Transition Initiation Sites (TIS) Recognition in DNA Sequence using Machine Learning

Muhammad Hossain and Kanaan Faisal
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
{mdimtiaz, Kanaan}@kfupm.edu.sa

Abstract— Transition Initiation Sites (TIS) prediction is a challenging problem in computational biology. In the literature TIS is predicted using various machine learning techniques such as Neural Network (NN), Support Vector Machine, etc. We have applied Principal Component Analysis (PCA) to remove highly correlated features which improves the performance in terms of time and accuracy. In this paper we have used Group Model of Data Handling (GMDH) based algorithm Abductive Network (AN) to predict TIS and got accuracy of 93%.

Keywords—Bioinformatics, Transition Initiation Sites (TIS), mRNA sequence, Machine Learning, Neural Network, Abductive Network, GMDH.

I. INTRODUCTION

Proteins are synthesized from mRNAs by a process called translation. The region at which the process initiates is called the Translation Initiation Site (TIS). The coding sequence is ranked by non-coding regions which are the 5' and 3' UnTranslated Region (UTR) respectively. The translation initiation site prediction problem is to correctly identify the TIS in a mRNA or cDNA sequence. This forms an important step in genomic analysis to determine protein coding from nucleotide sequences. In his research we have predicted TIS in human mRNA sequence.

In eukaryotes, the scanning model postulates that the ribosome attaches first to the 5' end of the mRNA and scans along the 5' to 3' direction until it encounters the first AUG. The problem of predicting the TIS is compounded in real-life sequence analysis by the difficulty of obtaining full-length and error-free mRNA sequences.

Machine learning techniques have been used successfully in TIS prediction using the mRNA or cDNA sequence.

In this research the feature dimension reduction is performed using PCA to select the most significant features and finally AN and Multi Layer Perceptron is used for TIS prediction.

The rest of the paper is organized as follows. Section 2 deals with recent literatures. Section 3 describes the proposed recognition system. Section 4 shows experimental results. Finally Section 5 mentioned conclusion and future work.

II. LITERATURE REVIEW

Pedersen and Nielsen [1] found that almost 40% of the mRNAs extracted from GenBank contain upstream AUGs. This accords with the scanning hypothesis that the ribosome operates in a linear fashion on the sequence to recognize the start site. They have trained an artificial neural network (ANN) on a 203 nucleotide window centered on the AUG. They obtained results of 78% accuracy on start AUGs and 87% accuracy on non-start AUGs on their vertebrate dataset, giving an overall accuracy of 85%. This system is available on the Internet as the NetStart 1.0 prediction server.

Zien et al. [2] obtain improved results on the same vertebrate dataset from Pedersen and Nielsen by using support vector machines (SVM). The same 203 nucleotide window is used as the underlying features to be learnt. They show how to obtain improvements by appropriate engineering of the kernel function - using a locality-improved kernel with a small window on each position, a codon-improved kernel using codon structure in the downstream sequence and a Salzberg kernel using conditional positional probabilities. With the nucleotide-based kernels [3], they obtain an accuracy of 69.9% and 94.1% on start and non-start AUGs respectively, giving an overall accuracy of 88.1%. The Salzberg kernel gives an overall accuracy of 88.6%.
Hatzigeorgiou [4] reports a highly accurate TIS prediction program, DIANA-TIS, using ANN trained on human sequences. Their dataset contains full-length cDNA sequences which has been altered for errors. An overall accuracy of 94% is obtained using an integrated method which combines a consensus ANN with a coding ANN together with the ribosome scanning model.

Zeng et al. [5] obtained 94% overall accuracy on the dataset used in [1, 2, 4] by using simple feature generation and selection on a variety of standard machine learning methods. In the work of Zien et al. [2] and Hatzigeorgiou [4], improved TIS prediction is obtained by a more complex method. Zeng et al. showed that the use of simple feature generation followed by correlation-based feature selection allows a variety of standard machine learning methods such as ANN, decision trees, SVM, Naive Bayes to obtain accurate TIS prediction. Feature selection results in only a very small number of features, at most 13, to get good results. The results from the simple TIS prediction are directly comparable with Zien et al. [2] and Pedersen and Nielsen [1]. The highest overall accuracy obtained is 89.4% which is better than previous results on this dataset. Incorporating distance as a feature improves this result. Finally with the use of a scanning model, they have obtained an overall accuracy of 94.4% which compares very favorably to Hatzigeorgiou [4].

III. DATASET

The dataset used is the vertebrate dataset created by Pedersen and Nielsen [1]. This dataset was also used by [2, 4, 5]. So our results can be compared directly with the two previous works. The original dataset of Pedersen and Nielsen [1] consists of a selected set of vertebrate genomic sequences extracted from GenBank [6]. It consists of sequences from Bos taurus (cow), Gallus gal-lus (chicken), Homo sapiens (man), Mus musculus (mouse), Oryctolagus cuniculus (rabbit), Rattus norvegicus (rat), Sus scrofa (pig), and Xenopus laevis (African clawed frog). It has been shown by Pedersen and Nielsen that these vertebrates have similar start codon contexts [1]. These sequences are processed by removing possible introns and joining the exons. This is analogous to the splicing of mRNA sequences. From these sequences, only those with an annotated translation initiation site, and with at least 10 upstream nucleotides as well as 150 downstream nucleotides are selected. The sequences are altered to remove those belonging to same gene families, homologous genes from Different organisms, and redundant sequences, so as to avoid over-optimistic performance resulting from biased data [7]. This resulting dataset consists of 3312 sequences. Since the dataset is processed DNA, the TIS site is ATG. In total, there are 13503 ATG sites. Of the possible ATG start sites, 3312 (24.5%) are the true start ATGs while the other 10191 (75.5%) are non-start ATGs. The dataset is available in

http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html

An example entry from this dataset is given below in Figure 2 and Figure 1 shows the basic TIS terminologies.

A. Feature Extraction

Frequency of k-gram amino acid. (k = 1, 2, 3.. Amino acid patterns)

- Count the frequency of amino acid X in upstream and downstream.

B. Feature Vector

After extracting feature we have a feature vector 13,310 X 927. We have used PCA to reduce the feature dimension from
927 to 70 which make the prediction faster by the machine learning techniques.

C. Feature Vector

After extracting feature we have a feature vector 13,310 X 927. We have used PCA to reduce the feature dimension from 927 to 70 which make the prediction faster by the machine learning techniques.

IV. MACHINE LEARNING TECHNIQUES

ANN, SVM, HMM, FUZZY LOGIC, Bayesian Network, Group Model of Data Handling Based Abductive Network (AN), etc are the well known techniques in machine learning. We have used ANN and AN to predict TIS.

A. Abductive Network (AN)

Abductive Networks approach based on the Group Method of Data Handling (GMDH) algorithm as an alternative learning tool. The GMDH approach to classification offers the advantage of simplified and more automated model synthesis. Abductive Inductive Mechanism (AIM) is a Machine Learning tool that automatically discover network solutions to complex decision, prediction, control and classification problems. The tool generate a mathematical models from relationships it finds in the training data. It does so by trying out all potential relationships of linear, multiple and polynomial on various combination of input variables. It iteratively build a network of functional elements based on prediction performance using Predicted Square Error (PSE).

\[
PSE = FSE + CPM \left(\frac{2K}{N}\right)\sigma^2
\]

- FSE: Fitted Square Error.
- CPM: Complexity Penalty Multiplier.
- K: # of Coefficients.
- N: # of Inputs.
- \(\sigma\): Estimation of predicted error.

The unique property of automatic selection of only the most relevant input features by abductive network models gives useful insight into the contribution of the various features in the dataset.

AN Functional Elements

- **Normaliser**: Transforms the original input into a normalized variable having a mean of zero and a variance of unity.
- **Unitizer**: Restores the result to the original problem space
- **Node**: The node has input(1, 2 or 3) and the polynomial equation is limited to the third degree, that is:

\[
y = z_0 + z_1x + z_2x^2 + z_3x^3
\]

V. NEURAL NETWORK

In this paper Multi Layer Perceptron (MLP) classifier is used which is one of the popular Artificial neural networks (ANN) consist of simple processing elements and a high degree of interconnection. The elements are organized into an initial input layer, intermediate “hidden” layers, and a final output layer (Figure 7). In MLP information proceeds from the input layer to the output layer through hidden layer(s). It uses back propagation algorithm makes to learn the weights within the elements and construct arbitrarily complex nonlinear decision boundaries to separate multiple classes.

VI. PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is used to reduce highly correlated features. PCA was first introduced by Pearson in 1901 and become a standard tool in modern data analysis.
PCA is actually a technique to find the directions in which a cloud of data points is stretched most. PCA perform linear transformation by choosing a new coordinate system in such a way that greatest variance by any projection of the data set comes to lie on the first axis (the first principal component). PCA can be used for reducing dimensionality by eliminating the later principal components.

The objective of PCA is to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible. PCA performs a linear mapping of the data to a lower dimensional space in such a way, that the variance of the data in the low-dimensional representation is maximized. At first, the correlation matrix of the data is constructed and the eigenvectors on this matrix are computed so that the eigenvectors that correspond to the largest eigenvalues (the principal components) can be used to reconstruct a large fraction of the variance of the original data. Moreover, the first few eigenvectors can often be interpreted in terms of the large-scale physical behavior of the system. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.

VII. FRAMEWORK OF TIS PREDICTION

At first the TIS features are reduced using the Principal Component Analysis (PCA) by removing highly correlated features. 70% of the digits used for training the AN and MLP to build the model and 30% were used for testing. In the following sections the steps are described in details.

VIII. EXPERIMENTAL RESULTS

In this paper, the performance of two standard machine learning classifiers on the selected features is evaluated. We have used the Abductive Network model and Neural Network. We have used Abductive Network as a classifier. And later on we have used the features that are chosen by the AN as an input for Neural Network. Each ATG is labeled whether or not it's a true TIS site. Thus, each ATG in a sequence from the set of training sequences contributes one training instance. Training and testing is performed with a random sampling method. 70% of the dataset is taken as training and 30% is kept for testing.

The results testing are evaluated using standard performance measures. To describe the performance, the results from testing a classifier can be arranged in the following matrix:

<table>
<thead>
<tr>
<th>Actual Yes Class</th>
<th>Classified as Yes</th>
<th>Classified as No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of True Positive</td>
<td>No. of False Negative</td>
</tr>
<tr>
<td>Actual No Class</td>
<td>No. of False Positive</td>
<td>No. of True Negative</td>
</tr>
</tbody>
</table>

We have the following measures:

True Positive Rate (also called Sensitivity) = \( \frac{TP}{TP + FN} \)

True Negative Rate = \( \frac{TN}{TP + FN} \)

Specificity = \( \frac{TP}{TP + FP} \)

Overall Accuracy = \( \frac{TP + TN}{TP + TN + FN + FP} \)

Adjusted Accuracy = \( \frac{TPRate + TNRate}{2} \)

Because the dataset consists of significantly more negative than positive examples, we have also used Adjusted Accuracy as a performance measure which gives a fairer comparison than overall accuracy for skewed datasets such as the one here where the number of non-start ATGs is disproportionately larger than the number of start ATGs.

For example, if 80% of the ATGs are non-start, then a trivial predictor which simply classifies every ATG as non-start would already obtain an overall accuracy of 80%. Adjusted accuracy, on the other hand, is less skewed giving...
50% accuracy. As the results in the literature do not give sufficient data to compare on the basis of adjusted accuracy, we continue to use overall accuracy in the comparisons with existing work.

<table>
<thead>
<tr>
<th>Table 2 Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classified as No</strong></td>
</tr>
<tr>
<td><strong>Actual No Class</strong></td>
</tr>
<tr>
<td><strong>Actual Yes Class</strong></td>
</tr>
</tbody>
</table>

True Positive Rate (Sensitivity) = 77.74%
True Negative Rate = 94.22%
False Positive Rate (Specificity) = 5.77%
Adjusted Accuracy = 85.98%
Overall Accuracy = 90.33%

A. ROC CURVE ANALYSIS for AN

<table>
<thead>
<tr>
<th>Table 3 RESULTS FOR AN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AUC</strong></td>
</tr>
<tr>
<td>0.95925</td>
</tr>
<tr>
<td>Standardized AUC</td>
</tr>
</tbody>
</table>

Cut-off point for best Sensitivity and Specificity (blu circle in plot) = 0.2536

In the ROC plot, the cut-off point is the closest to [0,1] point or, if you want, the closest to the green line

Table at cut-off point

<table>
<thead>
<tr>
<th>Table 4 CUT-OFF VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cut-off point</strong></td>
</tr>
<tr>
<td>867</td>
</tr>
<tr>
<td>81</td>
</tr>
</tbody>
</table>

Prevalence: 23.6%

Sensitivity (probability that test is positive on unhealthy subject): 91.5%

95% confidence interval: 89.7% - 93.2%
False positive proportion: 8.5%

Specificity (probability that test is negative on healthy subject): 86.2%
95% confidence interval: 85.0% - 87.4%
False negative proportion: 13.8%
Youden’s Index (a perfect test would have a Youden index of +1): 0.7765

Precision or Predictivity of positive test (probability that a subject is unhealthy when test is positive): 67.2%
95% confidence interval: 64.6% - 69.8%
Positive Likelihood Ratio: 6.6
Moderate increase in possibility of disease presence

Predictivity of negative test (probability that a subject is healthy when test is negative): 97.0%
95% confidence interval: 96.4% - 97.7%
Negative Likelihood Ratio: 0.1
Large (often conclusive) increase in possibility of disease absence

F-measure: 77.5%
Accuracy or Potency: 87.4%
Mis-classification Rate: 12.6%
B. **ROC CURVE ANALYSIS for ANN**

**TABLE 5 RESULTS OF ANN**

<table>
<thead>
<tr>
<th>AUC</th>
<th>S.E.</th>
<th>95%</th>
<th>C.I.</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97858</td>
<td>0.00333</td>
<td>0.97205</td>
<td>0.98510</td>
<td>Excellent test</td>
</tr>
</tbody>
</table>

**Standardized AUC**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>143.7515</td>
<td>1-tail p-value</td>
<td>0.000000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The area is statistically greater than 0.5

Cut-off point for best Sensitivity and Specificity (blu circle in plot)= 0.3439

In the ROC plot, the cut-off point is the closest to [0,1] point or, if you want, the closest to the green line

Table at cut-off point

**TABLE 6 CUT-OFF VALUES**

<table>
<thead>
<tr>
<th>cut-off point</th>
<th>877</th>
<th>211</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71</td>
<td>2853</td>
</tr>
</tbody>
</table>

Prevalence: 23.6%

Sensitivity (probability that test is positive on unhealthy subject): 92.5%

95% confidence interval: 90.8% - 94.2%

False positive proportion: 7.5%

Specificity (probability that test is negative on healthy subject): 93.1%

95% confidence interval: 92.2% - 94.0%

False negative proportion: 6.9%

Youden's Index (a perfect test would have a Youden index of +1): 0.8562

Precision or Predictivity of positive test (probability that a subject is unhealthy when test is positive): 80.6%

95% confidence interval: 78.3% - 83.0%

Positive Likelihood Ratio: 13.4

Large (often conclusive) increase in possibility of disease presence

Predictivity of negative test (probability that a subject is healthy when test is negative): 97.6%

95% confidence interval: 97.0% - 98.1%

Negative Likelihood Ratio: 0.1

Large (often conclusive) increase in possibility of disease absence

F-measure: 86.1%

Accuracy or Potency: 93.0%

Mis-classification Rate: 7.0%

Figure 10. ROC for ANN

Figure 11. PARTEST GRAPH of ANN

Figure 12.
IX. CONCLUSIONS

The experiments showed that the performance of Neural Network is better than Abductive Network. The overall accuracy of Neural Network is about 93.8% while the overall accuracy of Abductive Network is about 90.3%.

X. ACKNOWLEDGMENT

The Authors would like to acknowledge the support of King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

XI. REFERENCES


