Heliyon 6 (2020) e04825

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Global sequence features based translation initiation site prediction in human genomic sequences

Neelam Goel [a,*], Shailendra Singh [b], Trilok Chand Aseri [b]

[a] Department of Information Technology, University Institute of Engineering and Technology, Sector-25, Panjab University, Chandigarh 160014, India
[b] Department of Computer Science and Engineering, Punjab Engineering College (Deemed to be University), Sector-12, Chandigarh 160012, India

ARTICLE INFO

Keywords:
Computer science
Gene prediction
cDNA
mRNA
Genomic sequence
Support vector machine
Translation initiation site

ABSTRACT

Gene prediction has been increasingly important in genome annotation due to advancements in sequencing technology. Genome annotation further helps in determining the structure and function of these genes. Translation initiation site prediction (TIS) in human genomic sequences is one of the fundamental and essential steps in gene prediction. Thus, accurate prediction of TIS in these sequences is highly desirable. Although many computational methods were developed for this problem, none of them focused on finding these sites in human genomic sequences. In this paper, a new TIS prediction method is proposed by incorporating global sequence based features. Support vector machine is used to assess the prediction power of these features. The proposed method achieved accuracy of above 90% when tested for genomic as well as cDNA sequences. The experimental results indicate that the method works well for both genomic and cDNA sequences. The method can be integrated into gene prediction system in future.

1. Introduction

The process of translation is responsible for converting mRNAs into protein. The whole process is carried out in three different steps namely: initiation, elongation of polypeptide chain and termination. Translation initiation site is the position where this process initiates and it terminates at the first in-frame codon downstream. The non-coding sequences that present around coding sequences are known as 5′ and 3′ untranslated regions. The problem of translation initiation site prediction is to accurately recognize this site in genomic sequences (i.e mRNA, cDNA and uncharacterized DNA). By knowing the exact location of TIS in genomic sequences, the corresponding protein can be easily identified. Therefore, recognition of TIS is a central problem in gene prediction which further helps in understanding the gene structure and its product. Usually, a TIS consists of a tri-nucleotide ATG (in DNA or cDNA) and AUG (in mRNA) and used as a start codon. Translation can also occur at other codons, such as ACG and CTG, however, this is a rare event reported in eukaryotes and is not considered here.

The scanning model states that the ribosome, firstly, binds to the 5′ end of mRNA then migrates along the 5′-3′ end and stop at first AUG codon where a favorable context for initiating translation is present. Although the scanning model hypothesis of first AUG occurs in 90% of the cases, there are notable exceptions also. Some mechanism explaining these exceptions are: leaky scanning where the translation initiates from a downstream ATG after bypassing the first AUG codon due to the presence of poor context, reinitiation where a small open reading frame causes ribosome to continue with the scanning until another ATG is found to make the protein, internal initiation where the ribosome directly attaches near the actual ATG without scanning and this fact is reported in a number of viral mRNAs.

With the help of new technologies, more and more TISs have been verified experimentally. However, it has been reported that 40 % of the mRNAs taken from GenBank database include upstream AUGs. The problem turns out to be more difficult when unannotated genomic sequences or ESTs (small sequences derived from cDNAs) which usually contain more errors, are used. Moreover, the experimental approaches are very costly and time-consuming. Thus, prediction of correct TIS is a complex problem because the sequences are not complete, can contain errors and their underlying mechanisms are not fully understood. Therefore, efficient and accurate computational methods that automatically try to solve this problem are needed for TIS prediction. Moreover, the computational recognition of TISs is an important component of every gene prediction system and thus plays a crucial role in genome annotation projects. In past, several computational methods are

* Corresponding author.
E-mail address: erneelam@pu.ac.in (N. Goel).

https://doi.org/10.1016/j.heliyon.2020.e04825
Received 20 January 2019; Received in revised form 25 May 2020; Accepted 26 August 2020

developed to predict TIS in transcript data i.e. mRNA, EST, and cDNA. However, in gene prediction system, the TIS need to be predicted at genomic level. This makes the problem of TIS prediction even more challenging because at the genome level the number of false TIS is much more as compared to true TIS. Also, the prediction of TIS in genomic sequences is different from transcript sequences due to the following reasons: transcripts usually contain zero or one TIS which aid the prediction process; scanning models of translation cannot be applied to genomic sequences because of the presence of large number of candidate TISs; genomic sequences contain introns which disrupt the coding sequence downstream the TIS. The focus of this paper is to predict TIS in genomic sequences.

This section addresses the problem of TIS prediction in genomic sequences and its importance in gene prediction. The remaining of the paper is outlined as follows: The next section describes related work for TIS prediction. Thereafter, the datasets and the methods used in this work are presented. In section 4, the results of the proposed method are illustrated along with evaluation measures. The last section includes the discussion and conclusion of this work.

2. Related work

A TIS is dependent on the position and context of the ATGs. Kozak used probabilistic methods and gave the consensus motif GCCRCCatgG around TIS [1]. In this motif, the most highly conserved nucleotides were purine (usually, A) found at -3 position and G found at +4 position. Although the motif was frequently used in biological experiments for initial scanning to identify TIS, the consensus is only a rough guide and cannot predict TIS alone. Later, Kozak used the first AUG rule postulated by the ribosome scanning model to predict TIS [2]. However, in eukaryotes, the process of translation does not always initiate at the first AUG codon suggesting that context information also plays an important role in the prediction of TISs. Therefore, different methods based on biological approaches, machine learning, soft computing, and statistical models have been extensively studied for TIS prediction.

Pedersen and Nielsen, in 1997, proposed a method based on artificial neural networks to predict which AUG is the actual start codon [3]. The network used both local context around start codon and global sequence information. The method did not require any knowledge of the positions of start codons in relation to mRNA's end and can be useful in case of EST data and genome sequences. Though the method attained an accuracy of 85% on the vertebrate sequences dataset, did not optimally utilize the local information around TIS. A method based on a statistical model was developed by Salzberg et al. in 1997 [4] and considered the dependencies of adjacent bases in contrast to other techniques which treated each base independently. This method was appropriate for locating signals in uncharacterized genomic DNA and was showing better results than conventional matrices. It was further enhanced by using interpolated context Markov model to capture the coding potential of the region present downstream ATG and resulted in the overall improvement of 5% when incorporated in the gene finding system [5].

Another method based on a linear discriminant approach was developed by Salamov et al. in 1998 [6]. The probability of each ATG being the true initiation codon was determined on DNA sequences. Later on, Nishikawa et al. improved the accuracy of this method by combining statistical information with similarity to protein sequences [7]. A statistical model which generalized existing methods and took into account higher order dependencies of base position was given by Agarwal and Bafna [8]. The technique employed here is generalized second-order profile (GSP). The main problem with GSP is its less informative content which leads to false positives. To resolve this issue, the ribosome scanning model was investigated, which reduces the search space and accounts for its effectiveness [9].

The problem of TIS prediction can be treated as a classification task. Support Vector Machine (SVM)was used for classification in a number of biological problems and have the capability to ameliorate the prediction performance. Zien et al. manifested how to integrate biological knowledge in SVM by engineering a suitable kernel function [10]. Another attempt, where artificial neural networks were utilized to provide a performance guaranteed prediction was made by Hatzigeorgiou et al. [11]. The method used a modular approach, with one module sensitive to conserved sequences and other to coding/non-coding regions present around ATG. Also, it made use of ribosome scanning model and 94 % accuracy was achieved.

In a different approach, better performance was achieved by using feature generation and correlation-based feature selection with a variety of machine learning algorithms [12]. The method gave an overall accuracy of 90 % when only 7 features were selected and it increased to 94 % when selected features were retrained using ribosome scanning model. The same approach was repeated and features were generated from the translation of mRNA into the corresponding protein sequences rather than directly from mRNAs [13]. This method obtained better sensitivity than existing methods. This work was further enhanced by selecting top 100 features which were integrated using different methods including C4.5, SVM, and Naive Bayes and an accuracy of more than 92% was attained with SVM [14].

The presence of shorter first exon in eukaryotic genes makes the problem of TIS prediction more difficult. Thus, to deal with this problem Wang et al. developed a method where significant characteristics of shorter flanking fragments around TIS were analyzed and the expectation-maximization (EM) algorithm based on incomplete data was applied to recognize TISs [15]. The accuracy reported was 87.8% with a 6-fold cross-validation test. In 2004, five computational methods namely first_AUG, ESTScan, Diogenes, NetStart, and ATGpr were compared to find the most accurate method for TIS recognition [16]. The results of this study indicated that ATGpr is the best method to predict TISs.

Low-order Markov models are not able to capture hidden and complex features which are present in the proximity of TISs. To deal with this fact, Ho et al. proposed a neural network approach in which biological knowledge based lower-order models were combined with non-linearity to represent higher-order nucleotide dependencies at TISs and in the surrounding coding/non-coding regions [17]. With a 3-fold cross-validation test, the method attained 93.8 % of sensitivity and 96.9 % of specificity. Most of the efforts discussed above make use of local information for feature extraction.

The data encoding method greatly affects the performance of TIS prediction in genomic sequences. Moreover, Pederson and Nielsen advised that global information could improve the accuracy of prediction. In 2005, Li et al. presented a TIS prediction method where features based on both local and global information were used to produce numerical data from biological sequences [18]. Then, mixture Gaussian models were applied for TIS prediction. The method outperformed many existing methods in sensitivity while maintaining specificity high. In the same year, Li and Jiang introduced a class of sequence similarity kernels based on the concept of string edit distance [19]. The property of edit kernels is their simplicity and they have important statistical and biological interpretations. A discriminative approach was used in this method where SVM was applied to predict TISs. The results indicated that both these ideas can improve the accuracy of prediction and this method performed better than those methods based on SVM with the polynomial kernel or Salzberg kernel and neural networks.

The similar methodology of feature generation and selection was employed by Tzanis et al. for TIS prediction [20]. A combination of features was used in this study, some of them were taken from the previous work and some new were proposed. A number of classifiers were used here, namely C4.5, RIPPER, Naive Bayes. The experimental results indicate that the use of this new combination of features helps in improving the prediction accuracy. The work was further enhanced by combining this novel feature set with some proposed features and ribosome scanning model [21]. Further, some new features were proposed and combined with old features used in their previous studies [22]. The evaluation of this new feature set showed better prediction accuracy.

Ma et al. analyzed the effect of C + G content surrounding ATG codon on the features used for TIS prediction and they have found that some TIS and non-TIS features are heavily dependent on C + G content [23]. Based on this fact, 10 models were constructed and all were built using SVM with 11 basal features as input. A prediction program named TISKey was developed based on these models. In 2007, Tzanis et al. proposed a component-based data mining methodology named MANTIS for TIS prediction [24]. The methodology was modular in nature and contained three major components: a consensus component which was based on Markov-chain, a coding region classification component and a new component based on ATG location which taken into consideration the advantages of ribosome scanning model and surmounts its limitations. All these components were incorporated into a meta-classifier using a technique called stack generalization.

In 2007, Saeys et al. assessed the performance of several TIS prediction methods at the genomic level and compared them with other existing methods for TIS recognition in transcript data [25]. The results indicated that the proposed model obtained a sensitivity of 80% on a well-annotated human chromosome. Most of the algorithms to predict TIS were designed around the sequence context. These algorithms often failed if the sequence context is not present i.e. a purine at -3 position and guanine at +4 position. In 2008, Tikole et al. developed a neural network method to identify TIS in mRNA sequences that missed the preferred nucleotide at position -3 and +4 around the start site [26]. Sparks et al. developed a package MetWAMer to predict TIS in eukaryotes of non-viral origin [27]. All the methods implemented in METWAMer utilized a specific weight array matrix based on a start-methionine signal which contains base transition frequencies in protein coding sequences. The results demonstrated that improvements in the accuracy of TIS prediction can be attained by taking start-methionines into consideration and the software can be integrated into gene prediction systems with minor modifications.

The simple approach of feature generation and selection was used by Gao et al. with the aim of improving the performance of TIS prediction models [28]. Instead of applying the standard machine learning algorithm on selected features, a new pattern classification algorithm Universum SVM was used here. The highest accuracy achieved with this new algorithm was 96.51% and it was comparable to the best results obtained in the past. The similar approach was also used by Kongmanee et al. and the values of features were calculated using TF-IDF approach [29]. The proposed model showed better performance with less computation time.

Another attempt of predicting TIS, where the context conditions are in weak positions, was made by Husin et al. in 2011 [30]. The main focus of this research effort was to optimize the supervised learning methods to correctly predict TIS in a weak context with a minimal error of rates. The Bayesian model developed in this work outperformed Tikole model by increasing the sensitivity with 10% and specificity with 26%. The focus of previous research efforts for TIS prediction was on mRNA or cDNA sequences. Identifying DNA motifs in genomic sequences which correspond to actual TISs are much more difficult than identifying them in mRNA or cDNA sequences. Mora et al. developed a prediction tool to predict TIS in genomic sequences of plants [31]. Using the information of *Arabidopsis thaliana* (A. t.), the tool required only the genomic sequences, not expressed sequence tags. The accuracy of this TIS prediction method resulted in a sensitivity of 90.75% and specificity of 90.77%.

Although many computational methods were developed for TIS prediction, none of them taken into account global or long-range sequence-order effects of DNA due to which their prediction accuracy was limited. To deal with this kind of effects, Chen et al. developed new software, called iTIS-PseTNC by merging the physicochemical properties into the pseudo trinucleotide composition (PseTNC) [32]. The predictor has shown the overall success rate of 97% when its performance is evaluated using the jackknife test. In a similar attempt, three sequence representation methods namely dinucleotide composition (DNC), pseudo-dinucleotide composition (PseDNC) and trinucleotide composition (TNC) were used to obtain important sequence features in the form of numerical descriptors [33]. The evaluation results showed that this combined feature extraction method with SVM outperformed existing TIS prediction methods reported in the literature. Recently, several attempts are made to predict TIS using ribosome profiling data in both eukaryotes and prokaryotes [34, 35, 36, 37, 38, 39, 40].

Most of the methods developed in the past have focused on the prediction of TISs in cDNA or mRNA sequences. A method which can predict TISs in genomic sequences (i.e. uncharacterized DNA) can be used in the annotation of new genomes. Therefore, in this study, a method is proposed to predict TISs in both genomic as well as in cDNA sequences.

3. Materials and methods

To develop a TIS prediction system, first, the dataset is prepared then the features are extracted from the acquired dataset. The methodology used in the proposed method is given in Figure 1. The details of the dataset and the features used are discussed in the below subsections.

3.1. Dataset

One of the important steps for training and testing the prediction model is to prepare a suitable dataset. For this study, two different datasets are prepared out of them one is used for training and the other is used for the testing the proposed model. Both training and testing dataset includes human genomic sequences. All these sequences are taken from GenBank (a publically available database for nucleotide sequences which is created and maintained by National Centre of Biotechnology Information (NCBI)). The human genomic sequences are searched in GenBank and their Fasta file is downloaded. Among all the sequences downloaded from GenBank, the sequences having 150 bp upstream and downstream the true TIS are considered only. In this way, 755 genomic sequences are selected to form the training set. Then, more number of nucleotide sequences is downloaded from GenBank. After discarding the sequences having length <150 bp surrounding true TIS, 52 sequences are selected to form the testing set. In genomic sequences, false initiation sites are more prevalent than true initiation sites. For this reason, experiments were performed using a different number of randomly selected false initiation sites and two datasets are formed namely balanced and unbalanced. The balanced training dataset consists of 1510 sequences of TISs. Out of these, 755 sequences were taken as true TIS sequences and an equal number of sequences were taken as false TIS sequences. On the other side, the unbalanced training dataset consists of 3020 sequences of

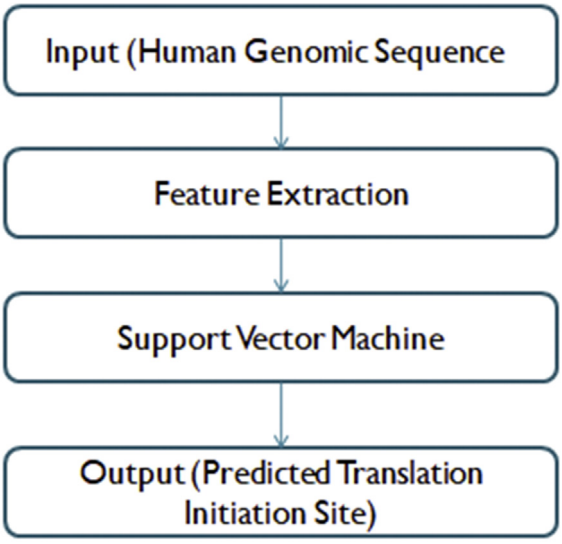

Figure 1. Methodology used.

TISs. Out of these, 755 sequences were taken as true TIS sequences and 2265 sequences were taken as false TIS sequences.

To further validate the performance of the proposed model on mRNA or cDNA sequences a benchmark dataset is taken which was used by two previous studies [32, 33]. The dataset consists of a total of 2318 sequences of TISs. Out of these, 1159 were true TIS sequences and 1159 were false TIS sequences. This dataset can be represented by the following equation:

$$S^{bench} = S^{+} \text{ U } S^{-} \quad (1)$$

where S^{+} contains only true TIS sequences and S^{-} includes only false TIS sequences. The union of these two sets S^{+} and S^{-} is denoted using symbol U. All these sequences were taken from the human genome and genome coordinates of their initiation sites were obtained from the TIS database.

3.2. Feature extraction

The fundamental step in designing a predictor is feature extraction from biological sequence so as to train and test the model in an effective manner. One of the most important and difficult tasks in bioinformatics is to extract useful features from the sequence with a feature vector. This is because most of the existing classifiers like SVM and neural networks make use of feature vector instead of taking the biological sequence directly.

A number of useful features were reported by previous studies in the sequence surrounding the ATG site for predicting TIS [14, 18, 22, 23, 25]. Most of these features were local due to their ability to take into account the context of the sequences surrounding TIS. In this study, the local feature set is extended by adding global features which are affected by nucleotides up to 150bp from the ATG site. Some of these global features were introduced by Mora et al. in a study done in 2013 [31]. In this study, the length of 150 bp upstream and 150 bp downstream the TIS are considered. The same length of 300bp is considered in the proposed work with ATG appearing at location 150–152 when counted in 5′-3′ direction. The length of 300 bp is suitable for extracting global features from the genomic sequences.

The features used in this study fit to several broad categories. The first of these is based on successive k-mers of nucleotides of TIS surrounding sequences and scores generated from position weight matrices (PWMs). The second class of features was formed using statistics related to codon biases. Information gain is also taken as a feature surrounding TIS sequences which form another important class of features used in this study. Finally, a number of motifs are identified in TIS surrounding regions by using Dragon Motif Finder, a tool that identifies groups of similar polynucleotide patterns from a set of genomic sequences and builds various families of short motifs from them [41]. Out of all the features considered, a total of 53 were chosen to design the TIS predictor. The important features employed in this work are described below:

- *K-mer Frequencies:* These represent the frequency of nucleotides and dinucleotides in TIS sequences. Both the regions surrounding TIS sites i.e. upstream and downstream are considered for generating these features.
- *Position Weight Matrices:* Here, two features represented as Pscore and Nscore are introduced, which are calculated in the following manner: By taking the frequencies of 16 dinucleotides i.e. the combinations of A, C, G, and T, two PWMs are created (one from positive samples and another from negative samples) which results in generating two features per sample. The values of Pscore and Nscore from these PWMs are calculated as follows:
  Let S (c_j) is a sequence of length L and P (p_{ij}) is a PWM of L-1 columns and 16 rows (r_1, r_2,.....r_{16}). The Nscore from negative data and Pscore from positive data are computed using the equation given below:

$$\left[\frac{P}{N}\right] Score = \sum_{i=1}^{16} \sum_{j=1}^{L-1} \log_2 \left(\frac{P_{ij} \otimes c_j c_{j+1}}{Pb_i} \right) \quad (2)$$

$$P_{ij} \otimes c_j c_{j+1} = \begin{cases} P_{ij}, & c_j c_{j+1} = r_i \\ 1, & c_j c_{j+1} \neq r_i \end{cases}$$

where P_{bi} denotes the background probability from a uniform distribution.

- *Kozak's Feature:* The Kozak's consensus sequence proposed in 1987 [1] is utilized for this feature. The feature was based on the observation that in TIS surrounding regions the probability of finding an A or G at position -3 and a G +4 at position is high. For each sequence in the training set, it is checked that if the sample sequence matches with the regular expression GCC [A/G]CCatgG then assign this binary feature value of one. Otherwise, a value of zero is assigned to it. Though the feature is important, it is not sufficiently discriminative alone to make an accurate prediction of TIS. In addition to this binary feature, based on Kozak consensus, the score based on this consensus is also used which represents the number of positions that coincide with the Kozak's consensus sequence. For the sample sequence GCCTCAatgG, the score would be 5 as the nucleotides which are underlined are not expected at their position as per Kozak's rule.
- *ATG Frequencies:* The following two features were derived from ATG frequencies:
  1. Total number of ATG nucleotide triplets in the entire sample sequence
  2. A number of ATG in-frame triplets of non-overlapping nucleotides upstream i.e. the number of occurrences of ATG triplets at positions that are aligned inframe. The inframe triplets can appear only downstream of the ATG signal, but to compute the value of this feature, the upstream equivalents are taken as in-frame. With reference to ATG signal starting at position 1, these would be the occurrences of ATG triplets at positions -3, -6,-9 and so on. Towards the 5′ end of the sample sequence counting the A in ATG as position zero.
- *Putative Coding Sequence:* This binary feature designates whether the sample sequence contains an in-frame stop codon (TGA, TAA or TAG) or not. The main idea behind including this feature is that most of the protein sequences are longer than 50 amino acids so it is very rare that a positive sample would contain a stop codon in-frame downstream of ATG. For a similar reason, the probability of finding stop codon in negative samples is higher.
- *G-quadruplets Frequencies:* The probability of having G-quadruplets in the downstream region for positive TIS sequences is high. This generates two features, one corresponds to the number of G-quadruplets in-frame and the other corresponds to the number of G-quadruplets out-frame both aligned to ATG segment.
- *Information Gain:* The analysis of the frequency of nucleotides at specific locations in 300 long positive and negative sequences was done and it was found that a higher level of entropy in negative sequences is more common than in positive sequences. In order to utilize this fact, the entropy is calculated for a given position P and nucleotide N in the training sequences by the following formula:

$$E(P,N) = -p/(p+n)\ \log2(p/(p+n)) - n/(n+p) \log2(n/(n+p)) \quad (3)$$

where p refers to the number of occurrences of nucleotide N at position P in positive sequences and n refers to the number of occurrences of the same nucleotide at the same position in negative sequences.

Another entropy measure at position P is also utilized in this work to adjust the proportion of positive and negative sequences and it is computed as follows:

$$E(P) = -p/(p + n) \log2(p/(p + n))-n/(n + p)\log2(n/(n + p)) \quad (4)$$

where p and n represent the number of positive and negative sequences in the training set.

The information gain for position P is defined by the following equation:

$$Gain(P) = E(P)-E(A,P)-E(C,P)-E(G,P)-E(T,P) \quad (5)$$

The sum of information gain of complete sequence i.e. maximum information gain is then taken as a feature for classification.

- *C and G Frequencies:* Another observation from the analysis was that the frequency of C and G nucleotides is higher in the upstream region of the positive sequences. In this study, C and G frequencies are taken from twenty positions that have the highest information gain resulting in two features i.e. frequencies of C and G.
- *In-frame Nucleotides Score:* This category of features was inspired by the position-specific k-gram approach. We first determined the in-frame triplets and within each of these triplets, we computed the number of As, Cs, Gs, and Ts that are present in each of three positions within these triplets. After this, we sum up for each nucleotide the number of its occurrences in positions 1, 2 and 3 within triplets which are then taken as feature values. There are four nucleotides and three positions so 12 features are formed in this way. Since these features are calculated separately for upstream and downstream in-frames, the total numbers of features generated from this process are 24. Out of these, 14 features are considered for this study. For example, by considering the sequence ATGattgcc we can identify two in-frame triplets i.e. att and gcc as it is downstream and then count the number of occurrences of each of the nucleotides at three different positions. At position 1, we have 1 a, 0 c, 1 g, and 0 t. For position 2, we have 0 a, 1 c, 0 g, 1 t. Similarly, for the last position, we have 0 a, 1 c, 0 g, 1 t. Thus, the feature value corresponding to the referenced sequence would be 1,0,1,0,0,1,0,1,0,1,0,1.
- *Motifs Present:* The dragon motif finder tool is used to identify a number of motif families of various lengths which are present in the upstream, downstream and central regions surrounding the TIS site in positive samples. The above-mentioned regions are defined as follows: the upstream region includes 150 nucleotides from the 5′ end of the sequence to the ATG triplet, the downstream region includes 147 nucleotides between ATG and 3′ end of the sequence and the central region includes 50 nucleotides upstream ATG and 50 nucleotides that exist in the downstream of ATG. The accumulative count of the presence of the sequences of these motif families is then taken as a feature in the prediction model. In this work, 3 features are considered i.e. for upstream, downstream and central region respectively.

A total of 53 features are selected for the proposed model. The procedure used to extract the above mentioned features is as follows: Initially, the genomic sequences are searched for true TIS sites. Then, the sequences surrounding true TIS sites having length 300bp are extracted. Thereafter, for all these sequences, the features mentioned above are computed one by one and stored in a feature metric of size 755 × 53. Here, 755 is the number of genomic sequences and 53 is the number of features considered. The same procedure is applied again to extract the features for false TIS sites. As mentioned earlier, the false TIS sites are more prevalent than true TIS sites in genomic sequences. Therefore, in this work false TIS sites are randomly selected to form balanced and unbalanced dataset. The selected false TIS sites are then used for extracting the above mentioned features one by one. Then, all the features extracted using above procedure are simply added to form the feature vector. The feature vector formed in this way is used further to train the proposed model. In this work, SVM is used to train the selected features and to predict TIS in genomic as well as in cDNA sequences.

3.3. Support Vector Machine (SVM)

SVM is one of the most popular machine learning methods which was introduced in 1995 by Vapnik and his group [42]. It is being used in various domains of machine learning, bioinformatics and speech recognition due to its powerful classification ability. Initially, SVM was designed for binary class prediction but after that, it was enhanced for multi-class prediction. In bioinformatics, classification and prediction problems are very common and many of these problems contain noisy data. SVM has been successfully used to handle such type of data.

The SVM aims to find the maximum margin between classes by transforming the data into high dimensional feature space. The kernel functions are used for mapping data into high dimensional space and for learning non-linearly separable functions [43]. The various kernel functions used in SVM are the linear kernel, polynomial kernel, radial based function kernel, and sigmoid kernel. The prediction accuracy of SVM is highly dependent on the kernel chosen and the value of its parameters. Researchers have proposed new kernels for solving specific problems but most of the problems are still solved using radial based function and polynomial kernel. In this work, SVM is designed using LIBSVM package and the polynomial kernel is utilized for mapping. The class label specifies a value yes for the true site and no for the false site.

4. Experimental setup and results

4.1. Performance evaluation measures

In this study Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), Area Under the ROC Curve (AUC), F-measure and Mathew's correlation coefficient (Mcc) have been used as performance measuring parameters. A large value of these parameters indicates better performance of the classification algorithm. The parameters are defined by the following equations:

$$Sn(\%) = \frac{TP}{TP + FN} \times 100 \quad (6)$$

$$Sp(\%) = \frac{TN}{TN + FP} \times 100 \quad (7)$$

$$Acc(\%) = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (8)$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}} \quad (9)$$

$$F - measure = 2 \times \frac{\frac{TP}{TP+FP} \times Sn}{\frac{TP}{TP+FP} + Sn} \quad (10)$$

where TP, FP, FN, TN denotes the number of true positives, false positives, false negatives and true negatives respectively. In receiver operating characteristic (ROC) curve, sensitivity is plotted against 1-specificity to see the performance of a binary classifier. The area under the ROC curve (AUC) is generally used to recapitulate the performance in a single number. The larger value of AUC indicates the accurate performance of the model.

4.2. Results

The 10-fold cross-validation has been used to evaluate the performance of the proposed method. For this, the dataset is divided into 10 equal sized parts (folds). Out of these, 9 of the folds are used as a training set and the fold which is left used as a testing set. This process is repeated 10 times with different testing set each time and the average of 10 independent testing set is taken as evaluation results.

Table 1. Performance evaluation of the proposed method on the balanced and unbalanced training dataset.

Dataset	Sn (%)	Sp (%)	Acc (%)	AUC	F-measure
Balanced Training	90.00	91.50	90.99	90.90	90.60
Unbalanced Training	82.10	95.80	92.35	88.90	84.20

Table 2. Performance evaluation of proposed method on unbalanced testing dataset.

Dataset	Sn (%)	Sp (%)	Acc (%)	AUC	F-measure
Highly Unbalanced Testing	76.92	97.88	97.76	88.30	28.90

Table 3. Performance evaluation of the proposed method on benchmark dataset.

Dataset	Sn (%)	Sp (%)	Acc (%)	AUC	F-measure	Mcc
Benchmark	97.67	98.18	97.92	97.80	97.80	0.958

Due to the existence of a large difference between true and false sites, two datasets namely balanced (1:1) and unbalanced (1:3) are created from the training set. The balanced dataset is created by taking all the true sites and by randomly choosing an equal number of false sites. On the other hand, for the unbalanced dataset, the number of false sites taken is three times the number of true sites. Here, the unbalanced dataset used for training is not critically imbalanced. In case of critically imbalanced dataset the ratio can be 1: 100 to 1:1,000 or 1: 10,000. The ratio between balanced and unbalanced class varies from one problem to other. This unbalanced dataset is only considered to analyze the effect of imbalance class on the evaluation parameters. The 10-fold cross-validation has been run on both balanced and unbalanced dataset. The independent dataset used for testing is critically imbalanced because the ratio between balanced and unbalanced class is 1:167. This ratio is in case of genomic sequences and it may vary for complete human chromosomes.

As mentioned before, SVM is used in this study for classification and for its implementation the LIBSVM package is used. In this work, C-SVC with the polynomial kernel is employed for TIS prediction. First, the experiments were performed using different values for cost and gamma on the balanced dataset. The optimal values obtained for cost and gamma are 2 and 1 respectively. Then, the experiments were performed using different values for cost and gamma on the unbalanced dataset. The values obtained for cost and gamma, in this case, are 4 and 0.5. The results of both balanced and unbalanced datasets are given in Table 1.

From the results given in Table 1, it is clear that the performance of the proposed method is better on the balanced dataset as compared to the unbalanced dataset. In case of balanced dataset, the values obtained for Sn, Sp, Acc, AUC, and F-measure are 90%, 91.5%, 90.99%, 90.90 and 90.60 respectively. On the other hand for the unbalanced dataset, these values are 82.1%, 95.8%, 92.35%, 88.9 and 84.2 respectively. The value for Sp and Acc is higher in case of unbalanced training dataset. In unbalanced training dataset, the number of false TIS sites is three times more than true TIS sites. These results show that the prediction of false sites dominates the prediction of true sites due to large number of false sites. However, in genome annotation the aim is to obtain high value of both sensitivity (Sn) and specificity (Sp) at the same time. Both these values are higher in case of balanced dataset. Also, the value of accuracy depends on both Sn and Sp and due to large value of Sp this value is higher in case of unbalanced dataset. But the value of AUC and F-measure is better in case of balanced dataset which demonstrates that the overall performance is better in case of balanced dataset. It is evident from the above discussion that the ratio between balanced and unbalanced dataset largely affects the performance of TIS prediction model.

After this, the experiments are conducted on the independent test dataset. The dataset contains a total of 8718 sites, out of these 52 are true sites and 8666 are false sites. The testing dataset is highly unbalanced due to the presence of a large number of false TIS sites. The ratio of true and false TIS sites is 1:167 approximately. Also here, the experiments are repeated using different values for cost and gamma. The optimal values obtained for cost and gamma on the testing set are 1 and 0.5. The evaluation results of the testing set are given in Table 2.

From the results, it is evident that the performance of the proposed method on testing set is not as good as it is for the training set and this is due to its highly unbalanced composition. The values for Sn, Sp, Acc, AUC, and F-measure, in this case, are 76.92%, 97.88%, 97.76%, 88.30 and 28.90 respectively.

An additional evaluation is performed on the benchmark dataset consisting of cDNA sequences to analyze the consistency of the proposed method for both cDNA as well as for genomic sequences. The dataset contains an equal number of true and false splice sites. After performing the experiments with different values of cost and gamma, the best results were obtained with values 2 and 1 respectively. The results are summarized in Table 3.

The proposed method has given promising results on the benchmark dataset. The values for Sn, Sp, Acc, AUC, F-measure, and Mcc, in this case, are 97.67%, 98.18%, 97.92%, 97.80, 97.80 and 0.958 respectively. The results of some existing methods for TIS prediction on benchmark dataset is taken from one published paper [33]. All of the methods considered here are designed to predict TIS in mRNA or cDNA sequences. The main purpose of including these results is to compare the performance of the proposed method with some popular and recent TIS prediction methods. The methods taken for comparison are StartScan [25], iTIS-PseTNC [32] and iTIS-PseKNC [33]. These results are shown in Table 4.

The prediction results of StartScan method were 95.32% of sensitivity, 96.43% of specificity, 96.02% of accuracy and 0.921 of Mcc. The

Table 4. Performance evaluation of existing methods on benchmark dataset.

Method	Sn (%)	Sp (%)	Acc (%)	Mcc
StartScan [25]	95.32	96.43	96.02	0.921
iTIS-PseTNC [32]	97.49	98.42	97.92	0.958
iTIS-PseKNC [33]	99.31	99.48	99.40	0.988
Proposed Method	**97.67**	**98.18**	**97.92**	**0.958**

prediction results of iTIS-PseTNC were 97.49% of sensitivity, 98.42% of specificity, 97.92% of accuracy and 0.958 of Mcc. One method named iTIS-PseKNC was proposed recently and its prediction results were 99.31.49% of sensitivity, 99.48% of specificity, 99.40% of accuracy and 0.988 of Mcc.

The overall performance of the proposed model suggests that it can be used in the annotation of new genomes. The proposed model only needs genomic sequences to make predictions not expressed sequences. The results of the proposed model are promising in case of genomic sequences. Also, in case of cDNA sequence the model shown comparable performance among recent TIS predictors. The accurate prediction of TISs can help in finding new protein-coding genes and improve annotation of new and existing genomes. Identifying TIS in genomic sequences that corresponds to actual TIS signals is more challenging than identifying them in cDNA/mRNA sequences and the proposed model addresses this challenge.

5. Discussions and conclusion

Identifying TIS is important for successful genome annotation. In this paper, a method is proposed for predicting TIS locations in human genomic DNA sequences by taking into consideration both local and global sequence features. The proposed method is based on SVM classifier. The performance results of the proposed method suggest that it can be used with good success for annotating TIS in eukaryotes. However, the data used in this work is not precisely comparable with those provided by previous studies [10, 11, 12, 13, 14, 15, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32, 33]. For this reason the results of these studies are incomaparable as well. This is due to differing experimental designs as most of these studies focused on cDNA sequences and some studies used genomic sequences of other genomes like *Arabidopsis thaliana, Vitis vinifera, Populus trichocarpa* etc. [31]. The comparison of proposed method with these published studies is often not practical.

We are only able to utilize StartScan [25], iTIS-PseTNC [32] and iTIS-PseKNC [33] to compare against the proposed method and that too on benchmark dataset of cDNA sequences. Among all these methods, StartScan is the only tool designed to work on genomic sequences. The proposed method demonstrates improved prediction accuracy over the reported TIS predictor which is achieved by 10-fold cross validation. The comparison of proposed method with iTIS-PseTNC and iTIS-PseKNC also helps in finding the best feature extraction technique when SVM is used for classification. Most of the feature extraction techniques adopted by previously published methods were suitable for TIS prediction in mRNA or cDNA sequences only. The comparison demonstrates that the features used in this study are appropriate for TIS prediction in both genomic as well as in cDNA sequences. Although, the proposed method is trained for human genomic sequences, it can be used to predict TIS in other eukaryotic genomes as well. Also, it overcomes the weakness of current TIS predictors to deal with high class imbalance between true and false TISs.

We hope that the proposed method will find its use in annotation of human genome and may provide insight to understand the biological mechanism of translation initiation. On the other line, the proposed method can be incorporated into existing gene prediction systems and can play a complementary role to existing methods in this field. In the future, it can also be used to predict TISs in newly sequenced genomes.

Declarations

Author contribution statement

N. Goel: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

S. Singh: Contributed reagents, materials, analysis tools or data.

T. Chand: Conceived and designed the experiments.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e04825.

References

[1] M. Kozak, An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs, Nucleic Acids Res. 15 (1987) 8125–8148.
[2] M. Kozak, The scanning model for translation : an update, J. Cell Biol. 108 (1989) 229–241.
[3] A.G. Pedersen, H. Nielsen, Neural network prediction of translation initiation sites eukaryotes : perspectives for EST and genome analysis, in: Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, AAAI, 1997.
[4] S.L. Salzberg, A method for identifying splice sites and translational start sites in eukaryotic mRNA, Comutational Appl. Biosci. 13 (1997) 365–376.
[5] M. Pertea, S.L. Salzberg, A method to improve the performance of translation start site detection and its application for gene finding, in: Algorithms in Bioinformatics, Lecture Notes in Computer Science, Springer, 2002.
[6] A.A. Salamov, T. Nishikawa, M.B. Swindells, Assessing protein coding region integrity in cDNA sequencing projects, Bioinformatics 14 (1998) 384–390.
[7] T. Nishikawa, T. Ota, T. Isogai, Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences, Bioinformatics 16 (2000) 960–967.
[8] P. Agarwal, V. Bafna, Detecting non-adjoining correlations within signals in DNA, in: RECOMB, 1998.
[9] P. Agarwal, V. Bafna, The ribosome scanning model for translation initiation : implications for gene prediction and full-length cDNA detection, in: Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology, AAAI, 1998.
[10] A. Zien, G. RatSch, S. Mika, et al., Engineering support vector machine kernels that recognize translation initiation sites, Bioinformatics 16 (2000) 799–807.
[11] A.G. Hatzigeorgiou, Translation inititation start prediction in human cDNAs with high accuracy, Bioinformatics 18 (2002) 343–350.
[12] F. Zeng, R.H. Yap, L. Wong, Using feature generation and feature selection for accurate prediction of translation initiation sites, Genome Inf. 13 (2002) 192–200.
[13] J. Li, S.-K. Ng, L. Wong, Bioinformatics adventures in database research, in: Database Theory, Lecture Notes in Computer Science, Springer-Verlag, 2003.
[14] H. Liu, L. Wong, Data mining tools for biological sequences, J. Bioinf. Comput. Biol. 15 (2003) 1–29.
[15] Y. Wang, H. Ou, F. Guo, Recognition of translation initiation sites of eukaryotic genes based on an EM algorithm, J. Comput. Biol. 10 (2003) 699–708.
[16] A. Nadershahi, S.C. Fahrenkrug, L.B.M. Ellis, Comparison of computational methods for identifying translation initiation sites in EST data, BMC Bioinf. 10 (2004).
[17] L.S. Ho, J.C. Rajapakse, High sensitivity technique for translation initiation site detection, in: Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, 2004.
[18] G. Li, T. Leong, L. Zhang, Translation initiation sites prediction with mixture Gaussian models in human cDNA sequences, IEEE Trans. Knowl. Data Eng. 17 (2005) 1152–1160.
[19] H. Li, T. Jiang, A class of edit kernels for SVMs to predict translation initiation sites in Eukaryotic mRNAs, J. Comput. Biol. 12 (2005) 702–718.
[20] G. Tzanis, C. Berberidis, A. Alexandridou, et al., Improving the accuracy of classifiers for the prediction of translation initiation sites in genomic sequences, in: Advances in Informatics, Lecture Notes in Computer Science, Springer-Verlag, 2005.
[21] G. Tzanis, C. Berberidis, I. Vlahavas, A novel data mining approach for the accurate prediction of translation initiation sites, in: Biological and Medical Data Analysis, Lecture Notes in Computer Science, Springer-Verlag, 2006.
[22] G. Tzanis, I. Vlahavas, Prediction of translation initiation sites using classifier selection, in: Advances in Artificial Intelligence, Lecture Notes in Computer Science, Springer-Verlag, 2006.
[23] C. Ma, D. Zhou, Y. Zhou, Feature mining and integration for improving the prediction accuracy of translation initiation sites in eukaryotic mRNAs, in: Fifth International Conference on Grid and Cooperative Computing Workshops, IEEE Computer Society, 2006.
[24] G. Tzanis, C. Berberidis, I. Vlahavas, MANTIS: a data mining methodology for effective translation initiation site prediction, in: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007.
[25] Y. Saeys, T. Abeel, S. Degroeve, et al., Translation initiation site prediction on a genomic scale : beauty in simplicity, Bioinformatics 23 (2007) i418–i423.

[26] S. Tikole, R. Sankararamakrishnan, Biochemical and Biophysical Research Communications Prediction of translation initiation sites in human mRNA sequences with AUG start codon in weak Kozak context : a neural network approach, Biochem. Biophys. Res. Commun. 369 (2008) 1166–1168.
[27] M.E. Sparks, V. Brendel, MetWAMer, Eukaryotic translation initiation site prediction, BMC Bioinf. 9 (2008) 1–16.
[28] T. Gao, Y. Tian, X. Shao, et al., Accurate prediction of translation initiation sites by Universum SVM, in: The Second International Symposium on Optimization and System Biology, Lijiang, China, 2008.
[29] T. Kongmanee, The TF-IDF and neural networks approach for translation initiation site prediction, in: 2nd IEEE International Conference on Computer Science and Information Technology, IEEE, 2009.
[30] N.A. Husin, N.S. Herman, B. Hussin, Comparative analysis using Bayesian approach to neural network of translational initiation sites in alternative polymorphic context, Int. I. Bioautomation 15 (2012) 251–260.
[31] A.M. Mora, H. Ashoor, B.R. Jankovic, et al., Dragon TIS Spotter : an Arabidopsis-derived predictor of translation initiation sites in plants, Bioinformatics 29 (2013) 117–118.
[32] W. Chen, P. Feng, E. Deng, et al., iTIS-PseTNC : a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83. Available from:.
[33] M. Kabir, M. Iqbal, S. Ahmad, et al., iTIS-PseKNC : identification of translation initiation site in human genes using pseudo k-tuple nucleotides composition, Comput. Biol. Med. 66 (2015) 252–257. Available from:.
[34] K. Reuter, A. Biehl, L. Koch, et al., PreTIS : a tool to predict non-canonical 5 ' UTR translational initiation sites in human and mouse, PLoS Comput. Biol. 1–22 (2016).
[35] B. Malone, I. Atanassov, F. Aeschimann, et al., Bayesian prediction of RNA translation from ribosome profiling, Nucleic Acids Res. 45 (2017) 2960–2972.
[36] S. Zhang, H. Hu, T. Jiang, et al., TITER : predicting translation initiation sites by deep learning, Bioinformatics 33 (2017) i234–i242.
[37] A. Giess, V. Jonckheere, E. Ndah, et al., Ribosome signatures aid bacterial translation initiation site identification, BMC Biol. 15 (2017) 1–14.
[38] P. Zhang, D. He, Xu, et al., Genome-wide identification and differential analysis of transslation initiation, Nat. Commun. 8 (2017) 1749.
[39] C.H. Na, M.A. Barbhuiya, M. Kim, et al., Discovery of non-canonical translation initiation sites through mass spectrometric analaysis of protein N termini, Genome Res. 28 (2018) 25–36.
[40] J. Clanwaert, G. Menschaert, W. Waegeman, DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns, Nucleic Acids Res. 47 (2019) e36.
[41] B. Marchand, V.B. Bajic, D.K. Kaushik, Highly scalable Ab initio genomic motif identification, in: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. New York, 2011.
[42] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. Available from: http://link.springer.com/10.1007/BF00994018.
[43] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1997) 121–167.