Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction

George Tzanis, *Member, IEEE*, Christos Berberidis, and Ioannis Vlahavas, *Member, IEEE*

Abstract—**The prediction of the translation initiation site in a genomic sequence with the highest possible accuracy is an important problem that still has to be investigated by the research community. Current approaches perform quite well, however there is still room for a more general framework for the researchers who want to follow an effective and reliable methodology. We developed a prediction methodology that combines ad hoc as well as discovered knowledge in order to significantly increase the achieved accuracy reliably. Our methodology is modular and consists of three major decision components: a consensus component, a coding region classification component and a novel ATG location-based component that allows for the utilization of the advantages of the popular Ribosome Scanning Model while overcoming its limitations. All three of them are combined into a meta-classification system, using stacked generalization, in a highly effective prediction framework. We performed extensive comparative experiments on four different datasets, showing that the increase in terms of accuracy and adjusted accuracy is not only statistically significant, but also the highest reported.**

I. Introduction

THE accurate identification of the Translation Initiation Site (TIS) in mRNA sequences has been extensively studied since the 80s. It was important that the expensive and slow in vitro methods should be replaced by computational tools that could deliver the desired knowledge not only cheap and fast but also accurately. However, the fact that the exact mechanism of translation initiation has not been discovered yet, has led the research community to a number of computational tools that perform quite well but the problem still seems to be far from trivial. What makes it even more challenging is the ongoing sequencing of a large number of organisms whose genome has not been annotated and studied yet.

Data mining is a field of research and application that aims to provide efficient computational tools to overcome the obstacles and constraints posed by the traditional statistical methods. Experience has shown that general purpose data mining approaches may perform well on the TIS prediction; however, we believe that we could be confident of an approach only if we identified the main components of the problem and then designed this approach so that it optimally maps each one of them separately, in a modular fashion. In this direction one must embed the knowledge of the expert (molecular biologist) along with the knowledge that is automatically discovered.

In this paper, we propose a component-based data mining methodology, called MANTIS (the Greek word for "diviner" or "prophet"), that can be applied on virtually any TIS dataset with optimal results. For this purpose, we developed a system that implements this methodology, utilizing a variety of tools and techniques and selecting those that not only were theoretically sound but also the most effective ones, in order to assemble the optimal system. We evaluated the results by extensive experiments over 4 different datasets. These experiments showed that with MANTIS the increase in prediction accuracy and adjusted accuracy is significant compared to a number of different approaches and combinations.

Since 1982, the prediction of TISs has been investigated using biological approaches, data mining techniques and statistical models. The perceptron algorithm was used in [16] in order to distinguish the TISs. Kozak developed the first weight matrix for the identification of TISs in cDNA sequences [5]. The consensus pattern derived from this matrix is GCC[AG]CCatgG (the underlined residues are the highly conserved positions). In [6] the scanning model of translation initiation was proposed, which was later updated by Kozak [4]. According to this model translation initiates at the first start codon that is in a particular context.

Various data mining methods have been utilized for the prediction of TISs, including artificial neural networks [2, 13], support vector machines [20], Gaussian mixture models [7], linear discriminant approaches [15], techniques based on statistical and similarity information [11]. In [9] feature generation (k-gram nucleotide patterns) and correlation based feature selection along with classification algorithms were used. Later, in [8] the same three-step method was used, but k-gram amino acid patterns were utilized instead. A comparative study of five methods for the prediction of TIS was presented in [10].

This paper is outlined as follows: In the next section we describe the MANTIS methodology in detail, explaining every step of the process. In section III we present the datasets we used and Section IV contains the results of our experiments. Finally, section V contains our conclusions.

Manuscript received April 16, 2007.

G. Tzanis is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece. (e-mail: gtzanis@csd.auth.gr).

C. Berberidis is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece. (e-mail: berber@csd.auth.gr).

I. Vlahavas is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece. (e-mail: vlahavas@csd.auth.gr).

II. THE MANTIS METHODOLOGY

The proposed methodology, called MANTIS covers the entire knowledge discovery process, from data preprocessing to the fusion of the decision components into the final prediction. MANTIS consists of three major decision components as shown in Fig. 1. These components contribute to the final decision by considering different aspects of a candidate TIS. The first one is a classifier that captures differences in the coding potential around an ATG, the second is a Markov-chain based consensus pattern discovery algorithm and the third is a model based on the location of the ATG inside the sequence and the ribosome scanning model.

A. The Coding Region Classification Component

The most complex component of our methodology is the coding region classification component, which involves training a model to classify an ATG as a TIS or not, based on a feature set that is extracted from a sequence window that contains 99 nucleotides upstream and 99 downstream of the ATG. The basic mission of this component is to recognize if the downstream region of an ATG is a coding region and consequently if the ATG is the TIS.

This task requires the data to be preprocessed as follows: First, a window of size 201 (99 + 3(ATG) + 99) centered at each candidate ATG is constructed. Each window represents an instance in the training set. Then, for each window, a set of features is extracted.

For the conduction of our experiments we have utilized the Weka library of machine learning algorithms [19].

1) Feature Reduction and Transformation

The original feature set consists of features that were used in previous studies [18, 8] and were found to produce good classifiers in terms of classification accuracy. However, some features are correlated to each other. In order to build a feature space with a smaller number of uncorrelated features we applied Principal Component Analysis (PCA) and selected those components that have an eigenvalue greater than the mean of the eigenvalues of all the components.

2) Classification Algorithms

For the TIS prediction, we used the following 4 classification algorithms, representative of 4 different categories: Naïve Bayes, C4.5, k-Nearest Neighbors (1 to 15 neighbors selected via leave-one-out cross-validation) and SVM.

B. The Consensus Component

Previous studies [2, 14] have shown that for the identification of the TIS, it is important to examine a narrow area around it. This component uses Markov chains to capture the consensus pattern starting from position -7 and ends at position +5, as shown in Figure 4. The use of a Markov-chain based technique allows capturing not only the probability of the occurrence of a nucleotide at a certain position (as consensus pattern mining algorithms usually do) but also how the occurrence of one affects the occurrence of another.

We experimented with three Markov chains:

- A 1st order homogeneous Markov chain
- A 2nd order homogeneous Markov chain
- A 1st order non-homogeneous Markov chain.

We train a Markov chain using examples of the positive class only and another Markov chain using examples of the negative class only. When a new instance arrives for classification, each of the two Markov chains produces a score for this instance. These scores are scaled so that their sum equals to 1, as follows:

$$S_+ = \frac{S_+}{S_+ + S_-} \text{ and } S_- = \frac{S_-}{S_+ + S_-} \quad (1)$$

C. The ATG Location Component

This is a novel component, based on the location of the ATG inside the sequence and the Ribosome Scanning Model (RSM), as described by Kozak [3]. According to this model, the ribosome scans the sequence until it finds the first ATG that is in an optimal nucleotide context. In previous studies, an ATG was chosen by the RSM as a TIS, when it was the one among those selected by a classifier as a positive example that was closest to the 5'. In that case, all other ATGs were assigned to the negative class, even those that had been

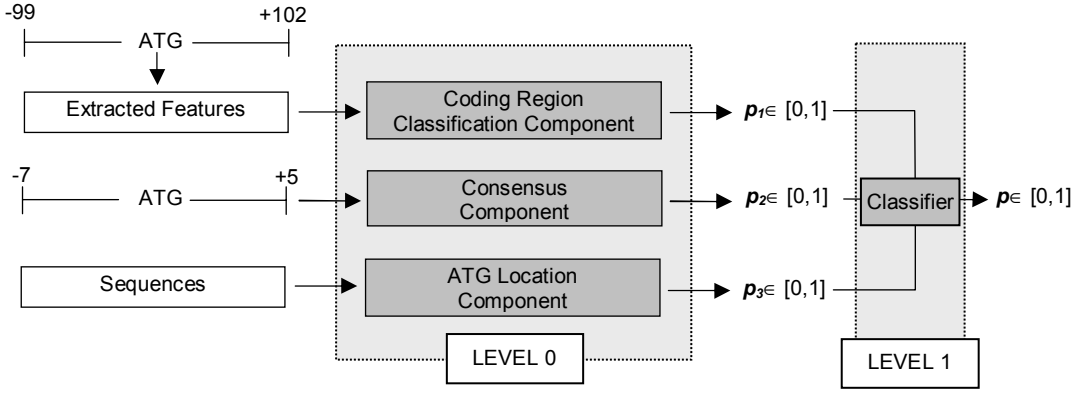

Fig. 1. The MANTIS methodology.

given a higher probability by the classifier. However, this rule has several exceptions that are explained in [3, 12].

In some cases, there is another positive ATG not long after and with a higher probability than the one selected by the RSM. Without the RSM, we would choose the ATG that has been classified as the most probable to be the TIS. The main parameters of this problem are the order of the ATG in the sequence (or distance from the 5' end) and the probability assigned by the classifier(s). Both of them must be incorporated into a model that learns how to combine them.

In MANTIS methodology we create two models, one which is order-based and one which is distance based. The first calculates the probabilities of an ATG to be the TIS according to its order in the sequence, whereas the second calculates the probability according to its distance from the 5'. Each of the two models is built on the positive and negative instances separately. The final scores are scaled as in equation 1. These two probabilities are the output of the ATG location component.

D. Stacked Generalization

The final stage in MANTIS is the fusion of the decision components described so far. In MANTIS we apply a popular classifier fusion technique called Stacked Generalization or Stacking. Stacking is a scheme for minimizing the generalization error rate of one or more models. According to that technique, a number of models (called level-0 models) that are trained on the original data (called level-0 data), produce the input (level-1 data) to a classifier, (level-1 classifier). In MANTIS, the level-0 models are produced by the algorithms of the three components, described earlier.

We tested 2 different level-1 classifiers, namely M5' and Multi-response Linear Regression (MLR). MLR is suitable for this task because in stacking it is necessary to use output class probabilities rather than class predictions (0/1) as shown in [17]. M5' is a continuous class, model tree classifier whereas MLR involves estimating several response variables using a common set of input variables. Stacking with M5' has been proposed as an extension of stacking with MLR in [1] and presented improved performance.

E. Candidate TIS Ranking

MANTIS output is a probability estimate of an ATG to be a TIS, instead of a single true/false decision. When probability estimates are available, one could rank the top scoring ATGs in order to consider alternatives.

III. Datasets

In our study we have used four datasets. Three of them (Vertebrates, H.sapiens and A.thaliana) were used in previous studies [13, 2, 7, 18] whereas one of them is new (A.aegypti). Vertebrates dataset consists of 3312 genomic sequences collected from various vertebrate organisms. H.sapiens dataset consists of 480 human sequences [2]. A.thaliana dataset contains 523 sequences collected from Arabidopsis thaliana, an organism that shows large deviation from vertebrates. The sequences of the Vertebrates and A.thaliana were extracted from GenBank, release 95. Only nuclear genes with an annotated start codon were selected. The DNA sequences have been processed and the interlacing non-coding regions (introns) have been removed. From the resulting data set, sequences containing at least 10 nucleotides upstream of the initiation point and at least 150 nucleotides downstream (relative to the A in ATG) were selected. All sequences containing non-nucleotide symbols in the interval mentioned above (typically due to incomplete sequencing) were excluded. Moreover, the datasets have been gone through very thorough reduction of redundancy [13]. The H.sapiens dataset was extracted from Swissprot protein database. All the human proteins whose N-terminal sites are sequenced at the amino acid level were collected and manually checked. Then, the full-length mRNAs for the proteins with TISs that had been indirectly experimentally verified were retrieved and the corresponding human cDNAs, completely sequenced and annotated, were found [2]. A.aegypti dataset contains 262 sequences retrieved from Ensembl (release 42) concerning Aedes aegypti also known as yellow fever mosquito. Based on the fact Ensembl gene annotations are based on experimental evidence, we selected the A.aegypti cDNA sequences from genes marked as "known" and contained a 5' UTR and a 3' UTR. Then, for our experiments we included only those, whose coding region begins with an ATG. Table I summarizes the information about the datasets.

TABLE I
DATASETS

Dataset Name	Sequences	ATGs	TISs/ATGs
A.aegypti	262	6453	4.0%
A.thaliana	523	2048	25.5%
H.sapiens	480	14108	3.4 %
Vertebrates	3312	13503	24.5%

IV. Results

We compared the results of MANTIS to a standard approach, which is common in TIS prediction literature [2, 8, 9]. This approach consists of combinations of a coding and/or a consensus component, followed by the RSM. In this paper, we present the results of our comparison of MANTIS to the best combination (coding + consensus + RSM). This approach from now on will be referred to as "*reference approach*". The coding and consensus components were combined using stacking with MLR and M5' as level-1 classifiers, in order to make it comparable to MANTIS.

Table II contains the results of previous approaches in terms of accuracy and adjusted accuracy, which is a skew-insensitive version of accuracy and is defined below:

$$adjusted\ accuracy = \frac{sensitivity + specificity}{2} \qquad (2)$$

We performed extensive experiments over the 4 datasets mentioned in section III, using stratified 10-fold cross-validation (CV), which is considered to be standard for classification model evaluation [22]. All stages of the knowledge discovery process (feature reduction and transformation, level-0 and level-1 model training and testing) were incorporated into the evaluation procedure. It is also important here to clarify that all the experimental runs were made over the same folds for all reference approaches, which makes the comparison as fair as possible.

TABLE II
RESULTS OF PREVIOUS STUDIES

Study	Dataset	Accuracy	Adjusted Accuracy
Pedersen & Nielsen [13]	A.thaliana	88.00%	88.50%
Pedersen & Nielsen [13]	Vertebrates	85.00%	82.50%
Zien et al. [20]	Vertebrates	88.10%	82.00%
Rajapakse and Ho [14]	Vertebrates	96.10%	95.35%
Liu et al. [8]	Vertebrates	92.45%	88.34%
Liu et al. [8]	H.sapiens	98.46%	85.34%
Hatzigeorgiou [2]	H.sapiens	94.00%	-
Li et al. [7]	H.sapiens	-	95.24%

After we applied MANTIS on the four datasets described in section III, in order to estimate its accuracy and adjusted accuracy, we considered the single, top-ranked ATG, for every sequence. Concerning the reference approach, we apply the RSM by scanning the sequence until the "ribosome" reads the first ATG that has been assigned a probability greater than or equal to 0.5. Given the high skewness of the datasets and the fact that most classifiers produce probability estimates, which do not always converge to the empirical class membership probabilities, it is important to stress out that 0.5 may not be the correct decision threshold. For that reason we calibrate the probability estimates of the classifiers. Given a probabilistic classifier C and a class c, C is said to be well-calibrated if the empirical class membership probability $P(c|s(x) = s)$ converges to the score value $s(x) = s$, as the number of examples goes to infinity [21]. The algorithm we used for the calibration is Pool Adjacent Violators (PAV), which performs isotonic regression.

Tables III and IV contain the results from the comparison of MANTIS against the reference approach, using MLR and M5' as level-1 classifiers, respectively. It is clear that MANTIS outperforms the reference approach on all datasets, both in terms of accuracy and adjusted accuracy. Moreover, MANTIS-M5' scores the highest reported accuracy for the three datasets that have been used in other studies. It is important to stress out that MANTIS' advantage over the reference study is even greater in terms of adjusted accuracy, which is a more appropriate measure, due to the data skewness. In particular, MANTIS-MLR outperforms the reference study in terms of accuracy by 2.19 percentage points and in terms of adjusted accuracy by 8.86 percentage points. Similarly, MANTIS-M5' outperforms the reference study in terms of accuracy by 2.87 percentage points and in terms of adjusted accuracy by 9.62 percentage points.

Table V contains the statistical comparison of the MANTIS against the reference approach. For this purpose we applied the non-parametric Wilcoxon signed-rank test, in order to perform a fold to fold comparison for each dataset. A +(a) denotes a statistically significant superiority (win) of MANTIS with confidence 1-a. Note that in all cases the superiority of MANTIS is statistically significant.

TABLE III
COMPARISON OF MANTIS VS. REFERENCE APPROACH, USING MLR AS A LEVEL-1 CLASSIFIER

	MANTIS		Coding + Consensus + RSM	
	Accuracy	Adjusted Accuracy	Accuracy	Adjusted Accuracy
A.aegypti	98.61%	91.05%	97.86%	79.15%
A.thaliana	96.58%	95.51%	92.82%	87.33%
H.sapiens	99.08%	92.99%	98.69%	86.05%
Vertebrates	95.16%	93.46%	91.32%	85.04%
Average	97.36%	93.25%	95.17%	84.39%
St. dev.	1.82%	1.83%	3.65%	3.62%

TABLE IV
COMPARISON OF MANTIS VS. REFERENCE APPROACH, USING M5' AS A LEVEL-1 CLASSIFIER

	MANTIS		Coding + Consensus + RSM	
	Accuracy	Adjusted Accuracy	Accuracy	Adjusted Accuracy
A. aegypti	98.64%	91.25%	97.99%	79.03%
A. thaliana	97.07%	96.15%	92.43%	87.00%
H. sapiens	99.14%	93.42%	98.89%	87.57%
Vertebrates	97.26%	96.30%	91.34%	85.04%
Average	98.03%	94.28%	95.16%	84.66%
St. dev.	1.02%	2.42%	3.83%	3.91%

TABLE V
STATISTICAL COMPARISON OF MANTIS VS. REFERENCE APPROACH

	Accuracy		Adjusted Accuracy	
	MLR	M5'	MLR	M5'
A. aegypti	+ (0.05)	+ (0.01)	+ (0.01)	+ (0.01)
A. thaliana	+ (0.01)	+ (0.01)	+ (0.01)	+ (0.01)
H.sapiens	+ (0.01)	+ (0.01)	+ (0.01)	+ (0.01)
Vertebrates	+ (0.01)	+ (0.01)	+ (0.01)	+ (0.01)
wins:losses (a = 0.01)	3:0	4:0	4:0	4:0
wins:losses (a = 0.05)	4:0	4:0	4:0	4:0

Fig. 2 and 3 display the percentage of the TIS missed when the n top-ranked ATGs were selected, using MLR and M5' respectively. What is really shown in these graphs is the quality of the alternative solutions provided by MANTIS. As we see, in all cases the percentage of missed TIS drops exponentially with respect to n. Additionally, the more balanced the dataset is (A.thaliana and Vertebrates) the faster their missed TISs converge to 0.

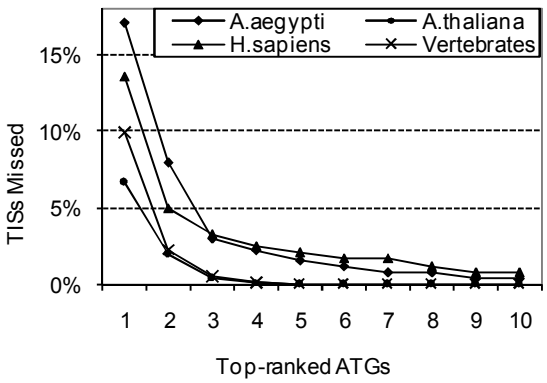

Fig. 2. TISs missed over the number of top-ranked ATGs using MLR.

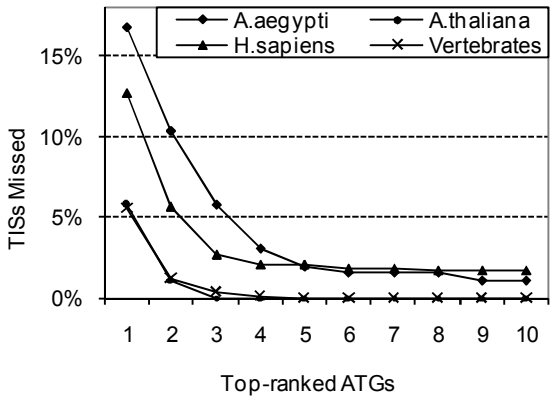

Fig. 3. TISs missed over the number of top-ranked ATGs using M5'.

V. Conclusions

In this paper we proposed a novel TIS prediction methodology, called MANTIS. MANTIS is an intuitive, component-based approach, consisting of three main components that map the biological sub-problems identified. It is worth mentioning that the utilization of the new ATG location component provides the advantages of the typical RSM model and overcomes its limitations. The three components are combined using a state-of-the-art method, namely stacking. The output of MANTIS is a (user-defined) number of ranked candidate TISs. Extensive experiments over 4 datasets (3 previously studied and 1 new) showed that MANTIS is a methodology that outperforms existing ones. We showed that the improvement in terms of accuracy and adjusted accuracy is statistically significant. Moreover, to the best of our knowledge, the overall accuracy is the highest reported in the literature.

We are currently working on a web-based version of MANTIS, which will be publicly available. Then, we aim to modify and extend MANTIS in order to apply it on other functional site prediction problems, such as splice site prediction and transcription start site prediction.

Acknowledgment

We are grateful to Dr. Anders Gorm Pedersen, Dr. Artemis Hatzigeorgiou and Dr. Guoliang Li, who kindly provided us with their datasets for our experiments.

References

[1] S. Dzeroski and B. Zenko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?", *Machine Learning*, vol. 54, pp. 255-273, 2004.

[2] A. Hatzigeorgiou, "Translation Initiation Start Prediction in Human cDNAs with High Accuracy", *Bioinformatics*, vol. 18, no. 2, pp. 343-350, 2002.

[3] M. Kozak, "Interpreting cDNA sequences: some insights from studies on translation", *Mamm. Genome*, vol. 7, pp. 563–574, 1996.

[4] M. Kozak, "The Scanning Model for Translation: An Update", *J. Cell Biol.*, vol. 108, no. 2, pp. 229-241, 1989.

[5] M. Kozak, "An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs", *Nucleic Acids Res.*, vol. 15, no. 20, pp. 8125-8148, 1987.

[6] M. Kozak and A. J. Shatkin, "Migration of 40 S Ribosomal Subunits on Messenger RNA in the Presence of Edeine", *J. Biol. Chem.*, vol. 253, no. 18, pp. 6568-6577, 1978.

[7] G. Li, T,-Y. Leong, and L. Zhang, "Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences" *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 17, pp. 1152-1160, 2005.

[8] H. Liu, H. Han, J. Li, and L. Wong, "Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites", *In Silico Biol.*, vol. 4, no. 3, pp. 255-269, 2004.

[9] H. Liu and L. Wong, "Data Mining Tools for Biological Sequences", *J. Bioinform. Comput. Biol.*, vol. 1, no. 1 pp. 139-168, 2003.

[10] A. Nadershahi, S. C. Fahrenkrug, and L. B. M. Ellis, "Comparison of computational methods for identifying translation initiation sites in EST data" *BMC Bioinformatics*, vol. 5, no. 14, 2004.

[11] T. Nishikawa, T. Ota, and T. Isogai, "Prediction whether a Human cDNA Sequence Contains Initiation Codon by Combining Statistical Information and Similarity with Protein Sequences", *Bioinformatics*, vol. 16, no. 11, pp. 960-967, 2001.

[12] V. M. Pain, "Initiaton of proteins synthesis in eukaryotic cells", *Eur. J. Biochem.*, vol. 236, no. 3, pp. 747–771, 1996.

[13] A. G. Pedersen, H. Nielsen, "Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis". in *Proc. 5th Int. Conf. Intelligent Systems for Molecular Biology*, pp. 226-233, 1997.

[14] J. C. Rajapakse and L.S. Ho. "Markov Encoding for Detecting Signals in Genomic Sequences", *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 2, no. 2, pp. 131-142, 2005.

[15] A. A. Salamov, T. Nishikawa, and M. B. Swindells, "Assessing Protein Coding Region Integrity in cDNA Sequencing Projects", *Bioinformatics*, vol. 14, no. 5, pp. 384-390, 1998.

[16] G. D. Stormo, T.D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. coli", *Nucleic Acids Res.*, vol. 10, no. 9, pp. 2997-3011, 1982.

[17] M.T. Ting and I.H. Witten, "Issues in Stacked Generalization", *J. Art. Intell. Res.*, vol. 10, pp. 271-289, 1999.

[18] G. Tzanis, C. Berberidis, and I. Vlahavas, "A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites", in *Proc. 7th Int. Symposium Biological Medical Data Analysis*, pp. 92-103, 2006.

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementations*. San Francisco: Morgan Kaufmann, 2000.

[20] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates" in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

[21] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K. R. Müller, "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites", *Bioinformatics*, vol. 16, no. 9, pp. 799-807, 2000.

[22] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", in Proc.14th Int. Joint Conf. Artificial Intelligence, pp. 1137-1143, 1995