
	NR-BL ∩ STRING	AZU1 - MPO CTSG - AZU1	AZU1 - MPO CTSG - AZU1	-	-
Model 2	R-BL ∩ STRING	MPO - CTSG AZU1 - MPO	MPO - CTSG AZU1 - MPO	MPO - CTSG AZU1 - MPO	MPO - CTSG AZU1 - MPO
	NR-BL ∩ STRING	AZU1 - MPO CTSG - AZU1	AZU1 - MPO CTSG - AZU1	-	-

Table 5.11: Overlapping protein-protein interactions between learnt Bayesian Networks from **M03 data** and STRING Database. *none* indicates that overlapping analysis revealed no interactions.

		S	D	A_S	A_D
Model 1	R-BL ∩ STRING	CTSG - ELANE	CTSG - ELANE	*none*	*none*
	NR-BL ∩ STRING	CTSG - ELANE	CTSG - ELANE	*none*	*none*
Model 2	R-BL ∩ STRING	CTSG - ELANE	CTSG - ELANE	*none*	*none*
	NR-BL ∩ STRING	CTSG - ELANE	CTSG - ELANE	*none*	*none*

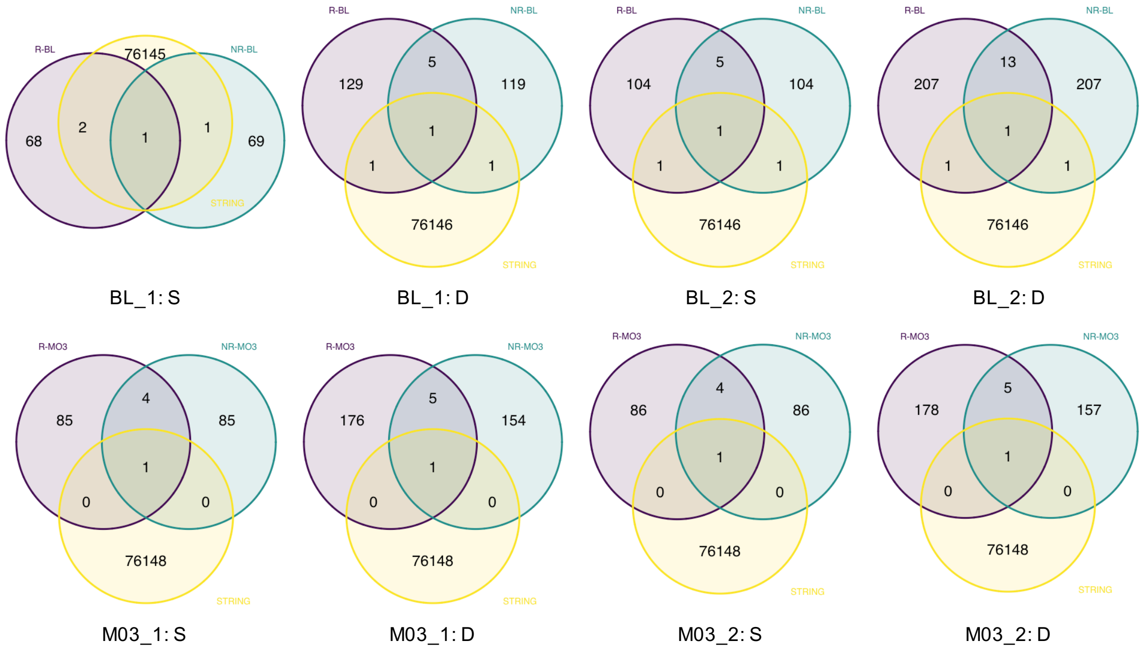
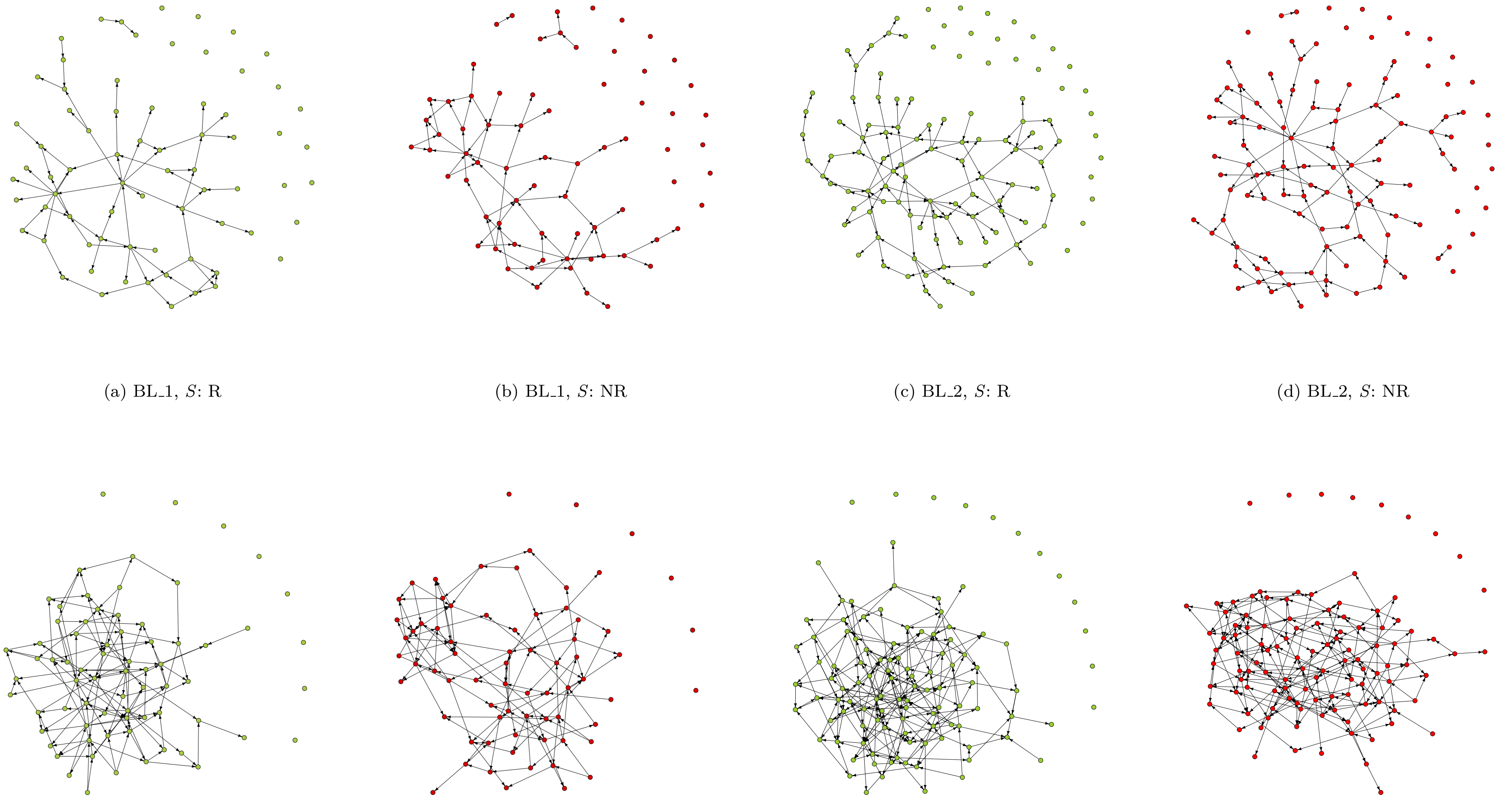


Figure 5.8: Venn diagrams showing common interactions between learnt BN from responders and non-responders groups and STRING database.



(a) BL_1, S: R

(b) BL_1, S: NR

(c) BL_2, S: R

(d) BL_2, S: NR

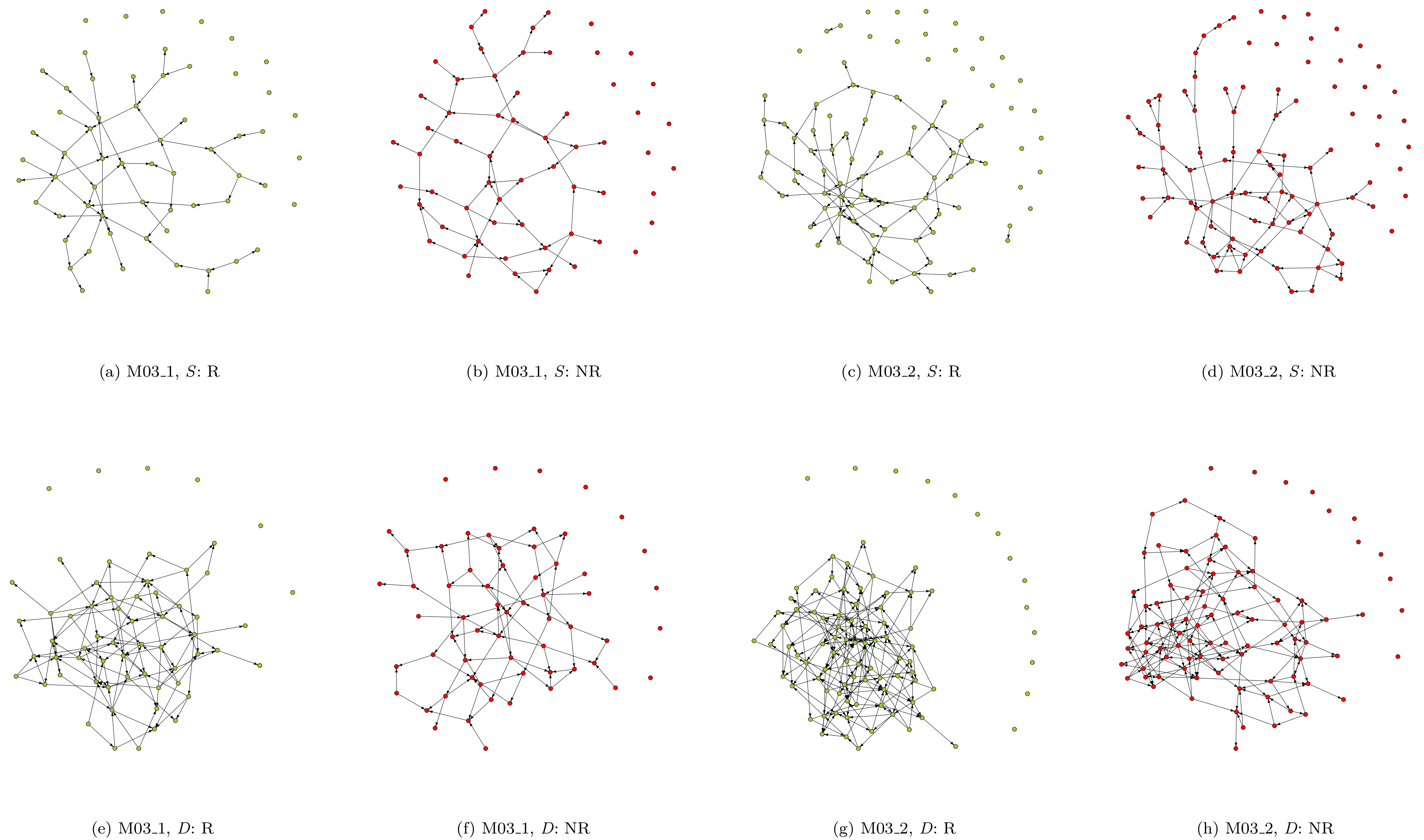
(e) BL_1, D: R

(f) BL_1, D: NR

(g) BL_2, D: R

(h) BL_2, D: NR

Figure 5.9: Bayesian Networks learnt from **BL data** when forcing the edge number to be the number of variables in the model (S; first row) and the double of that amount (D; second row). Model 1 and Model 2 correspond to first two and last two columns, respectively. Colors indicate the treatment response group: responders (green nodes) and non-responders (red nodes) groups.



(a) M03_1, S: R

(b) M03_1, S: NR

(c) M03_2, S: R

(d) M03_2, S: NR

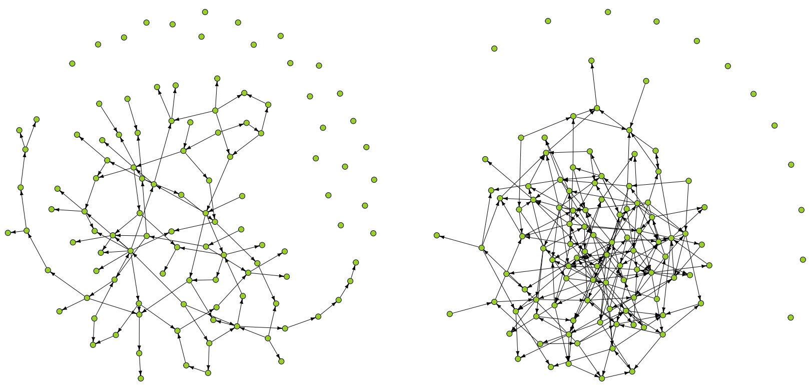
(e) M03_1, D: R

(f) M03_1, D: NR

(g) M03_2, D: R

(h) M03_2, D: NR

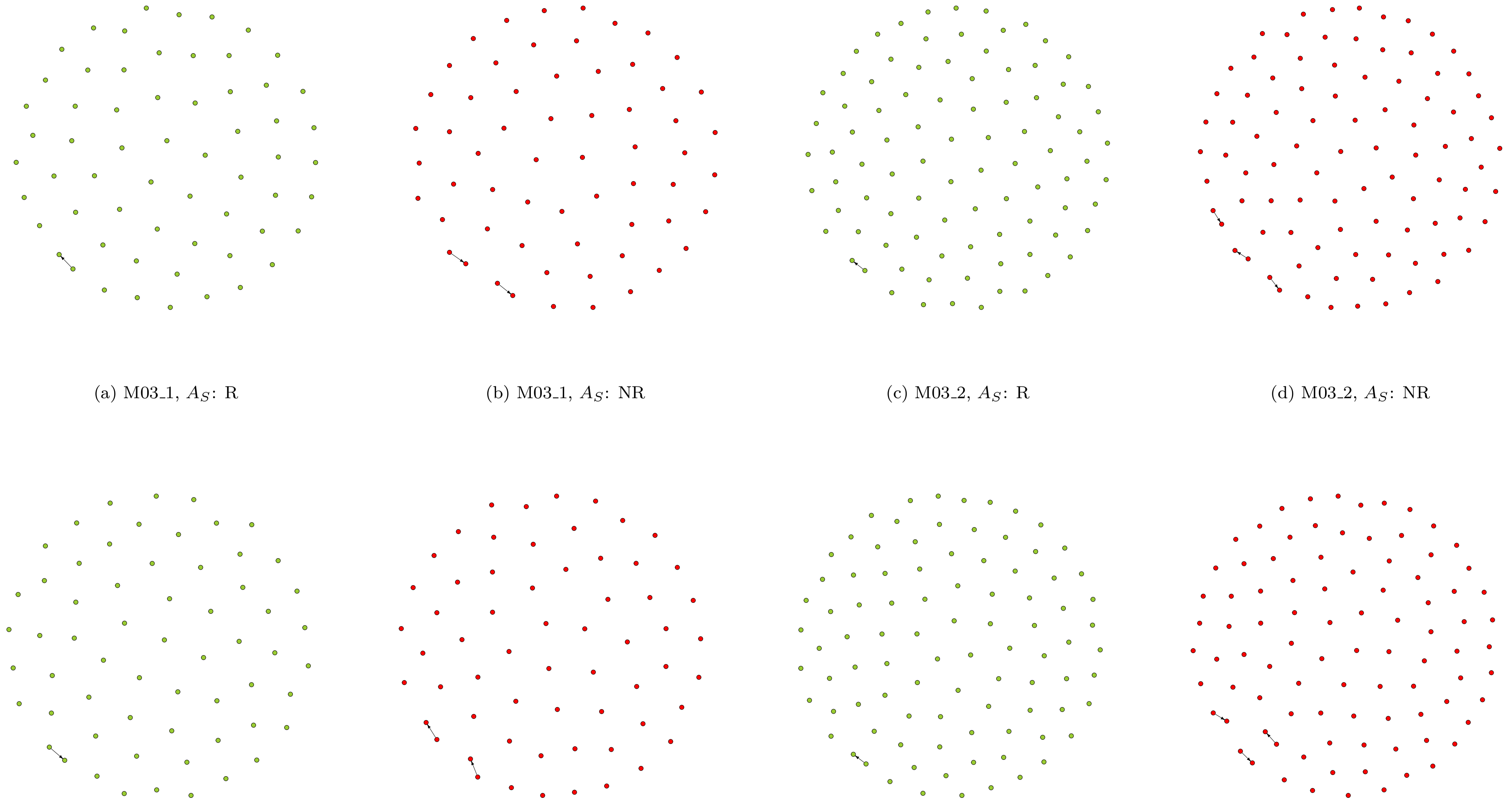
Figure 5.10: Bayesian Networks learnt from **M03 data** when forcing the edge number to be the number of variables in the model (S; first row) and the double of that amount (D; second row). Model 1 and model 2 correspond to first two and last two columns, respectively. Colors indicate the treatment response group: responders (green nodes) and non-responders (red nodes) groups.



(a) BL_2, A_S: R

(b) BL_2, A_D: R

Figure 5.11: Bayesian Networks learnt from **BL data** when allowing the algorithm to select the optimal solution from limiting the maximum edge number to be the number of variables in the model (A_S) and the double of that amount (A_D; second row). Algorithm only presented a different solution from 5.9 for model 2-responder groups.



(a) M03_1, A_S: R

(b) M03_1, A_S: NR

(c) M03_2, A_S: R

(d) M03_2, A_S: NR

(e) M03_1, A_D: R

(f) M03_1, A_D: NR

(g) M03_2, A_D: R

(h) M03_2, A_D: NR

Figure 5.12: Bayesian Networks learnt from **M03 data** when allowing the algorithm to select the optimal solution from limiting the maximum edge number to be the number of variables in the model (A_S;first row) and the double of that amount (A_D;second row). Model 1 and model 2 correspond to first two and last two columns, respectively. Colors indicate the treatment response group: responders (green nodes) and non-responders (red nodes) groups.

5.3 Classification algorithms analysis

The problem of classification will be analysed based on the results presented in Tables 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18, and 5.19. The performance of each classifier was evaluated based on the accuracy, sensitivity and specificity values obtained after performing the response prediction with the fitted models. Overfitting (which occurs when the algorithm "over-learns" the data from the training set and does not perform well on the unseen testing data) was a liability, specially when learning from the obtained sparse models whose sizes varied between 61 and 111 features *versus* 63/65 observations. Thus repeating the fitting and prediction over different partitions of the train and test sets allowed a better result interpretability. Training and testing from the raw data was the only exception to this matter: assessment was made one single time due to the computational time required.

Inexplicably, only one time it was possible to perform class label prediction using the fitted continuous-based classifiers when the pre-processing step applied was the *deseq-logcpm* normalization-transformation combination. Notwithstanding being less reliable, the decision was not to neglect those results (this regards Tables 5.14 and 5.18).

The plots from Figure 5.13 gather the classifiers according to their nature and type of pre-processing method (first two columns) and according to the feature selection competency (last column). Each bar corresponds to the accumulated accuracy obtained over the different sub-datasets used (as indicated in Figure 4.3), allowing an interpretation about the overall capacity of the classifiers and also an inference regarding the performance for the different data types explored.

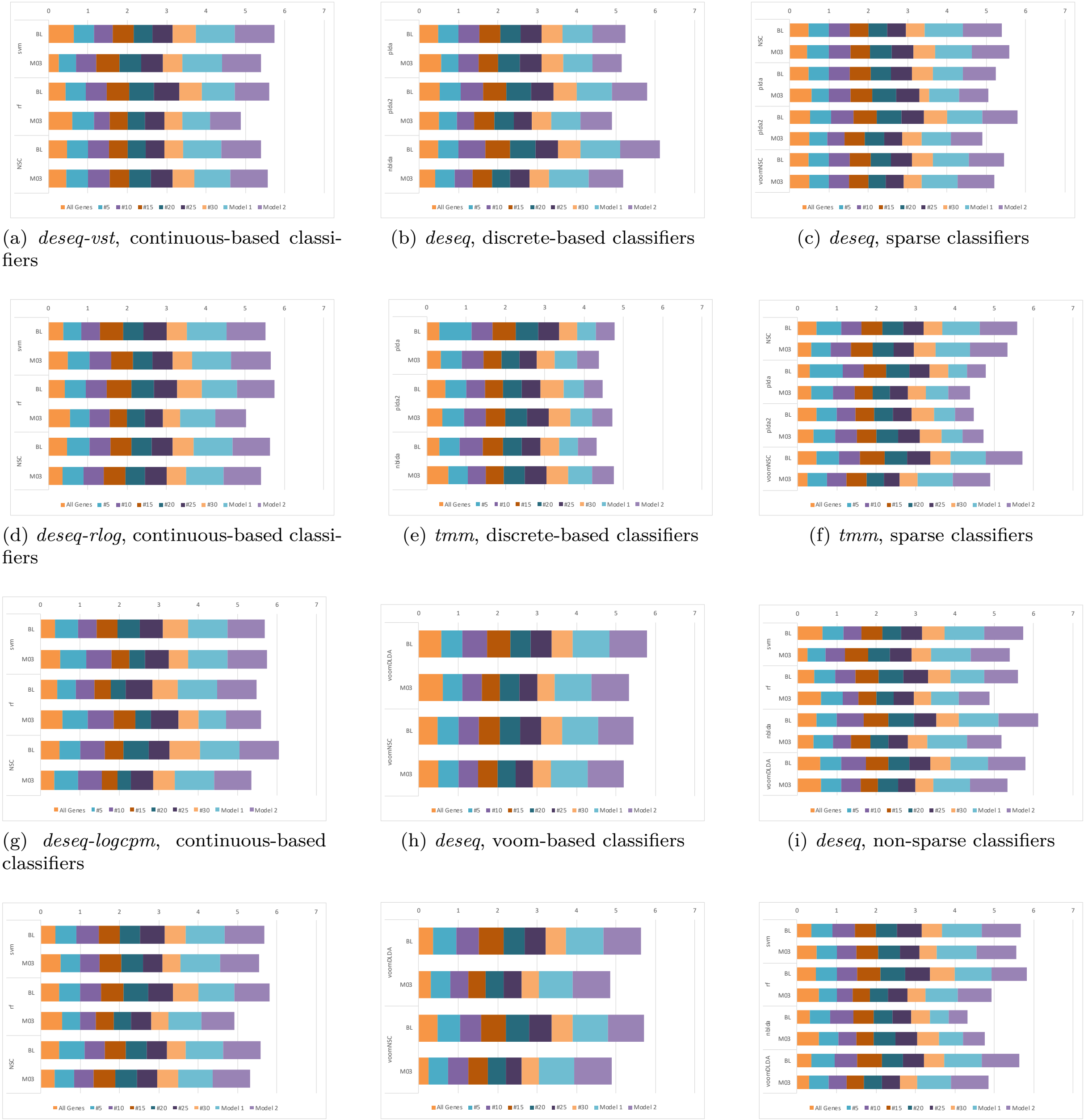
As it would be expected, the overall performance of the classifiers when learning from the complete datasets is poor (looking at the first two columns, the orange bar portion associated to *All Genes* corresponds to around 0.5 or less then the unity). Interestingly, increasing the number of high variance variables did not have a consistent positive impact on the overall testing accuracy. In fact, in some cases using the *top 5* to *top 15* of the high variant variables delivered better results then using the *top 20* to *top 30*, which may be related to the fact that models with lower complexity are less prone to overfit.

Being the *rf* models the only exception, all the classifiers reached a better prediction performance when constructed from the M03 data. The elastic net penalized models, however, revealed a better prediction accuracy at BL.

Despite the fact that only one *voom*-based classifier is sparse, they both delivered very similar accuracy when the data was transformed with *TMM* (Figure 5.13k). Moreover, when *deseq* was used, *voomDLDA* (non-sparse) slightly reached a better performance (Figure 5.13h).

Unquestionably the sparse data models obtained with the proposed methodology lead to more accurate classifiers and consequently the following observations will focus on them (last two portions of each plots bars). Witten has stated the little impact on choosing the normalization procedure on the classification performance (it is rather more important in differential expression analysis) [69]. However, concerning the two approaches used in this work, and looking at figure 5.13's second and third columns, *TMM* appears to impact negatively the algorithm's performance in relation to *deseq* in the case of discrete-based models. In the cases of other two types, it had no effect (case of *svm* models) or little positive effect (remaining models).

Data transformation on the other hand is considered to influence on classification results, by changing the distribution of data. Unfortunately, since there are no results available regarding *deseq-logcpm* combination, it is only possible to consider the influence of *vst* and *rlog* (Figures 5.13a and 5.13d). The latter did not seem to affect the *svm* models while it lead to a higher prediction accuracy in the *rf* and *NSC* models. Additionally, the transformation approach revealed to have a role on the number of variables selected, as it was previously observed [89]. In this study *vst* resulted in lower sparsity.



(a) *deseq-vst*, continuous-based classifiers

(b) *deseq*, discrete-based classifiers

(c) *deseq*, sparse classifiers

(d) *deseq-rlog*, continuous-based classifiers

(e) *tmm*, discrete-based classifiers

(f) *tmm*, sparse classifiers

(g) *deseq-logcpm*, continuous-based classifiers

(h) *deseq*, voom-based classifiers

(i) *deseq*, non-sparse classifiers

(j) *tmm-logcpm*, continuous-based classifiers

(k) *tmm*, voom-based classifiers

(l) *tmm*, non-sparse classifiers

Figure 5.13: Accumulated testing accuracy results for fitted classifiers. On the x-axis the classifiers are featured, whereas the y-axis indicates the added accuracy.

All sparse classifiers best performed when the data was normalized with *deseq*. Only *voomNSC* did not use all the features when given the elastic-net penalized models. The models obtained with *svm* outperformed the remaining, having *voom*-based classifiers and *rf* showed good results likewise.

Given the use of four sparse classification algorithms, it was compared which features were selected by each one to make the response predictions with the features given by the elastic net penalisation. This analyses concerned the fitting of the classifiers when it were the initial datasets given as input so that the fitting and consequent selection was accomplished from the raw data. More precisely, by applying *deseq* and *TMM* to the data, *NSC* (the transformation procedure used in this case was *vst* since it was the one with which best accuracy was achieved), *plda*, *plda2* and *voomNSC* algorithms performed feature selection. The common genes selected by these four sparse tools (further labelled as Z) were later compared to the BL and M03 models previously obtained. The resulting overlap showed no differences between comparing model 1 or model 2 of either the datasets. This observations allows to infer the diminutive relevance of the additional variables selected by the LOOCV approach. The analysis lead to the following findings:

- $BL_{models} \cap Z_{deseq}$: *SERINC2*, *CTSG*, *MPO* and *SERPINB10*;
- $BL_{models} \cap Z_{TMM}$: *RCAN3AS*, *SERINC2*, *EPHX4*, *SYT1*, *SKA3*, *CTSG*, *MPO*, *AZU1*, *ERICH6*, *IL2*, *SLC6A19*, *COBL* and *NTRK3*;
- $M03_{models} \cap Z_{TMM}$: *F3*.

The fact that the features selected by the sparse classifiers revealed genes selected by the initial implemented approach reinforces the first results. *MPO* and *CTSG* are relevant genes whose expression has an influence on the anti-TNF treatment response of each patient. It is then proposed that they may be of therapeutic value and represent important biomarkers which can be used in clinical practice. This analyses revealed an isolated gene in the M03 dataset which the LOOCV approach did not select but was present in the highest scored protein-protein interactions found through the BN learning: Coagulation Factor III, Tissue Factor or simply Tissue Factor (*F3*). It is an essential initiator of the extrinsic pathway of blood coagulation and it is also involved in the angiogenesis and the pannus formation of RA progression. In fact, it has been demonstrated that it is expressed not only in arthritic synovial tissue but also infiltrating macrophages, favoring extravascular coagulation and leading to inflammation in RA [90, 91].

To conclude, the sparse logistic approach used to obtain predictive models of anti-TNF treatment lead to the identification of genes consensually associated with therapy response, some known to be related with RA pathogenesis. The novel genes discovered are suggested to center further research regarding this subject. The BN learning analysis revealed protein-protein interactions both specific to the type of patients (responder and non-responder) and common to the whole group. The classification algorithms analysis allowed an heuristic evaluation of their performance predicting the treatment response, revealing several genes to be outcome predictors in accordance to the results from the sparse logistic regression methodology.

Table 5.12: Classifiers prediction performance for **BL** sub-datasets when using *deseq* method for regularization and *vst* method for transformation. Every performance value corresponds to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.632	0.526	0.474	0.526	0.474	0.526	0.579	1.000	1.000
	Sn	0.750	0.833	0.909	1.000	0.857	0.900	1.000	1.000	1.000
	Sp	0.429	0.167	0.000	0.000	0.000	0.000	0.000	1.000	1.000
rf	*Acc*	0.421	0.526	0.526	0.579	0.632	0.632	0.579	0.842	0.868
	Sn	0.417	0.778	0.571	0.700	0.818	0.636	0.636	1.000	1.000
	Sp	0.429	0.333	0.400	0.444	0.300	0.500	0.500	0.692	0.683
NSC	*Acc*	0.474	0.526	0.526	0.474	0.474	0.474	0.474	0.974	1.000
	Sn	0.667	0.917	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Sp	0.143	0.222	0.000	0.000	0.143	0.000	0.000	1.000	1.000
	Sparsity	0.000	0.800	0.300	0.933	0.650	0.200	0.367	1.000	1.000
plda	*Acc*	0.474	0.526	0.526	0.526	0.526	0.526	0.526	0.763	0.842
	Sn	0.417	0.400	0.417	0.500	0.444	0.444	0.429	0.707	0.809
	Sp	0.571	0.667	0.600	0.571	0.571	0.571	0.571	0.944	0.857
	Sparsity	1.000	1.000	0.900	0.600	0.500	0.400	0.233	1.000	1.000
plda2	*Acc*	0.526	0.526	0.579	0.579	0.632	0.579	0.579	0.895	0.895
	Sn	0.500	0.444	0.600	0.600	0.636	0.636	0.636	0.894	0.889
	Sp	0.571	0.556	0.545	0.571	0.600	0.556	0.545	0.894	0.866
	Sparsity	0.160	1.000	0.900	0.600	0.500	0.400	0.233	1.000	1.000
nblda	*Acc*	0.474	0.526	0.684	0.632	0.632	0.579	0.579	1.000	1.000
	Sn	0.417	0.400	0.571	0.615	0.571	0.636	0.667	1.000	1.000
	Sp	0.571	0.714	0.700	0.625	0.625	0.500	0.500	1.000	1.000
voomDLDA	*Acc*	0.579	0.526	0.632	0.579	0.526	0.526	0.526	0.947	0.947
	Sn	0.583	0.444	0.636	0.636	0.583	0.600	0.636	1.000	1.000
	Sp	0.571	0.636	0.600	0.500	0.500	0.455	0.444	1.000	0.955
voomNSC	*Acc*	0.474	0.526	0.526	0.526	0.526	0.526	0.526	0.921	0.895
	Sn	0.417	0.833	0.636	0.667	0.667	0.667	0.692	0.913	0.958
	Sp	0.571	0.333	0.400	0.444	0.429	0.429	0.375	0.888	0.903
	Sparsity	0.034	0.400	0.700	0.600	0.650	0.680	0.233	0.563	0.495

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.13: Classifiers prediction performance for **BL** sub-datasets when using *deseq* method for regularization and *rlog* method for transformation. Every performance value corresponds to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.368	0.474	0.474	0.579	0.526	0.579	0.526	1.000	1.000
	Sn	0.429	0.818	0.833	0.917	1.000	0.917	1.000	1.000	1.000
	Sp	0.200	0.143	0.143	0.143	0.000	0.000	0.111	1.000	1.000
rf	*Acc*	0.421	0.526	0.526	0.632	0.579	0.579	0.632	0.895	0.895
	Sn	0.429	0.667	0.636	0.750	0.625	0.667	0.692	1.000	1.000
	Sp	0.400	0.429	0.333	0.571	0.500	0.556	0.556	0.879	0.889
NSC	*Acc*	0.474	0.579	0.526	0.526	0.526	0.526	0.526	1.000	1.000
	Sn	0.500	0.917	1.000	0.778	0.778	0.786	0.818	1.000	1.000
	Sp	0.400	0.143	0.100	0.222	0.143	0.100	0.091	1.000	1.000
	Sparsity	0.016	1.000	0.600	0.800	0.800	0.440	0.300	1.000	1.000

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.14: Classifiers prediction performance for **BL** sub-datasets when using *deseq* method for regularization and *logcpm* method for transformation. Every performance value corresponds to a single repeat.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.368	0.579	0.474	0.526	0.579	0.579	0.632	1.000	0.947
	Sn	0.429	0.800	0.667	0.778	0.750	0.800	0.733	1.000	0.923
	Sp	0.200	0.333	0.143	0.300	0.286	0.333	0.250	1.000	1.000
rf	*Acc*	0.421	0.474	0.474	0.421	0.368	0.684	0.632	1.000	1.000
	Sn	0.429	0.500	0.500	0.444	0.417	0.700	0.600	1.000	1.000
	Sp	0.400	0.444	0.429	0.400	0.286	0.667	0.750	1.000	1.000
NSC	*Acc*	0.474	0.526	0.632	0.474	0.632	0.526	0.790	1.000	1.000
	Sn	0.500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Sp	0.400	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000
	Sparsity	0.012	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.15: Classifiers prediction performance for **BL** sub-datasets when using *TMM* method for regularization and *logcpm* method for transformation. Every performance value corresponds to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.368	0.526	0.579	0.526	0.526	0.526	0.526	1.000	1.000
	Sn	0.429	0.833	0.875	0.917	1.000	0.889	1.000	1.000	1.000
	Sp	0.200	0.100	0.286	0.000	0.000	0.000	0.000	1.000	1.000
rf	*Acc*	0.474	0.526	0.526	0.579	0.632	0.632	0.632	0.921	0.895
	Sn	0.500	0.636	0.727	0.700	0.750	0.700	0.750	1.000	1.000
	Sp	0.400	0.333	0.333	0.500	0.444	0.571	0.556	0.857	0.789
NSC	*Acc*	0.474	0.632	0.526	0.526	0.526	0.526	0.474	0.947	0.947
	Sn	0.500	0.923	1.000	0.833	0.778	0.818	1.000	1.000	1.000
	Sp	0.400	0.286	0.000	0.143	0.200	0.143	0.000	0.955	0.950
	Sparsity	0.032	0.800	0.200	0.533	0.400	0.320	0.200	1.000	1.000
plda	*Acc*	0.316	0.526	0.526	0.579	0.579	0.526	0.579	0.474	0.474
	Sn	0.286	0.833	0.917	1.000	1.000	1.000	1.000	0.442	0.551
	Sp	0.400	0.200	0.200	0.100	0.000	0.000	0.000	0.516	0.388
	Sparsity	1.000	0.800	0.700	0.467	1.000	0.160	0.133	1.000	1.000
plda2	*Acc*	0.474	0.526	0.474	0.474	0.474	0.474	0.474	0.526	0.474
	Sn	0.643	1.000	1.000	1.000	0.667	1.000	1.000	0.578	0.449
	Sp	0.000	0.000	0.000	0.000	0.143	0.000	0.000	0.429	0.500
	Sparsity	1.000	0.800	0.700	0.467	1.000	0.160	0.133	1.000	1.000
nblda	*Acc*	0.316	0.526	0.579	0.526	0.474	0.474	0.474	0.474	0.474
	Sn	0.286	0.667	0.667	0.444	0.625	0.625	0.500	0.442	0.551
	Sp	0.400	0.286	0.286	0.500	0.429	0.400	0.429	0.431	0.369
voomDLDA	*Acc*	0.368	0.579	0.579	0.632	0.526	0.526	0.526	0.947	0.947
	Sn	0.429	0.538	0.571	0.667	0.583	0.615	0.571	1.000	1.000
	Sp	0.200	0.545	0.571	0.571	0.500	0.444	0.444	1.000	0.955
voomNSC	*Acc*	0.474	0.579	0.526	0.632	0.579	0.579	0.526	0.895	0.921
	Sn	0.500	0.833	0.750	0.667	0.667	0.714	0.667	0.894	0.920
	Sp	0.400	0.300	0.400	0.571	0.444	0.400	0.400	1.000	0.905
	Sparsity	0.015	0.600	0.500	0.533	0.550	0.240	0.333	0.873	0.694

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.16: Classifiers prediction performance for **M03** sub-datasets when using *deseq* method for regularization and *vst* method for transformation. Every performance value corresponds to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.250	0.450	0.500	0.600	0.550	0.550	0.500	1.000	1.000
	Sn	0.000	0.778	0.875	0.818	0.833	0.917	0.846	1.000	1.000
	Sp	1.000	0.111	0.250	0.333	0.250	0.091	0.091	1.000	1.000
rf	*Acc*	0.600	0.550	0.400	0.450	0.450	0.500	0.450	0.700	0.725
	Sn	0.533	0.615	0.455	0.462	0.615	0.500	0.500	0.923	0.846
	Sp	0.800	0.429	0.375	0.333	0.429	0.375	0.375	0.389	0.597
NSC	*Acc*	0.450	0.550	0.550	0.500	0.550	0.550	0.550	0.925	0.938
	Sn	0.400	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Sp	0.600	0.000	0.000	0.000	0.000	0.000	0.000	0.806	0.819
	Sparsity	0.034	0.800	0.200	0.267	0.050	0.080	0.067	1.000	1.000
plda	*Acc*	0.550	0.450	0.500	0.500	0.550	0.550	0.550	0.750	0.725
	Sn	0.467	0.429	0.462	0.455	0.500	0.500	0.455	0.818	0.818
	Sp	0.800	0.556	0.667	0.667	0.625	0.667	0.667	0.690	0.667
	Sparsity	0.000	0.600	0.400	0.467	0.400	0.760	0.167	1.000	1.000
plda2	*Acc*	0.500	0.450	0.450	0.500	0.500	0.450	0.500	0.750	0.750
	Sn	0.467	0.455	0.455	0.462	0.455	0.462	0.462	0.794	0.772
	Sp	0.600	0.571	0.444	0.571	0.556	0.500	0.556	0.732	0.723
	Sparsity	1.000	0.800	0.600	0.667	0.400	0.640	0.267	1.000	1.000
nblda	*Acc*	0.400	0.500	0.450	0.500	0.450	0.500	0.550	1.000	0.975
	Sn	0.400	0.462	0.364	0.455	0.444	0.417	0.500	1.000	1.000
	Sp	0.400	0.667	0.571	0.571	0.571	0.556	0.500	1.000	0.944
voomDLDA	*Acc*	0.600	0.500	0.500	0.450	0.500	0.450	0.450	0.925	0.900
	Sn	0.533	0.500	0.500	0.462	0.455	0.417	0.364	0.962	0.916
	Sp	0.800	0.625	0.429	0.500	0.500	0.500	0.500	0.889	0.889
voomNSC	*Acc*	0.500	0.500	0.500	0.500	0.450	0.450	0.450	0.925	0.913
	Sn	0.467	0.636	0.636	0.636	0.556	0.538	0.500	1.000	1.000
	Sp	0.600	0.375	0.286	0.273	0.333	0.333	0.333	1.000	1.000
	Sparsity	0.003	0.400	0.500	0.267	0.350	0.280	0.267	0.693	0.670

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.17: Classifiers prediction performance for **M03** sub-datasets when using *deseq* method for regularization and *rlog* method for transformation. Every performance value corresponds to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.500	0.550	0.550	0.550	0.500	0.500	0.500	1.000	1.000
	Sn	1.000	0.786	0.769	0.750	0.727	0.833	0.778	1.000	1.000
	Sp	0.000	0.250	0.222	0.222	0.125	0.182	0.222	1.000	1.000
rf	*Acc*	0.550	0.500	0.500	0.450	0.450	0.450	0.450	0.900	0.900
	Sn	0.800	0.556	0.455	0.462	0.692	0.462	0.500	0.962	0.981
	Sp	0.300	0.375	0.375	0.333	0.333	0.375	0.333	0.857	0.857
NSC	*Acc*	0.350	0.550	0.500	0.550	0.500	0.550	0.500	0.950	0.950
	Sn	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Sp	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.903	0.910
	Sparsity	0.870	0.800	0.200	0.133	0.100	0.040	0.133	1.000	1.000

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.18: Classifiers prediction performance for **M03** sub-datasets when using *deseq* method for regularization and *logcpm* method for transformation. Every performance value corresponds to a single repeat.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.500	0.650	0.650	0.450	0.400	0.600	0.500	1.000	1.000
	Sn	1.000	0.917	0.917	0.875	0.857	0.909	0.818	1.000	1.000
	Sp	0.000	0.250	0.250	0.167	0.154	0.222	0.111	1.000	1.000
rf	*Acc*	0.550	0.650	0.650	0.550	0.400	0.700	0.500	0.700	0.900
	Sn	0.800	0.917	0.917	1.000	0.857	1.000	0.818	1.000	0.929
	Sp	0.300	0.250	0.250	0.250	0.154	0.333	0.111	0.000	0.833
NSC	*Acc*	0.350	0.600	0.600	0.400	0.350	0.550	0.550	1.000	0.950
	Sn	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Sp	0.500	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.833
	Sparsity	0.882	0.034	0.034	0.034	0.034	0.034	0.034	1.000	1.000

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Table 5.19: Classifiers prediction performance for **M03** sub-datasets when using *TMM* method for regularization and *logcpm* method for transformation. Every performance values correspond to the median over 16 repeats except the first column ones.

Classifier		All Genes	#5	#10	#15	#20	#25	#30	Model 1	Model 2
svm	*Acc*	0.500	0.500	0.500	0.550	0.550	0.500	0.450	1.000	1.000
	Sn	1.000	0.909	0.818	0.750	0.818	0.909	0.778	1.000	1.000
	Sp	0.000	0.111	0.250	0.182	0.182	0.143	0.125	1.000	1.000
rf	*Acc*	0.550	0.450	0.400	0.450	0.450	0.500	0.450	0.825	0.850
	Sn	0.800	0.545	0.462	0.545	0.636	0.500	0.583	1.000	1.000
	Sp	0.300	0.333	0.500	0.333	0.333	0.333	0.333	0.732	0.764
NSC	*Acc*	0.350	0.500	0.500	0.550	0.550	0.500	0.550	0.875	0.950
	Sn	0.200	1.000	1.000	1.000	1.000	1.000	1.000	0.909	1.000
	Sp	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.866	0.889
	Sparsity	0.886	0.600	0.300	0.133	0.100	0.040	0.067	1.000	1.000
plda	*Acc*	0.350	0.550	0.550	0.450	0.450	0.450	0.450	0.575	0.550
	Sn	0.000	0.833	0.875	0.643	0.538	0.429	0.417	0.523	0.583
	Sp	0.700	0.250	0.250	0.222	0.429	0.500	0.625	0.697	0.352
	Sparsity	1.000	0.200	0.100	0.067	0.050	0.040	0.033	1.000	1.000
plda2	*Acc*	0.400	0.550	0.550	0.500	0.550	0.550	0.550	0.550	0.525
	Sn	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.477	0.458
	Sp	0.800	0.000	0.000	0.000	0.000	0.000	0.000	0.670	0.604
	Sparsity	1.000	0.800	0.500	0.333	0.150	0.120	0.100	1.000	1.000
nblda	*Acc*	0.550	0.500	0.450	0.450	0.550	0.550	0.571	0.600	0.550
	Sn	0.700	0.333	0.444	0.538	0.625	0.500	0.550	0.523	0.840
	Sp	0.400	0.667	0.429	0.500	0.500	0.556	0.500	0.697	0.056
voomDLDA	*Acc*	0.300	0.500	0.450	0.450	0.450	0.450	0.450	0.850	0.950
	Sn	0.000	0.417	0.462	0.462	0.444	0.444	0.385	0.899	1.000
	Sp	0.600	0.571	0.500	0.444	0.556	0.500	0.429	0.882	0.889
voomNSC	*Acc*	0.250	0.500	0.500	0.500	0.450	0.400	0.450	0.900	0.950
	Sn	0.100	0.583	0.750	0.583	0.556	0.545	0.583	0.873	1.000
	Sp	0.400	0.375	0.167	0.429	0.250	0.286	0.333	0.958	0.889
	Sparsity	0.001	0.400	0.400	0.667	0.350	0.160	0.267	0.934	0.791

Acc: Accuracy; Sn: Sensitivity; Sp: Specificity.

Chapter 6

Conclusions

Over the last decade, several gene expression signatures associating with response to anti-TNF have been identified, but few replicated. A multiplicity of reasons may explain the inconsistencies. For example, the study design used (heterogeneous and small cohorts, different disease stages or time points considered and the analysis of distinct tissues or even cell types) or the technical/analytic approaches (different transcriptomic platforms, different computational analysis methods). Indeed, important cell-type specificities have been reported but can also be missed when whole tissue (such as blood or synovium) is tested. Nonetheless, transcriptomics has tremendous potential in the field of precision medicine.

This thesis' main goal was to identify biomarkers able to predict anti-TNF rheumatoid arthritis treatment response. The starting point was two publicly available datasets containing transcriptomic data, one from patients initiating that treatment and another from patients three-month into the treatment. A sparse logistic regression approach was used where elastic net regularization permitted a selection of the relevant features. Predictive models were achieved and considered to be reliable, having the models at M03 achieved a better prediction performance. Comparing the results in regard to the two time-points, the obtained genes which potentially may be able to predict the response were not the same. These changes in expression profile are consistent with a decrease in blood neutrophil counts and associated biology [17].

The protein-protein interactions found through the Bayesian network learning from the same transcriptomics datasets were later validated by the STRING database, revealing *CTSG – MPO* and *CTSG – AZU1* to be relevant at baseline for the prediction of, respectively, "responders" and "non-responders" patients. The results also suggested that interactions *MPO – AZU1* (prior to treatment initiation) and *CTSG – ELANE* (three month into treatment) were influential for both types of treatment response. The fact that these findings are in line with the known role of the proteins encoded by those genes argues that the created pipeline was favorable.

Regarding the second half of this study, several machine learning algorithms were applied to the data and different angles evaluated in order to expand the suite of the classification algorithms tested. The normalization methods applied to the data prior to classification revealed to have a smaller role in the model's accuracy then the transformation methods. The overall best performances were achieved by *svm*, *rf*, *voomDLDA* and *voomNSC* (being the latter the only one which performs feature selection). The four sparse classification approaches used selected from the raw data a number of genes in common with the ones obtained with the elastic net, such as *MPO*, *CTSG*, *AZU1* and *RCAN3AS*. Hence, the present study has led to the identification of all these genes whose expression is suggestive of being helpful for predicting response to anti-TNF therapy.

It is very encouraging that the second component of this thesis confirmed in part the conclusions taken from the sparse logistic regression and subsequent Bayesian network learning. Even though the eight machine learning algorithms were not further deeply investigated, evidence was shown that they have the potential to be used as basis to similar studies and thus should be explored in the context of RNA-Seq and RA treatment response.

The methodology created for obtaining the RA treatment response predictive models revealed to be satisfactory and thus it is suggested that it could be used to replicate the results with a dataset comprising a higher number of observations. In fact, that is one of the limitations of this work, since the number of patients used in each dataset was rather small. Alternative sparse tools should be considered, since the sparse logistic regression approach revealed a considerable variability in the results, and thus a stronger classifier is suggested, such as *voomDLDA* or *voomNSC*, which are two recent and promising tools specially devoted to the RNA-Seq field. Besides, individual gene signatures should be studied in order to validate the biomarkers and use them later in clinical practice. These biomarkers could be an important factor in modulating the response to anti-TNF or other biologic treatments and ultimately yield better treatment assignments to patients. In future, other factors besides transcriptomic data could be taken in regard, for example age, sex, disease duration and complete molecular profiling of plasma.

This is an exciting time for RA as the growth of big data in clinical research and advancements in computational approaches have opened up new avenues to study complex diseases. Hopefully in a near future the increasing efforts to support medical informatics standards and the enrichment of cohesive genome-wide transcriptional profiling for RA databases will result in more accurate and innovative insights and revolutionize RA healthcare.

Bibliography

[1] D. Aletaha, T. Neogi, A. J. Silman, J. Funovits, D. T. Felson, C. O. B. III, N. S. Birnbaum, G. R. Burmester, V. P. Bykerk, M. D. Cohen, B. Combe, K. H. Costenbader, M. Dougados, P. Emery, G. Ferraccioli, J. M. W. Hazes, K. Hobbs, T. W. J. Huizinga, A. Kavanaugh, J. Kay, T. K. Kvien, T. Laing, P. Mease, H. A. Ménard, L. W. Moreland, R. L. Naden, T. Pincus, J. S. Smolen, E. Stanislawska-Biernat, D. Symmons, P. P. Tak, K. S. Upchurch, J. Vencovský, F. Wolfe, and G. Hawker. 2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative. *ARTHRITIS RHEUMATISM*, 62 (9):2569–2581, 2010.

[2] S. P. de Reumatologia. Artrite reumatóide. URL `https://spreumatologia.pt/artrite-reumatoide/`. Acessed: 12 September 2020.

[3] D. A. Pappas, J. M. Kremer, G. Reed, J. D. Greenberg, and J. R. Curtis. Design characteristics of the corrona certain study: a comparative effectiveness study of biologic agents for rheumatoid arthritis patients. *BMC Musculoskeletal Disorders*, 15(1):113, 2014.

[4] J. S. Smolen, R. Landewé, J. W. Bijlsma, G. R. Burmester, M. Dougados, A. Kerschbaumer, I. B. McInnes, A. Sepriano, R. F. Van Vollenhoven, M. De Wit, et al. Eular recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. *Annals of the rheumatic diseases*, 70:685–699, 2020.

[5] C. Monaco, J. Nanchahal, P. Taylor, and M. Feldmann. Anti-tnf therapy: past, present and future. *International immunology*, 27(1):55–62, 2015.

[6] V. Maini and S. Sabri. *Machine Learning for Humans*. 2017.

[7] B. Aragam, J. Gu, and Q. Zhou. Learning large-scale bayesian networks with the sparsebn package. *Journal of Statistical Software*, 91(11):1–38, 2019.

[8] N. Friedman and D. Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50:95–125, 2003.

[9] O. Morozova, M. Hirst, and M. A. Marra. Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics*, 10:135–151, 2009.

[10] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[11] Y. Zheng. *Computational Non-Coding RNA Biology.* Academic Press, 1st edition, 2018.

[12] F. Vitali, Q. Li, A. G. Schissler, J. Berghout, C. Kenost, and Y. A. Lussier. Developing a 'personalome' for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Briefings in bioinformatics*, 20(3):789–805, 2019.

[13] K. Strimbu and J. A. Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

[14] B. V. J. Cuppen. Prediction of response to therapy in rheumatoid arthritis : Lost in validation. Master's thesis, Utrecht University, 2017.

[15] G. N. Goulielmos, M. I. Zervou, E. Myrthianou, A. Burska, T. B. Niewold, and F. Ponchel. Genetic data: The new challenge of personalized medicine, insights for rheumatoid arthritis patients. *Gene*, (583):90–101, 2016.

[16] C. A. Wijbrandts and P. P. Tak. Prediction of response to targeted treatment in rheumatoid arthritis. *Mayo Clinic Proceedings*, 2(7):1129–1143, 2017.

[17] V. Farutin, T. Prod'homme, K. McConnell, N. Washburn, P. Halvey, C. J. Etzel, J. Guess, J. Duffner, K. Getchell, R. Meccariello, B. Gutierrez, C. Honan, G. Zhao, N. A. Cilfone, N. S. Gunay, J. L. Hillson, D. S. DeLuca, K. C. Saunders, D. A. Pappas, J. D. Greenberg, J. M. Kremer, A. M. Manning, L. E. Ling, and I. Capila. Molecular profiling of rheumatoid arthritis patients reveals an association between innate and adaptive cell populations and response to anti-tumor necrosis factor. *Arthritis Research & Therapy*, 21(216):1–14, 2019.

[18] N. Yoosuf, M. Maciejewski, D. Ziemek, S. Jelinsky, L. Folkersen, M. Müller, P. Sahlström, N. Vivar, A. Catrina, L. Berg, L. Klareskog, L. Padyukov, and B. Brynedal. Molecular biomarkers of anti-tnf response in patients with rheumatoid arthritis. 2020.

[19] K.-J. Kim, M. Kim, I. E. Adamopoulos, and I. Tagkopoulos. Compendium of synovial signatures identifies pathologic characteristics for predicting treatment response in rheumatoid arthritis patients. *Clinical Immunology*, 202:1–10, 2019.

[20] Y. Guan, H. Zhang, D. Quang, Z. Wang, S. C. Parker, D. A. Pappas, J. M. Kremer, and F. Zhu. Machine learning to predict anti–tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. *Arthritis & Rheumatology*, 71(12):1987–1996, 2019.

[21] V. C. Romão, E. M. Vital, J. E. Fonseca, and M. H. Buch. Right drug, right patient, right time: aspiration or future promise for biologics in rheumatoid arthritis? *Arthritis Research Therapy*, 19 (239):90–101, 2017.

[22] S. Bek, A. Bojesen, J. Nielsen, J. Sode, S. Bank, U. Vogel, and V. Anderson. Systematic review and meta-analysis: pharmacogenetics of anti-tnf treatment response in rheumatoid arthritis. *The Pharmacogenomics Journal*, 17:403–411, 2017.

[23] G. S. Firestein, R. C. Budd, S. E. Gabriel, I. B. Mcinnes, and J. R. O'Dell. Kelly's textbook of rheumatology. In *Rheumatoid Arthritis*, volume II, pages 1059–1068. Saunders, 2013.

[24] A. J. Silman and J. E. Pearson. Epidemiology and genetics of rheumatoid arthritis. *Arthritis Research Therapy*, 4, 2002.

[25] S. Sakaguchi. Naturally arising cd4+ regulatory t cells for immunologic self-tolerance and negative control of immune responses. *Annual Review of Immunology*, 22(1):531–562, 2004.

[26] G. S. Firestein. Evolving concepts of rheumatoid arthritis. *Nature*, 423:356–361, 2003.

[27] L. C. Parish. An historical approach to the nomenclature of rheumatoid arthritis. *Arthritis and Rheumatism*, 6(2):138–158, 1963.

[28] D. G. de Saúde. Programa nacional contra as doenças reumáticas, 2004. URL `http://www.dgs.pt/upload/membro.id/ficheiros/i006345.pdf`. Acessed: 11 September 2020.

[29] D. M. Geriag, k. Raza, L. G. M. van Baarsen, E. Brouwer, C. D. Buckley, G. R. Burmester, C. Gabay, A. I. Catrina, A. P. Cope, F. Cornelis, S. R. Dahlqvist, P. Emery, S. Eyre, A. Finckh, S. Gay, J. H. Hazes, A. H. Mil, T. W. J. Huizinga, L. Klareskog, T. K. Kvien, C. Lewis, K. P. Machold, J. Rönnelid, D. van Schaardenburg, G. Schett, J. S. Smolen, S. Thomas, J. Worthington, and P. P. Tak. Eular recommendations for terminology and research in individuals at risk of rheumatoid arthritis: report from the study group for risk factors for rheumatoid arthritis. *Ann Rheum Dis 2012*, 71:638–641, 2012.

[30] C. Scheinecker. The role of t cells in rheumatoid arthritis. In *Rheumatoid Arthritis*, pages 91–96. Elsevier, 2009.

[31] L. Klimek, K. Bergmann, T. Biedermann, J. Bousquet, P. Hellings, K. Jung, H. Merk, H. Olze, W. Schlenter, P. Stock, J. Ring, M. Wagenmann, W. Wehrmann, R. Mösges, and O. Pfaar. Visual analogue scales (vas): Measuring instruments for the documentation of symptoms and therapy monitoring in cases of allergic rhinitis in everyday health care. *Allergo Journal International*, 26(1): 16–24, 2017.

[32] J. K. Anderson, L. Zimmerman, L. Caplan, and K. Michaud. Measures of rheumatoid arthritis disease activity: patient (ptga) and provider (prga) global assessment of disease activity, disease activity score (das) and disease activity score with 28-joint counts (das28), simplified disease activity index (sdai), clinical disease activity index (cdai), patient activity score (pas) and patient activity score-ii (pasii), routine assessment of patient index data (rapid), rheumatoid arthritis disease activity index (radai) and rheumatoid arthritis disease activity index-5 (radai-5), chronic arthritis systemic index (casi), patient-based disease activity score with esr (pdas1) and patient-based disease activity score

without esr (pdas2), and mean overall index for rheumatoid arthritis (moi-ra). *Arthritis care & research*, 63(S11):S14–S36, 2011.

[33] J. F. Fries, P. Spitz, R. G. Kraines, and H. R. Holman. Measurement of patient outcome in arthritis. *Arthritis and Rheumatism*, 23(2):137–145, 1980.

[34] B. Bruce and J. F. Fries. The health assessment questionnaire (haq). *Clinical and Experimental Rheumatology*, 23:S14–S18, 2005.

[35] H. A. Fuchs, R. H. Brooks, L. F. Callahan, and T. Pincus. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. *Arthritis and Rheumatism*, 32(5):531–537, 1989.

[36] J. S. Smolen, F. C. Breedveld, M. H. Schiff, J. R. Kalden, P. Emery, G. Eberl, P. L. van Riel, and P. Tugwell. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology*, 42(2):244–257, 2003.

[37] H. Singh, H. Kumar, R. Handa, P. Talapatra, S. Ray, and V. Gupta. Use of clinical disease activity index score for assessment of disease activity in rheumatoid arthritis patients: An indian experience. *Arthritis*, 2011:1–5, 2011.

[38] J. Fransen and P. L. C. M. van Riel. The disease activity score and the eular response criteria. *Clinical and experimental rheumatology*, (23):S93–S99, 2005.

[39] D. Felson. Defining remission in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 71:i86–i88, 2012.

[40] R. M. Shammas, V. K. Ranganath, and H. E. Paulus. Remission in rheumatoid arthritis. *Curr Rheumatol Rep.*, 12:355–362, 2010.

[41] eular. Defining remission in rheumatoid arthritis. URL `https://www.eular.org/myUploadData/files/RA%20Remission%20Slides-Web.pdf`. Accessed: 13 October 2020.

[42] H. Radner and D. Aletaha. Anti-tnf in rheumatoid arthritis: an overview. *Wiener Medizinische Wochenschrift*, 165:3–9, 2015.

[43] F. S. Paula and J. D. Alves. Non-tumor necrosis factor-based biologic therapies for rheumatoid arthritis: present, future, and insights into pathogenesis. *Biologics: Targets and Therapy*, 8:1–12, 2013.

[44] O. Benjamin, P. Bansal, A. Goyal, and S. L. Lappin. Disease modifying anti-rheumatic drugs (dmard). *StatPearls [Internet]y*, 2020.

[45] N. Farm. Sociedade portuguesa de reumatologia alerta para impacto social e económico da doença, 2019. URL `https://www.newsfarma.pt/noticias/8424-sociedade-portuguesa-de-reumatologia-alerta-para-impacto-social-e-econ%C3%B3mico-da-doen%C3%A7a.html`. Acessed: 30 October 2020.

[46] L. C. Miranda, H. Santos, J. Ferreira, P. Coelho, C. Silva, and J. saraiva Ribeiro. Finding rheumatoid arthritis impact on life (frail study): economic burden. *Acta Reumatológica Portuguesa*, 37(2):134–142, 2012.

[47] P. A. Laires and M. Gouveia. Association of rheumatic diseases with early exit from paid employment in portugal. *Rheumatology international*, 34(4):491–502, 2014.

[48] P. McCullagh and J. Nelder. *Generalized Linear Models.* Champman and Hall, 2nd edition, 1989.

[49] Y. Pua, H. Kang, J. Thumboo, R. A. Clark, E. S. Chew, C. L. Poon, H. Chong, and S. Yeo. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy*, 2019.

[50] J. C. Stoltzfus. Logistic regression: A brief primer. *Academic Emergency Medicin*, 18:1099–1104, 2011.

[51] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[52] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[53] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Champman and Hall, 2nd edition, 2015.

[54] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320, 2005.

[55] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, 1st edition, 2013.

[56] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[57] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*, volume 43. 2001.

[58] D. Goksuluk, G. Zararsiz, S. Korkmaz, V. Eldem, G. E. Zararsiz, E. Ozcetin, A. Ozturk, and A. E. Karaagaoglu. Mlseq: Machine learning interface for rna-sequencing data. *Computer Methods and Programs in Biomedicine*, 175, 2019.

[59] C. Evans, J. Hardin, and D. M. Stoebel. Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5):776–792, 2018.

[60] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(R106), 2010.

[61] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

[62] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), 2014.

[63] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(R29), 2014.

[64] S. Huang, N. Cai, P. P. Pacheco, S. Narandes, Y. W, and W. Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics Proteomics*, 15(1):41–51, 2018.

[65] W. S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[66] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[67] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[68] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

[69] D. M. Witten. Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, 5(5):2493–2518, 2011.

[70] K. Dong, H. Zhao, T. Tong, and X. Wan. Nblda: negative binomial linear discriminant analysis for rna-seq data. *BMC Bioinformatics*, 17(369), 2016.

[71] G. Zararsiz, D. Goksuluk, B. Klaus, S. Korkmaz, V. Eldem, E. Karabulut, and A. Ozturk. voomdda: discovery of diagnostic biomarkers and classification of rna-seq data. *PeerJ*, 5:e3890, 2017.

[72] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. v. Mering. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1): D447–D452, 2015.

[73] J. V. Brito. Unravelling breast and prostate common gene signatures by bayesian network learning. Master's thesis, Instituto Superior Técnico, 2018.

[74] C. Constantino, A. M. Carvalho, and S. Vinga. Sparse consensus classification for discovering novel biomarkers in rheumatoid arthritis. *Proceedings of the 6th International Conference on on Machine Learning, Optimization, and Data Science (LOD'20)*, volume 12514 of Lecture Notes in Computer Science, 2020 (accepted).

[75] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[76] R. M. Fernandes, N. P. da Silva, and E. I. Sato. Increased myeloperoxidase plasma levels in rheumatoid arthritis. *Rheumatology International*, 32(6):1605–1609, 2012.

[77] S. Gao, H. Zhu, X. Zuo, and H. Luo. Cathepsin g and its role in inflammation and autoimmune diseases. *Archives of Rheumatologyl*, 33(4):498–504, 2018.

[78] D. Trzybulska, A. Olewicz-Gawlik, K. Graniczna, K. Kisiel, M. Moskal, D. Cieślak, J. Sikora, and P. Hrycaj. Quantitative analysis of elastase and cathepsin g mrna levels in peripheral blood cd14(+) cells from patients with rheumatoid arthritis. *Cellular immunology*, 292(1-2):40–44, 2014.

[79] H. L. Wright, T. Cox, r. J. Moots, and S. W. Edwards. Neutrophil biomarkers predict response to therapy with tumor necrosis factor inhibitors in rheumatoid arthritis. *Journal of Leukocyte Biology*, 101(3):785–795, 2016.

[80] J.-S. Park, J.-H. Jeong, J.-K. Byun, M.-A. Lim, E.-K. Kim, S.-M. Kim, S.-Y. Choi, S.-H. Park, J.-K. Min, and M.-L. Cho. Regulator of calcineurin 3 ameliorates autoimmune arthritis by suppressing th17 cell differentiation. *The American Journal of Pathology*, 187(9):2034–2045, 2017.

[81] J.-z. Xu, J.-l. Zhang, and W.-g. Zhang. Antisense rna: the new favorite in genetic research. *Journal of Zhejiang University-SCIENCE B*, 19(10):739–749, 2018.

[82] L. Crocetti, M. T. Quinn, I. A. Schepetkin, and M. P. Giovannoni. A patenting perspective on human neutrophil elastase (hne) inhibitors (2014-2018) and their therapeutic applications. *Expert opinion on therapeutic patents*, 29(7):555—578, 2019.

[83] G. Celebi, H. Kesim, E. Ozer, and O. Kutlu. The effect of dysfunctional ubiquitin enzymes in the pathogenesis of most common diseases. *International Journal of Molecular Sciences*, 21(17):6335, 2020.

[84] L. Stangenberg, D. Burzyn, B. A. Binstadt, R. Weissleder, U. Mahmood, C. Benoist, and D. Mathis. Denervation protects limbs from inflammatory arthritis via an impact on the microvasculature. *Proceedings of the National Academy of Sciences*, 111(31):11419–11424, 2014.

[85] M. Leibovici, E. Verpy, R. J. Goodyear, I. Zwaenepoel, S. Blanchard, S. Lainé, G. P. Richardson, and C. Petit. Initial characterization of kinocilin, a protein of the hair cell kinocilium. *Hearing research*, 203(1-2):144–153, 2005.

[86] M. Monji, T. Nakatsura, S. Senju, Y. Yoshitake, M. Sawatsubashi, M. Shinohara, T. Kageshita, T. Ono, A. Inokuchi, and Y. Nishimura. Identification of a novel human cancer/testis antigen, km-hn-1, recognized by cellular and humoral immune responses. *Clinical Cancer Research*, 10(18 Pt 1): 6047–6057, 2004.

[87] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016, 2016.

[88] R. R. Schleef and T. L. Chuang. Protease inhibitor 10 inhibits tumor necrosis factor α-induced cell death: Evidence for the formation of intracellular highm r protease inhibitor 10-containing complexes. *Journal of Biological Chemistry*, 275(34):26385–26389, 2000.

[89] I. Zwiener, B. Frisch, and H. Binder. Transforming rna-seq data to improve the performance of prognostic gene signatures. *PloS one*, 9(1):e85150, 2014.

[90] N. Busso, C. Morard, R. Salvi, V. Péclat, and A. So. Role of the tissue factor pathway in synovial inflammation. *Arthritis & Rheumatism*, 48(3):651–659, 2003.

[91] L. Chen, Y. Lu, Y. Chu, J. Xie, F. Wang, et al. Tissue factor expression in rheumatoid synovium: a potential role in pannus invasion of rheumatoid arthritis. *Acta histochemica*, 115(7):692–697, 2013.

[92] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.

Appendix A

Complete names of referenced genes

Table A.1: List of gene names repeatedly found over all predictive models with Leave-on-Out Cross-Validation (referred in table 5.6) [92].

BL		**M03**	
Gene symbol	Gene name	Gene Symbol	Gene name
ALOX12B	Arachidonate 12-lipoxygenase	ADAM33	ADAM metallopeptidase domain 33
CAPN11	Calpain 11	CCDC110	coiled-coil domain containing 110
CAPNS2	Calpain small subunit	ELANE	Neutrophil elastase
CCDC108	Cilia- and flagella-associated protein 65	FBLIM1	filamin binding LIM protein 1
CTSG	Cathepsin G	GFAP	glial fibrillary acidic protein
EPHX4	Epoxide hydrolase 4	HYAL4	hyaluronidase 4
ERICH6	Glutamate Rich 6	KCNJ8	potassium inwardly rectifying channel subfamily J member 8
EVPLL	Envoplakin Like	KCNK4	potassium two pore domain channel subfamily K member 4
FAM133CP	family with sequence similarity 133 member C, pseudogene	KNCN	kinocilin
FOXD4L3	forkhead box D4 like 3	LOC100128076	Protein Tyrosine Phosphatase Pseudogene
HIST1H3J	H3 clustered histone 12	LOC101928222	*Uncharacterized* LOC101928222
IGF2BP1	insulin like growth factor 2 mRNA binding protein 1	LRRN4CL	LRRN4 C-terminal like
KCNH4	Potassium voltage-gated channel subfamily H member 4	MTRNR2L3	MT-RNR2 like 3
LINC00696	Long intergenic non-protein coding RNA 696	SERTM1	serine rich and transmembrane domain containing 1
LMOD3	Leiomodin 3	TMEM105	TMEM105 long non-coding RNA
LOC339975	*Uncharacterized* LOC339975	TPBG	trophoblast glycoprotein
LRGUK	Leucine-rich repeat and guanylate kinase domain containing	TRIM7	tripartite motif containing 7
MAG	myelin associated glycoprotein	TTC25	outer dynein arm docking complex subunit 4
MAGEC2	MAGE family member C2	UBE2QL1	ubiquitin conjugating enzyme E2 Q family like 1
MIR941-4	microRNA 941-4	VSTM2L	V-set and transmembrane domain containing 2 like
MPO	Myeloperoxidase	ZNF843	zinc finger protein 843
NUAK1	NUAK family kinase 1		
PMS2L2	PMS1 homolog 2, mismatch repair system component pseudogene 2		
PRKG1	Protein kinase cGMP-dependent 1		
PRSS30P	serine protease 30, pseudogene		
RAD21L1	RAD21 cohesin complex component like 1		
RCAN3AS	RCAN3 Antisense RNA		
RNU6-28P	RNA, U6 small nuclear 28, pseudogene		
ROPN1L-AS1	ROPN1L antisense RNA 1		
SKA3	Spindle and kinetochore-associated protein 3		
SLC6A19	Solute carrier family 6 member 19		
SYT1	Synaptotagmin-1		
TBX2	T-box transcription factor 2		
TGFB2	Transforming growth factor beta 2		

Table A.2: List of gene names belonging to the highest edges found in each BN [92].

BL		**M03**	
Gene symbol	Gene name	Gene Symbol	Gene name
ADAMTS9	ADAM Metallopeptidase With Thrombospondin Type 1 Motif 9	CCDC110	coiled-coil domain containing 110
BATF2	Basic Leucine Zipper ATF-Like Transcription Factor 2	CES1P1	Carboxylesterase 1 Pseudogene 1
CAPN11	Calpain 11	CTSG	Cathepsin G
CDC42EP4	CDC42 Effector Protein 4	C8B	Complement C8 Beta Chain
CYGB	Cytoglobin	ELANE	Neutrophil elastase
C1orf95	Stum, Mechanosensory Transduction Mediator Homolog	FBLIM1	Filamin Binding LIM Protein 1
DRD2	Dopamine Receptor D2	FSD2	Fibronectin Type III And SPRY Domain Containing 2
EPHX4	Epoxide hydrolase 4	F3	Coagulation Factor III, Tissue Factor
ERICH6	Glutamate Rich 6	HIST1H2AJ	H2A Clustered Histone 14
EVPLL	Envoplakin Like	KCNK4	potassium two pore domain channel subfamily K member 4
FGD5P1	FYVE, RhoGEF And PH Domain Containing 5 Pseudogene 1	KNCN	Kinocilin
IGF2BP1	Insulin Like Growth Factor 2 MRNA Binding Protein 1	LINC01361	Long Intergenic Non-Protein Coding RNA 1361
KCNH4	potassium voltage-gated channel subfamily H member 4	LOC100506071	*Uncharacterized*
LILRB4	Leukocyte Immunoglobulin Like Receptor B4	LOC101927468	*Uncharacterized*
LINC00696	*Uncharacterized*	LOC102467224	*Uncharacterized*
LOC339975	*Uncharacterized*	MIR3918	MicroRNA 3918
LRGUK	Leucine-rich repeat and guanylate kinase domain containing	MIR4271	MicroRNA 4271
MAG	myelin associated glycoprotein	MIR718	MicroRNA 718
MAGEC2	MAGE family member C2	MTRNR2L3	MT-RNR2 like 3
MIR941-2	MicroRNA 941-2	RSPH10B	Radial Spoke Head 10 Homolog B
MIR941-1	MicroRNA 941-4	RSPH10B2	Radial Spoke Head 10 Homolog B2
RCAN3AS	RCAN3 Antisense RNA	RS1	Retinoschisin 1
SLC25A52	Solute Carrier Family 25 Member 52	TMEM51-AS1	Transmembrane Protein 51 Antisense RNA1
SVLF1	Sulfatase 1	TRIM7	tripartite motif containing 7
TBX2	T-box transcription factor 2	UBE2QL1	ubiquitin conjugating enzyme E2 Q family like 1
TCN2	Transcobalamin 2	VWA1	Von Willebrand Factor A Domain Containing 1
		ZNF843	zinc finger protein 843