Sparse consensus classification for discovering novel biomarkers in rheumatoid arthritis*

Cláudia Constantino$^{1[0000-0003-4320-8437]}$, Alexandra M. Carvalho$^{2[0000-0001-6607-7711]}$, and Susana Vinga$^{1[0000-0002-1954-5487]}$

[1] INESC-ID, Instituto Superior Técnico, ULisboa, Portugal
susanavinga@tecnico.ulisboa.pt
[2] Instituto de Telecomunicações, Instituto Superior Técnico, ULisboa, Portugal
alexandra.carvalho@tecnico.ulisboa.pt

Abstract. Rheumatoid arthritis (RA) is a long-term autoimmune disease that severely affects physical function and quality of life. Patients diagnosed with RA are usually treated with anti-tumor necrosis factor (anti-TNF), which in certain cases do not contribute to reach remission. Consequently, there is a need to develop models that can predict therapy response, thus preventing disability, maintain life quality, and decrease cost treatment. Transcriptomic data are emerging as valuable information to predict RA pathogenesis and therapy outcome. The aim of this study is to find gene signatures in RA patients that help to predict the response to anti-TNF treatment. RNA-sequencing of whole blood samples dataset from RA patients at baseline and following 3 months of therapy were used. A methodology based on sparse logistic regression was employed to obtain predictive models which allowed to find 20 genes consensually associated with therapy response, some known to be related with RA. Gene expression levels at 3 months of therapy showed no added value in the prediction of response to therapy when compared with the baseline. The analysis using Bayesian network learning unveiled significant protein-protein interactions in both good and non-responders, further confirmed using the STRING database. Structured sparse regression coupled with Bayesian learning can support the identification of disease biomarkers and generate hypotheses to be further analysed by clinicians.

Keywords: Regularized optimization · Bayesian Networks · Protein-Protein Interaction Networks.

1 Introduction

As high-dimensional data becomes increasingly available in all the fields of research, effective analytic methods are fundamental to extract the maximum scientific understanding from the data. For instance, in genetic studies, in which

* Partially funded by FCT (PTDC/CCI-CIF/29877/2017, UIDB/50021/2020, UIDB/50008/2020)

the number of variables (genes) is particularly large compared with the number of samples (subjects), methods to extract knowledge from the data are essential.

Rheumatoid arthritis (RA) is a common systemic autoimmune disease that severely damages physical function and quality of life. This chronic disease affects about 1% of adult citizens [1]. RA cause remains unknown, although genetic factors are responsible for a part of disease predisposition.

Nowadays, RA therapy is processed by the administration of disease-modifying anti-rheumatic drugs (DMARDs), that have shown to slow down the disease progression. However, DMARDs may have no effect on patients, and therapy with anti-tumor necrosis factor (TNF) is then recommended [2, 3]. If the therapy with TNF inhibitors fails in a particular patient, an alternative agent is again chosen. It is usually very difficult to find an agent, or a combination of agents, that induce remission. So, RA patients may experience therapy with successive changes in the administration of these agents until disease remission, which can significantly worsen patient disability and can increase the cost of treatment.

The prediction of the patient's response to anti-TNF therapy is then of paramount importance, and it has been the object of study in several studies that take into account demographic, clinical, and genetic data [4, 7]. Interestingly, the studied clinical baseline biomarkers do not seem to add value in the prediction of treatment response [5]. Notwithstanding, gene expression profiling may provide insight into disease pathogenesis and already showed significant changes before and after anti-TNF treatment [6], which illustrates the potential of using molecular information for prognosis. At the baseline of anti-TNF treatment, innate/adaptive immune cell-type-specific genes revealed associations with the response to treatment within 3 months of therapy [7]. Recent studies have studied and demonstrated the role of molecular data to conduct robust predictions in different human diseases [8, 9]. Therefore, additional evaluation of the gene expression can be helpful to define biomarkers of outcome and response to therapy in RA.

The aim of this study is to find gene signatures in patients with RA before and after the beginning of the therapy, which may help to predict the response to anti-TNF treatment. We propose a predictive model, based on dimensionality reduction techniques and on a consensus approach, which uncover the most relevant genes to predict the response to therapy. We also use a Bayesian network methodology to discover relevant protein-protein interactions.

The paper is organised as follows. In Section 2 we present the RA data under study and the sparse logistic regression and Bayesian networks. Next, in Section 3 we present the experimental results and discuss them. Finally, we draw some conclusions and discuss future works.

2 Methods

2.1 Rheumatologic transcriptomic data

The dataset under study is constituted by RNA-sequencing of whole blood samples from biologic naïve RA patients. All the files are publicly available from the CORRONA CERTAIN registry [10] and at the NCBI-GEO database (GSE129705); these data were previously analysed by Farutin et al. [7]. These patients had no previous biologic agent treatment, and they are initiating therapy with anti-TNF. The transcriptomic data are composed of a set of 25,370 variables (gene expressions) measured from 63 patients at baseline (BL), and from 65 patients at 3 months (M3) after the beginning of anti-TNF treatment.

According to EULAR criteria for clinical response to therapy at 3 months [11], each patient is classified as *good responder* or *non-responder* (GR and NR, respectively). Under this classification, 36 patients were categorized as GR and the remaining 27 as NR to therapy with anti-TNF.

2.2 Sparse logistic regression

Binary logistic regression defines the relationship between n independent observations $\{\mathbf{X}_i\}_{i=1}^n$ and a binary outcome $\{Y_i\}_{i=1}^n$, where each observation is measured over p variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$.

Specifically in this work, $n = 63$, $Y_i = 1$ corresponds to a GR patient and $Y_i = 0$ to a NR patient. Then, the logistic regression is given by

$$P(Y_i = 1|\mathbf{X}_i) = \frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})}, \tag{1}$$

where $\boldsymbol{\beta}$ represents the regression coefficients related with the p variables and $P(Y_i = 1|\mathbf{X}_i)$ is the probability of observation i being a good responder (GR).

Logistic regression is a classical method that has shown to be competitive, with equally performing results, compared with alternative machine learning techniques in clinical and biological research [12–14].

To deal with datasets with a number of variables much higher than the number of observations, ($p \gg n$), an initial dimensionality reduction step is fundamental. Getting an adequate generalized model can be extremely difficult in a high dimensional dataset due to the number of variables to be considered in the final model and the few observations to support the model's hypothesis. To overcome this problem, additional constraints in the cost function can be applied. Regularization methods like Least Absolute Shrinkage and Selection Operator (Lasso), Ridge regression, elastic net, and other sparsity methods, provides a sparse estimate of the unknown regression coefficients and have become a classical approach to deal with the possible non-identifiability of the regression models.

For example, Lasso regression [15] enables shrinkage and variable selection in the solutions by penalizing the sum of the absolute values of the coefficients (L1-norm), defined as:

$$\Psi(\beta) = \sum_{i=1}^{p} |\beta_i|. \tag{2}$$

Ridge regression [16] considers the L2-norm (sum of the squared error of the coefficients) penalty instead, having:

$$\Psi(\beta) = \sum_{i=1}^{p} |\beta_i^2|. \tag{3}$$

The elastic net regularization combines L1 and L2 norms with the objective of limit solution space [17]. So, the elastic net is a combination of Lasso and Ridge regression:

$$\lambda\Psi(\beta) = \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2), \tag{4}$$

where α is a controller between L1 and L2 penalties, given a fixed λ. Penalization control of the weights is given by λ.

If $\alpha = 0$, Ridge regression is applied. On the other hand, if $\alpha = 1$, we are dealing with Lasso regression. Elastic net allows the balance of sparsity with the correlation between variables, which gives to this method high flexibility for different types of datasets.

Maximum likelihood are used to estimate β coefficients. In the case of elastic net regression, the penalized log-likelihood function, with L1 and L2 weights, is the following:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Big\{ y_i \log P(Y_i = 1|\mathbf{X}_i) + (1-y_i)\log[1 - P(Y_i = 1|\mathbf{X}_i)] \Big\} + \lambda\Psi(\beta) \\ &= \sum_{i=1}^{n} \big(y_i \mathbf{X}_i^T \boldsymbol{\beta} - \log(1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})\big) + \lambda\Psi(\beta), \end{aligned} \tag{5}$$

where the binary variable y_i indicates if the ith patient is a good responder (GR) ($y_i = 1$) or a non-responder (NR) ($y_i = 0$).

To fit the best predictive model to a dataset, cross-validation (CV) is an important step. It allows estimating parameters for the envisaged model to improve predictions. When regularization is being performed, the addition of a variable in the model may increase model performance. However, its inclusion may also have a high cost, and in that case, the variable should be disregarded of the final model. CV allows tuning the model parameter to perform the best feature selection and prevent overfitting. The penalty λ parameter is estimated using this CV strategy, i.e., we use the value that achieves the minimum mean cross-validation error.

Leave-one-out cross-validation (LOOCV) is another technique generally used that we apply in the present study. It is based on estimating a model by considering all observations except one, that is left out from the training set. That observation is then used to validate the predictive power of the estimated model. This procedure is run the same number of times as the number of the existing observations. In this case study, the objective is to predict whether a RA patient responds or not to the treatment with anti-TNF, i.e., a binary outcome. To evaluate the estimated model, the classifier's specificity and sensitivity trade-off in the validation set can be visualized through Receiver operating characteristic

(ROC) curves. The area under the ROC curve (AUC) is then calculated as a quantitative measure of the classifier performance.

2.3 Bayesian networks

Bayesian networks (BNs) are a rich framework to model domains with complex connections between thousands or millions of random variables. BNs are graphical models that represent a family of probability distributions defined in terms of a directed graph. The nodes of the graph contain the random variables, and the product of local connections defines a unique joint probability distribution. These local connections represent dependencies between a random variable and its parents in the graph.

Rigorously, let $\mathbf{Z} = (Z_1, \ldots, Z_p)$ be a p-dimensional random vector. A Bayesian network (BN) is a pair (G, θ) where $G = (V, E)$ is a directed acyclic graph (DAG), with nodes in V, coinciding with the random variables in $\mathbf{Z}$, and edges in E. The parameters describe how each variable relates probabilistically with its parents. Using the chain rule, we can then obtain a joint probability distribution, in a factored way, according to the DAG structure, defined as:

$$P(Z_1, \ldots, Z_p) = \prod_{j=1}^{p} P(Z_j|\mathrm{pa}(Z_j), \theta_j), \tag{6}$$

where $\mathrm{pa}(Z_j) = \{Z_i : Z_i \to Z_j \in E\}$ is the parent set of Z_j and θ_j encodes the parameters that define the conditional probability distribution (CPD) for Z_j. In the case of continuos data, Gaussian CPDs are considered.

When learning a BN, the challenge is in structure learning. With the structure fixed, parameters are quite easy to learn. Structure learning is accomplished through score-based learning, where a score is used to understand the network that best fits the data. A possible scoring criterion is the maximum likelihood. When overfitting occurs, penalisation factors are used to avoid it.

Aragam et al. (2017) developed an R package, called `sparsebn` [18], especially devoted to high dimensional data. For that, a sparse BN is outputted using a score-based approach that relies on regularised maximum likelihood estimation. The scoring criterion is given by:

$$\mathrm{LL}(B; \boldsymbol{X}) + \rho_\lambda(B), \tag{7}$$

where LL denotes the negative log-likelihood, B the BN, and ρ_λ is some regulariser. For continuous data, a Gaussian likelihood with L1 or minimax concave penalty is proposed.

Considering gene expression, from rheumatologic transcriptomic data, as observations of random variables, a sparse BN is a rich model to describe the underlying data. The nodes represent the genes, and connections between them correspond to gene interactions/edges. The interactions found are those that best explain the data.

3 Results and Discussion

To unravel the most relevant genes in RA patients for the response to anti-TNF therapy, two datasets of gene expression levels, from the CORRONA CERTAIN registry, were used: (i) at baseline (BL), and (ii) three months after starting the treatment with anti-TNF (M3).

Both BL and M3 datasets contain a set of 25370 variables/genes. The data were preprocessed as follows. In both datasets, the variables with zero standard deviation were firstly excluded. This resulted in a reduction to 21911 and to 22142 variables, respectively, for BL and M3 datasets. Then, the variables were log-transformed and normalised to unit variance. A vector with binary responses, with '1' for GR patients and '0' for NR patients, were further used in logistic regression (a vector for BL dataset and another for M3 dataset).

To ensure full reproducibility of our results, all the R code and data are available at https://github.com/sysbiomed/RA-CORRONA.

3.1 Identification of response biomarkers

Dimensionality reduction was achieved by applying sparse logistic regression using the `glmnet` R package [19]. For model validation, 70% of the dataset were randomly split for training the model, and the remaining 30% for the test. This procedure was repeated 100 times. The model is estimated in the training set with logistic regression, defining the λ and α parameter. CV is used to choose the λ parameter that better fit the model. The α parameters used varied between 0 and 1 with intervals of 0.1. Then, the fitted model was used to predict the response of the treatment of the test set. For each model, the ROC curve was accessed, and the AUC calculated. The ROC curve is obtained by using different values for the classification probability threshold. The median and interquartile range of the models estimated using each α in the BL data and over the 100 runs are presented in Table 1.

Table 1: Area Under the Curve (AUC) results from the sparse logistic regression at baseline data for different α parameters. For each α, the statistics of the AUC over all 100 runs are reported, namely the median values, the interquartile range (IQR), the maximum (Max) and the minimum (Min). The best median AUCs results are highlighted in bold.

	α										
AUC	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.00
Median	0.62	**0.64**	0.60	0.60	**0.63**	0.59	0.58	0.57	0.58	0.58	0.52
IQR	0.12	0.13	0.13	0.16	0.11	0.14	0.12	0.13	0.14	0.12	0.09
Max	0.83	0.82	0.89	0.90	0.85	0.78	0.84	0.86	0.81	0.86	0.76
Min	0.37	0.45	0.42	0.43	0.44	0.42	0.39	0.38	0.41	0.37	0.36

The AUC values obtained with sparse logistic regression at month 3 (M3) were not significantly different from those at baseline (BL), presented in Table 1 (t-test, p-values ranging from 0.07 to 0.49). Therefore, only the dataset from BL was further used.

In Farutin et al. [7], the top 100 genes over-expressed with the most significant contributions to the negative correlation with the effect of treatment are presented. We applied the same methodology as before considering those specific genes, to achieve a predictive model. In this case, only $\alpha = 0$ was used for the logistic regression, to ensure that all the genes emerge in the model. The median and interquartile range (IQR) of the AUCs was 0.57 and 0.09, respectively. These results show that although these specified genes are over-expressed, they do not have better predictive power than those identified with a sparse logistic regression using all the available genes and $\alpha \in [0, 0.9]$ (Table 1).

The better predictive models were achieved with an α of 0.1 and 0.4 applied to the BL dataset, with median AUC of 0.64 and 0.63 for $\alpha = 0.1$ and $\alpha = 0.4$, respectively. These results are satisfactory, taking into account the few observations (63) in the dataset. Hereupon, these were the parameters for logistic regression further used.

To detect genes strongly associated with the response, LOOCV was applied. The intersection of genes appearing in all the predictive models calculated with LOOCV correspond to those that may have a better predictive response at BL. Amongst the 21911 genes in the dataset, 20 were repeatedly selected for model prediction, both for α=0.1 and α=0.4. The 20 identified genes are the following:

- *ALOX12B*, *CAPNS2*, *CTSG*, *EPHX4*, *EVPLL*, *FAM133CP*, *FOXD4L3*, *HIST1H3J*, *IGF2BP1*, *LOC339975*, *LRGUK*, *MPO*, *NUAK1*, *ODF3L2*, *PRKG1*, *PRSS30P*, *RAD21L1*, *SLC6A19*, *SYT1*, and *TGFB2*.

Even though none of these discovered genes were present in the top 100 genes over-expressed presented by Farutin et al. [7], some are already known to be related to the RA disease.

CTSG was previously found to participate in the pathogenesis of some autoimmune diseases, as, for instance, RA [20]. When compared with healthy controls, *CTSG* activity and concentration are augmented in the synovial fluids of RA patients. Therefore, identification of *CTSG* in our model may be associated with the response to anti-TNF therapy in RA patients.

Karouzakis et al. [21] found that *EPHX4* gene was one of the top-ranked genes differentially expressed in human lymph node stromal cells (LNSCs) during the earliest phases of RA. LNSCs are decisive in shaping the immune response in lymphoid tissue (where is initiated the adaptive immunity). This suggests that *EPHX4* could be related to the immune response to treatment.

The *MPO* gene encodes myeloperoxidase. Myeloperoxidase serum levels are encountered in inflammatory diseases, like RA. Fernandes et al. [22] observed significantly higher MPO plasma levels in RA patients. Yet, no correlation between disease activity measured by EULAR criteria and MPO expression was found. Contrarily, our analysis led to the identification of this gene as helpful for the prediction of the response to therapy.

Also, *RAD21L1* is part of the 21 upregulated by tumor necrosis factor-like ligand 1A (TL1A) [23]. TL1A is a tumor necrosis factor that influences positively the pathogenesis of autoimmune diseases, including RA.

All the hypotheses above should be confirmed in further studies with the contribution of a rheumatologist. The relevance of the remaining genes in RA was not explored in previous studies. So, we propose a subsequent investigation of the 20 genes achieved in our analysis with the response to therapy with anti-TNF agents. Moreover, the presented methods can be further added to a future benchmarking study that comprehensively assesses the performance of different classifiers.

3.2 Identification of protein-protein interactions

The disclosure of gene networks regulating the response to anti-TNF treatment was performed through Bayesian network (BN) learning. The 239 variables (genes), resulting from the sparse logistic regression with $\alpha = 0.1$, were used. This parameter α was selected, taking into account the trade-off between the AUC medians (Table 1) and the identification of a reasonable number of genes to be further analysed. The baseline dataset with the 239 variables was split into two independent sets: responders at BL (R-BL; 36 observations) and non-responders at BL (NR-BL; 27 observations), each one described by two distinct Bayesian networks.

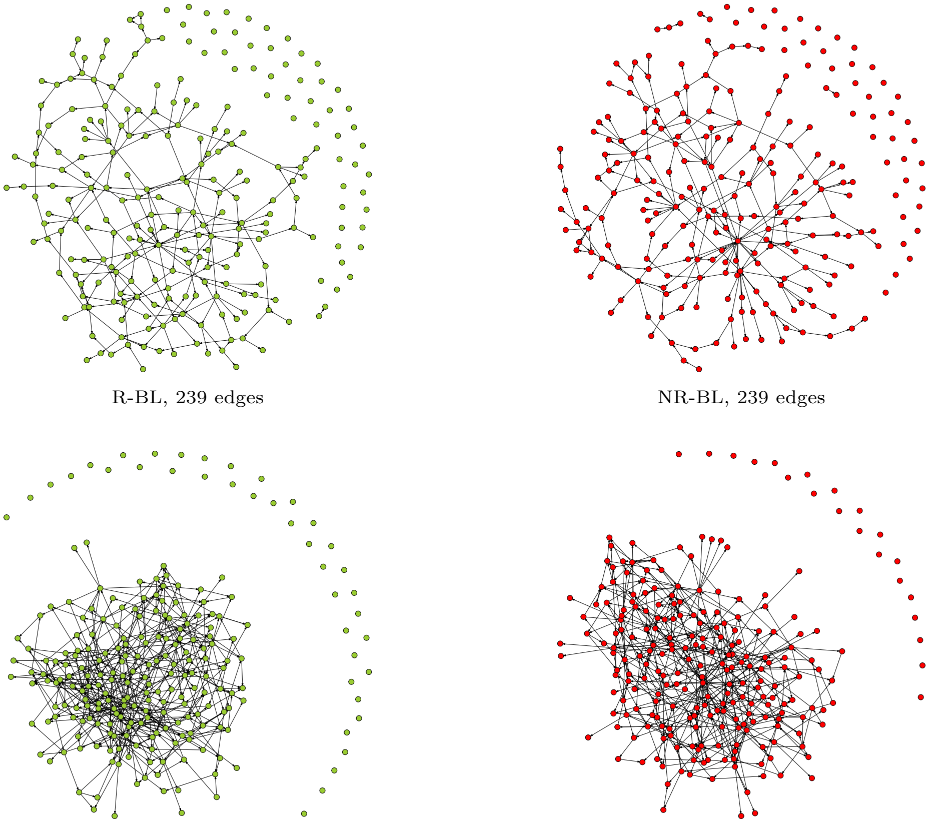
This methodology was applied using `sparsebn` R package [18]. The method `estimate.dag` was used to learn the two distinct Bayesian networks. Therein, we used default parameter settings; the `edge.threshold` parameter was specified to force the number of edges in the solution to be less or equal than the double of the nodes (239×2). The output is a set of different networks, each one with a distinct number of edges. From this set, the solution giving the number of edges equal to the number of nodes was chosen, both for R-BL and NR-BL datasets. The same was done in case the number of edges being twice the number of nodes. Figure 2 illustrates the obtained networks.

To verify if the obtained edges (pairwise gene connections) were previously identified, the STRING information was used. STRING is a database of known and predicted protein-protein interactions [24]. In our study, only protein-protein interactions with high combined score ($combined\_score > 0.7$) in the STRING database were considered. The highly connected genes from STRING were then compared with the edges given by BNs.

From the 239 edges, 4 in R-BL and 2 in NR-BL are reported in STRING. There was no improvement in the number of common edges found in STRING when increasing the number of edges from 239 to 478 edges.

The obtained Venn diagrams illustrate the overlap between the identified protein-protein interactions in R-BL, NR-BL, and STRING (Figure 4), and are as follows:

- **R-BL ∩ STRING:** *DEFA4—CTSG*; *AZU1—MPO*; *CTSG—SERPINB10*; *DEFA4—AZU1*



R-BL, 239 edges

NR-BL, 239 edges

R-BL, 478 edges

NR-BL, 478 edges

Fig. 2: Bayesian networks learnt from R-BL data (green nodes) and NR-BL data (red nodes). (a) BN for R-BL with 239 edges. (b) BN for NR-BL with 239 edges. (c) BN for R-BL with 478 edges. (d) BN for NR-BL with 478 edges.

- **NR-BL ∩ STRING:** *CTSG—AZU1*; *AZU1—MPO*

One overlap interaction between responders and non-responders was found, which suggests that protein-protein interaction (*AZU1—MPO*) may be relevant for both responders and non-responders. According to STRING, this interaction with a total combined score of 0.985, is associated with: 1) the co-expression of these two genes, 2) Database knowledge from the Reactome (in particular the Neutrophil degranulation pathway, in the Innate Immune System – R-HSA-6798751), and 3) the Textmining category. The remaining interactions stand out independently in the responders and non-responders. These known highly connected genes may be a strong indicator of how RA patients respond or not to anti-TNF therapy before the treatment initiation and therefore represent interesting biomarkers to be further explored.

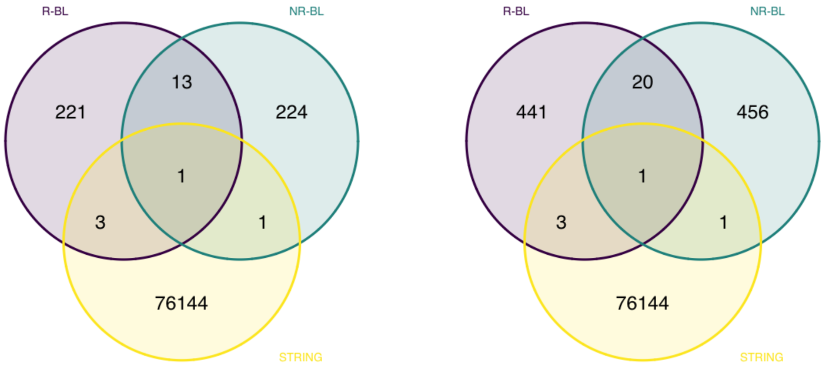


Fig. 4: Venn diagrams with protein-protein interactions in R-BL, NR-BL, and STRING. On the left, the overlap of STRING interactions with R-BL and NR-BL networks with 239 edges. On the right, the overlap of STRING interactions with R-BL and NR-BL networks with 478 edges.

4 Conclusions

Through transcriptomic data, we presented a satisfactory predictive model based on sparse logistic regression that may help to predict the response to anti-TNF therapy in RA patients prior to treatment initiation. A predictive model at BL showed an identical prediction performance compared with that at M3. Also, our methodology was able to unveil genes consistently associated with therapy response, which may be valuable in the expression profiling of RA patients. Some of these genes are already known to be related to RA disease. Moreover, the application of BN learning uncovered highly connected genes in responders and non-responders. The next challenge is to study promising gene signatures individually to validate biomarkers to be used in clinical practice.

References

1. Spector, T. D.: Rheumatoid arthritis. Rheum Dis Clin North Am. **16**(3), 513–37 (1990)
2. Radner, H., Aletaha, D.: Anti-TNF in rheumatoid arthritis: an overview. Wien Med Wochenschr., **165**(1–2), 3–9 (2015) https://doi.org/10.1007/s10354-015-0344-y
3. Smolen, J. S., Landewé, R., Breedveld, F. C., et al.: EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs. Annals of the Rheumatic Diseases, **69**(6), 964–975 (2010) https://doi.org/10.1136/ard.2009.126532
4. Wijbrandts, C. A., Tak, P. P.: Prediction of Response to Targeted Treatment in Rheumatoid Arthritis. Mayo Clin Proc., **92**(7), 1129–1143 (2017) https://doi.org/10.1016/j.mayocp.2017.05.009
5. Cuppen, B. V., Welsing, P. M., Sprengers, J. J., Bijlsma, J. W., Marijnissen, A. C., van Laar, J. M., Lafeber F. P., Nair S.C.: Personalized biological treatment for rheumatoid arthritis: a systematic review

with a focus on clinical applicability. Rheumatology, **55**(5), 826–839 (2016) https://doi.org/10.1093/rheumatology/kev421

6. Oswald, M., Curran, M. E., Lamberth, S. L., Townsend, R. M., Hamilton, J. D., Chernoff, D. N., Carulli, J., Townsend, M. J., Weinblatt, M. E., Kern, M., Pond, C. M., Lee, A., Gregersen, P. K.: Modular analysis of peripheral blood gene expression in rheumatoid arthritis captures reproducible gene expression changes in tumor necrosis factor responders. Arthritis Rheumatol., **67**(2), 344–351 (2015) https://doi.org/10.1002/art.38947
7. Farutin, V., Prod'homme, T., McConnell, K., Washburn, N., Halvey, P., Etzel, C. J., Guess, J., Duffner, J., Getchell, K., Meccariello, R., Gutierrez, B., Honan, C., Zhao, G., Cilfone, N. A., Gunay, N. S., Hillson, J. L., DeLuca, D. S., Saunders, K. C., Pappas, D. A., Greenberg, J. D., Kremer, J. M., Manning, A. M., Ling, L. E., Capila, I.: Molecular profiling of rheumatoid arthritis patients reveals an association between innate and adaptive cell populations and response to anti-tumor necrosis factor. Arthritis Res Ther., **21**(1), 216 (2019) https://doi.org/10.1186/s13075-019-1999-3
8. Barracchia, E. P., Pio, G., D'Elia, D., Ceci, M.: Prediction of new associations between ncRNAs and diseases exploiting multi-type hierarchical clustering. BMC Bioinformatics, **21**(1), 1–24 (2020) https://doi.org/10.1186/s12859-020-3392-2
9. Pio, G., Ceci, M., Prisciandaro, F., Malerba, D.: Exploiting causality in gene network reconstruction based on graph embedding. Machine Learning, 1–49 (2019) https://doi.org/10.1007/s10994-019-05861-8
10. Pappas, D. A., Kremer, J. M., Reed, G., Greenberg, J. D., Curtis J. R.: Design characteristics of the CORRONA CERTAIN study: a comparative effectiveness study of biologic agents for rheumatoid arthritis patients. BMC Musculoskelet Disord., **15**(1), 113 (2014) https://doi.org/10.1186/1471-2474-15-113
11. Fransen, J., van Riel, P. L.: The Disease Activity Score and the EULAR response criteria. Clin Exp Rheumatol., **23**(5 Suppl 39), S93–S99 (2005)
12. Pua, Y., Kang, H., Thumboo, J., Clark, R. A., Chew, E. S., Poon, C. L., Chong, H. C., Yeo, S. J.: Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc, (2019) https://doi.org/10.1007/s00167-019-05822-7
13. Faisal, M., Scally, A., Howes, R., Beatson, K., Richardson, D., Mohammed, M. A.: A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health Informatics J., **26**(1), 34–44 (2020) https://doi.org/10.1177/1460458218813600
14. Kuhle, S., Maguire, B., Zhang, H., Hamilton, D., Allen, A. C., Joseph, K. S., Allen, V. M.: Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC pregnancy and childbirth, **18**(1), 333 (2018) https://doi.org/10.1186/s12884-018-1971-2
15. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B, **58**(1) (1996)
16. Hoerl, A. E., Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, **12**(1), 55–67 (1970)
17. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, **67**(2), 301–320 (2005)

18. Aragam, B., Gu, J., Zhou, Q.: Learning Large-Scale Bayesian Networks with the sparsebn Package. Journal of Statistical Software, **91**(11), 1-38 (2019) https://doi.org/10.18637/jss.v091.i11
19. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, **33**(1), 1-22 (2010) www.jstatsoft.org/v33/i01/
20. Gao, S., Zhu, H., Zuo, X., Luo, H.: Cathepsin G and Its Role in Inflammation and Autoimmune Diseases. Arch Rheumatol., **22;33**(4), 498-504 (2018) https://doi.org/10.5606/ArchRheumatol.2018.6595
21. Karouzakis, E., Hähnlein, J., Grasso, C., Semmelink, J. F., Tak, P. P., Gerlag, D. M., Gay, S., Ospelt, C., van Baarsen, L. G. M.: Molecular Characterization of Human Lymph Node Stromal Cells During the Earliest Phases of Rheumatoid Arthritis. Front Immunol., **10**, 1863 (2019) https://doi.org/10.3389/fimmu.2019.01863
22. Fernandes, R. M., da Silva, N. P., Sato, E. I.: Increased myeloperoxidase plasma levels in rheumatoid arthritis. Rheumatol Int., **32**(6), 1605–1609 (2012) https://doi.org/10.1007/s00296-011-1810-5
23. Fukuda, K., Miura, Y., Maeda, T., Hayashi, S., Kuroda, R.: Expression profiling of genes in rheumatoid fibroblast-like synoviocytes regulated by tumor necrosis factor-like ligand 1A using cDNA microarray analysis. Biomed Rep., **1**(1), 1–5 (2019) https://doi.org/10.3892/br.2019.1216
24. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., von Mering, C.: STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res., **43**, D447–52 (2015) https://doi.org/10.1093/nar/gku1003