Increasing Splicing Site Prediction by Training Gene Set Based on Species

Beunguk Ahn, Elbashir Abbas, Jin-Ah Park and Ho-Jin Choi
Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea (South)
[e-mail: {elaborate, bashirsalah, jinah.park, hojinc}@kaist.ac.kr]
*Corresponding author: Ho-Jin Choi

Received April 7, 2012; revised May 29, 2012; accepted August 18, 2012; published November 30, 2012

Abstract

Biological data have been increased exponentially in recent years, and analyzing these data using data mining tools has become one of the major issues in the bioinformatics research community. This paper focuses on the protein construction process in higher organisms where the deoxyribonucleic acid, or DNA, sequence is filtered. In the process, "unmeaningful" DNA sub-sequences (called introns) are removed, and their meaningful counterparts (called exons) are retained. Accurate recognition of the boundaries between these two classes of sub-sequences, however, is known to be a difficult problem. Conventional approaches for recognizing these boundaries have sought for solely enhancing machine learning techniques, while inherent nature of the data themselves has been overlooked. In this paper we present an approach which makes use of the data attributes inherent to species in order to increase the accuracy of the boundary recognition. For experimentation, we have taken the data sets for four different species from the University of California Santa Cruz (UCSC) data repository, divided the data sets based on the species types, then trained a preprocessed version of the data sets on neural network(NN)-based and support vector machine(SVM)-based classifiers. As a result, we have observed that each species has its own specific features related to the splice sites, and that it implies there are related distances among species. To conclude, dividing the training data set based on species would increase the accuracy of predicting splicing junction and propose new insight to the biological research.

Keywords: Splice sites prediction, exon, intron, species, machine learning, DNA sequence

The preliminary version of this paper was presented in ICONI 2011 (International Conference on Internet 2011), December 15-19, 2011, Malaysia. This work was supported by the National Research Foundation (NRF) grant (No. 2012-0001001) of Ministry of Education, Science and Technology (MEST) of Korea.
The authors would like to acknowledge those who helped in the process of retreiving data and analyzing the results of the experiments reported in this paper. Special thanks should go to Professor Gwan-Su Yi and Mr. Kyu-Kwang Kim in the Department of Bio and Brain Engineering, and Mr. Min-Seok Hong in the Department of Biology, all at KAIST, Daejeon, Korea.

http://dx.doi.org/10.3837/tiis.2012.10.002

1. Introduction

Protein function prediction has been an important problem in molecular biology, genetics and bioinformatics, since proteins perform the most essential functions in an organism in such forms as structural proteins, enzymes, and transmembrane proteins. Thus understanding protein functions can help develop new drugs, better crops, biochemicals, etc. [1].

In molecular biology, proteins are known to be constructed by a process, called the *central dogma*, of converting a gene to protein via the transcription and translation phases, where transcription refers to the phase of producing ribonucleic acid, or RNA, copies from deoxyribonucleic acid, or DNA. During transcription, only exons and introns are copied from DNA to RNA. An *exon* is a nucleic acid sequence that is represented in the mature form of an RNA molecule, either after portions of a precursor RNA (i.e., introns) has been removed by splicing, or when two or more precursor RNA molecules have been ligated by splicing. An *intron* is any nucleotide sequence within a gene that is removed by RNA splicing to generate the final mature RNA product of a gene. After transcription, splicing occurs by modifying an RNA within which introns are removed and exons are joined. The next phase, translation, is the stage of protein biosynthesis. During translation, messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein.

Recognizing boundaries of exons and introns is an important problem for understanding the production of a protein, hence for the prediction of protein structure, for example, as used in alternative splicing [2][3]. Alternative splicing means the process by which exons of the RNA produced by transcription of a gene (e.g., a primary gene transcript or pre-mRNA) are reconnected in multiple ways during RNA splicing. As the result of this splicing, various forms of proteins are produced. In other words, predicting the boundaries of exons and introns leads to the possibility of predicting the protein structures [2]. This paper aims to address this problem of predicting splice sites.

Conventionally, approaches to splice site prediction have used machine learning techniques such as knowledge-based artificial neural networks (KBANN) [4][5], neural networks (NN) combined with rule-based systems [6], support vector machine (SVM) [7], and so on. We believe that this was a reasonable trend because splice sites prediction is basically a classification problem, just like many other problems in many engineering domains do use statistical classifiers for prediction (e.g., see [8][9][10][11]). In our domain, simply speaking, these conventional approaches tried to find answers to "Which machine learning algorithm is better-suited for this biological problem?" to achieve better accuracy in splice site prediction.

On the other hand, other school of researchers [2][12][13] observed that splicing patterns have species variations, suggesting that splice sites may be species dependent. Basically, their common findings are that for some particular genes or enzymes appearing in two species, even though their gene sequences are very similar, their splice sites are quite different species by species. Although they have dealt with only a few specific species and proteins focused individually, these observations may help us develop more general methods for splice site prediction if we can consider the relationships among species with respect to splicing patterns for thousands of splice sites. This idea has motivated our research, and this paper describes an initial attempt towards this direction.

In the preliminary version of this paper [14], we proposed a novel approach to splice site prediction and showed experimental results. There, we focused on sketching the experiments

and presenting the results, without explaining the detailed process. In this paper, we explicate our approach by providing biological background knowledge and describing in detail the representation scheme and the process of experimentation. Basically, our proposed approach handles together multiple sets of gene data from different species at the same time, with an aim to exploit the implicit relationships among various species in the process of predicting their corresponding splice sites. Based on the observations by [2][12][13] discussed earlier, we first assume that species belonging to the same class would present more similar characteristics to one another than to those belonging to other class, then we hypothesize that the accuracy of splice site prediction can be increased through a machine learning process which can utilize the implicit relationships among various species, inherently embedded in their gene data. Validating this hypothesis is the objective of our research in this paper.

As for the choice of machine learning algorithm, it does not matter which algorithm to use because our aim is not to develop the best prediction algorithm but to show the dependency of a prediction process by species characteristics. Nonetheless, we wish to be sure that our hypothesis can be validated by at least two standard machine learning algorithms. For this reason, we have chosen NN and SVM for our experiments.

The contribution of our work can be summarized as follows. First, we use about 4,000 genes from each species in training and testing the classifier in order to learn splicing patterns inherent to the particular species. This is different from existing approaches where one or two specific genes were studied to observe the species variance of splice sites. As for the number of species studied, we use four species whereas existing approaches used two or three. Second, we design an experiment process to validate the hypothesis of species variations by comparing the resulting species-specific classifiers. This is done by observing the difference in the prediction accuracy between the case when testing a species' data by that species' classifier and the case when testing the species' data by other species' classifier. We perform this test for all possible combinations among the four species chosen in our experiments. Third, we diversify our experiments to draw small but meaningful observations additionally. When choosing the four species, we diversity them in an "unbalanced" way by choosing two species from the same class (i.e., mammals) and the other two species each from totally different classes (i.e., one secernentea and one insect). This diversification will give us chance to observe whether similar species induce more similar splicing patterns to each other than to remote species in the taxonomy tree. In addition, we use two machine learning algorithms (i.e., NN and SVM), thus we will also have the chance to observe the difference in the performance of these two algorithms.

The rest of this paper is organized as follows. Section 2 summarizes existing research work which had similar objectives to ours. Section 3 introduces the basic knowledge about splice site and describes the method and process of our approach. Section 4 describes the detailed aspects of our experimentation performed, and the results are presented and analyzed in Section 5. Section 6 concludes.

2. Related Work

Several researches have been conducted to determine whether species differ in terms of their splice sites. In a global analysis of alternative splicing between human and chimpanzee [2], it was found that both species have similar protein sequences for enzyme GSTO2, but their splice sites were different, resulting in slightly varying enzymes. This difference in alternative splicing between human and chimpanzee causes variance of functionality such as signal transduction and cell death. Even if the RNA sequences are similar, splice site selections can

vary and this would cause species variation. Barash et al. [3] proposed a new way to find alternative splice sites, which looks for constant patterns that define splice sites by investigating the special meaning of sequence codes in 600 nucleotides around the plausible boundary of splice sites. This approach aimed to identify mutation-verified sequences as the biological rule of splicing patterns by consulting hundreds of RNA features in the assembly generation of biological complex. Greaser et al. [12] found the difference between rat and dog, in genes N2B and N2BA, that even though the gene sequences are similar, different splice sites exist, causing performance variation on passive tension in cardiac muscles. Bothwell et al. [13] observed species-specific difference in gene expression and splice-site choice in gene Inpp5b, having similar sequences but varying splice sites between human and rat. Overall, these researches provide evidence that splice sites are species dependent and the nucleotides that make up the boundary of the splice sites play an important role in RNA splicing.

On the other hand, some researches have used machine learning techniques to predict unclassified splice sites. Hebsgaard et al. [6] used artificial neural network combined with the concept of rule-based systems to predict the splice site in Arabidopsis thaliana. They applied factors such as confidence value and distance between potential splice sites, and as a result, managed to surpass the accuracy of other predicting models. Sonnenburg et al. [7] applied SVM in this same effort of splice site prediction. They employ weighted degree kernel method which turned out well suited for the genome-wide recognition of splice sites in Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Danio rerio, and Homo sapiens. This approach claimed higher accuracy prediction rate which was unparalleled in terms of splice site prediction upon its release.

KBANN [4][5] was also an earlier effort that attempted splice site prediction. It combined domain knowledge with neural network and managed to reduce the amount of time and training data required to reach optimal classification as compared to an unmodified neural network. It has been later enhanced to produce more comprehensible deliverables [15]. Although KBANN showed lower error rates even with small examples than the original neural networks, the resulting classifiers are not comprehensible because we cannot read the rules themselves but only see the output. To make the rules more comprehensible by human, NofM [15] extracts comprehensive rules from KBANN, but at the price of lower accuracy. Similarly, other classification algorithms such as k-nearest neighbor (k-NN), Markov chains, and feature-based classification might have been used. Deshpande and Karypis [16] compared the performance of these algorithms using the same data set, concluding that SVM was the best fitted one for gene sequence prediction.

In yet another direction, some approaches have tried to create databases for collecting information about splice sites and correcting errors on previous splice site records [17].

3. Methods

This section explains the problem of recognizing splice site and presents our approach to dealing with this problem. As the basic background knowledge, a brief expository discussion spanning the central dogma, gene expression and transcription is presented first.

3.1 Background – Central Dogma and RNA Splicing

The central dogma is the agreed upon framework for understanding the transfer of information within and in between living organisms. There are three major classes of building blocks (biopolymers) that govern this transfer of information: DNA, RNA and proteins. Although

there are nine conceivable direct information transfers among them, three of those general transfers are believed to occur naturally in most cells [18], called the central dogma of molecular biology (see **Fig. 1**).

The three transfers are (1) DNA replication where the DNA is copied, (2) transcription where DNA information is transcribed to mRNA, and (3) translation where proteins are synthesized using the information in mRNA. Transcription, simply speaking, is the process of making RNA copies from DNA. It uses the template strand, which is the non-coding DNA strand as a blueprint for the RNA molecule. An enzyme, RNA polymerase, performs this process by growing RNA molecule chain by adding one RNA nucleotide at a time. The product of transcription synthesizes premature messenger RNA (pre-mRNA). This pre-mRNA requires some extra processing in order to become mRNA, and eventually is translated into a final product. Among the several steps involved in the post-processing of the pre-mRNA, the one that is central to our research is RNA splicing.

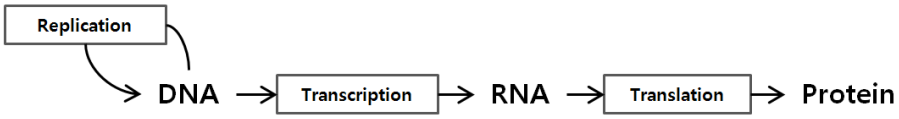

Fig. 1. The central dogma

RNA splicing consists of the removal of introns and the formation of the final mRNA molecule by joining the exons together (see **Fig. 2**). Introns, actually derived from in intragenic regions, refer to the regions in the gene which are believed not to be decoded to produce proteins. Exons are the regions which contain the important codes for producing proteins [19]. The removal of introns is achieved through a series of reactions which are facilitated by the spliceosome. There are two types of splicing: canonical splicing and non-canonical splicing. Canonical splicing is known to account for around 99% of all the splicing, and is defined by the removal of introns that contain 'GU' nucleotides at the acceptor sites and 'AG' at the donor sites. Non-canonical splicing, known to account for the remaining rare 1% of all splicing, occurs when the minor spliceosome excises introns with different splice sites [20]. After the introns are excised, exons are combined in order to form the mRNA molecule. This molecule is then ready to be translated into a protein [21].

Recognition of these splice sites, i.e., the boundaries that exist between the exons and introns, is an important factor in the production of proteins, hence in the prediction of protein structures, which is the key to the determination of the protein functions. A great deal of research has been conducted in this area, e.g., alternative splicing. Alternative splicing denotes the process in which pre-mRNA exons are joined in a number of alternative ways during the concluding part of RNA splicing. The result of this process is various forms of proteins that are in existence [22].

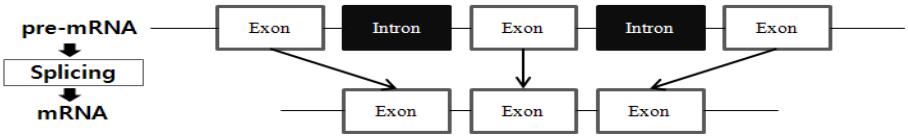

Fig. 2. RNA splicing of introns and exons

3.2 Hypothesis and Proposed Solution

Our aim is to show the difference in DNA sequence among various species by applying each data set for each species to the machine learning algorithms, and show the comparisons between algorithms and the difference in DNA sequence among various species for compensating the defect of the existing studies. Through experimentation, therefore, we wish to prove the hypothesis, "If the pattern of DNA sequence around the boundary sites are related to the species type, then the accuracy of boundary prediction will be increased by dividing the training sets according to the species". We choose the length of the gene sequence to be 60, because this number has been known to be sufficient enough to produce promising accuracy in the previous approaches [4][5][15].

Classification or supervised learning aims to produce a concept description inherent in data that accurately predicts a certain target or class for. In our experiment, we will use standard techniques such as NN and SVM, independently, in order to extract rules. We have chosen these two techniques among others because they have been the most popular baseline algorithms in many similar approaches, and it makes possible to confirm our approach is proper for predicting splice sites by having at least two techniques. With selected machine learning techniques, we will generate two kinds of classifiers: (1) classifiers of each species train-set, and (2) classifiers of mixed species train-set. Each train-set of one species is composed of the information about around 3,000 classified splice sites, and mixed species train-set is composed of the same ratio from each train-set of species. Similarly, each test-set is composed of the information about around 1,000 classified splice sites to verify after extracting classifiers. As the result, there are (n+1) train-sets (i.e., one set for each of n species plus one set of mixed species), and n test-sets (i.e., one set for each species). Mixed test-set is not required because valuable meaning cannot be inferred by testing with mixed species in proving our hypothesis. In this experiment, we generate 2n classifier models with train-sets, of which n models use NN and the other n models use sequential minimal optimization (SMO), which is a practical implementation of SVM on the WEKA tool [23], as shown in **Fig. 3**.

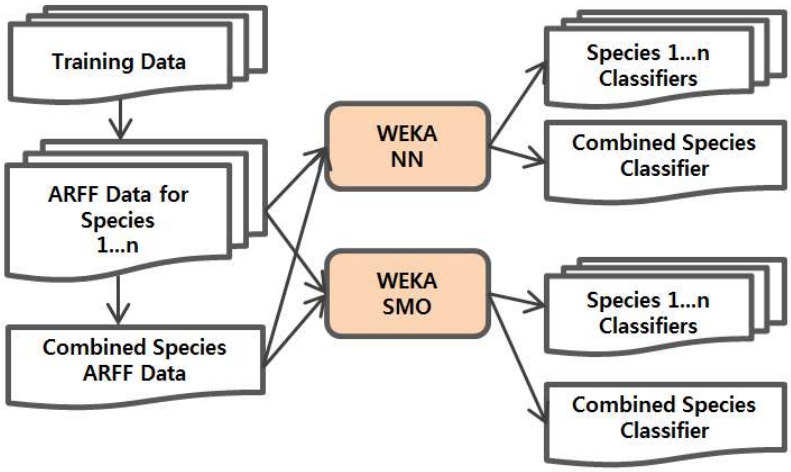

Fig. 3. The proposed process

3.3 Verification Model

To prove the hypothesis, we need to show two main points in our experiment. First, it is

required to see that using machine learning techniques (such as NN and SMO) is proper to predict splice sites in general by looking average values of accuracies in testing with their own species test-set (e.g., using human test-set for human classifiers). By showing high enough accuracy in testing with their own species, we can say that classifiers have been generated properly to predict splice sites for own species.

After confirming that each rule has enough accuracy for its own species, the next process is to check how accuracy becomes different when testing with other species test-sets (e.g., using mouse test-set for human classifiers). If the accuracy is decreased, it implies that there may be features in splice sites based on species. Having extracted accuracy for each case, it is possible to see overall changes to infer features on species in splice sites. This process is described in the upper box (indicated by '*') in **Fig. 4**. With combined species data, it also follows same process in **Fig. 3** to extract a combined species classifier. However, this classifier would be used to compare accuracies by testing with test-set of each species to look how much difference of predicting splice sites among selected species. The diagram of this process is described as the lower box (indicated by '**') in **Fig. 4**.

By comparing the two ways mentioned above, it is possible to look splice sites' features based on species. Consequently, comparing the result will directly verify our hypothesis.

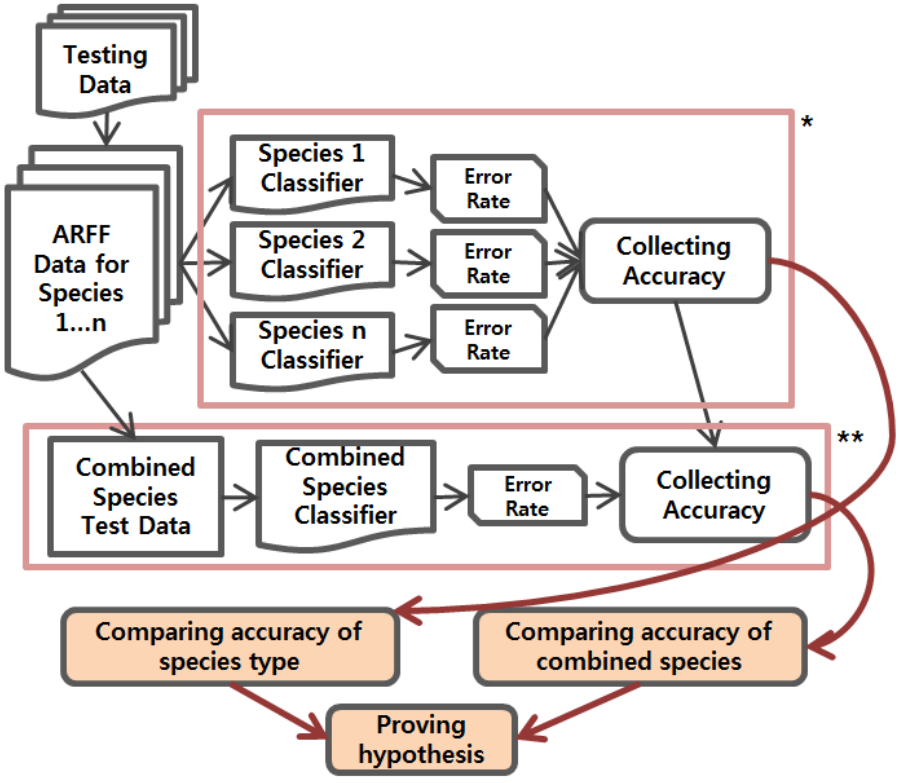

Fig. 4. Verification of the process

4. Experimentation

The experiment was conducted using WEKA (Waikato Environment for Knowledge Analysis), a machine-learning workbench implemented in Java [23]. It provides a collection of

algorithms for analyzing data and building predictive concept descriptions, and tools to visualize data. Two algorithms were used in our experiment: (1) SMO, which is a practical implementation of SVM on WEKA; (2) multilayer perceptron neural network.

4.1 Data Preparation

Our experimental data sets were obtained from University of California Santa Cruz (UCSC) data repository [24]. The data sets represent gene sequences of two mammals (human and mouse), one insect (D.melanogaster), and one nematode (C.elegans), that is, four species in total. 4,200 gene sequences were retrieved for each species, 3,150 of which were used for training and 1,050 used for testing.

The UCSC database is chosen for retrieving gene sequences. In this database, four species are selected that contains enough splice sites data: human, mouse, C.elegans, and D.malenogaster. Note that among the four species, human and mouse belong to the same class (i.e., mammal). This would make the experimental results to include more about species' features.

We have retrieved splicing boundary sites from refFlat.txt file in the UCSC database on each species which contains gene translation information. **Fig. 5** shows the scheme of refFlat.txt file, and **Fig. 6** illustrates an example.

Attribute	Type	Description
geneName	String	Name of gene as it appears in Genome Browser
name	String	Name of gene
chrom	String	Chromosome name
strand	char[1]	+ or – for strand
txStart	uint	Transcription start position
txEnd	uint	Transcription end position
cdsStart	uint	Coding region start
cdsEnd	unit	Coding region end
exonCount	uint	Number of exons
exonStarts	uint[exonCount]	Exon start position
exonEnds	unit[exonCount]	Exon end position

Fig. 5. The scheme of UCSC refFlat file

```
ITPRIPL1                          // geneName
NM_001008949                      // name
chr2                              // chrom
+                                 // strand
96354788                          // txStart
96357818                          // txEnd
96355304                          // cdsStart
96357764                          // cdsEnd
3                                 // exonCount
96354788,96355211,96356106,       // exonStarts
96355030,96355314,96357818,       // exonEnds
```

Fig. 6. Example of refFlat.txt file

The most important attributes in the scheme include "chrom", "strand", "exonCount", "exonStarts", and "exonEnds". "chrom" is the name of chromosome which gene in contained,

and “strand” means directionality of nucleic acid. “exonCounts” is the number exons that the gene contains, “exonStarts” is comma spliced array of position of each exon starts, and “exonEnds” is comma spliced array of position each exon ends. In other words, we can extract into specific data format by splicing gene sequence on each index of “exonStarts” and “exonEnds”. However, spliced gene sequence should be reversed and complemented (i.e., pairing A-T, G-C) if the value of “strand” is “-”.

There are several conditions while extracting splice site information from refFlat.txt file. It requires to include verified genes that are from real name of chromosomes on each species (e,g., Chr1, Chr2, etc.), because genes from other chromosomes (e.g., ChrU) has a possibility of containing unverified gene information. Moreover, the value of “exonCount” in refFlat scheme should be larger than 2, unless it is impossible to extract boundary information of intron-to-exon (I/E) and exon-to-intron (E/I) splice sites. Finally, it is also necessary to filter out genes when the length of exon is less than 60 and the number of nucleotides between exons in the gene is less than 120. The reason of this exclusion process is to avoid overlapping of gene sequence data when extracting 60 around genes on ‘E/I’, ‘I/E’, and nothing significant (N) splice sites. See **Fig. 7** for an illustration of the data format.

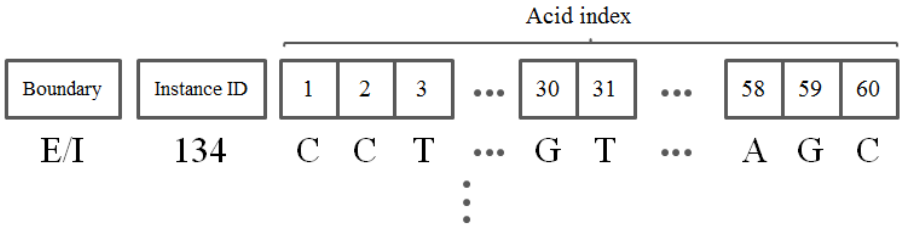

Fig. 7. Format of the instance in data set

The data format for each retrieved species contains three attributes: boundary type, instance ID, and gene sequence. Boundary type contains the class of the 30th position boundary location in the gene sequence, where it can be one of {I/E, E/I, N}. Again, ‘I/E’ represents the boundary from intron to exon, ‘E/I’ the boundary from exon to intron, and ‘N’ means “nothing significant”. Instance ID states identification number for each data for its uniqueness for experiment. Gene sequence contains 60 gene sequences and one boundary location, the 30th position and the classification of I/E, E/I, or N. As depicted in **Fig. 7**, this data set is an emulation of transcription, exon, and intron.

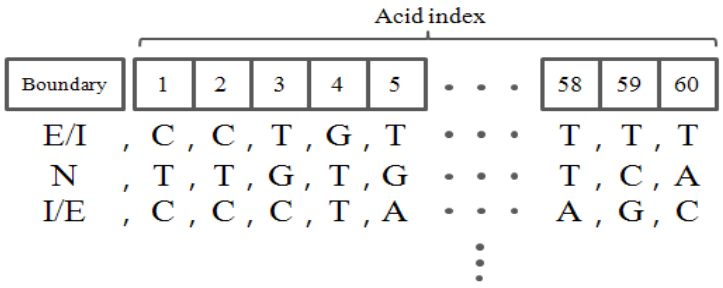

Fig. 8. Portion of an ARFF file

To select splice sites candidates based on the rules mentioned above, splice sites in a refFlat.txt file is sorted randomly for equal distribution on chromosomes by using Linux program called

"sort" with turning on random option flag. After selecting splice sites, they are sorted randomly again to make equal distribution.

With selected splice site candidates, we can generate train-set and test-set ARFF files composed with about 1,000 and 3,500 records, respectively. As a result, part of the ARFF files will look like as illustrated in **Fig. 8**. Here, each training data contains 61 columns. First column is boundary mark which can be one of 'I/E', 'E/I' and 'N'. Other 60-column attributes represent 60 nucleotides, which are deemed around splice sites. Each of these 60 columns contains one of {A, G, C, T}, where 'A' stands for Adenine, 'G' for Guanine, 'C' for Cytosine, and 'T' for Thymine.

4.2 Process Execution and Verification

According to the proposed process shown in **Fig. 3**, the training data is converted to the ARFF (attribute relation file format), which is accepted by WEKA, for each species. In addition, a combined species data set is generated from them, which is a data set that contains the same amount of data from each species. That is, this data set is composed of an "assortment" of all the species used, hence called the combined species data set. These training sets are fed into the two algorithms (i.e., NN and SMO) to generate predictive classification models (classifiers). That is, for each training set, we generate two classifiers based on the two algorithms, respectively, using WEKA [23].

After the predictive concept description is generated, we verify our hypothesis by the process shown in **Fig. 4**. We provide test data for each of the species used. To prove our hypothesis, we have to show two things. The first is that "the species classification is dependent on species type". We do this by providing each species predictive classification model with the different species test data. We then compare the accuracy to determine this. The second is to show that "the accuracy of prediction can be increased by separating the classifiers for each species". This is done by testing our merged species data set against test data for other species. We also test it against the combined species data set, which has the same characteristic as the merged species data set but composed of test data. We then compare the results with the previous species-specific accuracies, and then determine if this part of the hypothesis holds.

5. Results and Discussion

This section presents the results of the experiment, analyze the results, and discuss on the findings.

5.1 Results and Analysis

We measure the results in terms of the accuracies of the splice site prediction for the studied species. In the tables to follow, a shorthand convention is used: 'NN' for multilayer perceptron and 'SMO' for sequential minimal optimization in the captions. The captions read such as "Accuracy of <species>–<classifier> classifier", where <species> means one of the four species used to train the classifier, and <classifier> means either NN or SMO.

The results in **Table 1** to **Table 8** denote classifiers that were learned for the individual species, that is, **Table 1** and **Table 2** represent the classifiers learned from the human splice site data set, and the results contained in them represent the accuracies of the classifier being tested on the four species of our study (including human itself). **Table 3** and **Table 4** represent the C.elegans classifier and the results associated with testing the classifier on the four species. Similarly, **Table 5** and **Table 6** represent the D.melanogaster species, and **Table 7** and **Table**

8 represents the mouse species. On the other hand, Table 9 and Table 10 represent classifiers that were learned from a data set that contained a combination of all four species, whose results represent the accuracies of testing these classifiers on the species in our study.

Each table shows attributes that represent the statistics that govern the learned classifiers. True positive rate (TP rate) denotes the accuracy in the true identification of a splice site in a sequence. False positive rate (FP rate) denotes the false classifications of a splice site. Precision denotes the precise number of identified splice sites from the total number of splice sites. F-measure is a weighted average of the precision and recall. Finally, ROC area provides a confidence check on the accuracy of prediction of splice sites.Table 1 and Table 2 show that the learned human species classifier performs very well for predicting the splice sites in other species with the NN classifier, however it performs below 89% accuracy for the C.elegans species with the SMO classifier. Its highest splice site predictions are for the human and mouse species.

Table 1. Accuracy of Human-NN classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.958	0.02	0.958	0.958	0.994
C.elegans	0.91	0.058	0.915	0.91	0.987
D.melanogaster	0.947	0.031	0.947	0.947	0.994
Mouse	0.958	0.021	0.958	0.958	0.995

Table 2. Accuracy of Human-SMO classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.937	0.031	0.937	0.937	0.964
C.elegans	0.889	0.071	0.894	0.888	0.93
D.melanogaster	0.929	0.042	0.93	0.928	0.957
Mouse	0.944	0.028	0.944	0.944	0.969

Table 3 and Table 4 show the learned classifiers for the C.elegans species. The accuracy for splice site prediction is highest for C.elegans and D.melanogaster, and below 89% for the rest with the least accuracy being for the human species.

Table 3. Accuracy C.elegans-NN classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.863	0.088	0.875	0.863	0.979
C.elegans	0.966	0.017	0.966	0.966	0.996
D.melanogaster	0.909	0.058	0.913	0.907	0.99
Mouse	0.865	0.083	0.871	0.859	0.981

Table 4. Accuracy C.elegans-SMO classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.834	0.103	0.841	0.829	0.903
C.elegans	0.95	0.026	0.951	0.95	0.971
D.melanogaster	0.905	0.059	0.909	0.903	0.944
Mouse	0.85	0.089	0.853	0.846	0.907

Table 5 and **Table 6** show the D.melanogaster learned classifier and the results are above 90% for all species, with the least accuracy being for mouse and human species. Finally, **Table 7** and **Table 8** show the learned classifiers for the mouse species. The splice site prediction is above 90% except for the C.elegans species.

Table 5. Accuracy of D.melanogaster-NN classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.947	0.026	0.947	0.947	0.991
C.elegans	0.946	0.028	0.948	0.946	0.994
D.melanogaster	0.955	0.023	0.956	0.955	0.997
Mouse	0.937	0.028	0.94	0.938	0.993

Table 6. Accuracy of D.melanogaster-SMO classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.918	0.039	0.919	0.918	0.957
C.elegans	0.927	0.039	0.929	0.927	0.959
D.melanogaster	0.951	0.025	0.952	0.951	0.972
Mouse	0.91	0.042	0.913	0.91	0.956

Table 7. Accuracy of Mouse-NN classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.956	0.021	0.957	0.956	0.995
C.elegans	0.893	0.071	0.901	0.893	0.984
D.melanogaster	0.939	0.036	0.94	0.939	0.991
Mouse	0.95	0.024	0.95	0.95	0.994

Table 8. Accuracy of Mouse-SMO classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.927	0.036	0.927	0.927	0.962
C.elegans	0.874	0.084	0.884	0.873	0.923
D.melanogaster	0.913	0.052	0.915	0.913	0.948
Mouse	0.935	0.032	0.936	0.935	0.962

As for the combined species learned classifiers, **Table 9** and **Table 10** show that the accuracy results are above 90% for all species. The lowest species classified was the C.elegans, while the highest species classified was D.melanogaster.

Table 9. Accuracy of combined species-NN classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.951	0.025	0.951	0.951	0.993
C.elegans	0.95	0.026	0.951	0.95	0.991
D.melanogaster	0.96	0.022	0.96	0.96	0.995
Mouse	0.962	0.018	0.962	0.962	0.996

Table 10. Accuracy of combined species-SMO classifier

Test Species	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Human	0.924	0.038	0.924	0.924	0.958
C.elegans	0.94	0.033	0.94	0.94	0.966
D.melanogaster	0.936	0.034	0.936	0.936	0.962
Mouse	0.937	0.032	0.937	0.937	0.965

All the true positive results obtained from our experiment for species specific classifiers are at 80% true positive identification of splice sites for the worst case and best case being 96.6%. These high percentages are attributable to the inherent canonical features present in the genomic sequences, which account for about 95% of all the data. We believe this to be a reasonable and meaningful basis for comparison, from which to deduce several insights.

Regarding the accuracy in predicting splice sites, the NN classifiers generally outpreforms the SMO classifiers. This can be observed from the overall results in both the true positive and the ROC area results shown in the tables. That is, the NN classifier is more capable of generating better splice site concept description.

If we focus on the mammals (i.e., human and mouse), the true positive results obtained by their classifiers for the other species is similar to each other, that is, they exhibit similar patterns. They show 1%-difference in true positive classification. When looking at the other two species, we observe something different. When examining the C.elegans, for instance, the pattern seems to favor higher accuracy prediction for the same species and D.melanogaster, but poorer for mammals. For the case of D.melanogaster, the accuracy values are closer to the human overall results. From this, we can deduce that D.melanogaster is governed by a splice site pattern that is closer to humans than C.elegans. This can be further validated by the fact that 75% of human diseases genes and 50% of the protein sequence in D.Melanogaster are homologus [25]. Furthermore, when comparing **Table 1** and **Table 3**, we can clearly validate the hypothesis that a difference between species in terms of splice site exists. Therefore, the accuracy of prediction could be increased through the combination of species having similar splice sites.

5.2 Discussion

From the experiment, we could argue that features of splice sites may be affected by species. However, it seems necessary to complement some part of experiment to identify exact features in further study.

First, splice site data may not be equally distributed among all species in the world. In this paper, the four species consist of two mammals, one secernentea, and one insect. For this reason, our combined train-set could have been biased to have mammals splice site mainly. However, at the time of experiment, the UCSC database did not contain enough splice sites information in refFlat.txt file. Therefore, trying similar experiments with a variety of species cannot be done to specify features on splice sites.

Second, selecting 60 nucleotide acids around boundary sites may not be enough to have species features. On related works, mechanism on selecting alternative splice sites uses around 200 nucleotides to identify features. It is worth trying to generate classifiers with wider range of nucleotides around boundary sites.

Third, it is necessary to narrow the scope to specify where the features come from. Making variance of splice sites was not done in our experiment. We only considered that non-canonical splice sites would contain species features for splice sites. However, most of them are canonical splice sites and the number of splice sites is too small to use machine

learning techniques. If large amount of non-canonical splice sites has been discovered, importing similar techniques could be possible to identify features.

Finally, not only having high prediction for splice sites is, but finding human-understandable meaning is also important to meet the purpose of finding features. In this paper, the experiment used only the NN and SMO algorithms, which create complex models of "black box" that humans cannot interpret.

6. Conclusion

To conclude, the following points can be made from the experiment presented in this paper. First, although both NN and SMO show good accuracy results on splice sites, NN tends to show better performance than SMO. Second, most accuracy results show more than 95%, which justify that the method is appropriate to predict splice sites. Third, there is strong possibility that splice sites have features based on speices by comparing accuracy results among species classifiers. According to these points, prediction of splice sites can be enhanced, and finding new patterns of splice site that can solve more biological problems.

There are various ways to improve the experiment as the future work. One is to observe the differences in prediction accuracy by changing the range of boundary sites in order to find the best way to increase accuracy. Possible ranges can be some serial sequences around splice sites, or sparse gene sequences which have meaningful features in biology. Another way to make our hypothesis stronger is to conduct similar experiments with much larger volume of species data. If splice site classifiers discover many patterns from this large data set, showing different accuracy results depending on species, it will support our hypothesis more strongly. Finally, developing algorithms that extract human-readable patterns would also improve the analytical power of our approach.

References

[1] Gaurav Pandey, Vipin Kumar and Michael Steinbach, *Computational Approaches for Protein Function Prediction: A Survey*, Technical Report TR 06-028, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minneapolis, MN, USA, Oct. 2006. Available: http://www-users.cs.umn.edu/~kumar/papers/survey.php. Access date: Aug. 2012.

[2] John A. Calarco, Yi Xing, M. Caceres, Joseph P. Calarco, Xinshu Xiao, Qun Pan, Christopher Lee, Todd M. Preuss and Benjamin J. Blencowe, "Global Analysis of Alternative Splicing Differences between Humans and Chimpanzees," *Genes and Development*, pp. 2963-2975, Cold Spring Harbor Laboratory Press, Oct. 2007. http://dx.doi.org/doi:10.1101/gad.1606907

[3] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe and Brendan J. Frey, "Deciphering the splicing code," *Nature*, vol. 465, no. 7294, pp. 53-59, May 2010. Article (CrossRef Link)

[4] Michiel O. Noordewier, Geoffrey G. Towell and Jude W. Shavlik, "Training Knowledge-Based Neural Networks to Recognize Genes in DNA Sequences," Appears in *Advances in Neural Information Processing Systems*, vol. 3, R. Lippmann, J. Moody and D. Touretsky (eds.), Morgan Kaufmann, 1991. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4423

[5] Geoffrey G. Towell, Mark W. Craven and Jude W. Shavlik, "Constructive Induction in Knowledge-Based Neural Networks," Appears in *Machine Learning: Proceedings of the Eighth International Workshop*, L. Birnbaum and G. Collins (eds.), Morgan Kaufmann, 1991. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.6881

[6] Stefan M. Hebsgaard, Peter G. Korning, Niels Tolstrup, Jacob Engelbrecht, Pierre Rouzé and Søren Brunak, "Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information," *Nucleic Acids Research*, vol. 24, no. 17, pp. 3439–3452, July 1996.

Article (CrossRef Link)
[7] Sören Sonnenburg, Gabriele Schweikert, Petra Philips, Jonas Behr and Gunnar Rätsch, "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8, no. S-10, pp. 7-16 (supplementary), Dec. 2007. Article (CrossRef Link)
[8] Dongkyoo Shin, Dongil Shin, Quoc Cuong Nguyen and Seyoung Park, "Real-Time Tracking of Human Location and Motion using Cameras in a Ubiquitous Smart Home," *KSII Transactions on Internet and Information Systems*, vol. 3, no. 1, pp. 84-95, Feb. 2009. Article (CrossRef Link)
[9] Yi Zeng and Thomas M. Chen, "Classification of Traffic Flows into QoS Classes by Unsupervised Learning and KNN Clustering," *KSII Transactions on Internet and Information Systems*, vol. 3, no. 2, pp. 134-146, Apr. 2009. Article (CrossRef Link)
[10] Choong-Nyoung Seon, JinHwan Yoo, Harksoo Kim, Ji-Hwan Kim and Jungyun Seo, "Lightweight Named Entity Extraction for Korean Short Message Service Text," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 3, pp. 560-574, Mar. 2011. Article (CrossRef Link)
[11] Chankyu Park, Jaehong Kim, Joo-chan Sohn and Ho-Jin Choi, "A Wrist-Type Fall Detector with Statistical Classifier for the Elderly Care," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 10, pp. 1751-1769, Oct. 2011. Article (CrossRef Link)
[12] Marion L. Greaser, Mustapha Berri, Chad M. Warren and Paul E. Mozdziak, "Species variations in cDNA sequence and exon splicing patterns in the extensible I-band region of cardiac titin: relation to passive tension," *Journal of Muscle Research and Cell Motility*, vol. 23, pp. 473–482, July 2002. Article (CrossRef Link)
[13] Susan P. Bothwell, Leslie W. Farber, Adam Hoagland and Robert L. Nussbaum, "Species-specific difference in expression and splice-site choice in Inpp5b, an inositol polyphosphate 5-phosphatase paralogous to the enzyme deficient in Lowe Syndrome," *Mammalian Genome*, vol. 21, no. 9-10, pp. 458-466, Sep. 2010. Article (CrossRef Link)
[14] Beunguk Ahn, Jin-Ah Park, Elbashir Abbas and Ho-Jin Choi, "Increasing Splicing Site Prediction by Training Gene Set Based on Species," in *Proc. of 3rd Int'l Conf. on Internet (ICONI 2011)*, pp. 403-407, Sepang, Malaysia, Dec. 2011.
[15] Geoffrey Towell and Jude W. Shavlik, "Interpretation of Artificial Neural Networks: Mapping Knowledge-Based Neural Networks into Rules," Appears in *Advances in Neural Information Processing Systems*, vol. 4, J. Moody, S. Hanson and R. Lippmann (eds.), Morgan Kaufmann, 1992. Available: http://pages.cs.wisc.edu/~shavlik/abstracts/towell.nips4.abstract.html. Access date: Aug. 2012.
[16] Mukund Deshpande and George Karypis, "Evaluation of Techniques for Classifying Biological Sequences," Appears in *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '02)*, pp. 417-431, Taipei, Taiwan, 2002. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.9003
[17] M. Burset, I. A. Seledtsov and Victor V. Solovyev, "SpliceDB: database of canonical and non-canonical mammalian", *Nucleic Acids Research*, vol. 29, no. 1, pp. 255-259, Sep. 2001. Article (CrossRef Link)
[18] Francis Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258, pp. 561-563, Aug. 1970. Article (CrossRef Link)
[19] Gilbert Walter, "Why genes in pieces," *Nature*, vol. 271, no. 5645, pp. 501-501, 1978. Article (CrossRef Link)
[20] Bernard Ng, Fan Yang, David P. Huston, Yan Yan, Yu Yang, Zeyu Xiong, Leif E. Peterson, Hong Wang and Xiao-Fen Yang, "Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes," *Journal of Allergy and Clinical Immunology*, vol. 155, no. 6, pp. 1463-1470, December 2004. Article (CrossRef Link)
[21] Francis Crick, "On Protein Synthesis," *Symp. Soc. Exp. Biol*, vol. 12, pp. 138-163, 1958. Available: http://profiles.nlm.nih.gov/ps/access/SCBBZY.pdf. Access date: Aug. 2012.
[22] Douglas L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annual Reviews of Biochemistry*, vol. 72, no. 1, pp. 291-336, 2003. Article (CrossRef Link)
[23] http://www.cs.waikato.ac.nz/ml/weka/

[24] http://genome.ucsc.edu/
[25] http://en.wikipedia.org/wiki/Drosophila_melanogaster

Beunguk Ahn is an undergraduate student in the Dept. of Computer Science at KAIST. From 2008 to 2011, he served as a TA for Freshman Design Course (FDC) and as a consultant for developing and managing the Moodle e-learning system at KAIST. For these activities, he received an "enhancing education" award in 2010 and a "dedicated service to FDC" award in 2011. He won the first prize from ICWSM-11 Data Challenge by analyzing SNS data for finding sentiment movement on Egypt and Tunisia revolution. He published a paper at 2012 Annual Conference on American Society for Engineering Education for developing large scale grading system. His research interests include web content, database, and data mining.

Elbashir Abbas is currently a master candidate in the Dept. of Computer Science at KAIST. In 2009, he received a BS in Computer Science from the University of Medical Sciences and Technology, Khartoum, Sudan. From 2009 to 2010, he briefly worked as a sales manager and business consultant for SPC/GNOS, a software based solutions company, Suwon, Korea. His current research interests include data mining, machine learning, software engineering, information security and biomedical informatics.

Jin-Ah Park is currently a master candidate in the Dept. of Computer Science at KAIST, Daejeon, Korea. In 2007, she received a BS in Computer Science and Engineering from Hanyang University, Seoul, Korea. Her research interests include social network systems, big data analysis, data mining and database design.

Ho-Jin Choi is currently an associate professor in the Dept. of Computer Science at KAIST. In 1982, he received a BS in Computer Engineering from Seoul National University, Korea, in 1985, an MSc in Computing Software and Systems Design from Newcastle University, UK, and in 1995, a PhD in Artificial Intelligence from Imperial College, London, UK. From 1982 to 1989, he worked for DACOM, Korea, and between 1995 and 1996, worked as a post-doctoral researcher at Imperial College. From 1997 to 2002, he served as a faculty member at Korea Aerospace University, Korea, then from 2002 to 2009 at Information and Communications University (ICU), Korea, and since 2009 he has been with the Dept. of Computer Science at KAIST. Between 2002 and 2003, he visited Carnegie Mellon University (CMU), Pittsburgh, USA, and has been serving as an adjunct professor of CMU for the program of Master of Software Engineering (MSE). Between 2006 and 2008, he served as the Director of Institute for IT Gifted Youth at ICU. Since 2010, he has been participating in the Systems Biomedical Informatics National Core Research Center at the Medical School of Seoul National University. Currently, he serves as a member of the boards of directors for the Software Engineering Society of Korea, for the Computational Intelligence Society of Korea, and for Korean Society of Medical Informatics. His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics.