MINING PROTEIN REGULATORY RELATIONSHIPS USING NEURAL NETWORK METHODS FOR EARLY PREDICTION OF SARS

HONG-QIANG WANG  
Intelligent Computing Lab, Hefei Institute of Intelligent Machines,  
Chinese Academy of Science, Hefei, China  
Research Institute of Innovative Products and Technologies,  
The Hong Kong Polytechnic University, Hung Hom, Hong Kong  
horace.wang@yahoo.com.cn

HAI LONG ZHU  
Research Institute of Innovative Products and Technologies,  
The Hong Kong Polytechnic University, Hung Hom, Hong Kong  
rhltzhu@polyu.edu.hk

CHUNHOU ZHENG  
Intelligent Computing Lab, Hefei Institute of Intelligent Machines,  
Chinese Academy of Science, Hefei, China  
chzheng@iim.ac.cn

TIMOTHY T. C. YIP, WILLIAM C. S. CHO and STEPHEN C. K. LAW  
Department of Clinical Oncology, Queen Elizabeth Hospital,  
Hong Kong

This paper proposes to model protein regulation networks associated with severe acute respiratory syndrome (SARS) for early prediction of SARS. In the approach, specific to a patient group, a regulatory network is simulated using a fully-connected neural network and is optimized towards minimizing a novel energy function that is defined as a measure of disagreement between the input and output of the network. The nonlinear version of the network is achieved by applying a sigmoid function. Experimental results show that the proposed approaches can capture regulatory patterns associated with SARS and efficiently implement early prediction of SARS.

Keywords: Neural network; classification algorithms; regulatory network; SARS.

1. Introduction

A new strain of coronavirus (CoV) caused a pandemic outbreak of severe acute respiratory syndrome (SARS) in mainland China, Hong Kong, Singapore, Toronto, and Taiwan during 2002 and 2003, with 8,098 individuals being infected and 774 deaths. The pneumonia and accompanying adult respiratory distress syndrome can
be rapidly progressive and can lead to death. Despite many advances in virologic studies, early diagnosis of SARS has been based primarily on the clinical definitions released by WHO and CDC, which can be confusing or contradictory. Available serologic tests cannot guarantee an early diagnosis, and PCR-based molecular detection of the viral RNA suffers from unsatisfactory sensitivity and specificity. There is a need to establish a reliable diagnostic methodology for SARS-CoV, in particular, to distinguish the similar clinical manifestations of SARS and other respiratory tract infections.

Recently, there are some researchers who used ProteinChip array technology to investigate protein expression levels of SARS. The ProteinChip array technology is a novel high-throughput technique that can simultaneously interrogate thousands of proteins. The resulting protein expression data give important clues to studying molecules and their regulation relationship. However, analyzing ProteinChip data is a challenging task due to its small-sample character and unclear micromolecular activity mechanisms. Various computational approaches have been developed to analyze this type of data to discover biomarkers and their regulations. For example, Bayesian networks (BN) have been applied to analyze gene regulations in a probabilistic manner; Curtis et al. proposed a modular regulation analysis method; Yeung et al. presented the dominant spectral component technique for discovering transcriptional regulations. Qiu et al. proposed an ensemble dependence model (EDM) to model linear dependence relationships among gene clustering centers.

Exploring proteomic regulatory networks based on ProteinChip data is important to understanding and early prediction of SARS. In this paper, we propose a novel modeling approach to explore regulation networks associated with SARS. In the approach, underlying regulatory network (RN) is modeled as a fully-connected neural network. The RN is optimized towards minimizing a novel energy function that is defined as the disagreement between the network input and output. A nonlinear regulatory network (NLRN) is developed by applying a sigmoid function to the linear regulation network (LRN). Based on protein expression data provided by Yip et al., we apply our approach to construct regulatory networks associated with SARS and establish a RN-based classifier for early prediction of SARS. Experimental results suggest that the network models can effectively capture regulatory patterns associated with SARS and achieve the better performance of distinguishing SARS from normal or other respiratory diseases than the previous classification methods.

2. Construction and Optimization of Regulatory Networks

The regulatory network (RN) is constructed by employing a hopfield network-like structure to highlight the mutual relationships among biomarkers, as shown in Fig. 1. Each node represents a biomarker and arrow lines denote the regulatory relationships. Assume that there are $p$ biomarkers whose expression levels are denoted
by a vector $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$ and the regulation matrix of the RN is denoted as $A = \{a_{ij} \in \mathbb{R}; i, j = 1, 2, \ldots, p\}$ with the element $a_{ij}$ representing the regulatory coefficient from the $j$th node to the $i$th node. From a biological viewpoint, positive regulatory coefficients represent expression promotion while negative ones represent expression repression. Let $\mathbf{x}$ and $\mathbf{y} = [y_1, y_2, \ldots, y_p]^T$ be the input and output of the network, a linear regulatory network can be modeled as

\[
\begin{align*}
  y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p + b_1 \\
  y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p + b_2 \\
  &\vdots \\
  y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p + b_p,
\end{align*}
\]

or

\[\mathbf{y} = A\mathbf{x} + \mathbf{b}, \quad \mathbf{b} = [b_1, b_2, \ldots, b_p]^T, \quad (1)\]

where $b_i, i = 1, 2, \ldots, p$, is a constant bias.

For the matrix $A$, since regulation of a biomarker to itself is meaningless, the following constraint is imposed:

\[a_{ii} = 0, \quad i = 1, 2, \ldots, p. \quad (2)\]

Although some classic neural networks (such as Hopfield network) assume the connection matrix is symmetrical, no assumption is made to the regulation matrix of the RN because two biomarkers incline to regulate each other with different strengths from a biological viewpoint.
A critical issue of building the RN is how to characterize and analyze the RN. An energy function is defined as the measurement of the disagreement between the network input and output as follows:

\[ E = \frac{1}{2} (y - x)^T (y - x) . \] (3)

By substituting Eq. (1) into Eq. (3), the energy function can be rewritten as the following form:

\[ E = \frac{1}{2} (Ax + B - x)^T (Ax + B - x) . \] (4)

The definition of the energy function indicates that lower energy status corresponds to higher agreement between input and output. If high agreement state remains true for all of the samples in a specific patient group, then the regulation matrix of the network can reflect the consistent patterns of biomarker inter-regulations specific to the patient group. The unique patterns of regulations can be used for SARS classification. What follows presents the obtainment and optimization of the regulation matrix.

Let \( X = [x_1, x_2, \ldots, x_l] \) be a set of \( l \) biomarker expression observations of a patient group. The optimal \( A \) can be achieved through minimizing the energy sum over all the observations, i.e.,

\[
\text{Minimize } f = E(A, X) = \frac{1}{2} \sum_{j=1}^{l} ((A - I)x_j + B)^T ((A - I)x_j + B),
\] (5)

where \( I \) is the identity matrix. Let \( \hat{A} = (A - I) \), the objective function of Eq. (5) can be rewritten as

\[
f = \frac{1}{2} \sum_{x \in X} (\hat{A}x + B)^T (\hat{A}x + B),
\] (6)

where the diagonal elements \( \hat{a}_{ii} \) of \( \hat{A} \) satisfy

\[
\hat{a}_{ii} = -1 .
\] (7)

By rewriting \( \hat{A} \) as \([\hat{A}_1; \hat{A}_2; \ldots; \hat{A}_p]\), Eq. (6) is expanded to the following form:

\[
f = \frac{1}{2} (\hat{A}_1X + b_1e)(\hat{A}_1X + b_1e)^T + \frac{1}{2} (\hat{A}_2X + b_2e)(\hat{A}_2X + b_2e)^T
\]
\[
+ \cdots + \frac{1}{2} (\hat{A}_pX + b_pe)(\hat{A}_pX + b_pe)^T ,
\] (8)

where \( e \) is a \( l \)-dimensional row vector whose elements equal one. The present work only considers a simple case of zero bias, i.e., \( b_i = 0, i = 1, 2, \ldots, p \). In this case, the objective function becomes

\[
f = \frac{1}{2} (\hat{A}_1X)(\hat{A}_1X)^T + \frac{1}{2} (\hat{A}_2X)(\hat{A}_2X)^T + \cdots + \frac{1}{2} (\hat{A}_pX)(\hat{A}_pX)^T .
\] (9)
By rewriting the \(i\)th row vector of \(X\) as \(U_i = [x_{i1}, x_{i2}, \ldots, x_{il}]^T\) and creating \(Z_i = \{x_{jk}; k = 1, \ldots, i-1, i+1, \ldots, p, j = 1, \ldots, l\}\), Eq. (9) is then converted into

\[
f = \frac{1}{2}(U_1 - Z_1 \phi_1)^T(U_1 - Z_1 \phi_1) + \cdots + \frac{1}{2}(U_p - Z_p \phi_p)^T(U_p - Z_p \phi_p),
\]

(10)

where \(\phi_i = [a_{i1}, a_{i2}, \ldots, a_{i(i-1)}, a_{i(i+1)}, \ldots, a_{ip}]^T\). Let the derivatives of the objective function in Eq. (10) toward \(\phi_i\) be zero, we have

\[
\frac{\partial f}{\partial \phi_i} = -Z_i^T U_i + Z_i^T Z_i \phi_i = 0, \quad i = 1, 2, \ldots, p.
\]

(11)

By solving Eq. (11), the regulatory coefficients can be obtained as follows:

\[
\phi_i = (Z_i^T Z_i)^{-1} Z_i^T U_i, \quad i = 1, 2, \ldots, p.
\]

(12)

In case that the inverse of \(Z_i^T Z_i\) does not exist, the pseudo-inverse can be used to compute the solution.

To reduce noise in data, a nonlinear regulatory network (NLRN) is constructed by adding a sigmoid-kernel transformation unit to the linear RN (LRN). The unit takes \(x\) as input and outputs the following vector:

\[
v = [v_1, v_2, \ldots, v_p]^T, \quad v_i = f(x_i) = \left(1 + e^{-\beta(\frac{x_i - \mu_i}{\sigma_i})^2}\right)^{-1}, \quad i = 1, 2, \ldots, p,
\]

(13)

where \(\beta \in (0, 1]\) is a tunable sigmoid parameter, \(\mu_i = (\sum x_{ij})/n\) and \(\sigma_i = \sqrt{(\sum (x_{ij} - \mu_i)^2)/(n - 1)}\), are the mean and standard deviation of the expression data of \(i\)th biomarkers. For the NLRN, in terms of Eqs. (4) and (13), the corresponding energy function becomes

\[
E = \frac{1}{2}(Av + B - v)^T(Av + B - v).
\]

(14)

Similar to the linear network, the nonlinear network can be optimized through choosing proper network parameters, \(A\) and \(B\), and the sigmoid parameter \(\beta\).

3. SARS Diagnosis with the Regulatory Network

Since the regulatory network is optimized towards minimizing its energy function, the regulatory matrix reserves the unique regulatory pattern of a particular group of patients. Given an input belonging to the group, the network will approach a low energy status. So, the regulatory network can be used to predict and diagnosis SARS. Assuming that there are totally \(G\) groups, based on the corresponding \(G\) regulatory networks, a classification function can be designed as follows:

\[
D_u = \arg\{\min_g(E_g(u), \ g = 1, 2, \ldots, G)\},
\]

(15)

where \(u\) represents a test sample, \(E_g\) denotes the energy function of the \(g\)th network, and \(D_u \in \{1, 2, \ldots, G\}\) indicates the predicted class label. Figure 2 illustrates the framework of the classifier.
4. Experimental Results

Based on the ProteinChip array profiling dataset of SARS provided by Yip et al., the regulatory pattern of SARS is explored by our proposed approach, and a classifier is constructed for early prediction of SARS. The SARS dataset contains 74 samples, each consisting of 103 protein biomarkers. The 74 samples are labeled into three categories: 10 normal (NRL), 20 influenza-infected (IFZ), 44 SARS. In this experiment, the dataset is randomly split to a training set (48 samples) and an independent test set (26 samples).

Since the dataset contains too many protein biomarkers, we firstly used the regulation probability method to select significantly differentially expressed proteins to reduce the computation burden of seeking for best regulatory networks. As a result, twenty significant proteins with higher significance (<0.01) were picked. From the twenty proteins, different numbers $p$ of proteins were combined to construct RN classifiers to track their classification performances. Since smaller-scale networks are computationally more preferable, $p$ was initially set to 2 and then was gradually increased. For each combination, both the LRN and NLRN models were applied to implement linear and nonlinear RN classifiers, respectively. For the NLRN models, the parameter $\beta$ of the sigmoid function is optimized within 0.001, 0.01, 0.05, 0.1 and 0.5. Table 1 lists the best leave-one-out-cross-validation (LOOCV) accuracies on the training set and the corresponding accuracies on the testing set for different $p$. From Table 1, it can be seen that the accuracies of both LRN and NLRN increase with more proteins involved, suggesting that more biomarkers provide more comprehensive regulatory information which can strengthen the classification power. Table 1 also suggests that nonlinear classifiers (NLRN) tend to have better classification performances than linear classifiers (LRN), irrespective of the values of $p$. 

![Fig. 2. Classification framework based on regulatory networks.](image)
Table 1. Classification performance of linear and nonlinear regulatory networks.

<table>
<thead>
<tr>
<th></th>
<th>NLRN</th>
<th>LRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOCV</td>
<td>Test</td>
<td>LOOCV</td>
</tr>
<tr>
<td>$p = 2$</td>
<td>0.854</td>
<td>0.770</td>
</tr>
<tr>
<td>$p = 3$</td>
<td>0.937</td>
<td>0.808</td>
</tr>
<tr>
<td>$p = 4$</td>
<td>0.937</td>
<td>0.846</td>
</tr>
<tr>
<td>$p = 5$</td>
<td>0.958</td>
<td>0.885</td>
</tr>
<tr>
<td>$p = 6$</td>
<td>0.958</td>
<td>0.885</td>
</tr>
</tbody>
</table>

Fig. 3. Comparison of the expression levels of the 5 proteins used by the 5-biomarker NLRN classifier in NRL, IFZ and SARS groups.

The testing accuracies of the NLRN reach its peak value (0.885) when 5 biomarkers are used. Figure 3 illustrates the expression levels of the 5 proteins used by the 5-protein NLRN classifier, which are remarkably different in NRL, IFZ and SARS groups.

To gain deep insight into the regulatory patterns associated with the three groups, NRL, IFZ and SARS, Fig. 4 illustrates the three nonlinear regulatory networks encapsulated in the 5-protein NLRN classifier. From Fig. 4, it can be found that the connection metrics of the three networks are remarkably different, suggesting the regulation patterns change in the three groups. For example, in SARS group, regulatory coefficients between protein C05909_7 and C08989_3 are 0.23 and
0.30, suggesting mutual promotion between the two proteins; while in NRL and IFZ groups, their regulatory coefficients become negative, suggesting mutual repression between them. Further comparing NRN and IFZ groups, it can be found that the repressions in the two groups have different strengths: distinctly stronger in IFZ group than those in NRL group. Another significant example is, for SARS group, proteins C03278 and C08131 repress each other with the magnitudes of $-0.01$ and $0.03$; while for the other two groups, they have strong mutual promotions. Viewing the regulatory coefficient as the regulation level, Fig. 5(a) compares the difference of the regulation levels of the three classes on each regulatory channel in a bar diagram form, and Fig. 5(b) shows the corresponding three regulatory spectra.
Fig. 5. Comparison of the obtained regulatory spectra of the three classes: Normal, Influenza-infected, and SARS.

on the whole channel range. The two figures more clearly indicate the remarkable difference of the regulatory patterns of the three classes. In summary, these obtained regulatory patterns are crucial to the performance of the NLRN classifier by disclosing the differences of regulation patterns and, clinically, are promising to help in revealing the molecular mechanism of SARS.
Table 2. Classification performance of SVMs.

<table>
<thead>
<tr>
<th></th>
<th>Rbf-SVM LOOCV</th>
<th>Test</th>
<th>Linear-SVM LOOCV</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 2</td>
<td>0.833</td>
<td>0.704</td>
<td>0.833</td>
<td>0.731</td>
</tr>
<tr>
<td>p = 3</td>
<td>0.833</td>
<td>0.704</td>
<td>0.833</td>
<td>0.731</td>
</tr>
<tr>
<td>p = 4</td>
<td>0.791</td>
<td>0.808</td>
<td>0.771</td>
<td>0.731</td>
</tr>
<tr>
<td>p = 5</td>
<td>0.854</td>
<td>0.846</td>
<td>0.750</td>
<td>0.808</td>
</tr>
<tr>
<td>p = 6</td>
<td>0.818</td>
<td>0.846</td>
<td>0.771</td>
<td>0.808</td>
</tr>
</tbody>
</table>

SVMs including linear SVM and rbf SVM, which have proven to have superior classification performance in various application areas, were compared with our RN approaches. For the rbf-SVM, the regularization parameter and width parameter were optimized through two-dimension grid searching. Table 2 lists the classification accuracies of the two SVMs, showing that the best testing accuracy of the SVMs is 0.846 which is achieved by the rbf-SVM using 6 proteins. The result is lower than that (0.885) of our NLRN, as shown in Table 1. We also compared the computation cost of the SVMs and our algorithm. The comparison shows that the CPU times of the LRN and NLRN (0.017 and 0.066) are less than those of the linear-SVM and rbf-SVM (0.091 and 0.103). The running environment is Pentium® D CPU 3.4 GHz, 1GB RAM and Matlab 7.0.

5. Conclusions and Discussions

In this paper, we have proposed a novel approach for extracting regulation networks associated with SARS. In brief, the approach constructs the regulation network as a fully-connected neural network, in which nodes represent proteins and links indicate regulations between proteins. By minimizing an energy function which is defined as a measure of the disagreement between the input and output of the network, the regulation network model can be optimally solved. Based on the idea, initially, we modeled the regulation networks of SARS in terms of linearity. Using 6 proteins, the classification accuracy of the linear network on an independent test set is up to 80.8%. Although linear relationships are in general simple and easy to model, they are unnecessarily reliable and efficient to capture underlying regulatory patterns. By applying a nonlinear transformation unit, we construct nonlinear regulation networks based on the linear network. The experiments show that the nonlinear networks are more effective to capture the regulation networks associated with SARS. In particular, a 5-protein nonlinear network recognizes the regulation patterns of SARS that are significantly different from those of other groups. The corresponding nonlinear classifier classifies the three patient groups with the accuracy of up to 88.5% that is far better than those of the linear- and rbf-SVMs.

Another important advantage of our approach is that it can identify biological aberrance of SARS. From a biological viewpoint, SARS is associated with deviant
regulatory patterns of a set of related proteins/genes. The proposed RN model contributes toward the discovery of new proteomic or genomic biomarkers and the underlying abnormal regulatory patterns. Future works will include designing different nonlinear transformations and network optimization methods to obtain a better understanding of the underlying regulatory network of SARS.

Acknowledgments
This work was supported by the grants of the National Science Foundation of China, No. 30700161, China Postdoctoral Science Foundation, No. 20070410223.

References