A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites

George Tzanis, Christos Berberidis, and Ioannis Vlahavas[*]

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

Abstract. In an mRNA sequence, the prediction of the exact codon where the process of translation starts (Translation Initiation Site – TIS) is a particularly important problem. So far it has been tackled by several researchers that apply various statistical and machine learning techniques, achieving high accuracy levels, often over 90%. In this paper we propose a mahine learning approach that can further improve the prediction accuracy. First, we provide a concise review of the literature in this field. Then we propose a novel feature set. We perform extensive experiments on a publicly available, real world dataset for various vertebrate organisms using a variety of novel features and classification setups. We evaluate our results and compare them with a reference study and show that our approach that involves new features and a combination of the Ribosome Scanning Model with a meta-classifier shows higher accuracy in most cases.

1 Introduction

The last decades has seen a rapid progress in two major scientific areas, biology and computer science. Lately, the field of data mining and machine learning has provided biologists, as well as experts from other areas, a powerful set of tools to analyze new data types in order to extract various types of knowledge fast, accurately and reliably. These tools combine powerful techniques from different areas such as statistics, mathematics, artificial intelligence, algorithmics and database technology. This fusion of technologies aims to overcome the obstacles and constraints posed by the traditional statistical methods.

Translation is one of the basic biological operations that attract biologists' 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. 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

[*] This work was partially supported by the Greek R&D General Secretariat through a PENED program (EPAN M.8.3.1, No. 03E D 73).

the candidate TISs is not a trivial task as it requires the highest possible accuracy. Classification and meta-classification methods have been used in order to deal with this problem.

In this paper, we propose the use of a new feature set along with a combination of meta-classifiers with the *Ribosome Scanning Model* (RSM). We test the feature set with 2 different statistics and then use the extracted features to train 7 different classifiers and a meta-clasiffier. Then we estimate the prediction accuracy of our approach using a state of the art evaluation method, namely 10 times 10-fold cross validation. We also train the same classifiers and perform the same evaluation using the feature sets from a reference study [12] and compare the results. In most cases the proposed approach has a clear advantage against the reference study, showing that both our feature set and our classification setup are more effective in terms of accuracy.

This paper is organized as follows: In the next section, we provide a concise review of the literature on TIS prediction. Section 3 contains a brief introduction on the biological problem attacked in our study. In Section 4 we describe the mining approach we propose and in Section 5 we explain the experimental methodology we followed, show the results of the extensive experiments we performed and finally the evaluation and comparison of our work with a reference study. In the last sections we summarize our paper with our conclusions and directions for future research.

2 Related Work

Since 1982 the prediction of TISs has been extensively studied using biological approaches, data mining techniques and statistical models. Stormo et al. [20] used the perceptron algorithm to distinguish the TISs. In 1987 Kozak developed the first weight matrix for the identification of TISs in cDNA sequences [8]. The consensus pattern derived from this matrix is GCC[**AG**]CCatg**G** (the bold residues are the highly conserved positions). Meanwhile, Kozak and Shatkin [10] had proposed the scanning model of translation initiation, which was later updated by Kozak [9]. According to this model translation initiates at the first start codon that is in an appropriate context.

Pedersen and Nielsen [16] make use of artificial neural networks to predict the TISs achieving an overall accuracy of 88% in Arabidopsis thaliana dataset and 85% in vertebrate dataset. Zien et al. [25] studied the same vertebrate dataset, but instead of neural networks employed support vector machines using various kernel functions. Hatzigeorgiou [4] proposed an ANN system named “DIANA-TIS” 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 start codon. The method applied in human cDNA data and 94% of the TIS were correctly predicted. Salamov et al. [19] developed the program ATGpr, using a linear discriminant approach for the recognition of TISs by estimating the probability of each ATG codon being the TIS. Nishikawa et al. [15] presented an improved program, ATGpr_sim, which employs a new prediction algorithm based on both statistical and similarity information. In [11] Gaussian Mixture Models were used for the prediction of TISs improving classification accuracy.

Feature generation and correlation based feature selection along with machine learning algorithms has also been employed [13, 24]. In these studies a large number of *k*-gram nucleotide patterns were utilized. By using a scanning model an overall accuracy of 94% was attained on the vertebrate dataset of Pedersen and Nielsen. Later, in [12] the same three-step method was used, but k-gram amino acid patterns were considered, instead of nucleotide patterns.

Nadershahi et al. [14] compared five methods -firstATG, ESTScan, Diogenes, Netstart [16] and ATGPr [19]- for the prediction of the TIS. For the comparison a dataset of 100 Expressed Sequence Tag (EST) sequences, 50 with and 50 without a TIS, was created. ATGPr appeared to outperform the other methods over this dataset.

3 Background Knowledge

Translation is the second process of protein synthesis. In particular, after a DNA molecule has been transcribed into a messenger RNA (mRNA) molecule, an organelle called ribosome scans the mRNA sequence. The ribosome reads triplets, or *codons*, of nucleotides and “translates” them into amino acids. 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*.

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 $5' \rightarrow 3'$) 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 call *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 problem the reference point is an AUG codon. The above are illustrated in Fig. 1.

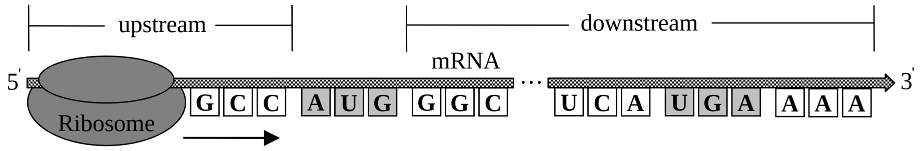

Fig. 1. Translation initiation – The ribosome scans the mRNA sequence from the $5'$ end to the $3'$ end until it reads an AUG codon. If the AUG codon has appropriate context, the translation initiates at that site and terminates when a stop codon (i.e. UGA) is read. An in-frame codon is represented by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

4 The Proposed TIS Prediction Approach

We propose a machine learning approach, focusing on all stages of the prediction, namely the data selection, the feature extraction, the training of the classifier and the evaluation of the effectiveness.

4.1 Datasets

The dataset we used in our study consists of 3312 genomic sequences collected from various vertebrate organisms, acquired from the Kent Ridge Biomedical Data Set Repository (http://sdmc.i2r.a-star.edu.sg/rp). Being DNA sequences, they contain only the letters A, C, G and T. Therefore, a candidate TIS is referred to as ATG codon instead of AUG codon.

The sequences of the dataset were extracted from GenBank, release 95 [2]. Only nuclear genes with an annotated start codon were selected. The DNA sequences have been processed and the introns have been removed. From the resulting dataset, the selected sequences contain at least 10 nucleotides upstream of the initiation point and at least 150 nucleotides downstream (with reference to A in the ATG codon). All sequences containing non-nucleotide symbols in the interval mentioned above (typically due to incomplete sequencing) were excluded. Moreover, the dataset has been gone through very thorough reduction of redundancy [16].

4.2 Features

One of the most important tasks in prediction is the extraction of the right features that describe the data. This is also a particularly crucial point in our approach in terms of novelty and performance. The basic features used in our approach are summarized in Table 1. Some of them (features 1, 2, 12-15) have been already studied in previous research works [12, 13, 24]. However, there is a number of new features that we propose and study in this paper. One set of features (features 3), that have been proposed in previous works of ours [21, 22], concern the periodic occurrence of particular nucleotides at a specific position inside an in-frame codon (Figure 2). Another set of features that has not been studied yet (features 4) includes features that count the amino acids that appear at each in-frame position. The numbering of each position of a sequence is presented in Figure 3. For numbering we used the same conventions as in other studies [12, 13, 21, 22, 24]. What’s more important, we propose a number of extra features based on the chemical properties of amino acids that haven’t been considered before. These features are counts of hydrophobic, hydrophilic, acidic, or basic amino acids, as well as counts of aromatic, aliphatic, or neither aromatic, nor aliphatic amino acids (features 5-11). We used a window size of 99 nucleotides upstream and 99 nucleotides downstream the ATG for calculating the values of the features.

position:		1 2 3	1 2 3		1 2 3	1 2 3	
	5′	G C C	A C C	**A T G**	G C A	T C G	3′

Fig. 2. The positions of nucleotides inside the in-frame codons

Table 1. The features used in our approach

	Features	Description
1	up_*x* down_*x*	Count the number of amino acid *x* in the upstream and downstream region respectively.
2	up-down_*x*	Counts the difference between the number of occurrences of amino acid (or set of amino acids) *x* in the upstream region and the number of occurrences of *x* in the downstream region.
3	up_pos_*k*_*x* down_pos_*k*_*x*	Count the number of occurrences of nucleotide *x* in the k^{th} position of the in-frame codons ($k \in \{1, 2, 3\}$) in the upstream and downstream region respectively.
4	pos_-3*k* pos_3(*k*+1)	Concern the presence of amino acids at in-frame positions in the upstream and downstream region respectively ($k \geq 1$).
5	up_hydrophobic down_hydrophobic	Count the number of hydrophobic amino acids in the upstream and downstream region respectively.
6	up_hydrophilic down_hydrophilic	Count the number of hydrophilic amino acids in the upstream and downstream region respectively.
7	up_acidic down_acidic	Count the number of acidic amino acids in the upstream and downstream region respectively.
8	up_basic down_basic	Count the number of basic amino acids in the upstream and downstream region respectively.
9	up_aromatic down_aromatic	Count the number of aromatic amino acids in the upstream and downstream region respectively.
10	up_aliphatic down_aliphatic	Count the number of aliphatic amino acids in the upstream and downstream region respectively.
11	up_non_aromatic/ aliphatic down_non_aromatic/ aliphatic	Count the number of amino acids that are not aromatic nor aliphatic in the upstream and downstream region respectively.
12	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]CCatgG).
13	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]CCatgG).
14	up_ATG	A Boolean feature that is true if there is an in-frame upstream ATG codon.
15	down_stop	A Boolean feature that is true if there is an in-frame downstream stop codon (TAA, TAG, TGA).

position:	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	
5′	G	C	C	A	C	C	**A**	**T**	**G**	G	C	A	T	C	G	3′

Fig. 3. The positions of nucleotides relative to an ATG codon

4.3 Feature Selection Algorithms

For the conduction of our experiments we have utilized the Weka library of machine learning algorithms [23]. We have used the following feature selection methods:

- *Chi-Squared*. Evaluates the worth of an attribute by computing the value of the X^2 statistic with respect to the class.
- *Gain Ratio*. Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

4.4 Classification Algorithms

For classification we have used the following classification algorithms:

- *C4.5*. Algorithm for generating a decision tree [18].
- *RIPPER*. This is a propositional rule learner called Repeated Incremental Pruning to Produce Error Reduction. [3].
- *Decision Table*. This algorithm implements a simple decision table majority classifier [7].
- *Naïve Bayes*. A Naive Bayes classifier [5].
- *SVM*. This is the John Platt's [17] sequential minimal optimization algorithm for training a support vector classifier. The Weka implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.
- *Multilayer Perceptron*. This algorithm implements a neural network that uses back-propagation to train.
- *k*-Nearest Neighbors classifier. The algorithm normalizes attributes by default and can do distance weighting. We have used this algorithm with 1-nearest neighbor. More information about this algorithm can be found in [1].

The idea of meta-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. We implemented and used two meta-classification algorithms of the second paradigm:

- *Simple Voting*. This algorithm combines the decisions of multiple classifiers and makes the final decision by considering the majority of the votes. Each classifier participates equally to the voting.
- *Weighted Voting*. This algorithm also applies voting but each classifier participates with a different weight to the voting procedure. The weight for each classifier is its

classification accuracy. In other words, the more accurate classifiers contribute more to the final result than the less accurate ones.

4.5 Evaluation Method

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 follows. 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 each time. The use of the 10-fold CV was based on the widely accepted study of R. Kohavi [6]. The results of this work indicate that for many real-word datasets, the best method to use for model selection is stratified 10-fold CV, even if computation power allows using more folds.

Furthermore, in order to increase the reliability of the evaluation, we have repeated each experiment 10 times, each time generating randomly different folds and we finally took into account the average of the 10 independent repeats.

5 Experimental Results

In this section we present the results of the experiments we conducted, compared to a reference study [12] on the same dataset. The first subsection describes the results of feature selection and the second subsection deals with the results of classification. Finally, a discussion about the results is given in the third subsection.

5.1 Feature Selection Results

The best features selected by the two feature selection methods are presented in Table 2. When using the X^2 statistic for feature selection the 8 out of the 10 top features are the ones we propose (5 are proposed in this paper and the other 3 have been proposed in [21]). When using the gain ratio measure, the 6 out of the 10 top features are the ones we propose (4 are introduced in this paper and the other 2 have been proposed in [21]). It should be noted here that the feature down_stop of the reference study is not Boolean (see Table 1). Instead, it counts the number of the in-frame downstream stop codons. Moreover, in this study, features of the form down_*xy* and up_*xy* are also considered, where *x* and *y* are amino acids, or a stop codon.

5.2 Classification Results

After extensive experiments, we present the results produced by three different setups:

- *Setup 1*. The use of a meta-classifier (simple voting or weighted voting) for predicting the TIS.
- *Setup 2*. The use of a meta-classifier (simple voting or weighted voting) incorporated with the *ribosome scanning model* (RSM). The first candidate TIS that appears inside a sequence and has received more than 50% of votes for being a TIS is selected as the true TIS. The remaining candidates of the same sequence are considered not to be TISs, even if they have received more than 50% of votes for being a TIS.
- *Setup 3*. We propose the use of the RSM based on the results of a meta-classifier (simple voting or weighted voting). Among all candidate TISs of a sequence, the one that has received the larger number of (positive) votes is considered as the true TIS. The remaining candidates of the same sequence are considered as non-TISs.

Table 2. The top features that were selected by the feature selection methods from the set of features we propose (middle column) and from the set of features used in the reference study [12] (last column). The features are ordered according to their ranking

FS Method	Our Features	Reference Features
X^2	up_ATG	up_ATG
	down_1_G	down_stop
	down_hydrophobic	up_M
	down_non_aromatic/aliphatic	down_A
	down_3_C	down_L
	down_stop	down_V
	down_aliphatic	down_E
	up-down_non_aromatic/aliphatic	down_D
	up-down_hydrophobic	up_-3_[AG]
	down_2_T	down_G
Gain Ratio	up_ATG	down_stop
	down_stop	up_RR
	up_M	up_NH
	up_-3_[AG]	down_MY
	down_non_aromatic/aliphatic	up_ATG
	down_1_G	up_M
	up-down_non_aromatic/aliphatic	down_Lstop
	down_hydrophobic	down_stopR
	up-down_hydrophobic	down_stopS
	down_2_C	down_Pstop

The results of the three setups using either simple voting, or weighted voting are presented in Tables 3 and 4. The first table contains the results produced using the set of features we propose, while the second one contains the results produced using set of features utilized in the reference study.

Table 3. Classification results using the set of features we propose. The grayed cells indicate the setup that achieves the highest accuracy

		Simple Voting			Weighted Voting		
FS Method	# Top Features	Setup 1	Setup 2	Setup 3	Setup 1	Setup 2	Setup 3
X^2	5	86.20%	86.34%	86.32%	86.20%	86.34%	86.38%
	10	90.36%	91.58%	93.30%	90.36%	91.58%	93.65%
	15	94.60%	95.19%	95.53%	94.60%	95.19%	96.01%
	20	94.71%	95.34%	95.89%	94.71%	95.34%	96.25%
	25	94.78%	95.41%	95.85%	94.79%	95.41%	96.19%
	30	94.79%	95.37%	95.79%	94.78%	95.37%	96.21%
Gain Ratio	5	93.09%	94.16%	94.87%	93.09%	94.16%	94.77%
	10	94.10%	94.99%	95.12%	94.10%	94.99%	95.46%
	15	94.58%	95.14%	95.27%	94.58%	95.14%	95.75%
	20	94.60%	95.16%	95.34%	94.60%	95.16%	95.79%
	25	94.66%	95.23%	95.53%	94.66%	95.23%	95.98%
	30	94.70%	95.28%	95.60%	94.70%	95.28%	95.99%

Table 4. Classification results using the set of features utilized in the reference study. The grayed cells indicate the setup that achieves the highest accuracy

		Simple Voting			Weighted Voting		
FS Method	# Top Features	Setup 1	Setup 2	Setup 3	Setup 1	Setup 2	Setup 3
X^2	5	87.98%	90.71%	91.64%	87.98%	90.71%	90.90%
	10	91.75%	92.57%	93.68%	91.75%	92.57%	93.70%
	15	91.93%	92.61%	93.76%	91.93%	92.61%	93.84%
	20	92.40%	92.90%	94.07%	92.40%	92.90%	94.25%
	25	92.40%	92.85%	94.02%	92.40%	92.85%	94.19%
	30	92.53%	92.91%	94.18%	92.53%	92.91%	94.32%
Gain Ratio	5	82.62%	86.56%	87.42%	82.62%	86.56%	84.79%
	10	85.33%	91.80%	91.67%	85.33%	91.80%	87.59%
	15	87.23%	91.00%	92.01%	87.23%	91.00%	89.45%
	20	87.97%	90.98%	92.27%	87.97%	90.98%	89.82%
	25	88.00%	90.98%	92.33%	88.00%	90.98%	89.90%
	30	88.02%	91.00%	92.35%	88.02%	91.00%	89.96%

Figure 4 depicts the results for the 3rd Setup, which is introduced in this paper. In particular, the results obtained using the features proposed here are compared to the results obtain using the features of the reference study [12]. The results of each of the seven classifiers are not presented here for brevity since only their output is considered in the meta-classification step.

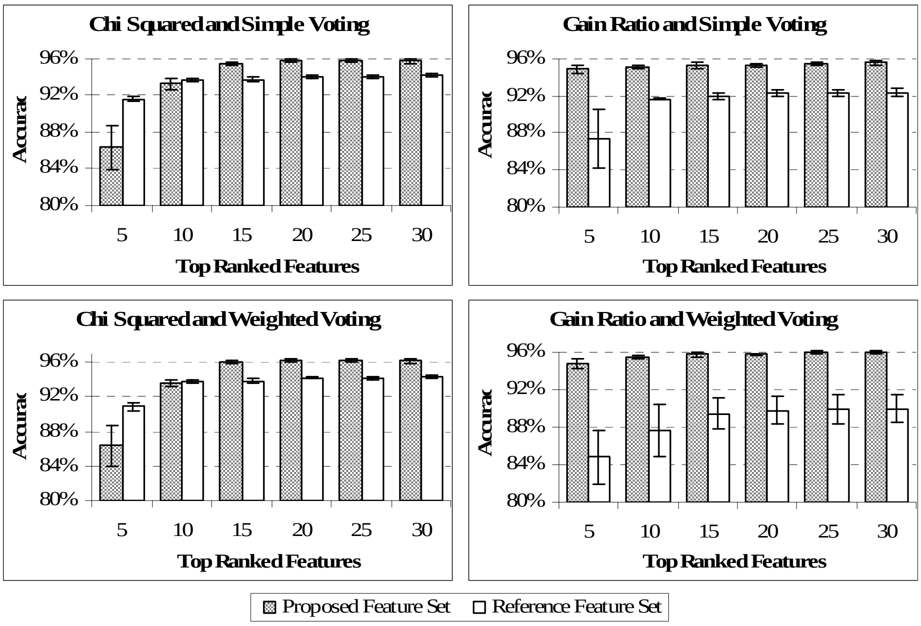

Fig. 4. Comparison of the results obtained using the proposed feature set against the reference feature set [17], for Setup 3. The error bars are calculated as ± 2 standard deviations

5.3 Discussion

The classification accuracy achieved by the setup proposed in this study (Setup 3) is higher in almost all cases when using the set of features we propose. When using the features of the reference study, Setup 3 outperforms Setups 1 and 2 with the X^2 statistic for feature selection. However, this is not the case when the Gain Ratio measure is used. Moreover, when the features we propose are used, the best accuracy is achieved with X^2 statistic for feature selection and the top 20 ranked features along with the weighted voting meta-classification algorithm and Setup 3 (96.25%). At the other hand, when the features of the reference study are used, the best accuracy is achieved with X^2 and the top 30 ranked features along with the weighted voting meta-classification algorithm and Setup 3 (94.32%).

The results of Setups 1 and 2 for both simple and weighted voting are identical. This happens because there are zero or very few decisions of the simple voting schema that are based on a majority near 50%. The percentage of majority is high enough, so that the weighting of the votes do not affect the results. However, in Setup 3 the results are better when using the weighted voting schema.

The incorporation of the ribosome scanning model (RSM) (Setup 2) provides better classification accuracy in almost all cases. Moreover, the improvement is much higher when RSM is used with the features of the reference study. Specifically, the improvement achieved is 6.48 percentage units. The improvement when the set of our proposed features are used is only 1.22 percentage units. However, when using the set of features we propose in Setup 1, the classification accuracy is in most cases higher than the results achieved using the reference study features enhanced with RSM (Setup 2). This implies that a large portion of the improvement provided by the use of RSM is incorporated in the features we use. In other words, although the use of RSM greatly enhances the effectiveness of the reference feature set, its effect on our feature set (used in Setup 3) is not that significant because the accuracy attained by the latter alone is already the highest so far.

Although previous works [13, 21, 24] have shown that the feature that counts the distance of a candidate ATG from the start of the sequence is very good for discriminating a TISs from non-TISs we excluded it from our study. The reason is that this feature is highly affected by the intrinsic characteristics of the sequences contained in the dataset we used. For example, for each sequence there are always 150 nucleotides downstream of the TIS and there is up to a maximum of 150 nucleotides upstream.

6 Conclusions and Future Work

In this paper we tackle the problem of the prediction of Translation Initiation Sites in genome sequences. We implement a machine learning approach that shows higher accuracy than previous approaches on a public vertebrate dataset. First, we provide a review of the literature on this task and a short section on biological background knowledge. By extensive experiments using two different statistics (X^2 and Gain Ratio) we propose the use of a novel feature set that leads to higher accuracy when compared to the feature sets of the reference study. Then, we introduce a new prediction setup that utilizes meta-classifiers and the ribosome scanning model in order to achieve higher accuracy. We support our claims by extensive experiments using 7 different classifiers along with 2 meta-classifiers. Then, we evaluated our results by performing 10 times 10-fold cross validation, in order to prove the reliability of our approach.

In the near future we are going to apply our approach on more datasets, run more experiments with more classifiers and new feature sets. We also plan to investigate the application of our approach on other functional site prediction problems, such as splice sites.

References

1. Aha, D., Kibler, D. Instance-based learning algorithms, Machine Learning (1991) 6, 37-66
2. Benson, D., Boguski, M., Lipman, D., Ostell, J. Genbank. Nucleic Acids Research 25, (1997) 1–6
3. Cohen, W.: Fast Effective Rule Induction. In Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann, Lake Tahoe, USA (1995) 80-89
4. Hatzigeorgiou, A.: Translation Initiation Start Prediction in Human cDNAs with High Accuracy. Bioinformatics (2002) 18(2) 343-350
5. 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, USA (1995) 338-345

6. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, In Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995
7. Kohavi, R. The Power of Decision Tables. In Proceedings of the 8th European Conference on Machine Learning (ECML'95), LNAI 914, 174-189. Springer Verlag, 1995.
8. Kozak, M.: An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs. Nucleic Acids Research (1987) 15(20) 8125-8148
9. Kozak, M.: The Scanning Model for Translation: An Update. The Journal of Cell Biology (1989) 108(2) 229-241
10. 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
11. 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
12. 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
13. Liu, H., Wong, L.: Data Mining Tools for Biological Sequences. Journal of Bioinformatics and Computational Biology, (2003) 1(1) 139-168
14. Nadershahi, A., Fahrenkrug, S.C., Ellis, L.B.M.: Comparison of computational methods for identifying translation initiation sites in EST data. BMC Bioinformatics (2004) 5(14)
15. 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
16. 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, USA (1997) 226-233
17. Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, A. Smola (Eds.), MIT Press, (1998)
18. Quinlan, J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, USA (1993)
19. Salamov, A.A., Nishikawa, T., Swindells, M.B.: Assessing Protein Coding Region Integrity in cDNA Sequencing Projects. Bioinformatics (1998) 14(5) 384-390
20. 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
21. 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
22. Tzanis, G., Vlahavas, I.: Prediction of Translation Initiation Sites Using Classifier Selection. In Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06), G. Antoniou, G. Potamias, D. Plexousakis, C. Spyropoulos (Eds.), Springer-Verlag, LNAI 3955, Heraklion, Greece (2006) 367 - 377
23. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco (2000)
24. 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
25. 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