Evaluation of Weight Matrix Models in the Splice Junction Recognition Problem

Leonardo G. Tavares, Heitor S. Lopes, Carlos R. Erig Lima
Bioinformatics Laboratory
Federal University of Technology Parana (UTFPR)
Av. 7 de setembro, 3165 80230-901, Curitiba (PR), Brazil
leonardo.tavares@up.edu.br, hslopes@pesquisador.cnpq.br, erig@utfpr.edu.br

***Abstract*—The amount of data produced by the several genomic sequencing projects has increased dramatically in recent years. One of the main goals of bioinformatics is to analyze biological data aiming at identifying genes. The splice junction recognition problem is an important part of the gene detection problem. This work evaluates the performance of two classification models, derived from the Weight Matrix Model, when applied to the splice junction recognition problem. Two splice junction data sets were used in this work and some measures of predictive accuracy were reported. Based on the experiments, classification thresholds were established, which can be useful for further implementation of an automatic gene detection system.**

***Keywords*-Gene detection; DNA; Weight Matrix Model;**

I. Introduction

In living organisms the DNA encodes the genetic information with a sequence of nucleotides, namely, Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Such information is organized in genes along DNA chains. Genes are sequences of nucleotides that are transcribed into mRNA (messenger RNA) and later translated into proteins. This process is called gene expression.

The number of genes in the three billion base pairs of the human DNA is controversial, and it is estimated between 20,000 and 42,000 [18], [23]. Supposing an average of 30,000 genes, this means that only 1.1% of the human DNA seems to contain useful coding information [11]. However, there might be a large number of unknown human genes that still remains to be identified. Because of this fact, several gene detection approaches have been proposed all over the world in the recent years. Today, the gene prediction problem is still an open issue, for which there is no definitive method.

The objective of this work is to study the performance of two Weight Matrix Models for the detection of donor and acceptor signals in DNA, and establishing optimal classification thresholds for them.

Although there are known methods with better results than the WMMs, they were chosen for this study due to its simplicity and, mainly, by the intrinsic parallelism of the model. These features allow their further exploration in massive parallel architectures, such as FPGAs (Field Programmable Gate Array) and GPUs (Graphic Processing Unity), thus opening new possibilities for high-performance gene prediction systems [12].

A. Gene detection

A relevant problem in Molecular Biology is the identification of genes in DNA sequences. At the DNA level, there are basically three strategies for identifying genes: methods based on the search for similarities, methods based on the search for biological signals, and content-based methods.

The search for similarities is the oldest approach used for identifying genes. The principle is based on the trend that some coding regions are conserved throughout the evolution. This method is summarized as a search for similar regions among sequences from a known database and the sequence under study. The disadvantage of this method is that the quality of results depends on the quality of the database used in the search.

The principle underlying the search for signals is the investigation for specific sequences involved in the gene expression process. The search for promoters, start and stop codons, splice sites are some examples in this approach. Techniques such as the search for a consensus sequence, weight matrix models, neural networks are some of the methods used for this task. The main difficulty for these methods is that the signals are not always present in DNA sequences, or when they are, they cannot be promptly recognized.

Content-based methods perform a search for segments that have the same statistical properties of regions of DNA that encode proteins. To discriminate coding from non-coding regions only statistics-based models are used. The search is independent of databases for comparison, as in the search for similarity methods, and this is the main advantage of content-based methods.

B. Splice junction recognition problem

There are some differences in the cell structure between eukaryote and prokaryote organisms. Eukaryotes are complex organisms (such as humans) that have a membrane surrounding the nucleus of their cells, and their genetic material is delimited within this nucleus. On the other hand, prokaryotes are simpler organisms (such as bacteria), and they have their genetic material dispersed in the cell. The main difference between eukaryotic and procaryotic gene expression is the splicing of some regions of the genetic material. The process of splicing is a result of the fact that eukaryote genes are composed of two types of segments: exons and introns. Exons are regions of the genetic material that encode proteins. Introns are regions that intermediate exons and do not encode proteins, and thus have to be removed from the mRNA to synthesize proteins. Overall, the function of introns is not yet fully clarified.

Splice junctions are the boundary points where the process of splicing occurs. The transition from an exon to an intron is commonly called exon/intron site (EI). Similarly, the transition from an intron to an exon is called intron/exon site (IE). The fact that causes the process of splicing is the presence of two signals in the chain: the donor signal for exon/intron sites and the acceptor signal for intron/exon sites.

Splice junction recognition is an important part of the eukaryote gene structure prediction process and, basically, it is a problem of detecting donor and acceptor signals. In the last years, several computational models have been proposed for detecting donors and acceptors. For instance, Towell [21] and Rampone [14] employed Artificial Neural Networks, Gelfand & Roytberg [8] used a dynamic programming approach, Cai et al. [5] used Bayes Networks, and Lopes et al. [12] used decision-trees. Other approaches based on statistical models were also proposed and, amongst them, possibly the most important model is the Weight Matrix Model (WMM), by Staden [17]. Derivations of this model are: Weight Array Model (WAM), proposed by Zhang & Marr [24], Windowed Weight Array Model (WWAM) and Maximum Dependence Decomposition (MDD), both proposed by Burge [3].

II. Weight Matrix Models

Position Weight Matrices (henceforth, PWM) are traditionally used to represent small sequence patterns that are related with some molecular functions. In recent years, many works have used PWMs for several problems in bioinformatics. For instance, Staden [17] used PWM for representing some nucleotide sequence signals; Senapathy [15] used them to predict splice sites signals; Bucher [2], Ficket [7] and Wingender [22] represented promoter elements; Gershenzon [9] used PWMs for detecting DNA/protein binding sites.

There are several approaches for building PWMs. However, the most widely used method was proposed by Staden [17]. Using a collection of aligned sequences by some signal, a nucleotide frequency table is constructed by counting the number of times that each base occurs at each position. This frequency table has four rows (one row for each nucleotide: A, C, G and T) and the number of columns are equal to the motif signal length.

The counting in the table can be converted into frequencies. For example, if 137 out of 303 sequences have a T in the first column, the frequency of T in this column is 0.45. Similarly, for the CAP Signal, a C occurs in the second column with frequency 1.00, and so on. Such frequencies matrix indicates the probability that a given base appears at each position of the signal. Equivalently, it can said that a PWM is a classical zero-order Markov model per position.

To measure the similarity of a new sequence to the constructed PWM it is necessary to multiply the probabilities of each nucleotide at each position in the matrix - see equation (1). The larger the similarity of X with the training data, the higher the $P_{WMM}(X)$ scores.

$$P_{WMM}(X) = \prod_{i=1}^{L} p_i(x_i) \tag{1}$$

The frequencies matrix of each nucleotide at each sequence position can be converted into log odds scores. The odds score is simply the frequency observed in the column divided by the frequency expected, that is, the background frequency of the base, usually averaged over the whole genome. Then, the odds score can be converted to a log odds score by taking the logarithm of the odds score, usually to the base 2 and, sometimes, to the natural logarithm. This operation results in the well-known Weight Matrix Model (WMM). An example of WMM is illustrated in Table I, corresponding to the data presented by [2].

	-2	**1**	**0**	**1**	**2**	**3**	**4**	**5**
W(A)	-1.14	-5.26	0.00	-1.51	-0.65	-0.55	-0.91	-0.82
W(C)	-1.16	0.00	-5.21	-0.41	-0.45	0.00	-0.29	-0.18
W(G)	-0.75	-5.26	-5.21	0.00	-4.56	-0.86	-0.38	-0.65
W(T)	0.00	-5.26	-2.74	-0.29	0.00	-0.36	0.00	0.00

Table I
Weight matrix model for the CAP signal [2].

For measuring the similarity of a new biological sequence using a WMM it is necessary to sum the log odds score of each nucleotide at each position in the matrix, using equation 2.

$$Score(X) = \sum_{i=1}^{L} s_i(x_i) \tag{2}$$

The WMM model does not consider the dependence between positions (nucleotides or amino acids), in a chain of DNA or protein. A very common WMM derivation called Weight Array Model (WAM), proposed by Zhang & Marr [24], takes into account the dependencies between adjacent positions in a sequence. In the WAM, a final score is assigned to each position in the sequence for each word with length k (when $k = 1$, the two methods are the same). A variant of WAM, called Windowed Second-Order WAM Model (WWAM), proposed by Burge [3], has some modifications when training the model in order to reduce the incidence of sampling error.

Another very popular derivation model of the WMM is the Maximum Dependence Decomposition (MDD), designed by Burge [3]. The MDD model consists of two components: a decision tree, which first finds the nucleotides at important positions with large amount of influence over others nucleotides, and a simple WMM model at each leaf of the tree.

Currently, WMMs and derivations are widely used in many signal searching applications. Softwares such as Genscan, Twinscan, N-Scan, TigrScan, GlimmerM, and databases, such as TRANSFAC [22], use WMMs and derivations to represent a large number of signals.

All the cited models allow searching DNA sequences to find sites similar to the original set of known sites, typically using a cutoff value or score threshold. However, such cutoff is not clearly defined and is strongly dependent of the data used.

The processing time of the WMMs can be drastically reduced when implemented in massive parallel architectures. Using such this approach, operations that have no dependence can be done at the same time, increasing the overall performance. Figure 1 shows a simple example of how this reduction can be achieved with a parallel implementation.

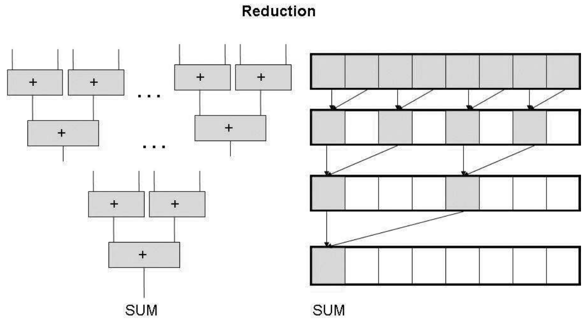

Figure 1. Parallel operations using a FPGA and a GPU.

III. Methodology

A. Data sets

Two data sets of splice junctions were considered in this work. The first one was taken from the well known UCI-Machine Learning Repository [10] and it is a database of sequences from primates. Although this database contains 3190 instances in total, some small changes were done in this work. First, all sequences containing values different from the standard nucleotide symbols 'A', 'C', 'G' and 'T' were removed. Next, sequences were randomly sampled and removed in order to obtain a database with the number of instances per class in a specific proportion: 25% for donors, 25% for acceptors and 50% to other sequences, so as to keep the data set approximately equal to the original one. Such balance between classes also avoids bias in the training of classifiers.

The second database used was taken from the Homo Sapiens Splice Sites Data set (HS3D) [13]. This database contains originally 5947 for human DNA sequences with known splice sites (donors and acceptors) and 635666 sequences with false splice sites. For this data set, 11184 sequences were randomly chosen, distributed in the same proportion as before: 25% for donors, 25% for acceptors and 50% for false splice sites. Table II summarizes the data sets used in this work, showing the total number of instances (# Instances) and the class distribution (# per Class). EI, IE and N stands for Exon-Intron junctions (donors), Intron-Exon junctions (acceptors) and false splice sites, respectively.

Data set	**# Instances**	**# per Class (%)**		
		EI (25%)	**IE** (25%)	**N** (50%)
UCI	3048	762	762	1524
HS3D	11184	2796	2796	5592

Table II
Summary of the two data sets used in the experiments.

B. Predictive accuracy measures

Sensitivity and specificity are two predictive accuracy measures that have been frequently used in the classification literature, especially in bioinformatics [4].

Basically, sensitivity measures the proportion of actual positives instances which are correctly identified as such, and specificity measures the proportion of negatives instances which are correctly identified. Eq. 3 shows how sensitivity (Sn) and specificity (Sp) are computed, based on four possible outcomes of a classifier, as follows:

- TP: true positive - number of positive instances that were correctly classified as positive;
- FN: false negative - number of positive instances that were wrongly classified as negative;
- FP: false positive - number of negative instances that were wrongly classified as positive;
- TN: true negative - number of negative instances that were correctly classified as negative.

$$Sn = \frac{TP}{(TP + FN)} \qquad Sp = \frac{TN}{(TN + FP)} \tag{3}$$

Both, sensitivity and specificity, are defined in the range [0..1], with perfect prediction occurring if and only if both (Sn and Sp) are equal to 1. However, it is possible to have a classifier with high sensitivity and a low specificity, or the opposite. It is easy to observe that, alone, either Sn or Sp, are not a good measure of accuracy. Therefore, it is necessary to devise a single value of overall accuracy to summarize both measures. Such a measure, used for two-class problems, is the Matthews Correlation Coefficient (MCC) (Eq. 4), regarded as a balanced measure and frequently used in bioinformatics [1],[19].

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{4}$$

Another way to evaluate a classifier is the ROC graph. A ROC (Receiver Operating Characteristics) graph is an useful technique for comparing classifiers and observing visually their performance. This kind of graph is commonly used not only in decision making, but also in machine learning, data mining and bioinformatics [16]. In a ROC graph axes x and y are defined, respectively, as $1 - Sp$ and Sn. These axes can be interpreted as the relative trade-offs between the benefits and costs of a classifier. Therefore, the ROC graph can be represented by a single ROC point for each non-parametric classifier, corresponding to their $(1 - Sp, Sn)$ pairs [6]. When comparing classifiers using a ROC graph, the best possible prediction method would be that lying as close as possible to the upper left corner (coordinates $(0, 1)$), representing 100% sensitivity and 100% specificity.

IV. Computational experiments and results

Two different models, proposed by Burge [3], were chosen and evaluated in this work: the Maximal Dependence Decomposition (MDD), for Donors signals; and the Windowed Weight Array Model (WWAM), for Acceptors signals. Both are part of several programs for gene prediction currently in use.

The matrices and the source code of the routines used in this work were taken from the N-Scan, which is a well known gene prediction software developed at Washington University (USA). N-Scan is open source and it is available for download at *http://mblab.wustl.edu*. Since these matrices are found with the N-Scan source code, the reproducibility of this work is possible. Furthermore, N-Scan is, possibly, the open software that currently achieves the highest accuracies in gene prediction.

Both models (MDD and WWAM) were run with the two data sets (UCI and HS3D) varying the prediction threshold from -80 to 100, in unity steps. For each threshold, the corresponding *TP*, *FP*, *FN* and *TN* values obtained from classification were stored. Then, *Sp*, *Sn* and *MCC* were computed.

Figure 2 shows two plots resulting from the evaluation of the MDD model applied to the UCI data set. The first one shows the behavior of *Sp*, *Sn* and *MCC* with different threshold values. The second is the ROC graph for the same model and data set.

Figure 3 shows the plots corresponding to the evaluation of the MDD applied to the HS3D data set. It is possible to observe that the behavior of *Sp*, *Sn* and *MCC* is similar to the first case (MDD applied to the UCI data set). Observing the curve of *MCC*, it is possible to note that the maximum value of *Sp*, *Sn* and *MCC* is approximately equal for the two cases. However, the ROC graphics shows that this model is more sensitive to the threshold when applied to the HS3D data set. This is due to the fact that this data set is more challenging for the classifier method.

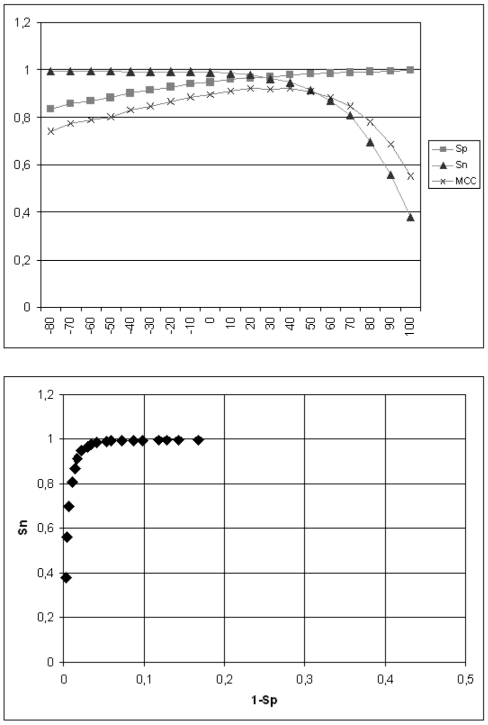

Figure 2. MDD applied to the UCI data set.

Data set	Best threshold	Sp	Sn	MCC
UCI	41	0.979	0.948	0.92330
HS3D	44	0.936	0.936	0.83948

Table III
Best thresholds for the donors problem (MDD model).

Table III shows values for the best cost-benefit for this model. The best classification threshold were very similar for both data sets (namely, 41 and 44).

The same methodology was employed using WWAM. Figure 4 shows two plots resulting from the evaluation of the WWAM model applied to the UCI data set. The first shows the behavior of *Sp*, *Sn* and *MCC*, and the second is the ROC graph. Similarly, figure 5 shows the plots corresponding to the evaluation of the WWAM applied to the HS3D data set.

Data set	Best threshold	Sp	Sn	MCC
UCI	31	0.969	0.936	0.89618
HS3D	50	0.933	0.877	0.79133

Table IV
Best threshold for the acceptors problem (WWAM model).

Results for the second model were somewhat similar to those of MDD, except for the best threshold. In this case the best thresholds were not so close each other than in the MDD evaluation. Table IV shows that the best situation occurs for a threshold around to 31 for the UCI data set, and around to 50 for the HS3D data set.

The comparison of the ROC graphs confirms the fact that the HS3D data set poses a more difficult challenging

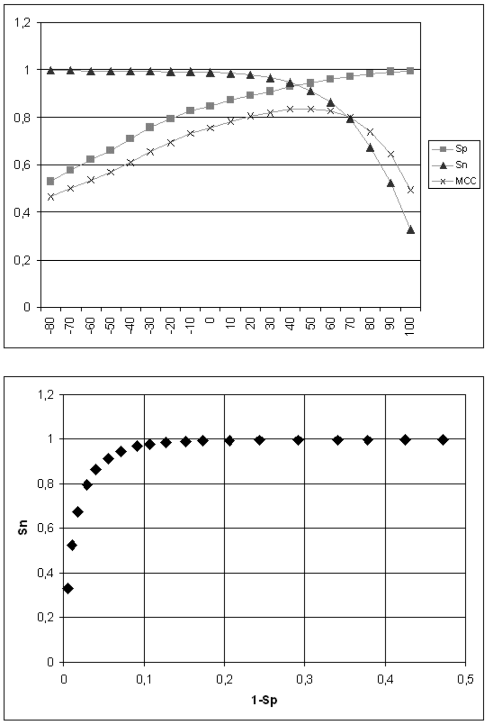

Figure 3. MDD applied to the HS3D data set.

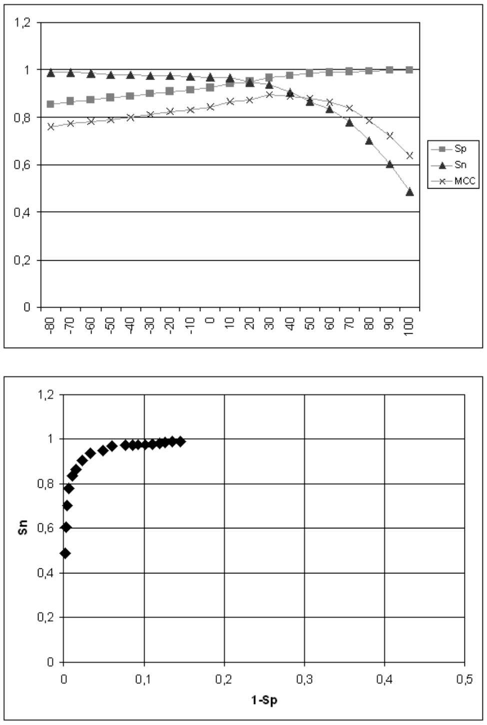

Figure 4. WWAM applied to the UCI data set.

classification problem than the UCI data set for both, donors and acceptors.

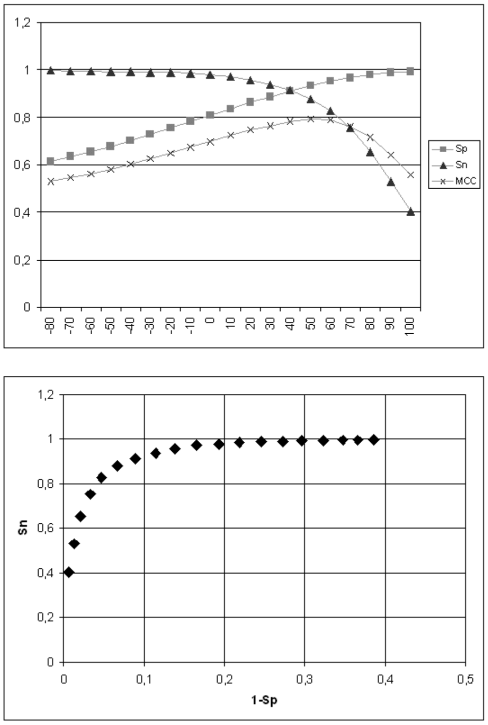

Figure 5. WWAM applied to the HS3D data set.

V. Conclusions

This paper investigated the application of two WMM-derived techniques for the splice junction recognition problem: Maximal Dependence Decomposition (MDD) and Window Weight Array Model (WWAM). The open issue of establishing a classification threshold was addressed for both models, evaluated with two different data sets.

The measures of predictive accuracy using *Sn*, *Sp* and *MCC*, as well as the ROC graph enable a good evaluation of different experiments. In particular, *MCC* offers a good trade-off between sensitivity and specificity, thus suggesting its suitability for other similar bioinformatics classification problems.

Our experiments showed a small difference in performance of the two methods, MDD and WWAM, in favor of the first one. However, results are not conclusive, since two sub-problems were addressed. Although it is clear that both methods showed satisfactory performance, more experiments shall be done.

The direct comparison of results obtained here with other classifiers for the same data sets (such as [21]) may lead to interpretation errors, since different partitions of the data sets and cross-validation procedures can be used. However, as shown in [19] probabilistic-like methods, such as WWAM and MDD, tend to achieve better predictive accuracy than other methods for biological sequences.

Although both data sets are related to the same problem, and are relative to the same subset of organisms (primates/humans), the original methodology for constructing them was, possibly, different each other. This fact is clearly reflected in the difference in performance of the classifiers for the two data sets. The UCI data set turned out to be easier to be classified than the HS3D for both, Acceptors and Donors signals.

The thresholds reported in this work are valuable information that can be later used in the splice junction problem, in the scope of an automatic gene detection system. It is important to recall that this work does not aim to be a complete gene prediction system. Instead, we evaluated the utility of different WMM methods for an important sub-problem of gene prediction and, based on our experiments, we established optimal classification thresholds. In fact, these important results will drive further research in improving the classifiers to be used in a gene prediction system. The present study encourages the development of new experiments using other data sets and instances, particularly aiming at some kind of data-independent automatic threshold prediction.

References

[1] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, pp. 412–424, 2000.

[2] P. Bucher, "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences," *Journal of Molecular Biology*, vol. 212, no. 4, pp. 563–578, 1990.

[3] C. Burge, S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of Molecular Biology*, vol. 268, pp. 78–94, 1997.

[4] M. Burset, R. Guigó, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, pp. 353–367, 1996.

[5] D. Cai, A. Delcher, B. Kao, S. Kasif, "Modeling splice sites with Bayes networks," *Bioinformatics*, vol. 16, pp. 152-158, 2000.

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[7] J.W. Fickett, "The gene identification problem: an overview for developers," *Computers in Chemistry*, vol. 20, 103–118, 1996.

[8] M.S. Gelfand, M.A. Roytberg, "Prediction of the intron-exon structure by a dynamic programming approach," *BioSystems*, vol. 30, pp. 173–182, 1993.

[9] N.I. Gershenzon, G.D. Stormo, I.P. Ioshikhes, "Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2290–2301, 2005.

[10] S. Hettich, C.L. Blake, C.J. Merz, "UCI repository of machine learning databases," *http://www.ics.uci.edu/ mlearn/MLRepository.html*, 1998.

[11] E.S. Lander, L.M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.

[12] H.S. Lopes, C.R. Erig Lima, N.J. Murata, "A configware approach for high-speed parallel analysis of genomic data," *Journal of Circuits, Systems, and Computers*, vol. 16, no. 4, pp. 527–540, 2007.

[13] P. Pollastro, S. Rampone, "HS3D, a dataset of Homo sapiens splice regions, and its extraction procedure from a major public database," *International Journal of Modern Physics*, vol. 13, no. 8, pp. 1105–1117, 2002.

[14] S. Rampone,"Splice-junction recognition on DNA sequences by BRAIN learning algorithm," *Bioinformatics*, vol. 14, pp. 676–684, 1998.

[15] P. Senapathy, M.B. Shapiro, N.L. Harris, "Splice junctions, branch point sites, and exons: sequence statistics identification, and applications to genome project," *Methods in Enzymology*, vol. 183, pp. 252–278, 1990.

[16] T. Sing, O. Sander, N. Beerenwinke, T. Lengauer, "ROCR: Visualizing classifier performance in R," *Bioinformatics*, vol. 21, pp. 3940–3941, 2005.

[17] R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucleic Acids Research*, vol. 12, pp. 505–519, 1984.

[18] L.D. Stein, "Human genome: end of the beginning," *Nature*, vol. 431, pp. 915–916, 2004.

[19] L.G. Tavares, H.S. Lopes, C.R. Erig Lima, "A comparative study of machine learning methods for detecting promoters in bacterial DNA sequences," *Lecture Notes in Artificial Intelligence*, vol. 5227, pp. 959–966, 2008.

[20] G. Towell, J. Shavlik, M. Noordewier, "Refinement of approximate domain theories by knowledge-based artificial neural networks," In: 8^{th} *National Conference on Artificial Intelligence* AAAI Press, pp. 861–866, 1990.

[21] G.G. Towell, "Symbolic Knowledge and Neural Networks: Insertion, Refinement, and Extraction," PhD thesis, University of Wisconsin, Madison, 1991.

[22] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, "TRANSFAC: an integrated system for gene expression regulation," *Nucleic Acids Research*, vol. 28, pp. 3431–3432, 2000.

[23] F. Wright, W.J. Lemon, W.D. Zhao et al., "A draft annotation and overview of the human genome," *Genome Biology*, vol. 2, pp. 1–18, 2001.

[24] M.Q. Zhang, T.G. Marr, "A weight array method for splicing signal analysis," *Computer Applications in Biosciences*, vol. 9, no. 5, pp. 499–509, 1993.