Gene Prediction of Prokaryotic Genome Using Neural Network

Nurezayana Zainal a,1,*, Chee Sim Lee b,2, Weng Howe Chan b,3

aDepartment of Software Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor.
bSchool of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor.
1nurezayana@uthm.edu.my*; 2cheesim95@gmail.com; 3cwenghowe@utm.my

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Gene prediction
Prokaryote
Genome
Artificial Neural Network

ABSTRACT

The need for gene prediction has been increasing throughout the years since the beginning of Human Genome Project. With the emergence of next-generation sequencing technologies, the sequencing of Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA) sequence can be done within lesser time and money. However, the annotation process is unable to keep up with the enormously growing sequence. Thus, a lot of DNA sequences still remains as characters as they have yet to be deciphered by scientist by analyzing and determining the location and the function of the gene within the sequence. However, finding a gene from a vast amount of character manually is nearly impossible and this is where gene prediction becomes helpful. Gene prediction allows the scientists to narrow down the scope from the DNA sequence to be further researched by predicting potential coding region in the sequence using computational methods. Researchers from Malaysia Genome Institute have recently sequenced the genome of Enterobacter sp. but the sequence has not been annotated yet. Therefore, artificial neural network is proposed to determine the position of the potential gene to save the researchers’ effort and time to find gene from Enterobacter sp. An artificial neural network (ANN) will be implemented in order to classify whether a certain region is a potential coding region. The DNA sequence from the Enterobacter sp. genome provided by Malaysia Genome Institute will be used as the subject data of the artificial neural network. It is concluded that the length of the genome does not have any significant correlation to the gene prediction performance while the minimum length of the ORF extracted as candidate for neural network have a significant to the gene prediction.

This is an open access article under the CC–BY-SA license.

1. Introduction

Deoxyribonucleic Acid (DNA) sequences without any processing are only mere characters; the sequence must first be analyzed and studied to determine the location and the function of the gene in the genome. A DNA sequence can be viewed as a string of characters or nucleotides: A (adenine), C (cytosine), G (guanine) and T (thymine) [1]. The DNA sequence is too large to be handled manually, that is when computational methods came in to help predicting the possible regions where the genes are located to narrow down the range of the sequence to be further researched by wet lab scientist. The need of gene prediction has rose since the beginning of Human Genome Project 3 decades ago; many methods/algorithms has been applied to model gene models which includes. Dynamic Programming, linear dis-criminate analysis, Linguist methods, Hidden Markov Model and Neural
Network Out of which neural network is implemented in this project as it can be achieved higher accuracy by grasping sequence characteristic (signals) after training with enough sample data.

The general structure of prokaryotic genes is shown as in Fig. 1. Genes from prokaryote generally have remarkably simple structures consist only a single continuous coding region and one or two transcription-factor binding sites (also known as promoter) at the upstream of the coding region. Due to having a single continuous coding region, the DNA sequences is transcribed into messenger ribonucleic acid (mRNA) and translated directly into proteins without significantly modification [2].

Fig. 1. Prokaryotic Gene Structure

The most basic way to find protein coding region in DNA sequence is to search for its opening reading frames (ORF) [2]. An ORF is a length of DNA sequence that containing contiguous set of codons, having a start codon at the beginning and a stop codon at the end of the sequence [1, 3]. Although, not all ORF can be classified as a coding region, in fact, majority of ORFs in a genome are non-coding, unless the ORF must fulfil a certain minimum length and has a specific composition, but all protein coding genes are ORFs. That’s why most ab initio prediction of genes from microbial data applies the conventional approach that based on the identification of ORFs [4]. Two different types of information are used in locating the genes in a genomic sequence, content sensors and signal sensors [5].

Intrinsic content sensors use statistical properties to differentiate between coding and non-coding segments [6]. As genes are uninterrupted coding regions in prokaryotes, the simplest approach would be searching for sufficiently long open reading frames [5]. Others measures that are used as intrinsic content sensors are nucleotide composition (G+C content especially), codon composition, hexamer frequency, base occurrence periodicity etc. Among the above, the usage of hexamer (6 nucleotide long words) is the most discriminative variable to determine coding and non-coding region and is used by large number of algorithms through different method.

Signal sensors are the presence of functional site which match with the consensus sequence determined by the multiple sequence alignment of functionally related documented sequences [5]. The signals are represented in positional weight matrices which are indications of probability of a given base appearing at each position of the signal. The examples of signals in a DNA sequence are splice sites, promoter, poly(A) sites and translation initiation codon.

In this article, we present a novel neural network (ANN) implementation to determine the position of the potential gene to save the researchers’ effort and time to find gene from Enterobacter sp. ANN is a type of model that can also be trained to predict survival based on the idea of neurons in the processing of information [7]. ANN will be implemented in order to classify whether a certain region is a potential coding region. Also, the DNA sequence from the Enterobacter sp. genome provided by Malaysia Genome Institute will be used as the subject data of the artificial neural network.

2. Data

2.1. Subject Data

The subject data which is the DNA sequence of the Enterobacter sp. is sequenced and provided by Malaysian Genome Institute (MGI). The data provided is in FASTA format and since it is a DNA sequence, it contains only 4 characters (A, C, G, T) corresponding to the 4 nucleotide bases. As the genome of Enterobacter sp. newly sequenced, there isn’t any more detail on the DNA sequence. As the DNA sequence does not contain any spoilt data, the cleaning of the DNA sequence is not
necessary and we can proceed to ORF extraction. The DNA sequence is 4566231 base pair long and it consist of 1010375 Adenine, 1267781 Cytosine, 1271951 Guanine and 1014621 Thymine. The CG content of the genome is 55.62%.

2.2. Benchmark Data

There are a total of 14 known genomes which are selected as benchmark data to determine the accuracy of this research flow. The genome sequences in FASTA format and the gene annotation files in GFF format of these genomes are obtained from the GenBank database. The name of the species respective to their GenBank accession number is as Table 1.

<table>
<thead>
<tr>
<th>GenBank Accession Number</th>
<th>Species</th>
<th>Length (Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_000909</td>
<td><em>Methanococcus jannaschii</em></td>
<td>1.7</td>
</tr>
<tr>
<td>NC_000913</td>
<td><em>Escherichia coli</em></td>
<td>4.6</td>
</tr>
<tr>
<td>NC_000917</td>
<td><em>Archaeoglobus fulgidus</em></td>
<td>2.2</td>
</tr>
<tr>
<td>NC_000921</td>
<td><em>Helicobacter pylori</em></td>
<td>1.6</td>
</tr>
<tr>
<td>NC_000964</td>
<td><em>Bacillus subtilis</em></td>
<td>4.2</td>
</tr>
<tr>
<td>NC_002516</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>6.3</td>
</tr>
<tr>
<td>NC_002528</td>
<td><em>Buchnera aphidicola</em></td>
<td>0.6</td>
</tr>
<tr>
<td>NC_002932</td>
<td><em>Chlorobium tepidum</em></td>
<td>2.2</td>
</tr>
<tr>
<td>NC_006350, NC_006351</td>
<td><em>Burkholderia pseudomallei</em></td>
<td>7.2</td>
</tr>
<tr>
<td>NC_006833</td>
<td><em>Wolbachia endosymbiont</em></td>
<td>1.1</td>
</tr>
<tr>
<td>NC_007164</td>
<td><em>Corynebacterium jeikeium</em></td>
<td>2.5</td>
</tr>
<tr>
<td>NC_007426</td>
<td><em>Natronomonas pharaonis</em></td>
<td>2.6</td>
</tr>
<tr>
<td>NC_007577</td>
<td><em>Prochlorococcus marinus</em></td>
<td>1.7</td>
</tr>
</tbody>
</table>

3. Method

Fig. 2 shows the overall process of gene prediction on the genome of prokaryote using artificial neural network (ANN). The process is divided into two stages. At the first stage, the prokaryotic genomes which were obtained from any database in FASTA format are sent to the ORF extraction program to select the ORF candidate from the genome sequence. During the second stage, the extracted ORF sequences will be classified as either coding sequence or non-coding sequence using the gene prediction tool. The focus of this project will be on the first stage which is the ORF extraction process as we are involved in the development of the ORF extraction program unlike the second stage where we only implemented the neural network model which was already trained and built.
3.1. ORF Extraction

Instead of making the entire genome as the input of the gene prediction tool, the open reading frame within the sequence will first be extracted. This is due to the fact that in prokaryotic genome, although majority of ORF are not encoding protein, but all protein encoded sequences are ORFs. Before the extraction of the ORF, the genome sequence is divided into fragments for every 1,000,000 nucleotide bases as the gene prediction tools has a buffer size limit of 2,000,000 nucleotide and thus unable to take in a single sequence longer than 2,000,000 nucleotide as input.

For each fragment, the nucleotide sequence is first translated into amino acid sequence and as DNA is interpreted as codon, a DNA strand will have 3 reading frames and since DNA are double stranded (forward and reverse), there will be a total of 6 reading frames. For each reading frame, the program will search for the start codon, ATG which will be translated in Methionine, M in the amino acid sequence codon by codon. Once the methionine is found, the program will search for the next stop codon (TGA, TAG, TAA) which are translated to '*' in the amino acid sequence. When the stop codon is located, the program will check whether the ORF contains more than minimum amino acids set. If it does, the ORF will be selected and transferred into the output file. As a default, the minimum amino acid required for the ORF to be selected is set to 100 so that only long and quality sequence will be selected.

3.2. Gene Prediction using Neural Network

Fig. 3 shows the architecture of the neural network in the gene finder tool for binary ORF classification. A standard multilayer perceptrons with one layer of 25 hidden nodes and a single logistic output function is used. The output of the neural network is the approximation of the posterior coding probability of the class representing the coding ORFs with can be interpreted as gene probability. This gene prediction tools has been supervised trained with characteristics of coding and non-coding regions from 131 fully sequenced prokaryotic genomes and their annotation for their genes.
A total of seven features based on the ORFs’ sequence characteristics are used as the input of the neural network. However, all these features can have a different preprocessing step which will be explain below. Codon and dicodon usage are perhaps the most crucial features to discriminate between coding and non-coding ORF. The features of the monodicon, $x_1$ and dicodon, $x_2$ usage are computed based on linear discriminant scores which were obtained by training with annotated genome data.

The features $x_3$ and $x_4$ were obtained by using the Gaussian probability density functions of the linear discriminant score derived positive and negative training examples of TIS. Length features is also one of the feature that discriminate between coding and non-coding ORF. $x_5$ and $x_6$ stands for incomplete length and complete length respectively. The feature $x_5$ exists in this gene finding tool because the tool was originally meant for metagenomic which frequently contain incomplete ORF. GC-content is the last feature, $x_7$ of the neural network which is a value between 0 to 1 that represents the GC-content proportion from the whole fragment the ORF occurs.

### 3.3. Result Analysis and Discussion

In this project, the ORF extraction process is done several times for a same genome but with different parameter. min_pro_len is a variable that determines the minimum amino acid in the ORF to be selected as a candidate. ORF extraction process is done with different min_pro_len of 100, 80, 60, 40 and 20. The ORF extraction is also done while considering alternative start codon such as ‘TTG’ and ‘GTG’ with again different min_pro_len of 100, 80, 60, 40 and 20. This is to determine the effect of minimum protein length and type start codon considered during ORF extraction process on the performance of the neural network.

### 3.4. Effect of Genome Size

Fig. 4 shows the graph of the recall and precision value (ORF extraction is done with minimum protein length of 100 amino acid and using only ‘ATG’ as start codon) against the genome size of the known genome. From the observation, we can notice that the value of precision and recall fluctuate as the genome size increases. The precision and recall value is the highest, at 0.76707 and 0.663165 respectively when the genome size is largest at 6.3 Mbp. However, the second highest precision and recall value is found when the genome size is smallest at 0.6 Mbp. Therefore, we can conclude that the genome size does not have significant influence in the performance of the prediction.
3.5. Effect of Minimum Protein Length

Fig. 5 shows the graph of the recall value of the 9 known genome against the different minimum protein length during ORF extraction. We can observe that with as the minimum protein length decreases, the recall value increases. This is because when the minimum protein length is large, there will be a lot of ORF that will not be extracted as a candidate for the gene prediction tool to classify, and thus genes which are short will not be able to be detected since they are not even selected in the first stage of the workflow. We can also observe that as the minimum protein length are getting smaller, the level of increment of recall is also getting smaller, the recall values for 20 and 40 minimum amino acid lengths are the same for most genome above. This is due to the fact that only very minor of genes have length shorter than 20 amino acids.

Fig. 6 shows the graph of the precision value of the 9 known genome against the different minimum protein length during ORF extraction. We can observe that the precision value of the genomes are very close to each other although have different minimum protein length during ORF extraction. In general, the precision increases as the minimum protein length decreases but for some genomes, their precisions reach their peak at minimum protein length of 60 and started to decrease slightly when the minimum amino acid length decreases. There are also genomes with same precision value for minimum protein length of 20 and 40, this is because the prediction of the gene finding tool does not give different predictions for both the ORF candidates provided. For genome NC_007577, we can see that even though although it has the same recall value for 20 and 40 minimum protein length, the precision of 40 minimum protein length of the value is slightly higher than 20 minimum protein length. This is due to the lesser amount of non-coding ORF being wrongly classified as coding gene as the ORF extracted with minimum length of 40 are lesser.
3.6. Effect of Start Codon used

Fig. 7 shows the graphs of the recall value when consider different start codon for genome NC_006833. We can observe that all the recall values when we considered ATG, GTG and TTG as start codon is higher than just considering ATG codon. This is because unlike eukaryotic genome where the ‘ATG’ codon is the start codon for almost every gene, prokaryotic genome has high percentage of GTG and TTG being used as alternative start codon. We also can see that all the curves share similar shape whereas the minimum protein length become smaller, the graph becomes less curved.

Fig. 8 shows the graphs of the precision value when consider different start codon for genome NC_006833. We observed that just like the recall values, all precision values when considering GTG and TTG codon as start codon is higher than only considering ATG codon. Unlike the recall value which either increases or maintains when the minimum protein length decreases, the precision curves may drop or fluctuate as the minimum protein length decreases, but both curves from each graph shares a similar pattern.
3.7. Implementation on Subject Data

The proposed workflow is implemented to predict the gene location in the genome of Enterobacter sp. However, due to the limitation of unforeseen error, the proposed workflow can only be run when the min_pro_len is set to 244. This has drastically decreased the number of predicted genes by the workflow. Since the species of this Enterobacter sp. provided is still yet to be identified and its gene is yet to be annotated, we can only compare it to the gene prediction result of other software. In this project, we are comparing the predicted result of the workflow with the result from Prodigal which stands for Prokaryotic Dynamic Programming Genefinding Algorithm. The comparison is as Table 2 below:

<table>
<thead>
<tr>
<th></th>
<th>Proposed Workflow</th>
<th>Prodigal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Predicted Gene</td>
<td>2552</td>
<td>4217</td>
</tr>
<tr>
<td>Number of Gene Matches Both Left and Right Location</td>
<td>2046</td>
<td></td>
</tr>
<tr>
<td>Number of Gene Matches on Either One Location</td>
<td>505</td>
<td></td>
</tr>
<tr>
<td>Unmatched Genes</td>
<td>1</td>
<td>1666</td>
</tr>
</tbody>
</table>

4. Conclusion

From the data analysis above, we can conclude a few things based on the observation. Firstly, we can say that the length of the genome does not have any significant correlation to the gene prediction performance. However, the minimum length of the ORF extracted as candidate for neural network does. As the minimum length of ORF extracted decreases, more ORF are selected as the candidate for the neural network, thus the shorter gene can be predicted using neural network. Therefore, the recall of the result has increased since more annotated genes are predicted by the neural network. However, when the minimum protein length is getting too small e.g. less than 40, there isn’t any obvious change in the recall as there are very few genes, sometimes even no gene which is shorter than 40 amino acids. In fact, when the minimum protein for ORF extraction is too short, the precision decreases as there are more negative ORF selected as candidate causing more false positive prediction. In overall, the precision value of the prediction does not have a vast difference throughout the different minimum protein length of ORF extraction.

As for the type of initiation codon considered when extracting the ORF, we have 2 types of observation. The performance (either recall or precision) are significantly higher when taking in alternative initiation codon such as ‘GTG’ and ‘TTG’ instead of considering only ‘ATG’ codon. The second observation is that even though the alternative initiation codon is considered during ORF extraction process, the performance of the prediction shows a similar pattern as the performance when only considered the ‘ATG’ codon. Thus, we can conclude that considering alternative start codon and having a low minimum protein length (around 40) during the ORF extraction process can increase the performance of the gene prediction using neural network.

Acknowledgment

References


