Bio-Medical Materials and Engineering 26 (2015) S1837–S1843
DOI 10.3233/BME-151486
IOS Press

Identification of 13 blood-based gene expression signatures to accurately distinguish tuberculosis from other pulmonary diseases and healthy controls

Hai-Hui Huang, Xiao-Ying Liu, Yong Liang*, Hua Chai and Liang-Yong Xia
Macau University of Science and Technology, Taipa 999078, Macau, China

Abstract. Tuberculosis (TB), caused by infection with mycobacterium tuberculosis, is still a major threat to human health worldwide. Current diagnostic methods encounter some limitations, such as sample collection problem or unsatisfied sensitivity and specificity issue. Moreover, it is hard to identify TB from some of other lung diseases without invasive biopsy. In this paper, the logistic models with three representative regularization approaches including Lasso (the most popular regularization method), and $L_{1/2}$ (the method that inclines to achieve more sparse solution than Lasso) and Elastic Net (the method that encourages a grouping effect of genes in the results) adopted together to select the common gene signatures in microarray data of peripheral blood cells. As the result, 13 common gene signatures were selected, and sequentially the classifier based on them is constructed by the SVM approach, which can accurately distinguish tuberculosis from other pulmonary diseases and healthy controls. In the test and validation datasets of the blood gene expression profiles, the generated classification model achieved 91.86% sensitivity and 93.48% specificity averagely. Its sensitivity is improved 6%, but only 26% gene signatures used compared to recent research results. These 13 gene signatures selected by our methods can be used as the basis of a blood-based test for the detection of TB from other pulmonary diseases and healthy controls.

Keywords: Tuberculosis, feature selection, early diagnostic, regularization, biomarkers

1. Introduction

Tuberculosis (TB), caused by infection with mycobacterium tuberculosis, is still a major cause of morbidity and mortality all around the world. Early diagnosis is significant to improve results of treatment. Current TB diagnostic methods encounter many limitations. For example, sample collection issues (Sputum Tests, LAM, Xpert MTB/RIF) [1], limited sensitivity and specificity problem (urinary lipoarabinomannan test) [2, 3]. Moreover, sarcoidosis, community acquired pneumonia and primary lung cancer present the similarly phenotype of TB. Identifying TB from these diseases, the invasive biopsy is required generally. Therefore, an efficient and non-invasive tool to distinguish TB from other pulmonary conditions phenotype and healthy controls is urgently needed.

* Address for correspondence: Yong Liang, Macau University of Science and Technology, Taipa 999078, Macau, China. Tel.: 0853-8897-2034; Fax: 0853-2882-3280; E-mail: yliang@must.edu.mo.

Gene expression microarray data of peripheral blood cells are useful supplementary information for insight into the biological mechanisms of TB infections and could be a promising tool for early TB diagnostic [4-6]. However, the number of observations is much smaller than the number of measured biomarkers in most of the genomic studies. Such limitations are known as high dimensional and low samples problem that may lead to over-fitting and negatively influence the diagnostic performance in traditional statistical models. Regularization methods have been widely used in microarray data analysis in order to deal with the problem of high dimensionality. They are an important embedded technique and perform continuous shrinkage and gene selection simultaneously [7]. Here, we focus on the three representative regularization approaches: Lasso [8], $L_{1/2}$ [9], Elastic Net [10]. Lasso is the most popular regularization method in practices, which estimates as the convex optimization problems. $L_{1/2}$ can be taken as a representative of the non-convex Lq ($0 < q < 1$) penalties and inclines to achieve more sparse solution than Lasso theoretically. Elastic Net is the method that encourages a grouping effect of genes in the results. Each mono-method performs in its own mechanism and leads to different sparse results generally.

In this paper, the logistic models with the three representative regularization methods including Lasso, $L_{1/2}$, Elastic Net are proposed to select gene expression signatures in the microarray data of peripheral blood cells. After that, SVM method [11] is used to fit the classifier based on the commonly selected gene signatures by the regularization methods. Since TB is involved in multi-biological pathway [4, 5], and different regularization approach may select the gene signatures from different aspects (or solution paths) in disease phenotype, the common selected gene may participate in many biological pathways and play a critical role in those biological processes that explain the activity of disease. The results of our proposed approach in this paper reveal the compound regularization methods are extremely useful for the diagnostic of TB.

2. Material and method

2.1. Regularization

Assuming that dataset D has n samples, $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is i^{th} sample with p genes and y_i is the corresponding variable that takes a value of 0 or 1. Define a classifier $f(x) = e^x/(1+e^x)$ and the logistic regression is defined as:

$$\mathrm{P}(y_i=1|X_i) = f(X'_i\beta) = \frac{exp(X'_i\beta)}{1+exp(X'_i\beta)} \quad (1)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the coefficients to be estimated. We can obtain β by minimizing the log-likelihood function of the logistic regression, and then the logistic regression model is presented as:

$$L(\lambda, \beta) = -\sum_{i=1}^{n}\{ y_i log[\mathrm{f}(X'_i\beta)] + (1 - y_i)\log[1 - f(X'_i\beta)] \} \quad (2)$$

In high dimensional application with the number of genes p >> the sample size n, solving Eq. (2) directly is ill-posed and may lead to over-fitting problem. The regularization approaches are widely applied to address this high dimensional problem. When adding a regularization term in Eq. (2), the sparse logistic regression can be written as:

$$L(\lambda_1, \lambda_2, \beta) = -\sum_{i=1}^{n}\{ y_i log[\mathrm{f}(X'_i\beta)] + (1 - y_i)\log[1 - \mathrm{f}(X'_i\beta)] \} + \lambda \sum_{j=1}^{p} P(\beta_j) \quad (3)$$

where $\lambda > 0$ is a regularization parameter. Many regularization terms have been proposed in the recent years. For example, Lasso (L_1), which has the regularization term $P(\beta) = |\beta_j|^1$. Elastic Net ($L_1 + L_2$) with $P(\beta) = |\beta_j|^2 + |\beta_j|^1$, $L_{1/2}$ with $P(\beta) = |\beta_j|^{1/2}$.

Each regularization method has its own merits. For example, Lasso is a popular gene selection method, which enjoys convex, and some of the attractive properties of both subset selection and ridge regression [8]. $L_{1/2}$ can be taken as a representative of non-convex Lq ($0 < q < 1$) penalties and has demonstrated many attractive properties, such as unbiasedness, sparsity and oracle properties [9, 12]. Elastic Net emphasizes a grouping effect, where strongly correlated genes tend to be in or out of the model together, such as, able to select groups of correlated genes [10]. In other hands, TB diseases are complex and often associated with multi-biological pathway. When the different regularization methods try to capture the most significant genes to TB disease, the results often vary greatly. This means that the different mechanisms of the regularization methods may relate to the different biological pathway of TB disease and the selected genes by the different regularization methods may significantly express in different disease-related biological pathway. The common part of the selected genes is occupying important positions in those paths, which mean these common genes may play a critical role in those biological processes that explain the activity of disease. In this paper, we use the three representative regularization methods including Lasso, $L_{1/2}$, Elastic Net to select the common gene expression signatures in the microarray data of peripheral blood cells for TB disease.

2.2. Data descriptions

We follow Bloom, et al. [4] to organize and preprocess gene expression microarray data of peripheral blood cells and describe as following: i) Training set (n = 95) (GEO: GSE42830) includes TB patients (n = 16) and other samples (n = 79); ii) Test set (n=102) (GEO: GSE42826) includes TB patients (n = 11) and other samples (n = 91); iii) Validation set (n = 42) (GEO: GSE42825) includes TB patients (n = 8) and other samples (n = 34). These three datasets come from the same microarray, Illumina HumanHT-12 V4.0 expression beadchip. The pathological characteristics of these three datasets were summarized in Table 1. Each datasets was divided into TB group (TB patient; Label 1) and non-TB group (includes sarcoidosis, pneumonia, lung cancer, healthy controls; Label 0).

Table 1

The pathological characteristics of datasets in this paper

		Other Lung Diseases				
	TB	Sarcoidosis	Pneumonia	LC	HC	TOTAL
Training Set	16	25	8	8	38	95
Test Set	11	25	6	8	52	102
Validation Set	8	11	-	-	23	42
Summation	35	61	14	16	113	239

Note: LC: Lung cancer; HC: Healthy controls.

3. Results and discussion

3.1. Gene signature selecting

The gene signatures were selected from the training set using the logistic models with Lasso, $L_{1/2}$, Elastic Net approaches respectively. 10 fold cross-validation was estimated for tuning the penalty parameters in Lasso and $L_{1/2}$ approaches. The penalty parameters of Elastic Net were selected using two dimensional surfaces cross-validation [10]. The numbers of gene signatures selected by these three regularization models are: Lasso (g=88), $L_{1/2}$ (g = 63), Elastic Net (g = 875) respectively. In the literature, Blo**om, et al. [4] reported that they proposed 144 gene signatures to distinguish TB from other lung dise**ase and healthy controls. Similar works include that 76 genes (recognized by the officer gene symbol) signatures in Maertzdorf, et al. [5] and 50 genes in Koth, et al. [6] were proposed to discriminate TB from other lung diseases.

The regularized logistic methods were compared to the above three studies [4-6] and the results were summarized in Table 2. All these methods were developed on the training set and evaluated on the test and the validation sets. The results of Bloom, et al. [4] come from their publication, in which the classifier was built by support vector machines (SVM) with 144 gene signatures. For comparison, the classification models of Maertzdorf, et al. and Koth, et al. were also built by SVM approach using their reported gene signatures in the literature [5, 6].

As showed in Table 2, the signatures selected by regularized logistic methods outperform the other three sets of signatures, because they achieved the best performance in every dataset. Such as, the $L_{1/2}$ method achieved 100% sensitivity in the training set. Elastic Net achieved the highest sensitivity in the test set. Lasso achieved 100% sensitivity in the validation set. Moreover, the Lasso approach achieved the best specificity in the training set. The $L_{1/2}$ approach achieved 95.60% and 94.12% specificity in test and validation sets respectively, which are the best performance amongst these two datasets compared to other methods. The classifier with the 144 gene signatures proposed by Bloom *et al.* to distinguish TB from other lung disease, showed a lower sensitivity (82–88%), though similar specificity (> 90%). The classifier with 76 genes (recognized by the officer gene symbol) signatures suggested by Maertzdorf *et al.* achieved a much lower sensitivity (45–56%), though similar specificity (> 90%). The classifier with 50 genes was shown to be differentially expressed in TB and sarcoidosis studied by Koth, et al. also resulted in a much lower sensitivity (45–75%) and lower specificity (87-92%).

Table 2

The discrimination results of all the methods

Method	Sensitivity				Specificity			
	Training	Test	Validation	Average	Training	Test	Validation	Average
Lasso	100.00%	72.73%	100.00%	90.91%	100.00%	90.11%	91.18%	93.76%
$L_{1/2}$	100.00%	81.82%	87.50%	89.77%	97.47%	95.60%	94.12%	95.73%
Elastic Net	93.75%	90.91%	87.50%	90.72%	94.94%	90.11%	91.18%	92.07%
Bloom, et al.	88.00%	82.00%	88.00%	86.00%	94.00%	91.00%	92.00%	92.33%
Maertzdorf, et al.	56.00%	45.00%	75.00%	58.67%	96.00%	92.00%	92.00%	93.33%
Koth, et al.	75.00%	45.00%	50.00%	56.67%	92.00%	87.00%	92.00%	90.33%

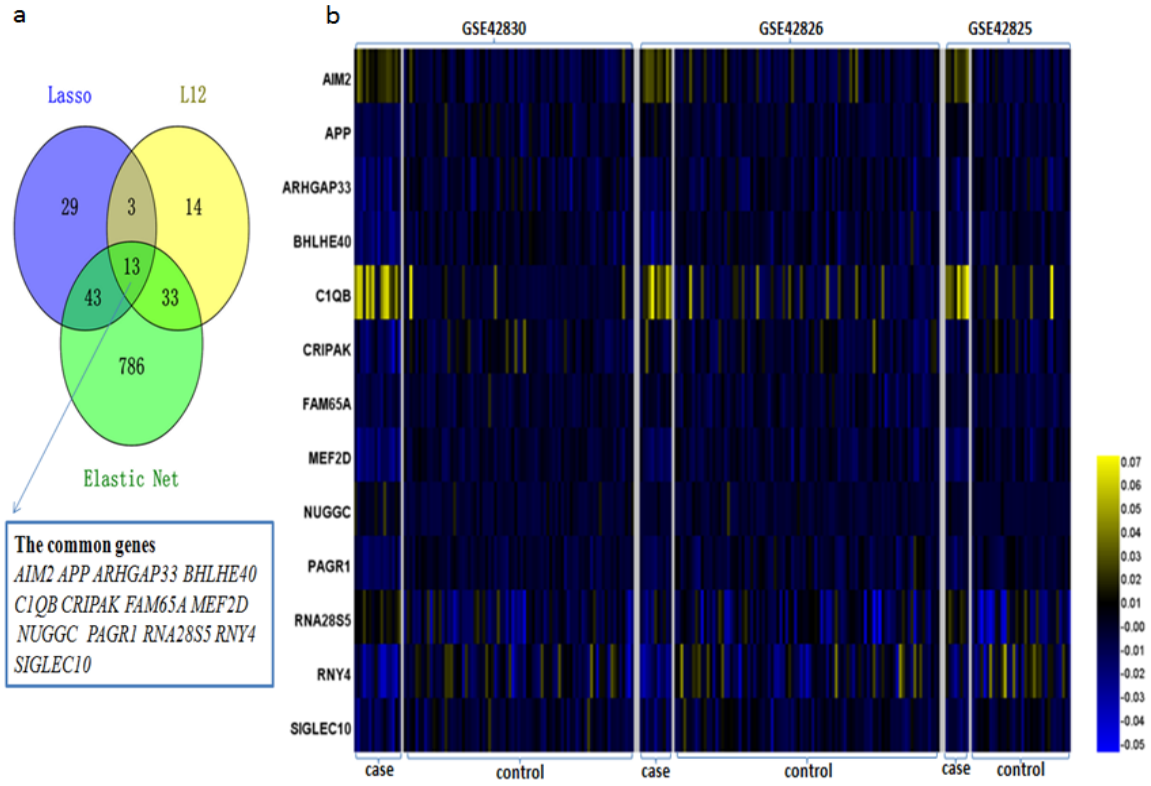

Fig. 1. The Venn diagram and heat map of the gene signatures selected by Lasso, $L_{1/2}$ and Elastic Net approaches. (a) The Venn diagram analysis of the results of Lasso, $L_{1/2}$ and Elastic Net methods. (b) The heat map of the common genes showed in (a) for datasets GSE42830, GSE42826 and GSE42825. Case: tuberculosis, control: non- tuberculosis.

3.2. Common gene signatures analyses

In this section, we consider the common gene signatures selected by the logistic models with Lasso, $L_{1/2}$, Elastic Net regularization methods, which are the most relevant signatures of TB disease. Hence, 13 common genes were found in these methods and described in Figure 1(a). Figure 1(b) showed a significant difference of the 13-signature gene expression between TB and non-TB in training set (GSE42830), test set (GSE42826) and validation set (GSE42825) using heat map analysis. For example, the expression of gene AIM2 in TB (case) patients are much higher than the expression in non-TB (control) patients.

The classifier based on these 13 commonly selected gene signatures was built by the SVM approach to fit the gene expression data in training set firstly. Then the model was evaluated on the test and validation sets. The results of the classifier performance were represented in Table 3. Overall, the classifier with 13 gene signatures achieved 91.86% sensitivity and 93.48% specificity averagely (the cutoff point is 0.075), in which its sensitivity (91.86%) is much better than the average sensitivity of other methods: 86.00% of Bloom, et al. model with 144 gene signatures, 58.67% of Maertzdorf, et al. model with 76 gene signatures and 56.67% of Koth, et al. model with 50 gene signatures. The overall specificity of the classifier with 13 gene signature is 93.48% and also outperforms the three methods (90.33-93.33%). On the other hand, the classifier with 13 gene signatures also competitive when

Table 3

The discrimination results of 13 gene signatures using the SVM approach

Method	Sensitivity				Specificity			
	Training	Test	Validation	Average	Training	Test	Validation	Average
13-gene	93.75%	81.82%	100.00%	91.86%	96.20%	90.11%	94.12%	93.48%

compared to the three regularization methods (Lasso, $L_{1/2}$, and Elastic Net). As showed in Tables 2 and 3, the average sensitivity (91.86%) and the average specificity (93.48%) of the 13-gene signature classifier are higher than the average sensitivity 90.72% and the average specificity 92.07% obtained by the Elastic Net approach. Moreover, the number of gene signatures selected by the Elastic Net was 875, which is over 67 times more than the 13 gene signatures. The average specificity 95.73% of the $L_{1/2}$ approach is better than the 93.48% of the 13-gene classifier; however, the average sensitivity 91.86% of the 13-gene classifier is better than the sensitivity 89.77% obtained by the $L_{1/2}$ approach. Besides, 63 gene signatures selected by the $L_{1/2}$ approach is much more than the 13 gene signatures used in the classifier. Hence, these common gene signatures selected by Lasso, $L_{1/2}$ and Elastic Net approaches were the core factors of molecular diagnostics for TB disease.

There are several biological studied that associated with these 13 gene signatures. For example, absent in melanoma 2 (AIM2) is an indicator of cytology DNA that is accounted for host immune responses to DNA viruses and intracellular bacteria. And AIM2 have been proved plays an important role in Mycobacterium tuberculosis infection [13]. Complement 1qb (C1qb) gene could be a potential diagnostic marker to discriminate active tuberculosis from latent tuberculosis infection as well as tuberculosis pleurisy from non-tuberculosis pleurisy [14]. Mef2d, be known as a key factor in muscle development, energy storage and immune responses, and have been suggested may play a part in the process of tuberculosis infection [15].

4. Conclusion

In this paper, the logistic models with three representative regularization approaches including Lasso, $L_{1/2}$, and Elastic net adopted together to select the common gene signatures in microarray data of peripheral blood cells. As the result, 13 common gene signatures were selected, and sequentially the classifier based on them is constructed by the SVM approach, which can be accurately distinguishing tuberculosis from other pulmonary diseases and healthy controls. This 13-gene signature model achieved an impressive performance of discriminating TB from other pulmonary diseases and healthy controls. The results in this paper have moved forward the clinical use of blood-based TB diagnostic.

Acknowledgment

The work described in this paper was supported by STDF of Macau 099/2013/A3.

References

[1] P.K. Mehta, A. Raj, N. Singh and G.K. Khuller, Diagnosis of extrapolating tuberculosis by PCR, FEMS Immunology & Medical Microbiology **66** (2012), 20-36.
[2] M.X. Rangaka, K.A. Wilkinson, J.R. Glynn, D. Ling, D. Menzies, J. Mwansa-Kambafwile and M. Pai, Predictive value of interferon-γ release assays for incident active tuberculosis: A systematic review and meta-analysis, The Lancet Infectious Diseases **12** (2012), 45-55.
[3] L. Fan, Z. Chen, X.H. Hao, Z.Y. Hu and H.P. Xiao, Interferon−gamma release assays for the diagnosis of extrapulmonary tuberculosis: A systematic review and meta−analysis, FEMS Immunology & Medical Microbiology **65** (2012), 456-466.
[4] C.I. Bloom, C.M. Graham, M.P. Berry, F. Rozakeas, P.S. Redford, Y. Wang and A. O'Garra, Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers, PloS One **8** (2013), e70630.
[5] J. Maertzdorf, J. Weiner, H.J. Mollenkopf, T. Network, T. Bauer, A. Prasse and S. Stenger, Common patterns and disease-related signatures in tuberculosis and sarcoidosis, Proceedings of the National Academy of Sciences **109** (2012), 7853-7858.
[6] L.L. Koth, O.D. Solberg, J.C. Peng, N.R. Bhakta, C.P. Nguyen and P.G. Woodruff, Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis, American Journal of Respiratory and Critical Care Medicine, **184** (2011), 1153-1163.
[7] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research **3** (2003), 1157-1182.
[8] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B (Methodological) **58** (1996), 267-288.
[9] Z.B. Xu, H. Zhang, Y. Wang, X.Y. Chang and Y. Liang, L1/2 regularization, Science China Series F **40** (2010), 1-11.
[10] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), 301-320.
[11] T. Joachims, Making large scale SVM learning practical, Technical Report 0943-4235, Universität Dortmund, 1999.
[12] B. Zhang, H. Chai, Z. Yang, Y. Liang, G. Chu and X. Liu, Application of L 1/2 regularization logistic method in heart disease diagnosis, Bio-Medical Materials and Engineering **24** (2014), 3447-3454.
[13] H. Saiga, S. Kitada, Y. Shimada, N. Kamiyama, M. Okuyama, M. Makino and K. Takeda, Critical role of AIM2 in mycobacterium tuberculosis infection, International Immunology **24** (2012), 637-644.
[14] Y. Cai, Q. Yang, Y. Tang, M. Zhang, H. Liu, G. Zhang and X. Chen, Increased complement C1q level marks active disease in human tuberculosis, PloS One **9** (2014), e92340.
[15] R.I. Clark, S.W. Tan, C.B. Péan, U. Roostalu, V. Vivancos, K. Bronda and M.S. Dionne, MEF2 is an in vivo immune-metabolic switch, Cell **155** (2013), 435-447.