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

Joana Alter Palhinha
joana.palhinha@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2020

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.

1. Introduction

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 [1; 2].

Biologic agents efficacy and safety have been clearly demonstrated, having revolutionized the RA treatment over the last decades [3]. However the patients response to the medications is not yet fully predictable and their effects are felt late in time after being initiated. The European League Against Rheumatism (EULAR) recommendations indicates the therapeutic options available and the course of action which should be applied when a medicine or a combination of medicines does not sort the pretended effect or has negative effects on the patient's health [4]. Thus the possibility of predicting the patient's response to a specific treatment is a therapeutic goal as it would prevent irreversible health damage caused by the trial-and-error approach currently applied.

Specifically in the medical departments, data mining, the field which focus on data analysis and consequent information extraction, aided to machine learning have been in vogue as they provide powerful tools in clinical trials and practice [5].

On another note, as high-dimensional data becomes increasingly available, sparse methods allow to go from an abounding number of features included in the data to a selection of the relevant and informational ones. Although the scientific and medical communities have seen great development in what biologic processes are underlying RA, the unresolved heterogeneity of its patients still constitutes a big barrier.

In this sense, the exploration of the associations between the disease processes and the clinical response to therapy has been extensively reviewed. For example, it has been shown in a research using the same dataset in this present study, that 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 [6]. A meta-analysis of RA synovial transcriptomic data has evidenced differences in the activation of genes involved in several key and targetable signalling pathways which could predict the response to infliximab, an anti-TNF drug, with high accuracy [7]. However, the identification of biomarkers has yet to reproducibly manifest carnally relevant predictive.

The main goal of this thesis was to identify gene signatures (biomarkers) from transcriptomic data in patients undergoing RA treatment with biologics. Transcriptome sequencing or gene expression profiling can be achieved by RNA sequencing (RNA-Seq), a technology which uses next-generation sequencing to quantify RNA in a sample [8; 9]. With those biomarkers the intention was to distinguish which patients would fall under the good-responder (further labeled as "R") or the non-responder ("NR") types, according to the EULAR criteria [10]. The transcriptomic data used was subject to regularisation methods which perform feature selection and subsequently Bayesian networks (BN) were learned from them in order to analyse which protein-protein interactions were underlying the patients in terms of their response to the treatment. At last, different machine learning algorithms with distinctive nuances were trained with the same data and their prediction performance evaluated.

RA is first described in Sec. 2, alongside with a review of the methods applied and a description of the data. The work methodology is presented in Sec. 3 and the subsequent results and corresponding discussion are analysed in Sec. 4. Finally Sec. 5 concludes about the overall work achievements and proposes future directions.

2. Background

2.1. Rheumatoid Arthritis

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].

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) [11]. Being an autoimmune disease, the case is these autoantibodies attack the self organism leading to abnormal immune reactions.

The exact pathogenesis leading to this immune system deregulation is still unknown, but evidence has been shown that certain genetic predispositions, such as class II major histocompatibility complex (MHC) genes, specifically the HLA (human leukocyte antigen) DRB1 alleles, and tumour necrosis factor (TNF) alleles, play an important role [12]. Furthermore, T and B cells, which are vital in the adaptive immune response, have long been implicated in mediating many joint inflammation aspects [13; 14].

RA's diagnose is a complicated task due to the several causes leading to joint stiffness and inflammation. Classification criteria is the usual approach to define RA and to assess its severity in the patient's health. This criteria is based on the disease activity, which in its turn includes different variables and quantitative evidence such as pain scales, questionnaires regarding functional damage, information from swollen joints, autoantibodies tests, erythrocyte sedimentation rate or C-reactive protein level. Tab. 1 indicates examples of different existing formulas used for disease scoring.

Table 1: Formulas for calculation of RA disease activity scores: DAS, DAS28, SDAI and CDAI [15; 16]

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.

The *DAS* (disease activity score) is a clinical index of RA 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 (or 3.7 when using DAS) it is considered high disease activity.

SDAI (Simplified disease activity index) and *CDAI* (Clinical disease activity index) commonly combine single measures into an overall continuous measure of the disease activity, differing only in the inclusion of C-reactive protein level (CDAI does not include it). High disease activity is defined if $SDAI > 26$ or $CDAI > 22$ [17; 18].

Due to varying definitions of what constitutes remission, 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 [10]. 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 Tab. 2).

In clinical trials, the treatment response is often assessed via the EULAR criteria, which is based on change from baseline and the individual change in DAS

Table 2: ACR/EULAR boolean and index based definition of remission for clinical trials and clinical practice [10]

	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)

Table 3: EULAR response criteria using DAS28 [16]

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

reached during followup. The patients are statified in good, moderate or non-responders according to Tab. 3.

The most recent update of the EULAR recommendations for RA treatment occurred in 2019. Nowadays the main therapeutic target is to reach clinical remission, being low disease activity considered the best possible alternative.

The patient should initiate treatment with disease-modifying antirheumatic drugs (DMARDs) which slow the progression of joint damage. Methotrexate is the most common prescribed conventional synthetic DMARD (csDMARD), occasionally combined with other DMARDs or glucocorticoids. If there is no improvement by at most 3 months after treatment initiation, adjustments should be made: if the patient does not presents with poor prognostic factors, other csDMARDs should be considered; otherwise it is recommended adding a biologic DMARD (bDMARD) or a targeted synthetic DMARD (tsDMARD). When in persistent remission, the therapeutics should be gradually and thoroughly tapered [4].

bDMARDs are engineered to act like a natural human protein and interrupt immune system signals. Depending on their target, the therapies may be TNF inhibitors (anti-TNF), which block the TNF alpha (TNF-α), a cytokine that induces local inflammation and pannus formation. Alternatively, they may target interleukin-1 or interleukin-6 receptors; may be produced in order to destroy B cells or even prevent T-cell activation [19; 20].

2.2. 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. Logistic regression is the special case of a generalized linear model in which the response variable is binary (and thus the vector of observations reflects binomial distribution).

Logistic regression is the special case of a generalized linear model which defines the relationship between n independent observations $\{\mathbf{X}_i\}_{i=1}^n$, each measured over p variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, and a binary outcome $\{\mathbf{Y}_i\}_{i=1}^n$. It is given by:

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

where $\boldsymbol{\beta}$ is the vector of unknown regression coefficients related with the p variables and p_i is the probability of success.

Logistic regression is a very commonly used tool for applied statistics and discrete data analysis, having even shown equally performing results when compared to alternative machine learning techniques in clinical and biological research. Nevertheless, there are important considerations when being conducted, such as including a careful variable selection [21; 22].

2.3. Regularization Methods

Sparsity can be encouraged by constraining the regression problems with regularization methods. *Elastic net* regularization is a weighted combination of *ridge* regression and *lasso* (Least absolute shrinkage and selection operator) regression [23]. The former imposes a ℓ_2 constraint (sum of the squared error of the coefficients) whereas the latter imposes an ℓ_1 constraint (sum of the absolute values of the coefficients) [24; 25]. It can be defined as:

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

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.

2.4. Bayesian Networks

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. BNs are intuitive directed acyclic graphs, commonly defined as *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.

Let $\mathbf{X} = \{X_1, ..., X_p\}^T$ be a p-dimensional vector of random variables X_j, where $X_j \subset \Re$, that coincides with the nodes *V* from $G = (V, E)$. *G* represents a joint probability distribution *P*(**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

$$P(X_1, ..., X_p) = \prod_{j=1}^{p} P(X_j | Pa_{X_j}^G) = \theta, \qquad (3)$$

where $Pa_{X_j}^G = \{Z_i : Z_i \rightarrow Z_j \in E\}$ denotes the set of parents of X_j in G and the factors θ_j are called conditional probability distributions.

To find the best network representing the data is usually the core of a BN learning problem. `sparsebn` R package learns a BN from data using a score-based approach relying on regularized maximum likelihood estimation. The criterion considered in that algorithm is:

$$\min_{\mathbf{B} \in \mathbb{D}} \quad l(B; \mathbf{X}) + \rho_\lambda(B), \qquad (4)$$

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 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 $\{\hat{\boldsymbol{\beta}}(\lambda_{\max}), \hat{\boldsymbol{\beta}}(\lambda_1), ..., \hat{\boldsymbol{\beta}}(\lambda_{\min})\}$ for a predetermined set of lambdas $\lambda_{\max} > \lambda_1 > ... > \lambda_{\min}$. Since the focus is on sparse graphs, the algorithm is terminated when the number of edges exceeds some user-defined threshold. From the solution path the preferred solution can either be selected by the user or automatically by the algorithm (this *optimal solution* is based on a trade-off between the increase in log-likelihood and the increase in complexity between solutions).

2.5. Classification Algorithms

The discrete nature of RNA-Seq data does not allow the use of microarray-based classifiers. Thus 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.

The continuous-based classifiers tested in this work were *svm* (Support vector machine, which creates a decision boundary between 2 classes [26]), *rf* (Random forests, an ensemble learning method based on decision trees [27]) and *NSC* (Nearest shrunken centroids, which constrats from the standard nearest centroid classification by shrinking each class centroid towards the overall centroid [28]).

As for the discrete-based classifiers, the nonnegative nature of RNA-Seq makes it more appropriate to model the data with discrete-count distributions, such as the poisson and the negative binomial. Therefore the models used were *plda* (poisson linear discriminant analysis [29]), *plda2* (its power transformation) and *nblda* (negative binomial linear discriminant analysis [30]).

voom transformation aims at dealing with sample quality variability, often encountered in small RNA-Seq experiments, by finding the compromise of using all available data, but to down-weight the observations from more variable samples [31]. Novel classification methods integrating voom transformation have been developed specificaly for RNA-Seq analysis, such as *voomDLDA* or *voomNSC* (extensions of the diagonal linear discriminant analysis and NSC, respectively) [32].

2.6. Data Description

The data used in this thesis consists of RNA-Seq of whole blood samples from biologic naïve patients from the CORRONA CERTAIN registry [3] 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 (anti-TNF therapies) in conjunction with methrotrexate.The files contained 25,370 variables (gene expressions) measured from 63 patients at BL and 65 patients at M03. Each patient was clinically evaluated based on EULAR criteria for clinical response at the third month of treatment as good responder (further denoted as "R") or non-responder ("NR") [16].

STRING is a database of known and predicted protein-protein interactions. The data used in this work corresponded to the interactions at highest confidence interval (a score each association is given which indicates the estimated likelihood that a given interaction is biologically meaningful, specific and reproducible, given the supporting evidence) [33].

3. Proposed methodology

3.1. Finding Biomarkers

Prior to any analysis, a pre-processing step was carried out, which removed the variables with zero standard deviation and performed log-transformation and normalization. Sparse logistic regression with elastic net regularization was performed by means of the `glmnet` R package [34]. For model validation, the data was split in 70% for training the model and 30% for testing it. This procedure was repeated 5,000 times for each dataset. In each run, the model was estimated from the training data with logistic regression using method *cv.glmnet*, where the parameter α (Eq. 2) varied between 0 and 1 with 0.1 intervals. The penalty λ (Eq. 2) was optimized by 10-fold cross-validation (CV, [35]): the chosen λ was the largest one with which the error was within one standard error of the minimum [34]. Lastly, the fitted model was used to predict the treatment response of the test set. For each model, the receiver operating characteristic (ROC) curve was estimated and the Area under the curve (AUC) calculated.

Two predictive models were chosen for each dataset (a pair of α values was selected for BL and for

M03). Afterwards, Leave-one-out cross-validation (LOOCV, [35]) 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.

BN learning was performed using the `sparsebn` R package [36] to uncover the gene networks. Each of the four models was split *a priori* into other two regarding the type of treatment response. At last the protein-protein interactions found were compared to the STRING database. Fig. 1 schematizes the steps until this point.

Figure 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.

At this point, eight different BN were to be obtained. However, different network's architectures were experimented in respect to the number of edges (n_edges) they contained in the solution and how they were learnt. Consequently, each of those eight BN actually evolved to 4 different networks with the following names and characteristics:

- *S*: Forced solution in which $n\_edges = n\_var$;
- *D*: Forced solution in which $n\_edges = 2 \times n\_var$;
- A_S: Trade-off solution in which *maximum* $n\_edges = n\_var$;
- A_D: Trade-off solution in which *maximum* $n\_edges = 2 \times n\_var$;

These four steps applied to both models of each dataset and to each group of patients are represented in Fig. 2.

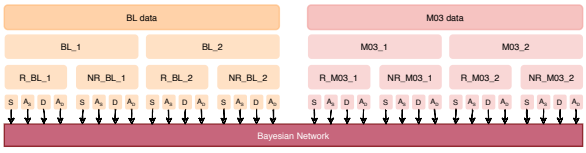


Figure 2: Complete set of BN obtained when adjusting the number of edges allowed in the solution. *S*, *D* and *A* refer to the number of edges in the solution (*S*: $n\_edges = n\_var$; *D*: $n\_edges = 2 \times n\_var$; A_S and A_D: trade-off solution chosen by the algorithm when given maximum number of edges $n\_edges = n\_var$ and $n\_edges = 2 \times n\_var$, respectively).

3.2. Classification Analysis

Different machine learning algorithms were further exploited. In order to inspect how each performed when given different portions of the same data, besides the initial datasets (after removal of variables with zero standard deviation), it was used six newly created sub-datasets with a maximum variance filter and the sparse models previously obtained. Fig. 3 illustrates these different extractions. Every step described henceforth was conducted with `MLSeq` R package [37].

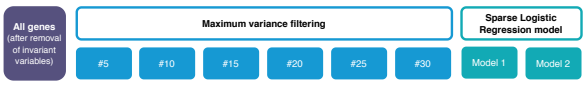


Figure 3: Sub-datasets used as starting point for classification analysis for each data group (BL and M03). "*#5*" indicates the sub-dataset with the top 5 features in terms of variable variance, and so on until "*#30*". Model 1 and 2 refer to the models obtained in the previous pipeline (Fig. 1) which resulted in selecting two α values for each data group.

The eight classifiers selected fit the data and predict the patient's response were the ones previously described: continuous-based (*svm* with radial basis function as kernel method, *rf* and *NSC*), discrete-based (*plda*, *plda2*, *nblda*) and voom-based (*voomDLDA* and *voomNSC*). The normalization approaches used were *deseq* [38] and *TMM* [39]. The transformation methods used in the continuous-based models were *vst* [38], *rlog* [40] and *logcpm* [39]. Note that the voom-based algorithms perform the *voom* transformation with 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 Fig. 4 describes the overall approach. 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.

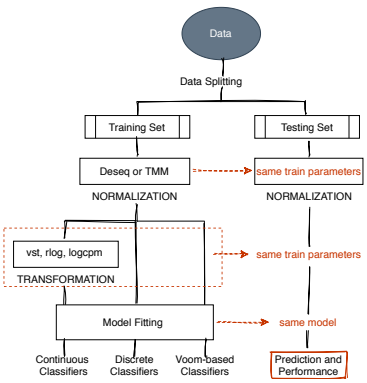


Figure 4: Flowchart of procedure used to fit data into classifiers and to compare model's performance after prediction of class labels.

4. Results

4.1. Sparse Logistic Regression

The pre-processing step applied to each dataset resulted in a reduction of the original 25,370 variables to 21,911 at BL and 22,142 at M03. Over the 5,000 runs, the median AUC values obtained for each α were rather similar and around 0.6. Considering the number of observations in each dataset, this number was considered satisfactory. Given that there was no obvious choice about which α resulted in a better model, two were chosen. Regarding BL: $\alpha = 0.3$ (titled by *BL_1*) and $\alpha = 0.2$ (*BL_2*); and M03: $\alpha = 0.4$ (*M03_1*) and $\alpha = 0.3$ (*M03_2*).

To each model, LOOCV was applied in order to find the biomarkers possibly in strong association with the treatment response. The ROC curves obtained revealed the M03 models to be more accurate than the BL models, which argues that the best prediction is achived from the data retrieved at the third month. Furthermore Tab. 4 shows that models 1 of each dataset obtained better accuracy values.

Table 4: 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 repeatedly selected variables in each iteration of the LOOCV was regarded as the relevant genes related to the treatment response. As expected, decreasing the α parameter resulted in a higher number of variables selected. Furthermore, all the variables in each *model 2* included the variables in the corresponding *model 1*. Tab. 5 lists the 24 genes for BL and the 12 found for M03.

The analysis was narrowed to the genes with a minimum reading count of 20, revealing that at BL the genes which expressions stand-out were *MPO* (encodes Myeloperoxidase), *PRSS30P*, *RCAN3AS* (regulators of calcineurin 3 antisense) and *CTSG* (Cathepsin G) and at M03 *ELANE* (elastase, neutrophil expressed) and *TRIM7* (Tripartite Motif Containing 7).

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 in nonresponder patients [41]. *RCAN3* has shown to modulate T cell development in murine models and suggested to be an effective treatment for RA [42]. The fact that it was its antisense identified motivated a further exploration: if a higher expression of an antisense is detected, it means that the complementary mRNA (in this case the *RCAN3* gene) is being under-expressed [43]. It was then hypothesized that by observing that in "NR" *RCAN3AS* has a higher median expression than in "R", than in the former this protein is prevented from being translated, and ultimately in "R" its expression levels will be higher and thus considered a biomarker. However, the results showed exactly the opposite *i.e.*, the "NR" group had median higher counts. Nevertheless, this is a study involving a small number of patients and so a more complex investigation focusing on gene *RCAN3* should not be disregarded. *PRSS30P* is a pseudogene related to a serine protease but of unknown function and with no allusion of it being related to inflammation. However, 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.

Table 5: List of predictive genes in RA treatment after applying Leave-one-Out Cross-Validation in each dataset's model 1.

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*
M03_1	*ADAM33*, *CCDC110*, *ELANE*, *KCNJ8*, *LOC101928222*, *LRRN4CL*, *MTRNR2L3*, *TMEM105*, *TRIM7*, *UBE2QL1*, *VSTM2L* and *ZNF843*

Regarding the two disclosed genes with a meaningful expression at M03, *ELANE* is a neutrophil serine protease involved in, amongst others, the killing of pathogens and regulation of inflammation. Regarding RA, it can directly degrade the matrix, destroying cartilage components [44]. In a research it was noticed the significantly different expression of ELANE, although the focus was only in the patients' transcriptomic data prior to the treatment initiation [41]. *TRIM7* encodes a member protein of a family implicated in a multitude of biological processes, having gained much attention in cancer studies [45]. However, other findings suggested 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 [7].

4.2. Bayesian Network Models

Applying the elastic net regularization to the datasets originated 4 new sparse models. Their number of variables was BL_1 with 71; BL_2 with 111; M03_1 with 61 and M03_2 with 91. 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 Fig. 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 BN corresponding to the optimal solution, it chose the one with the given number of edges *i.e.*, $S = A_S$ and $D = A_D$. This was observed in 3 cases, all regarding the BL data: model 1, "R"; model 1, "NR"; and model 2, "NR".

It was assessed the 3 interactions with the highest weight value for every BN in order to disclose which gene networks may regulate the response to anti-TNF treatment. The results are presented in Tables 6 and 7.

Table 6: BN interactions obtained (showing only 3) with highest edge weight for **BL** data. Note that the symbol "-" simply indicates the cases where the obtained network were repeated.

		S	D	A_S	A_D
Model 1	R	weight2EPHX4 - LRGUK MIR941-4 - MIR941-2 BATF2 - EVPLL	weight8TBX2 - CYGB weight2EPHX4 - LRGUK weight9LOC100507156 - LINC00696	- - -	- - -
	NR	weight3EVPLL - IGF2BP1 RCAN3AS - KCNH4 ERICH6 - SULF1	MAG - MAGEC2 weight3EVPLL - IGF2BP1 LOC339975 - LILRB4	- - -	- - -
Model 2	R	CDC42EP4 - TCN2 weight2EPHX4 - LRGUK weight9LOC100507156 - LINC00696	weight9LOC100507156 - LINC00696 weight8TBX2 - CYGB DRD2 - CAPN11	CDC42EP4 - TCN2 weight2EPHX4 - LRGUK weight9LOC100507156 - LINC00696	weight9LOC100507156 - LINC00696 weight8TBX2 - CYGB DRD2 - CAPN11
	NR	MIR941-4 - FGDSP1 C1orf95 - MAGEC2 SLC25A52 - ADAMTS9	MIR941-4 - FGDSP1 SLC25A52 - ADAMTS9 weight3EVPLL - IGF2BP1	- - -	- - -

Table 7: BN interactions obtained (showing only 3) with highest edge weight for **M03** data. *none* indicates that overlapping analysis revealed no interactions.

		S	D	A_S	A_D
Model 1	R	weight1KCNK4 - MIR718 weight7RSPH10B2 - RSPH10B CTSG - ELANE	weight1KCNK4 - MIR718 weight7RSPH10B2 - RSPH10B MTRNR2L3 - ZNF843	weight7RSPH10B2 - RSPH10B *none* *none*	weight7RSPH10B2 - RSPH10B *none* *none*
	NR	F3 - LOC101927468 C8B - LOC102467224 weight4KNCN - CCDC110	C8B - LOC102467224 F3 - LOC101927468 weight6FBLIM1 - UBE2QL1	weight4KNCN - CCDC110 weight7RSPH10B2 - RSPH10B *none*	weight4KNCN - CCDC110 weight7RSPH10B2 - RSPH10B *none*
Model 2	R	MTRNR2L3 - MIR4271 TRIM7 - TMEM51-AS1 weight1KCNK4 - MIR718	MTRNR2L3 - MIR4271 weight1KCNK4 - MIR718 FSD2 - RS1	weight7RSPH10B2 - RSPH10B *none* *none*	weight7RSPH10B2 - RSPH10B *none* *none*
	NR	VWA1 - LINC01361 weight6FBLIM1 - UBE2QL1 VWA1 - CES1P1	VWA1 - LINC01361 LOC100506071 - HIST1H2AJ weight6FBLIM1 - UBE2QL1	weight4KNCN - CCDC110 MIR3918 - VWA1 weight7RSPH10B2 - RSPH10B	weight4KNCN - CCDC110 MIR3918 - VWA1 weight7RSPH10B2 - RSPH10B

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. 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. 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); kinocilin, *KNCN*, has a role in stabilizing dense microtubular networks or in vesicular trafficking [46]; *CCDC110* has been identified as novel cancer/testis antigen recognized by cellular and humoral immune responses [47]; *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 [48] and finally *VWA1* belongs to a superfamily of extracellular matrix proteins and appears to play a role in cartilage structure and function [48]. The possible relation of these protein-protein interactions to RA is not evident in the literature.

It was essential to compare the protein-protein interactions with STRING database [33] in order to validate the results. Regarding the BN learnt from BL data: on the one hand, the *CTSG – MPO* interaction, which was found in the "R" group, is given a total score of 0.989 in STRING database. Given that both genes 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, an known important multifunctional inflammatory mediator for recruitment of monocytes in the second wave of inflammation. The Venn diagrams regarding model 2 (see Fig. 5) highlight one interaction common to both "R" and "NR": *MPO – AZU1*, suggesting that it might be be relevant for both types of patients. 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 with the latter, this protein-protein interaction being found in both types of patients indicates its importance in the mechanisms of anti-TNF treatment.

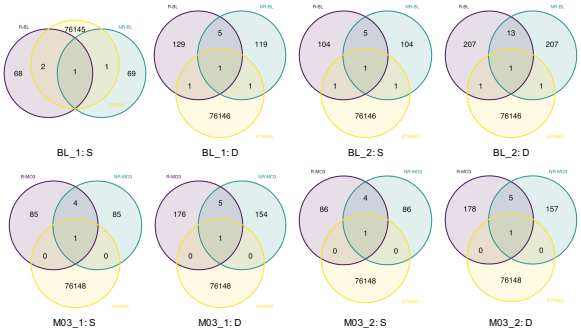


Figure 5: Venn diagrams showing common interactions between learnt BN from "R" and "NR" groups and STRING database.

4.3. Classification algorithms analysis

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. 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.

The plots from Fig. 6 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 Fig. 3).

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.

Despite the fact that only one *voom*-based classifier is sparse, they both delivered very similar accuracy when the data was transformed with *TMM* (Fig. 6k). Moreover, when *deseq* was used, *voomDLDA* (non-sparse) slightly reached a better performance (Fig. 6h).

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). It has been stated the little impact on choosing the normalization procedure on the classification performance (it is rather more important in differential expression analysis) [29]. However, concerning the two approaches used in this work, and looking at Fig. 6'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. Since there are no results available regarding *deseq-logcpm* combination, it is only possible to consider the influence of *vst* and *rlog* (Figures 6a and 6d). 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 [49]. In this study *vst* resulted in lower sparsity.

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.

Being sparse algorithms, *NSC*, *plda*, *plda2* and *voomNSC* performed feature selection. The common features selected by these 4 classifiers, both when using *deseq* and *TMM* as normalization procedure, were compared to the features given by the elastic net penalisation, BL_1 and M03_1 (since using models 2 did not produce any changes in the overlap analysis). In the case of *NSC* the transformation procedure used was *vst* since it was the one with which best accuracy was achieved. The common genes selected by these four sparse tools (further labelled as Z) were later compared to the BL and M03 models previously obtained, leading to the following results:

- $BL_{models} \cap Z_{deseq}$: *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*.

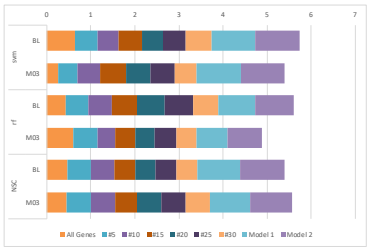
The fact that the features selected by the sparse classifiers revealed genes selected by the initial implemented approach reinforces the first results. *MPO* and *CTSG* are relevant genes whose expression has an influence on the anti-TNF treatment response of each patient. It is then proposed that they 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: 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 [50].

5. Conclusions

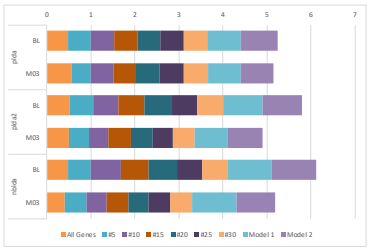
This thesis' main goal was to identify biomarkers able to predict anti-TNF RA treatment response. Through transcriptomic data, a sparse logistic regression approach was used in order to obtain the best predictive models for each dataset (BL and M03) leading to a selection of genes regarded as relevant in predicting the treatment response to the cited drug. The protein-protein interactions found through BN learning and validated by STRING database revealed genes to be consistently associated with the therapy response. Besides, highly connected associations were uncovered at baseline not only independently in "R" and "NR" patients but also in common in both types of patients.

Regarding the analysis of the different machine learning tools, the overall best performances were achieved by *svm*, *rf*, *voomDLDA* and *voomNSC*, being only the latter a sparse classifier. Attention is given to genes *MPO*, *CTSG*, *AZU1* and *RCAN3AS* which have shown to be involved in the RA modulation.

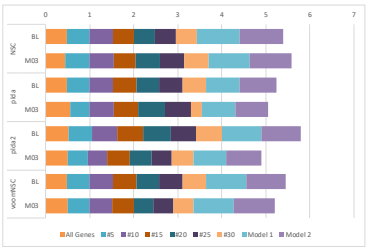
It is suggested to further investigate the potential of the classification algorithms in the context of RNA-Seq and RA treatment response. A critical limitation of this work is the datasets' sizes, as there were little observations. Thus a study involving the transcriptomic data of more individuals is further recommended. Lastly other



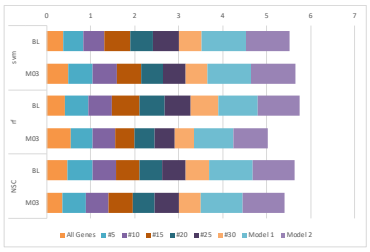
(a) *deseq-vst*, continuous-based classifiers



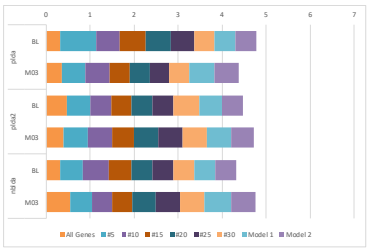
(b) *deseq*, discrete-based classifiers



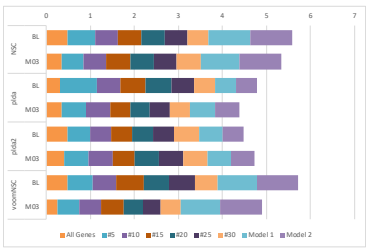
(c) *deseq*, sparse classifiers



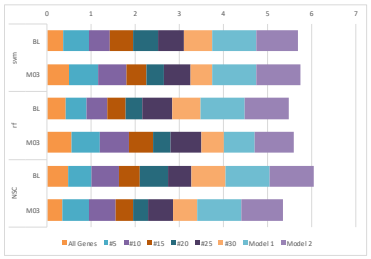
(d) *deseq-rlog*, continuous-based classifiers



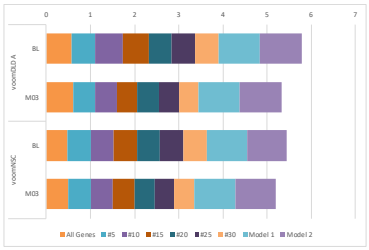
(e) *tmm*, discrete-based classifiers



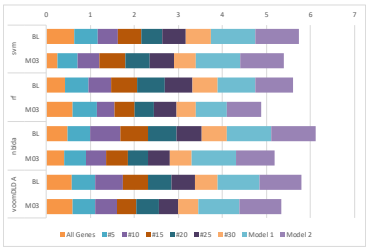
(f) *tmm*, sparse classifiers



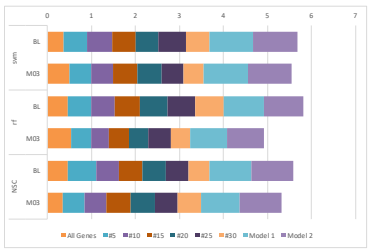
(g) *deseq-logcpm*, continuous-based classifiers



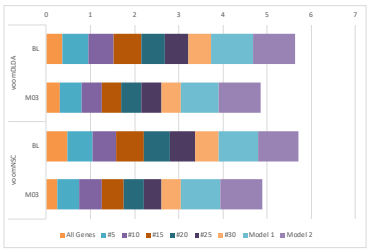
(h) *deseq*, voom-based classifiers



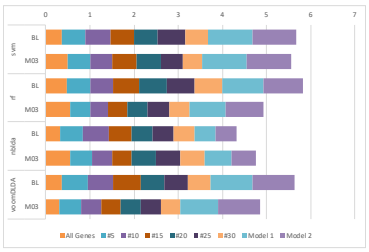
(i) *deseq*, non-sparse classifiers



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



(k) *tmm*, voom-based classifiers



(l) *tmm*, non-sparse classifiers

Figure 6: Accumulated testing accuracy results for fitted classifiers. On the x-axis the classifiers are featured, whereas the y-axis indicates the added accuracy.

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.

Acknowledgements

Great appreciation is given to Professors Susana Vinga and Alexandra M. Carvalho for the consistent support and valuable guidance.

References

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

[6] 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.

[7] 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.

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

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

[10] eular. Defining remission in rheumatoid arthritis. Accessed: 13 October 2020.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

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

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

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

[26] 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.

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

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.