Prediction of Translation Initiation Sites Using Classifier Selection

George Tzanis and Ioannis Vlahavas

Department of Informatics, Aristotle University of Thessaloniki,
Thessaloniki 54124, Greece
{gtzanis, vlahavas}@csd.auth.gr - http://mlkd.csd.auth.gr

Abstract. The prediction of the translation initiation site (TIS) in a genomic sequence is an important issue in biological research. Several methods have been proposed to deal with it. However, it is still an open problem. In this paper we follow an approach consisting of a number of steps in order to increase TIS prediction accuracy. First, all the sequences are scanned and the candidate TISs are detected. These sites are grouped according to the length of the sequence upstream and downstream them and a number of features is generated for each one. The features are evaluated among the instances of every group and a number of the top ranked ones are selected for building a classifier. A new instance is assigned to a group and is classified by the corresponding classifier. We experiment with various feature sets and classification algorithms, compare with alternative methods and draw important conclusions.

1 Introduction

The rapid technological advances of the last years have assisted the conduct of large scale experiments and research projects in biology. The completion of these efforts has lead to a giant collection of biological data. The development and use of methods for the management and analysis of these data is necessary. As a consequence to this need, a new research area called bioinformatics has emerged. Bioinformatics is an interdisciplinary area positioned at the intersection of biology, computer science, and information technology.

A large portion of biological data is represented by sequences. These sequences characterize a large molecule that is a succession of a number of smaller molecules. The study of the structure and function of such large molecules (macromolecules) is the mission of molecular biology. The scientists intend to discover useful biological knowledge by analyzing the various genomic sequences. The utilization of exploratory techniques in order to describe the vast amount of data is required. However, the use of traditional analysis techniques is not adequate and novel, high performance tools have to be developed. The field of data mining aims to provide efficient computational tools to overcome the obstacles and constraints posed by the traditional statistical methods.

Translation is one of the basic biological operations that attract biologist's attention. Translation along with replication and transcription make possible the transmission and expression of an organism's genetic information. The initiation of translation plays an important role in understanding which part of a sequence is translated and consequently what is the final product of the process. When the way that each of these operations takes place is explained, biologists will be one step closer to the unraveling of the mystery of life, which is the final objective of biology.

A sequence contains a number of sites where the translation might initiate. However, only one of them is the true *translation initiation site* (*TIS*). The recognition of the true TIS among the candidate TISs is not a trivial task and requires the use of data mining tools. Classification methods have been extensively used in order to deal with this problem. The idea of multiple classifier systems is an attempt to construct more accurate classification models by combining a number of classifiers. Classifier combination includes two main paradigms: classifier selection and classifier fusion. In the first case a new instance is classified by selecting the appropriate classifier, while in the second case a new instance is classified according to the decisions of all the classifiers.

In this paper we have followed an approach for classifier selection to tackle the problem of the prediction of TISs in DNA sequences. The traditional data mining methods are not directly applicable to sequence data. Thus, we had to transform the initial set of raw sequences to a new dataset consisting of a number of feature vectors that describe the initial data. In particular, all the sequences are scanned and the candidate TISs are detected. The candidate TISs are grouped according to the length of the sequence compartment upstream and downstream them and a number of features is generated for each one. The features are evaluated among the instances of every group according to their impact in the accuracy of classification. Then, a number of the top ranked features are selected for building a classifier. A new instance is assigned to one of the groups and is classified by the corresponding classifier. We experiment with various feature sets and classification algorithms, we compare with alternative methods and draw important conclusions.

This paper is outlined as follows: In the next section we briefly present the relative work in the area of TIS prediction. In section three we provide the necessary background knowledge. In section four our approach is presented in more detail. Section five contains the description of the dataset, the algorithms and the evaluation method we have used as well as the results of our experiments. Finally, in section six we present our conclusions and some directions for future research.

2 Related Work

The prediction of TISs has been extensively studied using biological approaches, data mining techniques and statistical models. In 1978 Kozak and Shatkin [8] proposed the ribosome scanning model, which was later updated by Kozak [7]. According to this model, translation initiates at the first candidate TIS that has an appropriate context. Later, in 1987 Kozak developed the first weight matrix for the identification of TISs in cDNA sequences [6]. The following consensus pattern was derived from this matrix: GCC[**AG**]CCatg**G**. Bold letters denote the highly conserved positions. Meanwhile, Stormo et al. [16] had used the perceptron algorithm to distinguish the TISs.

Pedersen and Nielsen [13] used artificial neural networks (ANNs) to predict which AUG codons are TISs achieving an overall accuracy of 88% in Arabidopsis thaliana dataset and 85% in vertebrate dataset. Zien et al. [20] studied the same vertebrate dataset, employing support vector machines. Hatzigeorgiou [3] proposed "DIANA-TIS", an ANN system consisting of two modules: the consensus ANN, sensitive to the conserved motif and the coding ANN, sensitive to the coding or non-coding context around the initiation codon. The method was applied in human cDNA data and 94% of the TIS were correctly predicted. ATGpr, developed by Salamov et al. [15], is a program that uses a linear discriminant approach for the recognition of TISs. Nishikawa et al. [12] presented an improvement of ATGpr, named ATGpr_sim, which employs a new prediction algorithm based on both statistical and similarity information and achieves better performance in terms of sensitivity and specificity. Li et al. in [9] utilized Gaussian Mixture Models for the prediction of TISs.

In [11] and [19] the researchers have utilized feature generation and feature selection methods with various machine learning algorithms. In their studies, they used a large number of features concerning the frequency of nucleotide patterns. Using a ribosome scanning model along with the best selected features they achieved an overall accuracy of 94% on the vertebrate dataset of Pedersen and Nielsen. Later, in [10] the same approach was used, but instead of nucleotide patterns, amino acid patterns were generated.

3 Background Knowledge

The main structural and functional molecules of an organism's cell are *proteins*. The information concerning the synthesis of each protein is encoded by the genetic material of the organism. The genetic material of almost every living organism is *deoxyribonucleic acid* (*DNA*). There are exceptions of some viruses that have *ribonucleic acid* (*RNA*) as genetic material. Moreover, RNA has many other functions and plays an important role in protein synthesis. DNA and RNA belong to a family of molecules called nucleic acids. Both proteins and nucleic acids are sequences of smaller molecules, *amino acids* and *nucleotides* respectively. A sequence can be represented as a string of different symbols. There are twenty amino acids and five nucleotides. Every nucleotide is characterized by the nitrogenous base it contains: adenine (A), cytosine (C), guanine (G), thymine (T), or uracil (U). DNA may contain a combination of A, C, G, and T. In RNA U appears instead of T. DNA and RNA sequences have two ends called the $5'$ and the $3'$ end and are directed from the $5'$ to the $3'$ end ($5' \rightarrow 3'$).

Proteins are synthesized by the following process. DNA is transcribed into a messenger RNA (mRNA) molecule (transcription). Then mRNA is used as template for the synthesis of a protein molecule (translation). In our setup, we focus on the process of translation, which is further explained below.

Translation takes place by an organelle called ribosome. The mRNA sequence is scanned by the ribosome, which reads triplets, or *codons*, of nucleotides and "translates" them into amino acids. Thus, a protein consisting of *n* amino acids is encoded by a sequence of 3*n* nucleotides. Since there are 64 different triplets formed from an alphabet of four nucleotides and the total number of amino acids is 20, it is obvious that some amino acids are encoded by more than one codon. Moreover, the triplet AUG, that encodes amino acid methionine is also used as a translation initiation codon. Finally, there are three stop codons for the termination of translation (UAG, UAA and UGA).

An mRNA sequence can be read in three different ways in a given direction. Each of these ways of reading is referred to as *reading frame*. The reading frame that is translated into a protein is named *Open Reading Frame* (*ORF*).

Translation, usually, initiates at the AUG codon nearest to the $5'$ end of the mRNA sequence. However this is not always the case, since there are some escape mechanisms that allow the initiation of translation at following, but still near the $5'$ end AUG codons. Due to these mechanisms the recognition of the TIS on a given sequence becomes more difficult.

After the initiation of translation, the ribosome moves along the mRNA molecule, towards the $3'$ end (the direction of translation is $5' \to 3'$) and reads the next codon. This process is repeated until the ribosome reaches a stop codon. For each codon read the proper amino acid is brought to the protein synthesis site by a transfer RNA (tRNA) molecule. The amino acid is joined to the protein chain, which by this way is elongated.

A codon that is contained in the same reading frame with respect to another codon is referred to as *in-frame codon*. We name *upstream* the region of a nucleotide sequence from a reference point towards the $5'$ end. Respectively, the region of a nucleotide sequence from a reference point towards the $3'$ end is referred to as *downstream*. In TIS prediction problems the reference point is an AUG codon. The above are illustrated in Fig. 1.

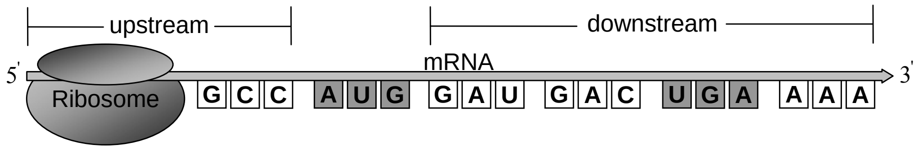

Fig. 1. Translation initiation – The ribosome scans the mRNA sequence from the $5'$ end to the $3'$ end until it reads an AUG codon. If the AUG codon has appropriate context, the translation initiates at that site and terminates when a stop codon (i.e. UGA) is read. An in-frame codon (in relation with AUG) is represented by three consecutive nucleotides that are grouped together

4 Our Approach

In this section we describe the approach we have followed in order to construct a multiple classifier system for the prediction of TISs in genomic sequences. Our approach consists of a number of steps. Each of these steps is described in detail in the following lines.

- **Step 1**: All sequences are scanned and every candidate TIS is detected as shown in Fig. 2 (In the rest of the paper we use the DNA alphabet, since the original dataset we have used contains DNA sequences. See section 5.1).
- **Step 2**: The candidate TISs found in step 1 are grouped according to the length of the sequence compartment upstream and downstream them. By this way the initial dataset of candidate TISs is divided into a number of smaller datasets (Fig. 3). In our setup we have divided the initial dataset in 4 smaller datasets. Table 1 lists the portion of the whole dataset that each of the four data subsets constitutes. We name $D_{m\text{-}n}$ a dataset that contains candidate TISs, that their feature values are calculated by considering *m* upstream and *n* downstream nucleotides.
- **Step 3**: For each of the candidate TISs the value of a number of features is calculated. More details about these features are listed in Table 2. Some of them (updown_*x*, up_pos_*k*_*x*, down_pos_*k*_*x*) have been proposed in our previous work [17] and have been found to present good performance in terms of classification accuracy.
- **Step 4**: The features are evaluated among the instances of every group according to their impact in the accuracy of classification. In our setup we have used the information gain measure.
- **Step 5**: A number of the top ranked features is selected and a classifier is built for each of the data subsets.

Finally, a new instance, namely a new candidate ATG, is assigned to one of the groups according to the length of its upstream and downstream regions' length and is classified by the corresponding classifier.

5' AGCC**ATG**GCATTCCGT**ATG**TTCTG**ATG**TTAA 3'

↓

1, upstream length: 4, downstream length: 24
2, upstream length: 16, downstream length: 12
3, upstream length: 24, downstream length: 4

Fig. 2. A sequence is scanned and every candidate TIS (ATG codon) is detected. Then, its upstream and downstream length is calculated in order to decide in which group belongs

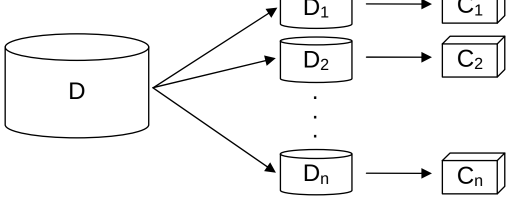

Fig. 3. The initial dataset *D* is divided into a number of smaller datasets D_i and finally a classifier C_i is built separately for each D_i

Table 1. The four data subsets used in our setup

Dataset	Portion of Initial Dataset
$D_{99\text{-}99}$	12.3 %
$D_{99\text{-}120}$	28.3 %
$D_{120\text{-}99}$	52.5 %
$D_{120\text{-}120}$	6.9 %

Table 2. The features used in our approach

Feature	Description
up_*x*	Counts the number of occurrences of amino acid *x* in the upstream region
down_*x*	Counts the number of occurrences of amino acid *x* in the downstream region
up-down_*x*	Counts the difference between the number of occurrences of amino acid *x* in the upstream region and the number of occurrences of amino acid *x* in the downstream region
up_pos_*k*_*x*	Counts the number of occurrences of nucleotide *x* in the k^{th} position of the upstream in-frame codons ($k \in \{1, 2, 3\}$)
down_pos_*k*_*x*	Counts the number of occurrences of nucleotide *x* in the k^{th} position of the downstream in-frame codons ($k \in \{1, 2, 3\}$)
up_-3_[AG]	A Boolean feature that is true if there is an A or a G nucleotide three positions before the ATG codon, according to Kozak's pattern (GCC[**AG**]CCatg**G**)
down_+1_G	A Boolean feature that is true if there is a G nucleotide in the first position after the ATG codon, according to Kozak's pattern (GCC[**AG**]CCatg**G**)
up_ATG	A Boolean feature that is true if there is an in-frame upstream ATG codon
down_stop	A Boolean feature that is true if there is an in-frame downstream stop codon (TAA, TAG, TGA)

5 Experiments

In this section we describe the dataset, the algorithms and the evaluation method we have used along with the results of our experiments.

5.1 Dataset

The original dataset we have used consists of 3312 genomic sequences collected from various vertebrate organisms. These sequences were extracted from GenBank, the US NIH genetic sequence database [2]. Only the sequences that contain an annotated TIS are included. The dataset is publicly available in [5]. The DNA sequences have been processed and the interlacing non-coding regions (introns) have been removed. Since they are DNA sequences, they contain only the letters A, C, G and T. Thus, a candidate TIS is referred to as ATG codon instead of AUG codon. Almost 25% of the ATGs in these sequences are true TISs.

In order to compare our approach we have used two datasets -derived from the original- each of them containing the entire set of candidate TISs. The candidate TISs in the first dataset are described by feature values calculated for 99 positions upstream and 99 downstream ($D_{99\text{-}99}$), while in the second dataset are described by feature values calculated for 120 positions upstream and 120 downstream ($D_{120\text{-}120}$). Note that $D_{m\text{-}n}$ here refers to a dataset containing the complete set of candidate TISs, that their feature values are calculated by considering *m* upstream and *n* downstream nucleotides and is different from the corresponding $D_{m\text{-}n}$ dataset of our approach, that contains only a portion of candidate TISs (see step 2 in section 4).

5.2 Algorithms

For the conduction of our experiments we have utilized the Weka library of machine learning algorithms [18]. We have used the following three classification algorithms:

- C4.5, that is a decision tree construction algorithm [14].
- Naïve Bayes classifier [4].
- PART, a rule learner [1].

5.3 Evaluation

In order to evaluate the results of our experiments we have used stratified 10-fold cross-validation (CV). In particular, the performance of a classifier on a given dataset using 10-fold CV is evaluated as following. The dataset is divided into 10 non-overlapping almost equal size parts (folds). In stratified CV each class is represented in each fold at the same percentage as in the entire dataset. After the dataset has been divided, a model is built using 9 of the folds as a training set and the remaining fold as a test set. This procedure is repeated 10 times with a different test set. The evaluation procedure of our approach is depicted in Fig. 4.

In order to increase the reliability of the evaluation, we have repeated each experiment 10 times and we finally took into account the average of the results.

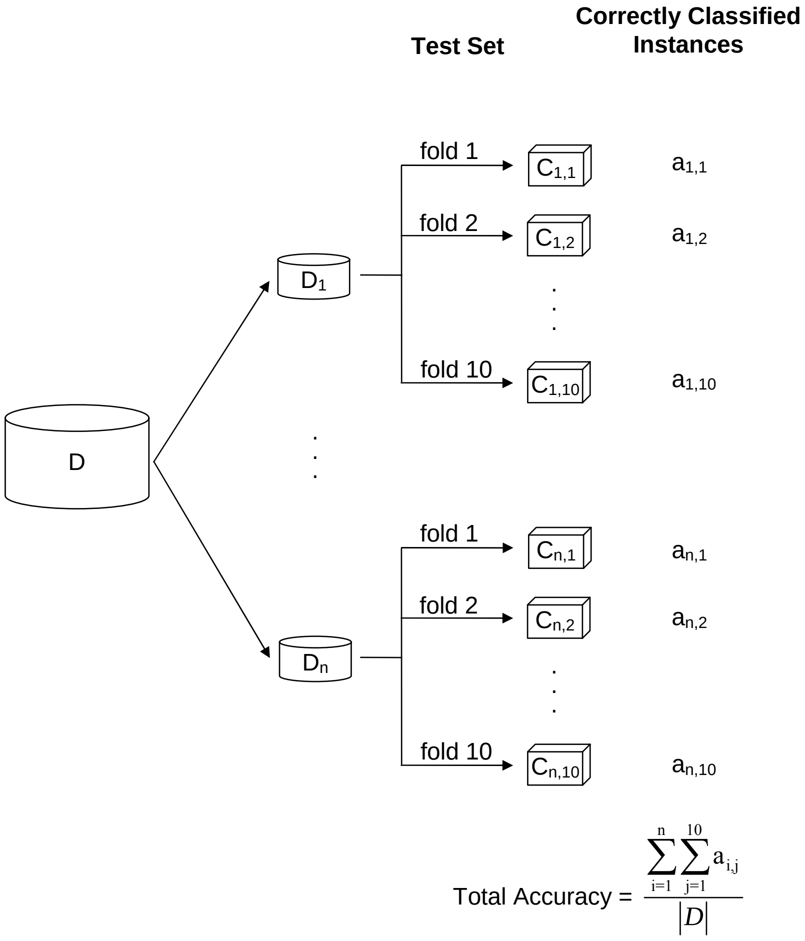

Fig. 4. The initial dataset *D* is divided into a number of smaller datasets D_i. For each D_i 10 classifiers are built and evaluated according to the 10-fold CV procedure. At the end, each instance of the initial dataset *D* will have been tested exactly once (in our setup $n = 4$)

5.4 Results

We have built classifiers by using various numbers of the top ranked, according to information gain measure, features. The results concerning the classification accuracy of each classifier are listed in Table 3. As shown in the table our approach performs better in almost every case. In particular, when C4.5 was used the difference between the best accuracy of our method and the best accuracy of anyone of the other approaches was 1.25%. When the Naïve Bayes classifier was used, this difference increased to 2.35% and when PART was used the difference was 1.11%.

Table 3. Classification accuracy of our multiple classifier system (*MCS*) and the classifiers built on datasets $D_{99\text{-}99}$ and $D_{120\text{-}120}$

Algorithm	Top Features	MCS	$D_{99\text{-}99}$	$D_{120\text{-}120}$
C4.5	50	90.30 %	89.82 %	84.15 %
	30	90.70 %	90.10 %	89.77 %
	20	91.43 %	92.82 %	90.21 %
	15	91.97 %	92.95 %	90.43 %
	12	92.26 %	93.01 %	92.79 %
	9	92.98 %	92.63 %	92.65 %
	7	93.27 %	92.21 %	92.36 %
	5	94.26 %	91.44 %	91.98 %
	3	93.59 %	91.12 %	91.92 %
Naïve Bayes	50	91.69 %	88.93 %	83.13 %
	30	92.89 %	88.37 %	89.37 %
	20	92.35 %	90.54 %	89.27 %
	15	91.73 %	90.08 %	88.56 %
	12	88.55 %	89.04 %	88.22 %
	9	88.49 %	88.00 %	86.94 %
	7	87.91 %	87.44 %	85.92 %
	5	86.98 %	85.24 %	84.30 %
	3	85.88 %	82.37 %	81.37 %
PART	50	90.56 %	89.60 %	83.60 %
	30	91.29 %	89.87 %	88.89 %
	20	92.31 %	92.86 %	89.14 %
	15	92.71 %	92.84 %	90.21 %
	12	92.56 %	93.08 %	92.86 %
	9	92.98 %	93.02 %	92.53 %
	7	93.32 %	92.22 %	92.06 %
	5	94.19 %	91.45 %	91.67 %
	3	93.84 %	91.18 %	91.46 %

We have also conducted experiments for datasets $D_{99\text{-}99}$ and $D_{120\text{-}120}$, using the features proposed in [10] (up_ATG, down_stop, up_-3_[AG], down_A, down_V, up_A, down_L, down_D, down_E, up_G). The results are presented in Table 4.

Using the same reasoning as in the comparisons above we can say that the differences of the best cases for each algorithm range from 3.51% to 3.67%, concluding that our approach performs better.

Table 4. Classification accuracy of classifiers built on datasets $D_{99\text{-}99}$ and $D_{120\text{-}120}$ using the features proposed in [10]

Algorithm	$\mathbf{D_{99\text{-}99}}$	$\mathbf{D_{120\text{-}120}}$
C4.5	90.29 %	90.59 %
Naïve Bayes	88.24 %	89.00 %
PART	90.34 %	90.68 %

6 Conclusions and Future Work

Translation is one of the basic biological processes and the accurate prediction of the translation initiation site in a genomic sequence is crucial for biologists. However, this is not a trivial task. First of all, the knowledge about the process of translation is limited. It is known that translation initiates at the first AUG codon of mRNA in more than 90% of eukaryotic organisms, but some escape mechanisms prevent this. The exact way that each of these mechanisms works, has not been explained up till now. Moreover, the available sequences are not always complete and contain errors.

In this paper, we considered the utilization of a large number of features. We constructed a multiple classifier system and used classifier selection in order to classify a new instance. For this purpose we developed a method for separating the candidate TISs according to the length of the sequence compartment upstream and downstream them. Then, a classifier is built for each data subset. We applied our approach on a real-world dataset that contains processed DNA sequences from vertebrates. We used various classification algorithms and after extensive experimentation we discovered that the use of our method improves the accuracy of classification.

The study of different ways of separation of the candidate TISs is involved in our future plans. Additionally, we aim to use more datasets and possibly from different kind of organisms. Finally, the experimentation with novel features is always under consideration.

Acknowledgements

This work was partially supported by the Greek R&D General Secretariat through a PABET-NE program (EPAN M.4.3.2, No. 04BEN51).

References

1. Frank, E., and Witten, I.H.: Generating Accurate Rule Sets Without Global Optimization. In Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconson, USA, (1998) 144-151
2. GenBank Overview. `http://www.ncbi.nlm.nih.gov/Genbank/index.html`

3. Hatzigeorgiou, A.: Translation Initiation Start Prediction in Human cDNAs with High Accuracy. Bioinformatics (2002) 18(2) 343-350
4. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, California, USA (1995) 338-345
5. Kent Ridge Biomedical Data Set Repository. `http://sdmc.i2r.a-star.edu.sg/rp/`
6. Kozak, M.: An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs. Nucleic Acids Research (1987) 15(20) 8125-8148
7. Kozak, M.: The Scanning Model for Translation: An Update. The Journal of Cell Biology (1989) 108(2) 229-241
8. Kozak, M., Shatkin. A.J.: Migration of 40 S Ribosomal Subunits on Messenger RNA in the Presence of Edeine. Journal of Biological Chemistry (1978) 253(18) 6568-6577
9. Li, G., Leong, T-Y, Zhang, L: Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences. IEEE Transactions on Knowledge and Data Engineering (2005) 8(17) 1152-1160.
10. Liu, H., Han, H., Li, J., Wong, L.: Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites. In Silico Biology (2004) 4(3) 255-269
11. Liu, H., Wong, L.: Data Mining Tools for Biological Sequences. Journal of Bioinformatics and Computational Biology, (2003) 1(1) 139-168
12. Nishikawa, T., Ota, T., Isogai, T.: Prediction whether a Human cDNA Sequence Contains Initiation Codon by Combining Statistical Information and Similarity with Protein Sequences. Bioinformatics (2000) 16(11) 960-967
13. Pedersen, A.G., Nielsen, H.: Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis. In Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California, USA (1997) 226-233
14. Quinlan, J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California, USA (1993).
15. Salamov, A.A., Nishikawa, T., Swindells, M.B.: Assessing Protein Coding Region Integrity in cDNA Sequencing Projects. Bioinformatics (1998) 14(5) 384-390
16. Stormo, G.D., Schneider, T.D., Gold, L., Ehrenfeucht, A.: Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. coli. Nucleic Acids Research (1982) 10 (9) 2997-3011
17. Tzanis, G., Berberidis, C., Alexandridou, A., Vlahavas, I.: Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In Proceedings of the 10th Panhellenic Conference on Informatics (PCI'2005), Volos, Greece, (2005) 426 – 436
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco (2000)
19. Zeng F., Yap H., Wong, L.: Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. In Proceedings of the 13th International Conference on Genome Informatics, Tokyo, Japan (2002) 192-200
20. Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., Müller, K.R.: Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. Bioinformatics (2000) 16(9) 799-807