AN ARTIFICIAL NEURAL NETWORK EVALUATION OF TUBERCULOSIS USING GENETIC AND PHYSIOLOGICAL PATIENT DATA

William O. Griffin[1], Josh Hanna[2], Svetlana Razorilova [4], Mikhael Kitaev [4], Avtandiil Alisherov [4], Jerry A. Darsey[1], and Olga Tarasenko [3]
(1-Department of Chemistry, 2-Department of Bioinformatics, 3-Department of Biology
University of Arkansas at Little Rock, Little Rock, AR, USA;
4-National Center of Tuberculosis, Bishkek, Kyrgyz Republic)

ABSTRACT: When doctors see more cases of patients with tell-tale symptoms of a disease, it is hoped that they will be able to recognize an infection and administer treatment appropriately, thereby speeding up recovery for sick patients. We hope that our studies can aid in the detection of tuberculosis by using a computer model called an artificial neural network. Our model looks at patients with and without tuberculosis (TB). The data that the neural network examined came from the following: patient' age, gender, place of birth, blood type, Rhesus (Rh) factor, and genes of the Human Leukocyte Antigens (HLA) system (9q34.1) present in the Major Histocompatibility Complex . With availability in genetic data and good research, we hope to give them an advantage in the detection of tuberculosis. We try to mimic the doctor's experience with a computer test, which will learn from patient data the factors that contribute to TB.

INTRODUCTION

An artificial neural network (ANN) is modeled like a biological neuron; it takes many numbers and can use pattern recognition to predict an output from those numbers. It uses a training dataset to learn, and an unseen value(s) set to test the ANN's ability to predict (Devillers, 1996; Jepson et al., 1993; El-Solh et al., 1999). The program Nets we used for the prediction was developed by NASA's Johnson Space Center (Shelton and Baffes, 1989; Sumpter and Noid,1996). Cross-validation is needed and is performed using the leave-one-out technique. Leave-one-out is where one or several test cases are removed from the dataset in order to see if the neural network has learned from repeated learning cycles of the dataset. Passing the value which was withheld to the neural network, improves the predictive ability over several learning cycles.

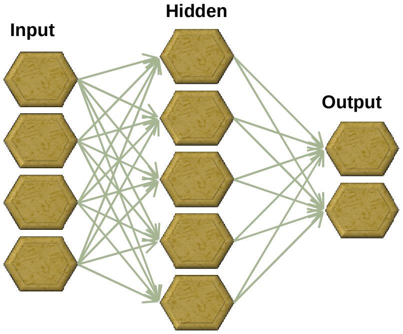

FIGURE 1. Data is sent through a network as inputs; it passes through a hidden layer and exits through an output layer. The layers are connected by weights which are represented as arrows in the figure. The weights start off as random values, but change as the

CP1229, *Proceedings of the 4th BioNanoTox Conference,* edited by O. Tarasenko

network “learns.” The output through the learning cycles will become closer to the actual value, by minimizing error, through adjusting the weights. The feed-forward backpropagation neural network is the most common type of ANN used. 95% of ANN are of this configuration (Matsuki et al., 2002).

The neural network works better than conventional programs, because it can see patterns in data which are not readily recognizable. However, the nature of this data had many clinical types of tuberculosis (TB). To predict outcome, we first needed to see if it could predict whether or not TB was predictable given the input data. We next would predict whether or not the patient had developed, the infiltrative (primary) and chronic type of TB.

We use a fully-connected, feed forward, back-propagation, and neural network model. The neural network passes information forward (feed-forward) through the three layers of the neural network. As shown in Figure 1, starting at the input layer, it moves through a hidden layer, and finally passes to the output layer. After each learning cycle, it evaluates the error of the data, and returns to the connections of the layers (weights) and adjusts them in order to generate a lower error on the next learning cycle (back-propagation).

METHODS

We prepared data for the study by normalizing to values between 0.95 and 0.05, utilizing a data transformation equation:

$$x''_i = [(y''_{max}-y''_{min})(x_i - x_{min})/(x_{max} - x_{min})] + y''_{min} \qquad 1$$

This equation converts the data to a desired value range of $y''_{max} = 0.95$ and $y''_{min} = 0.05$. The x_{max} and x_{min} values are calculated from the dataset maximum and minimum values respectively. The majority of the data is of a yes/no variety (i.e. yes, the gene is present, yes the patient is of an AB blood type), which facilitates the majority of the data to values of maximum 0.95 or minimum 0.05. The output values are also scaled to this range.

There were 149 healthy individuals used as control with no TB whose genetic makeup had been identified. There were 145 patients with diagnosed TB. Of the 145 TB patients, 71 had chronic, and 63 had infiltrative TB based on clinical and laboratory results. The data of the patients which was fed into the neural network is given in Table 1. Each patient has 103 values to give, be it physiological or genetic with an output (TB+, TB-, or a type of TB). There were 90 nodes in the hidden layer.

TABLE 1. Data type for inclusion in the artificial neural network

Physiological Data	Age (above or below 40) Place of birth (South or North region of Kyrgyz Republic)
Genetic Data	Gender (male or female) Blood type (A,B, AB, and O) Rhesus (Rh) factor 95 HLA genes from the Major Histocompatibility Complex (presence or absence)
Output (TB type)	Chronic or infiltrative type of TB (presence of one didn’t necessarily preclude the other, and there were cases with neither)

This study looked at the data through two neural networks. The first neural network used trained on a set of data in order to predict the patients' TB status. A group of 30 patients' data was withheld from the training, and the neural network attempted to predict whether they were from the TB positive set or the control group with no TB. This groups both patients' results, but we wanted to attempt to find more detail in the data. This prompted the design for the second neural network.

The second neural network trained on the same patients' data, but was used for a different purpose. It found which input factors had the most bearing on the prediction of infiltrative or chronic TB. After finding a stable, high-learning range for the training data from the first neural network, the second neural network was used to predict the input factors driving selection of chronic and infiltrative TB. This selection was based on the weight values of the connections between the layers of the neural network.

Once the network had been trained, the weights of the network were extracted. For each input, the weight of the connection from it to a hidden node was multiplied with the weight of the connection from the same hidden node to an output node. Once this was done for each hidden node connected to that particular input, the values were summed. This result is the total importance the network placed on the input for that particular output.

Once that had been done for each input-output pair, the weights were extracted. The lower half of the weights that were less than their median value were removed in order to show the most important selection factors for and against susceptibility to infiltrative or chronic TB.

RESULTS

The graphical results are shown in Figures 2 and 3. The neural network trained to predict the patients with and without TB in Figure 2. The training performance is shown in the Y-axis as the percentage correct and the cycle is shown in the x-axis in cycle/100 which is plotted on the x-axis. In Figure 3, the results of weights extracted from patients who had infiltrative and chronic TB, the selection factors have been identified (this is not clearly stated).

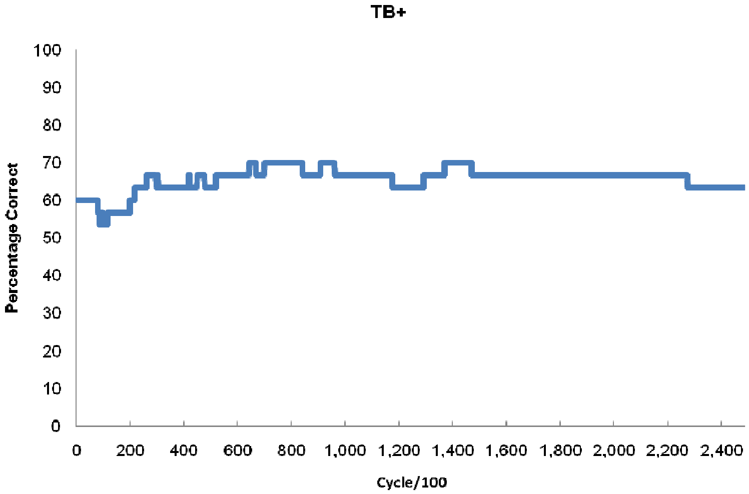

FIGURE 2. The neural network data shown here includes both TB (+) and (-) cases

Identifying where to select for boundaries of the output is a critical part of evaluating a neural network output. Since an output close to 0.95 is a yes and 0.05 is a no, we have to define a range for yes, no, and fuzzy. We decided where the values are in a "fuzzy" range, and where it gives clear yes/no values by distinguishing the data output from each other(for example in the scaled range of our data between 0.05 and 0.95, 0.05-0.3 is a no, and 0.7-0.95 is yes, 0.3-0.7 is fuzzy.) The majority gave clear values, but ambiguous or fuzzy values were not reportable, and contributed to the ambiguity of the network.

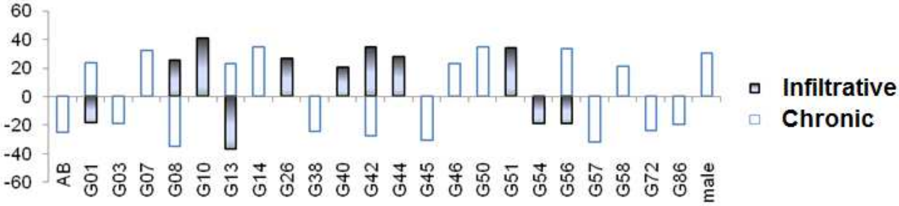

FIGURE 3. The Neural Network identified that the presence or absence of blood type (AB), the gender, and several HLA genes (G) contributed to whether the patients would develop infiltrative or chronical TB. The y-axis is the susceptibility factor, which has no units, but is derived from the weights of the neural network and shows by sign whether it is contributing or detracting from susceptibility to infiltrative (primary) or chronic TB.

Several selection factors for the neural network may prove useful to researchers trying to identify which genes may predispose a person to a type of TB. Susceptibility may have something to do with behaviors also, and we did not have that much data on factors that lead to behavior to give the network, but being a male was shown as a positive influence for chronic TB.

DISCUSSION

Figure 2 demonstrates that the network has fairly consistent learning for almost the entire range of training. The plateau area from 150,000 to 220,000 appears to be a particularly stable area where the learning has stabilized. However, 1,400,000 appears to be it's peak of learning percentage at 70%. This learning value was used to perform the learning of the second network which had weights extracted to use for figure 3.

Several genes, gene 1, gene 8, gene 13, gene 42, and gene 56 show they have opposing contributing and detracting effects to chronic and infiltrative TB. Other factors show strongly contributing to or detracting from susceptibility to primary and chronical TB. Larger datasets, and more input factors for the patients make the predictions of neural network better.

Additionally, the work that will come from the initial study will be evaluated against other studies for testing this methods which fall in the realm of quantitative structure-property relationship analysis.

CONCLUSION

TB affects millions of people and can be fatal. This study used artificial neural networks to predict susceptibility of patients to clinical TB using two configurations. The accuracy of the first network was as high as 70%, while the other extracted selection factors for and against the susceptibility for chronical and infiltrative tuberculosis. It is hoped that more use of the predictive power of neural networks can aid doctors in identifying genes as well as conditions that put others at risk of developing TB.

ACKNOWLEDGEMENTS

NIH grant number P20 RR-16460 (INBRE program)

REFERENCES

1. Devillers, J. 1996. “Strengths and Weaknesses of the Backpropagation Neural Network in QSAR and QSPR Studies” Neural Networks in QSAR and Drug Design. Academic Press Limited, San Diego, CA.
2. El-Solh, A. A., Hsiao, C-B., Goodnough, S., Serghani, J., and Grant, B.J.B. 1999. “Predicting Active Pulmonary Tuberculosis Using an Artificial Neural Network*” Chest.* 16, 4, 968-973.
3. Jepson, B., Collins, A., and Evans, A. 1993. “Post-Neural Network Procedure to Determine Expected Prediction Values and their Confidence Limits” *Neural Comput. and Applic.* Vol. 1, 224-228.
4. Matsuki, Y., Nakamura, K., Watanabe, H., Aoki, T., Nakata, H., Katsuragawa, S., and Doi, K. 2002 “Usefulness of an Artificial Neural Network for Differentiating Benign from Malignant Pulmonary Nodules on High-Resolution CT” *Am. J. Roentgenol.* 178:657-663.
5. Shelton R. and Baffes, P.T. 1989 “Nets Back-Propagation ver. 4.0: Software Technology Branch”, NASA, Johnson Space Center, Houston, TX.
6. Sumpter, B. and Noid, D. 1996. “On the design, analysis, and characterization of materials using computational neural networks” *Annu. Rev. Mater. Sci.* 26: 223-277.