Technology and Health Care 30 (2022) S451–S457
DOI 10.3233/THC-THC228041
IOS Press

Integrating molecular interactions and gene expression to identify biomarkers to predict response to tumor necrosis factor inhibitor therapies in rheumatoid arthritis patients[1]

Min-Fan He[a,b], Yong Liang[a] and Hai-Hui Huang[c,*]
[a]*Macau Institute of Systems Engineering and Collaborative Laboratory of Intelligent Science and Systems, Macau University of Science and Technology, Macau, China*
[b]*School of Mathematics and Big Data, Foshan University, Foshan, China*
[c]*Provincial Demonstration Software Institute, Shaoguan University, Shaoguan, China*

Abstract.
BACKGROUND: Targeted therapy using anti-TNF (tumor necrosis factor) is the first option for patients with rheumatoid arthritis (RA). Anti-TNF therapy, however, does not lead to meaningful clinical improvement in many RA patients. To predict which patients will not benefit from anti-TNF therapy, clinical tests should be performed prior to treatment beginning.
OBJECTIVE: Although various efforts have been made to identify biomarkers and pathways that may be helpful to predict the response to anti-TNF treatment, gaps remain in clinical use due to the low predictive power of the selected biomarkers.
METHODS: In this paper, we used a network-based computational method to identify the select the predictive biomarkers to guide the treatment of RA patients.
RESULTS: We select 69 genes from peripheral blood expression data from 46 subjects using a sparse network-based method. The result shows that the selected 69 genes might influence biological processes and molecular functions related to the treatment.
CONCLUSIONS: Our approach advances the predictive power of anti-TNF therapy response and provides new genetic markers and pathways that may influence the treatment.

Keywords: Rheumatoid arthritis, anti-TNF, network-based

1. Introduction

Rheumatoid arthritis (RA) is a complex autoimmune disease for which there is no cure. However, to relieve symptoms and prevent the disease from progressing, a variety of powerful treatments are available. TIn order to prevent permanent loss of function associated with structural damage to the joint, early therapeutic intervention is recommended [1]. For 90% of biologically untreated patients with RA,

[1] This article received a correction notice (Erratum) with the reference: 10.3233/THC-229002, available at https://content.iospress.com/articles/technology-and-health-care/thc229002

[*] Corresponding author: Hai-Hui Huang, Provincial Demonstration Software Institute, Shaoguan University, Shaoguan, China; Macau Institute of Systems Engineering and Collaborative Laboratory of Intelligent Science and Systems, Macau University of Science and Technology, Macau, China. E-mail: tomyhwang@163.com.

anti-TNF therapy provides the first effective treatment option if conventional synthetic disease-modifying antirheumatic drugs such as methotrexate do not work [2]. However, of these RA patients, *70% do not gain meaningful clinical change with anti-TNF treatment [3]. To predict which patients will not benefit from anti-TNF therapy, clinical tests should be performed prior to treatment beginning.

As genomic technologies have advanced, we have better understood inflammatory diseases and developed new treatments. Through the transcriptome, we can view specific genes over-expressed or under-expressed in diseases as a way to gain insight into a cellular response. Although various efforts have been made for identifying biomarkers and pathways [3], the specific response to anti-TNF therapy still remains unraveled. The statistical framework in most of these studies is based on a single set of data and does not take into account the knowledge in protein-protein interactions, biological regulatory networks and signaling pathways. In such a framework, the lack of biological information leads to the stability of prediction factors and reduces the predictive ability of the model [4]. In order to introduce modern precision medicine to autoimmune diseases, an advanced computational method combining genetic data with biological processes is needed.

There are many types of biological network information, such as functional interaction networks [5], protein-protein interactions (PPI) [6], correlations between genes [7,8], KEGG pathways [9]. There are several studies that use biological knowledge, including those by Li and Li [10], Huang et al. [11], Wang et al. [12] and Chen et al. [13]. They described genomic knowledge as a graph that encoded genetic relationships (edges) among genes (nodes). Following that, they implemented linear and classification models with penalties based on Laplace matrices. Models that exploit biological information a priori are known as network-based approaches.

The hypothesis that complex diseases such as RA arise and develop due to interactions between several interrelated pathogenic genes, is supported by a growing body of evidence, indicating that the evaluation of the influence of any single variant is complicated [14]. This study hypothesizes that combining biological interaction information with gene expression data would help identify more robust biomarkers to predict the clinical response to anti-TNF treatment. Therefore, we tried to select the predictive biomarkers by using a network-based computational method to guide the therapy of RA patients. Our results have provided new candidate genes and pathways that may be predicting the response the anti-TNF therapy.

2. Method

In order to integrate the analysis of gene expression data with biological networks, we propose using the Laplace constraint method [10]. Let a network $G = (V, E)$, where V represents the genes with p dimensional, E represents the connections between genes. w_{uv} denotes the weight of edge $e_{u \sim v}$. The typical Laplacian form L for G is:

$$L_{uv} = \begin{cases} 1 - w_{uv}/d_u & \text{if } u = v \text{ and } d_u \neq 0, \\ -w_{uv}/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases}$$

where d_u, d_v denote the degrees (including in and out) of features u and v, respectively. Then, the network-based model can be expressed as:

$$L(\lambda, \beta) = l(\beta) + \lambda \beta^T L \beta. \tag{1}$$

The first term in Eq. (1) represents the loss function, secondly, network-based penalty provides a chance to capture interactive biological knowledge. Parameter λ is used to control the strength of the penalty.

Equation (1) struggles in high-dimensional applications where the number of genes, is larger than the sample size [15–18]. To solve the problem of large p and small n, the regularization approach is widely applied. When the regularization term is added to Eq. (1), the sparse network constraint regression is expressed as Eq. (2).

$$L\left(\lambda_1, \lambda_2, \beta\right) = l\left(\beta\right) + \lambda_1 P\left(\beta\right) + \lambda_2 \beta^T L \beta \tag{2}$$

where λ_1 and λ_2 are parameters, which play a role in balancing the trade-off between complexity and fitness. The most widely used regularization method is Lasso (L_1), which has the form $P(\beta) = \sum_{j=1}^{p} |\beta_j|$. However, in the lasso method, λ_1 needs to be adjusted very carefully, because if λ_1 is too large, the model β may be heavily biased, and if λ_1 is too small, the model β may not be sparse enough. To avoid this issue, Fan el al. [19] proposed SCAD method, which is shown as follows:

$$P_{\lambda,\text{SCAD}}(\beta) = \begin{cases} \lambda\left|\beta\right|, & \text{if } 0 \leqslant |\beta| < \lambda, \\ -\frac{\beta^2 - 2\alpha\lambda|\beta| + \lambda^2}{2(\alpha-1)}, & \text{if } \lambda \leqslant |\beta| < \alpha\lambda, \\ \frac{(\alpha-1)\lambda^2}{2}, & \text{otherwise.} \end{cases} \tag{3}$$

Its higher estimation accuracy and Oracle property make it more advantageous than the lasso method. Therefore, we use the SCAD method to penalize the network-based methods as proposed in Eq. (5). Finally, the model we adopted in this article is defined as:

$$L\left(\lambda_1, \lambda_2, \beta\right) = l\left(\beta\right) + P_{\lambda_1,\lambda_2,\textit{SCAD-Net}}(\beta) \tag{4}$$

Where

$$P_{\lambda_1,\lambda_2,\textit{SCAD-Net}}(\beta) = P_{\lambda_1,\textit{SCAD}}(\beta) + \lambda_2 \sum_{1 \leqslant i < k \leqslant p;} w_{ik} \left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}} \right)^2, \tag{5}$$

and $l\left(\beta\right)$ is defined as a logistics regression model.

To solve Eq. (4), we use the following coordinate descent method. More detailed information can be found in Eq. (5).

```
Algorithm:
Input: Training dataset $\{X_{n\times p}, y_n\}$, $\lambda_1$, $\lambda_2$ and $L$.
Output: Model parameter $\beta$
  Step 1: Update $\beta\left(i\right)^t, i = 1, \ldots, p$.
  Step 2: Let $t \leftarrow t + 1$, if $t < E$, then repeat Step 1.
```

3. Results

3.1. Data description

To identify the key clinical predictive biomarkers for RA, 46 samples with RA, including 24 response to anti-TNF therapy and 22 no response to anti-TNF therapy, were included in the study. Expression data from peripheral blood from these subjects were collected [20].

We mapped the dataset to an official gene symbol, and we calculated average expression levels for multiple probe sets mapped to the same gene. BioGrid provides the biological interaction network L, which includes 14,621 genes or proteins and 327,721 interactions. After combining the gene expression into the L, 215,054 edges, and 15891 genes remain.

Table 1
The selected 69 genes from chronic obstructive pulmonary disease gene expression data

Genes						
CTBP1-AS	ARHGAP22-IT1	RAD9A	NAGPA-AS1	RPS10P7	ELOC	AGAP6
CYP4F8	ATRAID	DDX19A-DT	ALG1	SLC4A5	TMEM178B	APIP
DHDDS	BRD7P3	HLA-DQA1	NAT9	SLIT1	TMEM45A	ARF5
DPY19L1P1	CALM1	LOC101927699	PPM1J	SMARCB1	TUSC2	LAMC1
ERV9-1	CASP5	DTX2	PRMT7	SMCO4	UGT2A2	LINC01477
EXOC4	CD300LG	LTB	RAD9B	SNF8	UGT2A1	LOC100130987
FAM223A	CLTB	MIGA2	RARG	SPINT2	URB1	STOML2
FAM223B	HIKESHI	MIR7114	RIC3	SSNA1	URB1-AS1	SSX5
FAM3B	HOXA10-HOXA9	NSMF	RNF150	SSX7	VASN	SRSF9
GATC	MIR196B	MXRA7	INPP5E	SSX3	HOXA9	

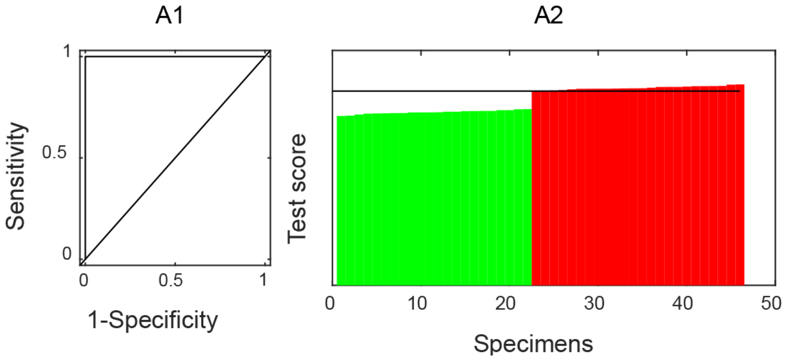

Fig. 1. Training Performance. A1: ROC curve analysis; A2: test scores to be a case of all samples from the dataset were ranked. No response anti-TNF therapy cases are colored in green and response cases in red.

3.2. Construct model and select biomarkers

Tenfold cross-validation on multiple dimensions was used to find the optimal regularization parameters of the model. A classifier model was constructed with the estimated tuning parameters and all the training data with 69 genes (Table 1) and perfect classification performances (Fig. 1). Among all the cutoff points, the one with the highest sum of sensitivity and specificity was chosen.

Among the 69 genes, there are some interesting findings. For example, Rui et al. [21] examined the contribution of *CASP5* gene polymorphisms to RA risk in a Chinese population. They confirmed that *CASP5* was related to the development of inflammation, which is the main feature of RA. Thus, through its role in mediating inflammation, CASP5 may play a role in RA pathogenesis. CD300LG is a novel O-glycosylated member of the CD300 antigen-like family. Besides a classical mucin-like domain, it contains a V-type Ig domain. CD300LG binds lymphocyte L-selectin via its Ig domain and supports lymphocyte rolling via its mucin-like domain. The unique structure and function of CD300LG suggest it may play an important role in inflammation [22].

These findings imply that the selected genes may contribute to or be a marker of the pathophysiology of RA treatment.

3.3. Brief biological analysis

We then perform GO and KEGG enrichment analyses to the 69 genes, as shown in Figs 2 and 3. The results of GO analysis shows the selected 69 genes are involved in 74 significant pathways (with $p <$

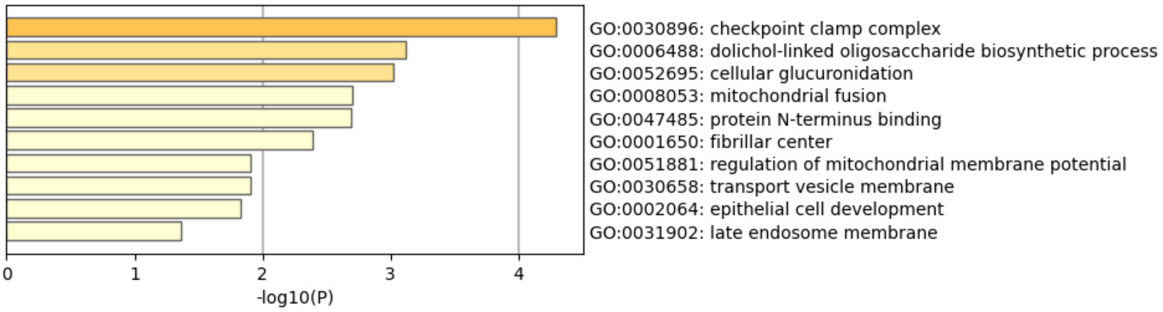

Fig. 2. GO enrichment analyze.

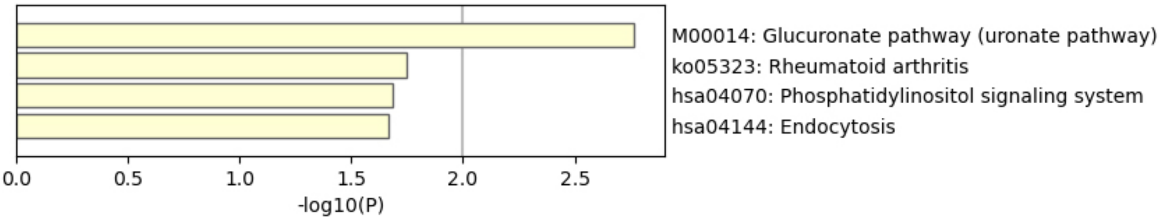

Fig. 3. KEGG enrichment analyses.

0.05), including checkpoint clamp complex, DNA replication checkpoint signaling, mitotic intra-S DNA damage checkpoint signaling, cellular response to abiotic stimulus, cellular response to environmental stimulus, cellular response to ionizing radiation, condensed nuclear chromosome, mitotic DNA damage checkpoint signaling, nuclear chromosome, mitotic DNA integrity checkpoint signaling, DNA damage checkpoint signaling, DNA integrity checkpoint signaling, mitotic cell cycle checkpoint signaling, response to ionizing radiation, dolichol-linked oligosaccharide biosynthetic process, oligosaccharide-lipid intermediate biosynthetic process, protein N-linked glycosylation, cellular glucuronidation, uronic acid metabolic process, glucuronate metabolic process, glucuronosyltransferase activity, bile acid metabolic process, hexosyltransferase activity, glycosyltransferase activity, organic hydroxy compound metabolic process, UDP-glycosyltransferase activity, mitochondrial fusion, mitochondrion organization, organelle fusion, and protein N-terminus binding.

The enriched pathways may role in RA treatment. It is becoming increasingly recognized that immune checkpoint inhibitors can result in inflammatory arthritis among patients treated with these drugs [23]. Checkpoint clamp complex pathway may play an important role in RA development. The genes in the cellular response to ionizing radiation may affect the effectiveness of anti-TNF therapy.

These pathways might offer a unique time-lapse window into the inflammatory arthritis process by which immune-related adverse events occur and predict or prevent them. They may also provide a unique window into the early occurrence of inflammatory arthritis in humans.

Table 1 and Figs 1–3 suggested that selected 69 genes might reveal the biological process of the treatment.

4. Discussion

A systemic inflammatory disease, RA is manifested by destructive distal polyarthritis. It can cause progressive joint damage, affect other organs, and even lead to cardiovascular disease unless diagnosed and treated. Targeted therapy using anti-TNF is the first option for patients with RA. Anti-TNF therapy, however, does not lead to meaningful clinical improvement in many RA patients. To predict which patients will not benefit from anti-TNF therapy, clinical tests should be performed prior to treatment beginning. Although various efforts have been made to identify biomarkers and pathways that may be helpful to predict the response to anti-TNF treatment, gaps remain in clinical use due to the low predictive power of the selected biomarkers. In this paper, we used a network-based computational method to identify the select the predictive biomarkers to guide the treatment of RA patients. We select 69 genes from peripheral blood expression data from 46 subjects using a sparse network-based method. The result shows that the selected 69 genes might influence biological processes and molecular functions related to the treatment.

5. Conclusion

Our approach advances the predictive power of anti-TNF therapy response and provides new genetic markers and pathways that may influence the treatment. One limitation of this paper is the lack of deep verification of the selected genes and network module.

Acknowledgments

This work was partially funded by the National Natural Science Foundation of China (62102261, 61976052), the Science and Technology Development Fund, Macau SAR (0056/2020/AFJ, 0158/2019/A3), the Jihua laboratory scientific project (X210101UZ210), the Foshan scientific project (2018AB003621), the School Moral Education Research project of Guangdong Education Department (2019GXSZ059), the Special Innovation Projects of Universities in Guangdong Province (2018KTSCX205), and the Science and Technology Project of Shaoguan City (200811104531028).

Conflict of interest

None to report.

References

[1] Singh JA, Saag KG, Bridges SL, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis & Rheumatology. 2016; 68(1): 1-26.

[2] van de Putte LBA, Atkins C, Malaise M, Sany J, Russell AS, van Riel PLCM, et al. Efficacy and safety of adalimumab as monotherapy in patients with rheumatoid arthritis for whom previous disease modifying antirheumatic drug treatment has failed. Annals of the Rheumatic Diseases. 2004; 63(5): 508-516.

[3] Mellors T, Withers JB, Ameli A, Jones A, Wang M, Zhang L, et al. Clinical Validation of a Blood-Based Predictive Test for Stratification of Response to Tumor Necrosis Factor Inhibitor Therapies in Rheumatoid Arthritis Patients. Network and Systems Medicine. 2020; 3(1): 91-104.

[4] Huang HH, Liang Y. A novel Cox proportional hazards model for high – dimensional genomic data in cancer prognosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021; 18(5): 1821-1830.

[5] Huang HH, Liang Y. An integrative analysis system of gene expression using self-paced learning and SCAD-Net. Expert Systems with Applications. Pergamon; 2019; 135: 102-112.

[6] Zhang W, Wan YW, Allen GI, Pang K, Anderson ML, Liu Z. Molecular pathway identification using biological network-regularized logistic models. BMC Genomics. 2013; 14(Suppl 8): S7-2164-14-S8-S7.

[7] Huang HH, Liang Y, Liu XY. Network-Based Logistic Classification with an Enhanced L1/2 Solver Reveals Biomarker and Subnetwork Signatures for Diagnosing Lung Cancer. BioMed Research International. 2015; 713953.

[8] Zhou Z, Huang H, Liang Y. Cancer classification and biomarker selection via a penalized logsum network-based logistic regression model. Technology and Health Care. 2021; 29(S1): 287-295.

[9] Huang HH, Liu XY, Li HM, Liang Y. Molecular pathway identification using a new L1/2 solver and biological network-constrained mode. International Journal of Data Mining and Bioinformatics. 2017; 17(3): 189.

[10] Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24(9): 1175-1182.

[11] Huang HH, Peng XD, Liang Y. SPLSN: An efficient tool for survival analysis and biomarker selection. International Journal of Intelligent Systems. 2021; (36): 5845-5865.

[12] Wang R, Su C, Wang X, Fu Q, Gao X, Zhang C, et al. Global gene expression analysis combined with a genomics approach for the identification of signal transduction networks involved in postnatal mouse myocardial proliferation and development. International Journal of Molecular Medicine. 2018; 41(1): 311-321.

[13] Chen J, Zhang S, Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics. 2016; 32(11): 1724-1732.

[14] Sharma A, Kitsak M, Cho MH, Ameli A, Zhou X, Jiang Z, et al. Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module. Scientific Reports. 2018; 8(1): 14439.

[15] Huang HH, Liu XY, Liang Y. Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2+2 Regularization. 2016; 11(5): e0149675.

[16] Liang Y, Liu C, Luan XZ, Leung KS, Chan TM, Xu ZB, et al. Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC Bioinformatics. 2013; 14: 198.

[17] Huang HH, Liang Y. Clinical drug response prediction by using a lq penalized network-constrained logistic regression method. Cellular Physiology and Biochemistry. 2018; 51(5): 2073-2084.

[18] Huang HH, Liang Y. Hybrid L1/2+2 method for gene selection in the Cox proportional hazards model. Computer Methods and Programs in Biomedicine. 2018; 164: 65-73.

[19] Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association. 2001; 96(456): 1348-1360.

[20] Bienkowska JR, Dalgin GS, Batliwalla F, Allaire N, Roubenoff R, Gregersen PK, et al. Convergent random forest predictor: Methodology for predicting drug response from genome-scale data applied to anti-TNF response. Genomics. 2009; 94(6): 423-432.

[21] Rui H, Yan T, Hu Z, Liu R, Wang L. The association between caspase-5 gene polymorphisms and rheumatoid arthritis in a Chinese population. Gene. 2018; 642: 307-312.

[22] Jiang X, Wang H, Li Z, Wei D, Yang Y, Zheng X, et al. A Monoclonal Antibody Against a Novel Sialomucin CD300LG. Monoclonal Antibodies in Immunodiagnosis and Immunotherapy. 2013; 32(2): 91-97.

[23] Braaten TJ, Brahmer JR, Forde PM, Le D, Lipson EJ, Naidoo J, et al. Immune checkpoint inhibitor-induced inflammatory arthritis persists after immunotherapy cessation. Annals of the Rheumatic Diseases. 2020; 79(3): 332-338.