Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS) 2 (1): 74-78
© Scholarlink Research Institute Journals, 2011 (ISSN: 2141-7016)
jeteas.scholarlinkresearch.org

Estimating An Optimal Backpropagation Algorithm for Training An ANN with the EGFR Exon 19 Nucleotide Sequence: An Electronic Diagnostic Basis for Non–Small Cell Lung Cancer(NSCLC)

[1]E. Adetiba, [1]J. C. Ekeh, [1]V. O. Matthews, [1]S.A. Daramola, [2]M.E.U. Eleanya

[1]Department of Electrical and Information Engineering,College of Science and Technology,School of Engineering, Covenant University,Ota, Nigeria.
[2]Chevron, 2,Chevron Drive, Lekki Peninsula, Lagos, Nigeria.
Corresponding Author: E. Adetiba

Abstract
One of the most common forms of medical malpractices globally is an error in diagnosis. An improper diagnosis occurs when a doctor fails to identify a disease or report a disease when the patient is actually healthy. A disease that is commonly misdiagnosed is lung cancer. This cancer type is a major health problem internationally because it is responsible for 15% of all cancer diagnosis and 29% of all cancer deaths. The two major sub-types of lung cancer are; small cell lung cancer (about 13%) and non-small cell lung cancer (NSCLC- about 87%). The chance of surviving lung cancer depends on its correct diagnosis and/or the stage at the time it is diagnosed. However, recent studies have identified somatic mutations in the epidermal growth factor receptor (EGFR) gene in a subset of non-small cell lung cancer (NSCLC) tumors. These mutations occur in the tyrosine kinase domain of the gene. The most predominant of the mutations in all NSCLC patients examined is deletion mutation in exon 19 and it accounts for approximately 90% of the EGFR-activating mutations. This makes EGFR genomic sequence a good candidate for implementing an electronic diagnostic system for NSCLC. In this study aimed at estimating an optimum backpropagation training algorithm for a genomic based ANN system for NSCLC diagnosis, the nucleotide sequences of EGFR's exon 19 of a non-cancerous cell were used to train an artificial neural network (ANN). Several ANN back propagation training algorithms were tested in MATLAB R2008a to obtain an optimal algorithm for training the network. Of the nine different algorithms tested, we achieved the best performance (i.e. the least mean square error) with the minimum epoch (training iterations) and training time using the Levenberg-Marquardt algorithm.

Keywords: ANN, NSCLC, EGFR, Lung Cancer, Diagnosis, Exon 19

INTRODUCTION

Since the cells in the human body need oxygen to work and grow, the rate at which human breathes is controlled by the brain. The brain is quick to sense changes in oxygen concentration and it does this because it is its biggest user and the first to suffer if there is a shortage. On the average, a human breathes nearly 25,000 times daily which provides around 11,000 liters of air, 21% of which is oxygen (Human Lung, 2010). In order to purify the air intake, the lungs have cleaning systems that traps, ejects or destroys irritants and other harmful substances that travel in with the air (Lung, 2010). Based on this, the genome within lung cells is exposed to mutagens and suffers mistakes in replication. This results in the divergence of the DNA sequence in each lung cancer cell from that constituted in the original cell. The somatic mutations alter the function of a critical gene, confers growth advantage to the cell in which it occurred and results in an expanded clone derived from this cell. These mutinous cells evolve as a result of waves of clonal expansions and forms a mass of tissue called a growth or tumor. Tumor cells can be benign (not cancer) or malignant (cancer). Benign tumor cells are usually not as harmful as malignant (or cancer) tumor cells (Pleasance et al., 2010; WTSI, 2010).

In recent years, the development of massively parallel sequencing technologies makes it feasible to catalogue all classes of somatically acquired mutation in a cancer. These mutations may include base substitution, insertions & deletions (indels), copy number changes and genomic rearrangement (Pleasance et al., 2010). However, existing literatures on oncogenomics (the genomics of lung cancer) suggest that Epidermal Growth Factor Receptor (EGFR) gene's mutations can be used to identify patient with non-small cell cancer in whom EGFR is essential to tumor growth (Thomas, 2004). Figure 1 shows the position of mutations in exon 19 and missense mutations in exon 21 of the EGFR gene in seven patients with non-small cell lung cancer (NSCLC).

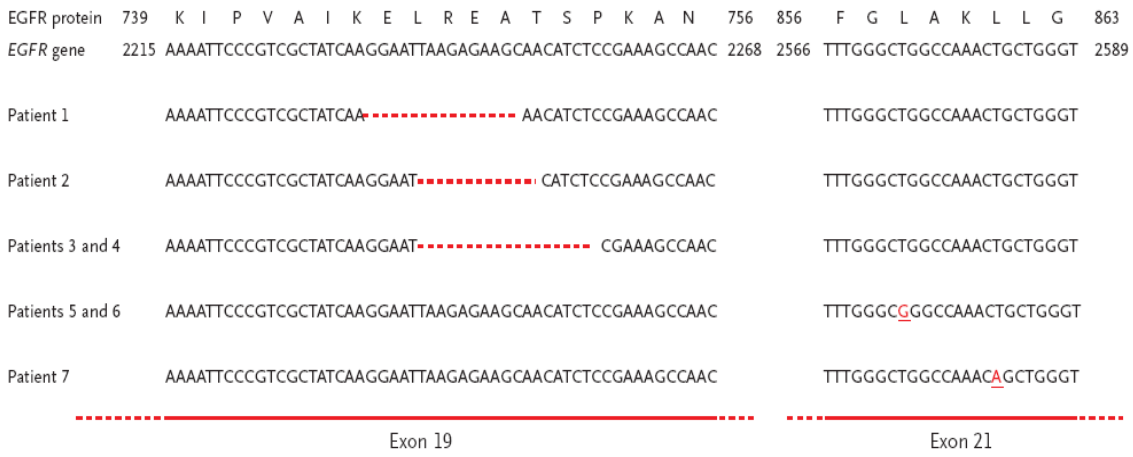

Figure 1: Mutations in the EGFR nucleotide sequence of non-small cell lung cancer (NSCLC) patients (Source: The New England Journal of Medicine, May 20, 2004, Vol. 350, No. 21)

These mutational sequences can form a reliable basis for the medical diagnosis of non-small cell lung cancer using artificial neural network (ANN).

ANN is a massively parallel computational model that simulates the structure and the functional aspects of biological nervous system. The brain which is central to the nervous system is made up of discrete units called neurons (the Greek word for nerves). Neurons are polarized cells that receive signals via highly branched extensions, called dendrites and send information along un-branched extensions, called axons. In total, the human brain contains approximately 10^{14} to 10^{15} interconnections of neurons. All neurons process information in much the same way and information within neurons are transmitted in the form of electrical impulses called action potentials via the axons from other neuron cells. When the action potential arrives at the axon terminal, the neuron releases chemical neurotransmitter which effects the interneuron communication at specialized connections called synapses.

In biological systems, learning involves adjustments to the synaptic connections between these neurons. Figure 2 shows the anatomy of a biological neuron.

Artificial neuron was inspired from the structure and functions of the biological neuron and figure 3 shows the structure and components of an artificial neuron. An artificial neuron functionality and properties such as its flexibility and ability to approximate functions to be learned depend on its activation function. Some important activation functions are linear, sigmoid, threshold and hyperbolic tangent activations. A neuron learns through an iterative process of adjustment of its synaptic weights and a neuron become more knowledgeable after each iteration of the learning process.

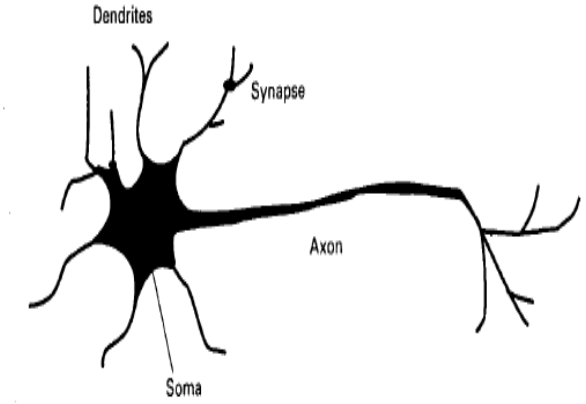

Figure 2: Biological Neuron Anatomy

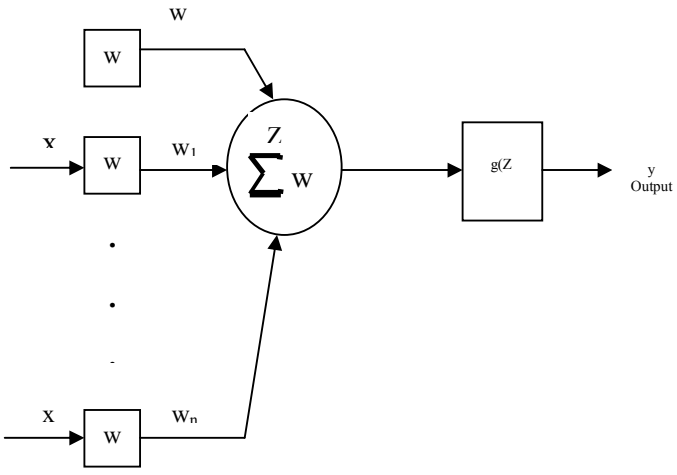

Figure 3: The Structure of an Artificial Neuron

The functionality of a single artificial neuron which is the basic unit of an artificial neural system is very limited. To learn how to solve a problem that cannot be learned by a single neuron, an interconnection of multiple neurons called Neural Network (NN) or

Artificial Neural Network (ANN) must be employed. Figure 4 shows the simplest Artificial Neural Network (ANN).

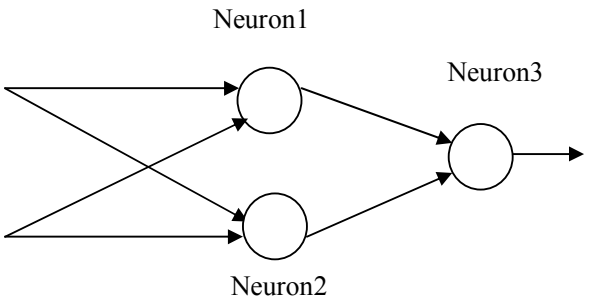

Figure 4: The Simplest Artificial Neural Network (ANN)

Currently, ANN is being hugely applied in medicine in such areas as radiology, cardiology, oncology and etc. Specifically, Zhi-Hua Zhou et al (2002) designed an automatic pathological diagnosis procedure named Neural Ensemble based Detection (NED) which utilizes ANN ensemble to identify cancer cells in the images of the specimens of needle biopsies obtained from the patient to be diagnosed. A very recent work by Gutte et al (2007) developed a completely automated method based on image processing techniques and artificial neural networks (ANN) for the interpretation combined fluorodeoxyglucose (FDG) positron emission tomography(PET) and computed tomography(CT) images for the diagnosis and staging of lung cancer. Several other ANN based diagnostic systems for lung cancer by (Chiou et al.,1993; Mori et al.,1996; Lin et al.,1995; Hayashibe et al.,1996; Cooper et al., 1991) utilize complex image processing techniques which requires high computational power but are grossly limited in precision. However in our study, a novel approach which is a genomics based ANN is utilized.

MATERIALS AND METHODS

MATLAB R2008a Neural Network Toolbox is utilized in this work because it contains various functions and backpropagation training algorithms for implementing feed forward neural networks. There are various backpropagation training algorithms with diverse capability based on the nature of problem the network is designed to solve. Generally, ANN can be trained for tasks such as function approximation (regression) or pattern recognition (discriminant analysis). Our problem falls under pattern recognition in which the network is trained with the nucleotide sequence of a non-cancerous patient (see fig 1). This is aimed at obtaining the optimal algorithms out of the nine available backpropagation training algorithms in MATLAB R2008a. It is actually very difficult to know which training algorithm will be the fastest for a given problem. It depends on many factors,

Table 1.0: Normalized EGRP Gene (Exon 19 Nucleotide Sequences)

S/N	Nucleotide Positions	Nucleotide Codes	ASCII Equivalent	Normalized Codes
1	2215	A	65	-1.0000
2	2216	A	65	-1.0000
3	2217	A	65	-1.0000
4	2218	A	65	-1.0000
5	2219	T	84	1.0000
6	2220	T	84	1.0000
7	2221	C	67	-0.7895
8	2222	C	67	-0.7895
9	2223	C	67	-0.7895
10	2224	G	71	-0.3684
11	2225	T	84	1.0000
12	2226	C	67	-0.7895
13	2227	G	71	-0.3684
14	2228	C	67	-0.7895
15	2229	T	84	1.0000
16	2230	A	65	-1.0000
17	2231	T	84	1.0000
18	2232	C	67	-0.7895
19	2233	A	65	-1.0000
20	2234	A	65	-1.0000
21	2235	G	71	-0.3684
22	2236	G	71	-0.3684
23	2237	A	65	-1.0000
24	2238	A	65	-1.0000
25	2239	T	84	1.0000
26	2240	T	84	1.0000
27	2241	A	65	-1.0000
28	2242	A	65	-1.0000
29	2243	G	71	-0.3684
30	2244	A	65	-1.0000
31	2245	G	71	-0.3684
32	2246	A	65	-1.0000
33	2247	A	65	-1.0000
34	2248	G	71	-0.3684
35	2249	C	67	-0.7895
36	2250	A	65	-1.0000
37	2251	A	65	-1.0000
38	2252	C	67	-0.7895
39	2253	A	65	-1.0000
40	2254	T	84	1.0000
41	2255	C	67	-0.7895
42	2256	T	84	1.0000
43	2257	C	67	-0.7895
44	2258	C	67	-0.7895
45	2259	G	71	-0.3684
46	2260	A	65	-1.0000
47	2261	A	65	-1.0000
48	2262	A	65	-1.0000
49	2263	G	71	-0.3684
50	2264	C	67	-0.7895
51	2265	C	67	-0.7895
52	2266	A	65	-1.0000
53	2267	A	65	-1.0000
54	2268	C	67	-0.7895

including the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, the error goal and the task the network is to perform (Howard et.al, 2010). The steps we adopted are detailed as follows:

Step1: Encoding of the EGRP Gene Exon 19 nucleotide sequence (54 nucleotides). The original sequence is alphabetical and since ANN takes numerical input and perform best with inputs in the range [-1 1], the alphabets were first encoded using ASCII and subsequently normalized between [-1 1] with MATLAB m-file. The normalized code is shown in Table 1.0.

Step 2: Creation of the sample input and target vectors for training the network from the normalized sequences in Table 1.0.
Step 3: Selection of the set of parameters to train the network. The main parameters are number of epochs (iterations), training time, performance goal (i.e. Mean Square Error (MSE)**)** and gradients. Other parameters are the defaults for each of the training algorithms.
Step 4: The network was trained based on a set of activation functions(tansig and purelin), number of neurons and training algorithm.
Step 5: Each of the nine backpropagation training algorithms was used for training the network and the changes in the network parameters recorded.
Step 6: The effectiveness of each of the backpropagation training algorithms for the task was obtained based on the least possible Mean Square Error (MSE) and training time.
The network architecture for our genomic based ANN as created by the MATLAB R2008a wizard is shown in fig. 5.

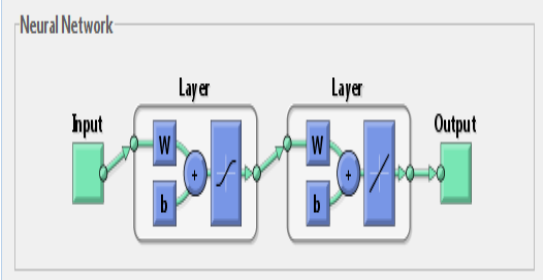

Figure 5: Generated ANN Architecture in MATLAB R2008a Wizard

RESULTS AND DISCUSSION

The network weights(W) and biases(b) are first initialised before the ANN can be ready for training. The initialisation is done automatically by MATLAB. The training process requires a set of network inputs(p) and target outputs(t). During training, the weights and biases of the network are iteratively adjusted to minimize the network performance function. The default performance function for feedforward network which is the adopted topology for this work is Mean Square Error(MSE). MSE is the averaged squared error between the network outputs(a) and the target outputs(t). All the backpropagation training algorithms that are used for training our network utilises the gradient of MSE to determine how to adjust the weights to minimize performance. Our ANN is a multi-layered feed forward network and we utilised 20 neurons in the hidden layer for all the nine algorithms. The results/outputs of the network which comprises of the epoch, time, performance (mse) and gradient are shown in Table 2.0.
From the table, it can be observed that Levenberg-Marquardt training algorithm with the training function "trainlm" has the best perfomance of 2.3e-23 with 3 epochs and 3 seconds training time. The performance of the other algorithms obtained from the training platform(MATLAB R2008a) are as shown on the table.

Table 2.0: Training Results for the Different Backpropagation Algorithms.

S/N	Training Algorithm	Training Function	Epoch (iteration)	Time (s)	Performance (mse)	Gradient
1	Levenberg-Marquardt	trainlm	3	3	2.3e-23	7.5e-12
2	BFGS Quasi-Newton	Trainbfg	9	4	4.95e-20	4.77e-10
3	Resilient Backpropagation	Trainrp	77	29	2.56e-22	6.64e-11
4	Scaled Conjugate Gradient	Trainscg	23	9	6.67e-13	2.70e-07
5	Conjugate Gradient with Powell/Beale Restarts	Traincgb	9	4	1.98e-10	3.56e-05
6	Fletcher –Powell Conjugate Gradient	Traincgf	16	8	6.20e-09	3.85e-04
7	Polak-Ribiere Conjugate Gradient	Traincgp	15	7	2.55e-08	1.6e-03
8	One Step Secant	Trainoss	1000	4m 38s	7.52e-13	4.17e-06
9	Variable Learning Rate Backpropation	Traingdx	21	8	0.202	0.642

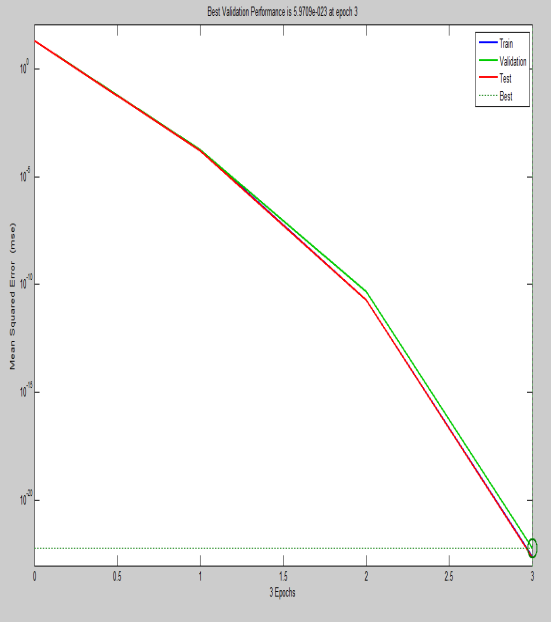

Figure 6: Performance Graph for Levenberg Marquardt(trainlm) ANN Backpropagation Training Algorithm

CONCLUSION

Existing literatures reveal that hitertoe, computer based lung cancer diagnostics systems are implemented with image processing techniques. This requires a lot of computational power and several researches are ongoing to enhance this approach. This work lays the foundation for a novel approach in implementing lung cancer diagnostics system which is based on genomics and ANN. Since the efficient training of an ANN within an optimum time frame is always a major requirement specification, we have been able to empirically obtain a backpropagation training algorithm that meets this requirement(i.e. the Levenberg-Marquardt(trainlm) algorithm).

REFERENCES

Chiou YSP, Lure YMF, Ligomerides P.A. (1993) 'Neural network image analysis and classification in hybrid lung nodule detection(HLND) system', Proceedings of the IEEE-SP Workshop on Neural Networks for Signal Processing, Pp.517-526.

Cooper L.N. (1991) 'Hybrid neural network architectures: equilibrium systems that pay attention', Neural Networks: Theory and Applications, San Diego, CA: Academic Press, Pp. 81-96.

Gutte Henrik, Jacobsson David, Olofsson Fredrik, Ohisson Mattias, Valind Sven, Loft Annika, Edenbrandt Lars, Kjaer Andreas(2007) 'Automated Interpretation of PET/CT images in patients with lung cancer' Nuclear Medicine Communications, Vol.28, No.2.

Hayashibe R., Asano N., Hirohata H., Okumura K., Kondo S., Handa S., Takizawa M., Sone S., Oshita S.(1996) 'An automatic lung cancer detection from X-ray images obtained through yearly serial mass survey', Proceedings of the International Conference on Image Processing, 1996.

Howard Demuth, Mark Beale, Martin Hagan (2010), Neural Network Toolbox™ 6 User's Guide. Available online at www.mathworks.com (accessed on 5 October, 2010).

Human Lung (2010). Available online at: http://www.wikipedia.com/wiki/Human_lung (accessed on 7 October,2010).

Lin J.S., Lo S.C.B., Hasegawa A., Freedman M.T., Mun S.K.(1995). Reduction of false positives in lung nodule detection using a two-level neural classification. IEEE Trans. Medical Imaging, Vol.15, No. 2, Pp. 206-217.

Lung(2010). Available online at: http://www.wikipedia.com/wiki/Lung. (accessed on 7 October, 2010).

Mori K., Hasegawa J., Toriwaki J., Anno H., Katada K.(1996) 'Recognition of bronchus in three-dimensional X-ray CT images with applications to virtualized bronchoscopy system', Proceedings of the 13th International Conference on Pattern Recognition, Vol. 3, Pp. 528-532.

Pleasance E.D., Stephen P.J., O'Meara S. et al.(2010) 'A Small-Cell Lung Cancer genome with complex signatures of tobacco exposure', Nature, Vol. 463, Pp. 184-190.

Thomas J. L.(2004) 'What you need to know about lung cancer', U.S. Department of Health and Human Services', National Institute of Health.

WTSI (Wellcome Trust Sanger Institute) (2010) 'Cataloguing cancer codes; International Cancer Genome Consortium plans to sequence 25,000 cancer genomes'. Available online at: http://www.sanger.ac.uk/about/press/2010/100414.html(accessed on 8 October,2010).

Zhi-Hua Zhou, Yuan Jiang, Yu-Bin Yang, Shi-Fu Chen(2002) 'Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles', Artificial Intelligence in Medicine, Vol.24, No.1, Pp.25-36.