Comparative Analysis using Bayesian Approach to Neural Network of Translational Initiation Sites in Alternative Polymorphic Context

Nurul Arneida Husin[1,2*], Nanna Suryana Herman[1], Burairah Hussin[1]

[1]*Faculty of Information and Communication Technology*
Universiti Teknikal Malaysia Melaka (UTeM)
Durian Tunggal, Melaka, Malaysia
Email: nsuryana@utem.edu.my, burairah@utem.edu.my

[2]*Genetics Laboratory, School of Science*
Monash University Sunway Campus
Jalan Lagoon Selatan, Bandar Sunway, 46150
Selangor Darul Ehsan, Malaysia
Email: nurul.arnieda@monash.edu

[*]*Corresponding author*

Received: November 21, 2011 **Accepted: December 27, 2011**

Published: January 31, 2012

Abstract: *Widely accepted as an important signal for gene discovery, translation initiation sites (TIS) in weak context has been the main focus in this paper. Many TIS prediction programs have been developed for optimal context, but they fail to successfully predict the start codon if the contexts conditions are in weak positions. The objectives of this paper are to develop useful algorithms and to build a new classification model for the case study. The first approach of neural network includes training on algorithms of Resilient Backpropagation, Scaled Conjugate Gradient Backpropagation and Levenberg-Marquardt. The outputs are used in comparison with Bayesian Neural Network for efficiency comparison. The results showed that Resilient Backpropagation have the consistency in all measurement but performs less in accuracy. In second approach, the Bayesian Classifier_01 outperforms the Resilient Backpropagation by successfully increasing the overall prediction accuracy by 16.0%. The Bayesian Classifier_02 is built to improve the accuracy by adding new features of chemical properties as selected by the Information Gain Ratio method, and increasing the length of the window sequence to 201. The result shows that the built model successfully increases the accuracy by 96.0%. In comparison, the Bayesian model outperforms Tikole and Sankararamakrishnan (2008) by increasing the sensitivity by 10% and specificity by 26%.*

Keywords*: Bayesian classifier, Neural network, Translation initiation sites, Weak context, Information Gain Ratio, Classification algorithms.*

Introduction

Translation initiation site (TIS) is a complex and regulated gene annotation process that involving a large number of polypeptides and a network of protein-protein interactions that can be occurred in mRNA, cDNA or other types of genomic sequences. It is part of protein biosynthesis that requisite for the accurate delineation of protein sequences from transcript data [8, 13].

In this paper, research efforts are focused on optimizing the supervised automated learning methods to correctly predicted TIS in weak contexts of mRNA sequences with a minimal error of rates. The specialty of the TIS in the weak or suboptimal context is that it proves to contribute as an alternative way of initiation of protein translators that could be sites for new

functional genes [3, 12]. At the translation stage of the protein synthesis process, in eukaryotic mRNA, the context of the start codon (normally "AUG") and the sequences around it are crucial for recruitment of the small ribosome subunit [1, 4].

However, in certain TIS cases, the initiation sites will occur in a weak (suboptimal) context especially in eukaryotic of higher complex organisms [5]. The characterization of the features around TIS will be helpful with a better understanding of translation regulation and accurate gene prediction of coding region in genomic and mRNA or cDNA sequences. This is an important step in genomic analysis to determine protein coding from nucleotide sequences [6, 7]. Study conducts by Tikole & Sankararamakrishnan [14] has successfully predicted the human mRNA sequences in weak contexts using neural networks (NNs) with 83% sensitivity and 73% specificity.

Materials and methods

The dataset

Two data sets are collected for this study. The first list of data set (Data set I) is provided by Dr. R. Sankararamakrishnan from Laboratory of Computational Biology, Indian Institute of Technology, India. It consists of 526 human mRNA sequences of Ref Seq which is originally extracted from NCBI GenBank database (Release 14). The second data set (Data set II) is formed on EST data by self-extracting a number of well-characterized and annotated sequences from NCBI GeneBank database. The resulting set consists of 400 non-redundant EST data from human sequences. 200 sequences are extracted for negative samples while another 200 sequences for positive samples.

Learning of neural network

The total set of 526 instances composing the supervised training set (input vectors and target outputs), have been randomly divided into 70% training and 30% testing data set. Inputs are presented to the networks by encoding the sequences into a binary string to encode the sequences. Each position in the vector corresponds to unique nucleotides that is set to bit 1 and all other bits are cleared to 0, and in spatial pattern where: A = 1000, T = 0100, G = 0010 and C = 0001. Table 1 introduces the sliding window frames used for this study. The number indicates the number of the DNA nucleotides in the respected frame. n represents any arbitrary of DNA nucleotides (A or T or G or C). While, UP represents the upstream position and DOWN represents the downstream positions of the context. In order to prevent the overfeeding of the built network, the window frames in this study are restricted to be 13, 14, 19, 23 and 24.

Table 1. Window frame specification

Window frame	Sequence pattern
13 (–5 to +8)	5(UPn)ATG5(DOWNn)
14 (–10 to +4)	10(UPn)ATG1(DOWNn)
19 (–15 to +4)	15(UPn)ATG1(DOWNn)
23 (–10 to +13)	23(UPn)ATG10(DOWNn)
24 (–20 to +4)	20(UPn)ATG1(DOWNn)

MATLAB toolbox using the feed forward nets of various architectures are designed and tested on the available data set. The input size of the input layer of the neural network varied depending on the number of nucleotides that are used as the context surrounding the

translation initiation sites. The sizes of the hidden layer are varied, and for model selection different hidden units are examined along with the implementation of the network. The output layer contains one neuron. A binary output of the output neuron is used to indicate the recognition site. An output close to 1 indicates the putative translation initiation sites. Training data are used to train the networks exclusively. The performance of the networks is recorded each time neurons are introduced in the hidden layer. The log sigmoid transfer function is the choice of the activation function. Table 2 shows the selected learning algorithms which are incorporated with the feed forward neural network. The programming environment is MATLAB 7.11 (R2010b).

Table 2. The applied learning algorithms

Label	**Learning algorithm**	**Algorithm**
[Bp-RP]	Resilient Backpropagation	trainrp
[Bp-SCG]	Scaled Conjugate Gradient Backpropagation	trainscg
[Bp-LM]	Levenberg-Marquardt	trainlm

To improve the generalization of the built network, early stopping and regularization is employed in this case study. In early stopping, the dataset is divided into three sets which are training, test and validation. In each epoch or training step, the network is equipped with training and validated with a test set. Validation sets are used to determine the performance of the network after the generalization process.

In the stopping criterion, caution should be taken in account to prevent early convergence in fast training algorithms such as in Levenberg-Marquadrt (LM). Thus, the parameters are set to 1 for Marquardt adjustment parameter (*mu*), decrease *mu* factor for 0.8 and increases *mu* factor for 1.5.

In regularization, the method is employed to improve the generalization process. The performance functions are modified, which are normally in the mean sum of squares of the network error (Eq. (1)) and the mean sum of squares of the network weights and biases (Eq. (2)).

$$F = mse = \frac{1}{N}\sum_{i=1}^{N}\left(e_i\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(t_i - a_i\right)^2 \quad (1)$$

$$msereg = \gamma mse + \left(1-\gamma\right)msw, \quad (2)$$

where γ is the performance ratio and *msereg* is the performance function for regularization.

By employing the performance function, the network has smaller weights and biases, and these forces the network to respond to be smoother and less overfitting.

$$msw = \frac{1}{N}\sum_{j=1}^{N} w_j^{\,2}, \quad (3)$$

where *msw* is the mean of the sum of squares of the network weight and biases.

Bayesian learning

Two classifiers of Bayesian Network is developed namely Bayesian Classifier_01 and Bayesian Classifier_02. All the dataset used in Bayesian learning are presented in the high level of features. The Bayesian Classifier_01 is a Bayesian neural network with 5 hidden units trained with different window frames of 13, 14, 19, 23 and 24. This classifier includes the main features used which a positional context and the features of stop codons. The activation functions applied is sigmoid function.

The Bayesian Classifier_02 is a Bayesian neural network with 5 hidden units trained with window frames of 201 and an additional of new feature sets of chemical properties as selected by information gain ratio (IGR) method. The applied IGR method that is based on feature ranking measures the attributes with many values than those with few values. IGR solves this problem by introducing an extra term on how the features split the data.

Performances

The programming environment is simulated using MATLAB 7.11 (R2010b). All machine learning calculations is performed on a personal computer running Windows XP with Pentium (R) Dual-Core CPU T4200 @ 2.00 GHz and 2GB of RAM. For target output comparison, confusion matrices, Matthews correlation coefficient (MMC), and mean square error (MSE) are used as the error function and performance value for the neural networks. The evaluation of the performance of a NN system can be measured by its sensitivity, its specificity, its positive predictive values, it negative predictive value and its accuracy. They are defined as follows in Eqs. (4) - (6).

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{4}$$

$$Specificity = \frac{TN}{TN + FN} \times 100 \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\,, \tag{6}$$

where *TP* is the number of true positive, *TN* is the number of true negative, *FP* is the number of false positive, and *FN* is the number of false negative.

True positive means pattern correctly assigned to ATG.
True negative means pattern correctly assigned to non-ATG.
False positive means pattern incorrectly assigned to ATG.
False negative means pattern incorrectly assigned to non-ATG.

The MCC can be calculated directly from the confusion matrix using the formula in Eq. (7).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{7}$$

Results and discussion

To study the effectiveness of the learning prediction from different aspects, the selected designed series of experiments are considered and outputs from the approaches are discussed.

a. To conduct the experiment of three different learning algorithms to Data set I are used to observe its performances.
b. From the output of a, the best learning algorithm is selected and further compared with Bayesian Classifier_01.
c. From the output of b, the performance of Bayesian Classifier_01 is compared with Bayesian Classifier _02 in additional to chemical properties features and increase in the length of window sequences.
d. Applying the well built in model to self-extract EST sequences of Data set II.

Experiment based on three different Backpropagation algorithms

In this experiment, three different learning algorithms namely Resilient Backpropagation (Bp-RP), Scaled Conjugate Gradient Backpropagation (Bp-SCG), and Levenberg-Marquardt (Bp-LM) are trained using several network architectures and different parameters to observe its performance. In the early stopping approach, data is divided into three parts such as 70% for training, 15% for testing and 15% for validation. Besides that, the regularization approach is also introduced in the network model to bring its effect to network generalization. Each of the number of hidden layers and hidden neurons are varied to find the optimal network architecture.

The activation function applied for hidden layer is sigmoid function (*logsig*). The epochs of the training progress is set to 1,000. The learning rate (LR) is set to 0.03 and the regularization ratio (RR) is set to 0.5 which gives an equal balance to its weights. The training goal is set to be 0.0001. All three applied training algorithms, show significantly better performances in terms of its sensitivity to others.

In experiments on resilient Backpropagation algorithms applied to the early stopping approach, window frame of 23 with one hidden layer of 15 hidden neurons show the highest prediction accuracy of 83.0% with 94.0% and 72.0% sensitivity and specificity, respectively. The precision output is 77.0% and MSE of error function is 1.7×10^{-1}. The good combination of sensitivity and specificity makes the [Bp-Rp] the best algorithm described in the case study. In comparative experiments with resilient Backpropagation applied with the manual regularization approach, window frame of 24 shows the prediction accuracy with 76.8%. The reg MSE value of 8.7×10^{-2} shows that it is quite low in comparison to the early stopping approach. On the overall, training with [Bp-RP] algorithms, it gives a high sensitivity value, and a least specificity value.

Experiment based on Bayesian Neural Network

In this experiment, the best applied algorithms are compared with Bayesian Classifier_01. Most of the prediction utilizes standard Backpropagation algorithms with their neural networks. This study is carried out to examine the Bayesian Learning approach to TIS prediction in a weak context. In a Bayesian training figure, SSE stands for Sum of Squared Errors. Sum of squared weights or weight decay is the term of regularization to penalize the network for large values of weights. The effective number of network parameters after training (γ) is also shown in the training figure. Bayesian learning prunes the redundant parameters of a feed forward neural network, which is the parameter with large variances with respect to others. These are set aside, so the cross validation is not employed in the learning

and it reduces the time needed for testing different number of hidden neurons [2, 9-11]. Table 3 and 4 show the standard parameter and measurement for comparative studies for both, Bayesian Classifier_01 and resilient Backpropagation [Bp-RP] algorithms. This second level of approach examines the efficiency of Bayesian learning in comparison to selected Backpropagation algorithm.

Table 3. Standard parameters and measurement using Bayesian Classifier_01, HU = 5, WF = 13, 14, 19, 23, 24

Network architecture	SENS (%)	SPEC (%)	ACC (%)	SSE	MCC	Time (m:s)
[13-5-1]	88.0	65.0	76.5	61.52	0.54	0:10
[14-5-1]	91.0	87.6	82.8	46.12	0.66	0:07
[19-5-1]	90.5	79.5	85.0	44.90	0.70	0:08
[23-5-1]	94.0	78.5	86.3	41.90	0.73	0:15
[24-5-1]	95.5	87.5	91.5	35.90	0.83	0:10

Table 4. Standard parameters and measurement using resilient Backpropagation Learning Algorithms (Bp-RP) with manual regularization, LR = 0.03, RR = 0.5, HU = 5, WF = 13, 14, 19, 23, 24

Network architecture	SENS (%)	SPEC (%)	ACC (%)	Reg MSE	MCC	Time (m:s)
[13-5-1]	79.0	63.0	71.0	0.119	0.43	0:17
[14-5-1]	89.0	62.0	75.5	0.110	0.53	0:19
[19-5-1]	89.0	59.0	74.0	0.103	0.45	0:15
[23-5-1]	84.5	64.5	74.5	0.098	0.53	0:18
[24-5-1]	88.0	62.5	75.3	0.097	0.54	0:18

The measurements are carried out using five different window frames of 13, 14, 19, 23 and 24 trained with single hidden layer of 5 hidden neurons. The high value of 91.5% accuracy is achieved when trained with Bayesian Classifier_01, while low accuracy of 75.5% is achieved when trained with [Bp-RP]. The high accuracy is obtained for Bayesian Classifier_01 and [Bp-RP] along with 24 and 14 window frames, respectively. Results of the accuracy suggest that Bayesian learning need to consider a long sequence of window frame to be used as an input to train the network. The sensitivity for Bayesian Classifier_01 and [Bp-RP] shows a high value of 95.5% and 89.0%, respectively. For specificity, high value of 87.6% is demonstrated by Bayesian Classifier_01 and 64.5% for [Bp-RP]. This indicates that although [Bp-RP] detects ATG sites quite well, it fails to provide same performance to detect the rejected ATG sites via its low specificity value of 65%.

On the overall, Bayesian Classifier_01 proves that the algorithm has better performance in terms of accuracy and a consistency in its value of sensitivity and specificity. The low value of specificity for all window frames demonstrated by [Bp-RP] also make the Bayesian Classifier_01 better algorithms to employ for TIS prediction task. To improve the accuracy performances of Bayesian Classifier_01, a few steps as follows are considered in Bayesian Classifier_02:

a. To add features of chemical properties as selected by IGR method.
b. Count the frequency values of features by 3 bases of amino acid.
c. A sequence window length of 201 is set for both upstream and downstream positions.

The new Bayesian model is compared with standard Backpropagation to observe its performance value. Table 5 and 6 show the standard parameter and measurement for comparative studies for both applied algorithms. This second level approach examines the improvement of learning by introducing the new features of chemical properties (hydrophilic, ambivalent, acidic, internal, non polar, polar, basic and external) and a lengthy window frame of 99 bases upstream and 99 bases downstream. The high value of 96.0% is achieved when trained with Bayesian Classifier_02 with 93.1% and 99.6% of sensitivity and specificity value, respectively. This accuracy of 96.0% outperforms the previous result with Bayesian Classifier_01 of 91.5% accuracy with no added features of chemical properties. The difference in term of accuracy between two classifiers gives a value of 4.5.

Table 5. Standard parameters and measurement using Bayesian Classifier_02, WF = 201

SENS (%)	SPEC (%)	ACC (%)	MSE	MCC	Time (m:s)
93.1	99.6	96.0	0.4703	0.89	2:47

Table 6. Standard parameters and measurement using standard Backpropagation, HU = 5, WF = 201

SENS (%)	SPEC (%)	ACC (%)	MSE	MCC	Time (m:s)
92.7	98.7	95.4	0.1869	0.84	2:67

For the comparison purposes, the standard Backpropagation is run again with additional new chemical features. The accuracy of 95.4% outperforms the previous results of [Bp-RP] with an increase of 19.9% in the value of accuracy. Overall performances of Bayesian Classifier_02 successfully proves that the chemical properties is important in helping to recognize the TIS sites in weak context with a lengthy window frame of 201. Fig. 1 demonstrates the efficiency comparison between two Bayesian Classifier and standard Backpropagation.

Testing accuracy of EST data sequences

In order to further evaluate the performance of algorithm and the feasibility of the method, the own prepared data set is applied on a well trained model built in the classification model.

The best network architecture gives the best combination of 92.2% sensitivity, 98.7% specificity and 95.5% accuracy. The training time taken to complete the prediction task is 2 min and 57 second.

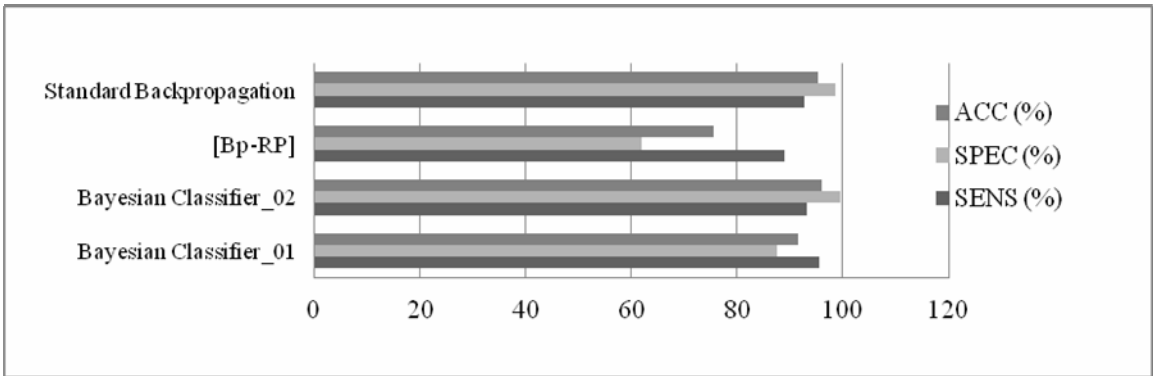

Fig. 1 The efficiency comparison between two Bayesian Classifiers and standard Backpropagation

Conclusion

We demonstrate that by using Bayesian approach with the introduced features of chemical properties are successfully helped to increase recognition by achieving a prediction accuracy of 96.0% for the case study. The results also suggests that lengthier window of input sequences are necessary when training with Bayesian model to incorporate the essential features.

As a conclusion, the new Bayesian model built in outperforms the Tikole and Sankararamakrishnan (2008) Neural Network's results by an increase of 21% of overall accuracy prediction.

Acknowledgement

The author is grateful to all lab members for discussion.

References

1. Fickett J. W. (1996). The Gene Identification Problem: An Overview for Developers, Computer Chem, 20, 103-118.
2. Gianla D., H. Okut, K. A. Weigek, G. J. M. Rosa (2011). Predicting Complex Quantitative Traits with Bayesian Neural Networks: A Case Study with Jersey Cows and Wheat, BMC Genetics, 12(87), 1-14.
3. Kochetov A. V. (2005). AUGs Codons at the Beginning of Protein Coding Sequences are Frequent in Eukaryotic mRNAs with a Suboptimal Start Codon Context, Bioinformatics, 21(7), 837-840.
4. Komberg R. D. (1999). Eukaryotic Transcriptional Control, Trends Cell Biol, 9, 46-49.
5. Kozak M., A. J. Shatkin (1978). Migration of 40 S Ribosomal Subunits on Messenger RNA in the Presence of Edeine, Journal of Biological Chemistry, 253(18), 6568-6577.
6. Kozak M. (1989). The Scanning Model for Translation: An Update, The Journal of Cell Biology, 108(2), 229-241.
7. Kozak M. (1987). An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs, Nucleic Acids Research, 15(20), 8125-8148.

8. Kochetov A. V. (2008). Alternative Translation Initiation Start Sites and Hidden Coding Potential of Eukaryotic mRNA, BioEssays, 30, 683-691.
9. Mackay D. J. C. (1992). A Practical Bayesian Framework for Backpropagation Networks, Neural Computing, 4(3), 448-472.
10. Marzban C., A. Witt (2001). A Bayesian Neural Network for Severe-hail Size Prediction, Wea Forecasting, 16, 600-610.
11. Neal R. M. (1993). Probabilistic Inference using Markov Chain Monte Carlo Methods, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 144.
12. Rogozin I. B., A. V. Kochetov, F. A. Kondrashov, E. V. Koonin, L. Milanesi (2001). Presence of ATG Triplets in 5' Untranslated Regions of Eukaryotic cDNAs Correlates with a "Weak" Context of the Start Codon, Bioinformatics, 17(10), 890-900.
13. Sparks M. E., V. Brendel (2008). MetWAMer: Eukaryotic Translation Initiation Site Prediction, BMC Bioinformatics, 9, 381.
14. Tikole S., R. Sankararamakrishnan (2008). Prediction of Translation Initiation Sites in Human mRNA Sequences with AUG Start Codon in Weak Context: A Neural Network Approach, Biochem Biophys Res Commun, 369(4), 1166-1168.

Mrs. Nurul Arneida Husin, B.Sc.
Email: nurul.arnieda@monash.edu

Nurul Arneida has received her B.Sc. in Biotechnology specialized in molecular biology from University Malaysia Sabah (UMS). Currently, she is working as a Senior Research Officer at Monash University. Her current research interest is data mining and predictive analysis in bioinformatics research.

Prof. Nanna Suryana Herman, Ph.D.
Email: nanna@utem.edu.my

Prof. Dr. Nanna Suryana has received his B.Sc. Soil & Water Eng. (Bandung, Indonesia), MSc Comp. Assisted for Geoinformatics& Earth Science, (Enschede, Holland), Ph.D. GIS (Wageningen, Holland). He is currently holding a position of Director of International Office and lecturer at Faculty of Information Technology and Communication (FTMK) of Universiti Teknikal Malaysia Melaka (UTEM). His current research interest is in field of Geographic Information System (GIS) and Data Mining.

Prof. Madya Burairah Hussin, Ph.D.
Email: burairah@utem.edu.my

Prof. Madya Dr Burairah has received his Ph.D. in Management Science from University of Salford, UK, M.Sc. Numerical Analysis Programming from University of Dundee, UK, and B.Sc. Computer Science and Diploma in Computer Science from Universiti Teknologi Mara, Malaysia. He is currently holding a position of deputy dean of research and a lecturer at Faculty of Information Technology and Communication (FTMK) of Universiti Teknikal Malaysia Melaka (UTEM). His current research interest is in field of Networking and Numerical Analysis.