# Boundary Exon Prediction Using External Information

**Homesh Kaplesh[1], Aman Kaushik[2], Neelam Goel[3]**

1 Homesh kaplesh, CSE/IT department BUEST, Baddi(HP)

2 Aman kaushik, CSE/IT department BUEST, Baddi(HP)

3 Neelam Goel, CSE department, PEC, Chandigarh

## ABSTRACT

***With numerous of genomes sequenced, gene prediction has become a challenging problem in bioinformatics. Gene prediction helps in identifying physical and mental features of different organisms. A large number of gene prediction tools have been developed in the past two decades. External information plays an important role in gene prediction. With the help of this information approximately more than half of genes can be predicted. Moreover it helps in improving the accuracy of ab-initio gene prediction methods. In this paper, a method for predicting boundary exons using external information is proposed. The proposed method uses grammatical rules to utilize the external evidences obtained from multiple alignment of cDNA, EST and Protein sequences.The experimental results show that the application of this new approach helps in the prediction of boundary exons. The method predicted most of the exons along with boundary exons.***

**Keywords:-** Boundary exon finder, bioinformatics, DNA

## 1.INTRODUCTION

Since the end of the Human Genome Project in 2003, a human DNA sequence has been determined to nearly 100%. Nowadays, we have also many sequences from different organisms. These sequences are very long strings over the alphabet {A; C; T; G} and to use them effectively in research, further analysis is required to and parts of the sequence called genes which are of major importance. Genes are protein coding parts of a sequence, but they are not very easy to distinguish from surrounding non coding sequence. Due to a vast size of DNA sequences, there is a need for an automated solution of this problem, since a manual gene finding would be very time consuming. Many gene finding programs have arisen to cope with this task using only genomic sequences; such an approach is called ab-initio from the beginning. Soon, they were expanded with the use of external information, and the gene prediction accuracy took a huge step forward. The external information gained from a variety of sources is typically in a preprocessing stage cut to simple intervals (or points), because gene finders cannot process more complex types of external information. In this paper, we propose a method able to predict boundary exons using external information . The method is based on the same principle as existing gene finders, but it uses a different approach in the processing of the external information. In particular, we first devised a solution to simple problem of finding hints. These hints are further processed based on gene structure prediction rules to generate the indices of predicted exons. Additionally, we experimentally evaluate the contribution of complex hints compared to simple interval hints and point wise hints. We built a simple gene finder based on our algorithm, and created several hint sets based on both real and artificial data. The rest of the paper is organized as follows. In section 2 the proposed method is described. After this in 3 section the results are discussed. Finally the conclusion of the work is provided.

## 2.PROPOSED METHOD

The implementation of proposed methodology runs in different modules. Each module is of equal importance. The modules are interdependent i.e. the result of one modules goes as input for the processing of next module. The three modules of Bondary exon detection model are shown in Figure 1.

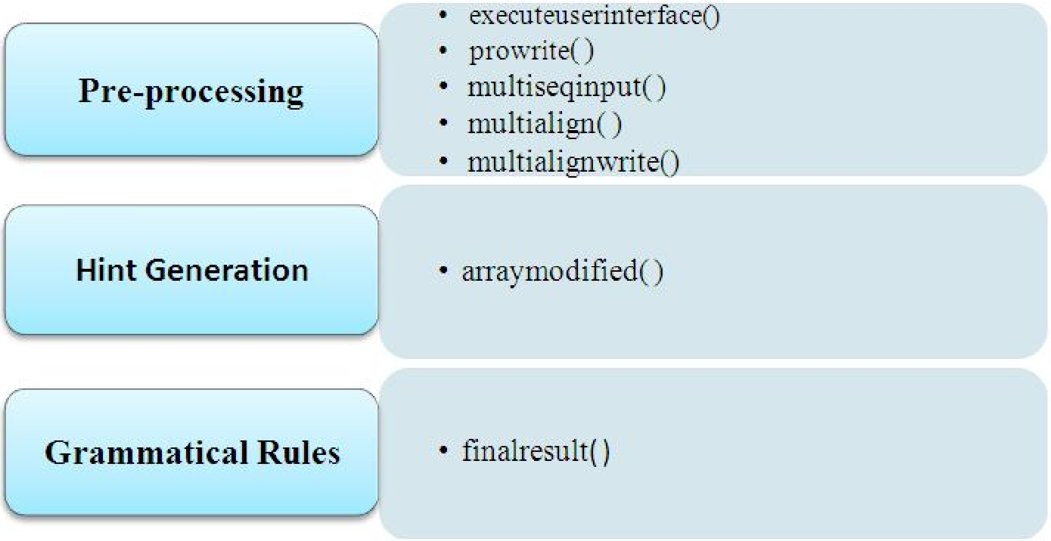


**Figure 1** Modules of Bondary Exon model

**These modules are discussed below**:

### 2.1 PRE-PROCESSING MODULE

In this module the initial processing take place with the help of following functions:

**executeuserinterface ( ):** This is the main function of implementation module. This function is used to take input from user. The user will provide DNA,cDNA ,EST and Protein sequences through graphical user interface in fasta format. The input is then read into different variables.

**prowrite ():** The input received from the executeuserinterface() is in nucleotide format. But the protein sequence is in amino acid format. To convert this amino acid sequence into nucleotide prowrite() is used. In this mat lab inbuilt aa2nt is used. It also writes the converted sequence into a file.

**multiseqinput( ):**This function takes as input the Dna ,cDNA,Est and converted Protein sequence. It writes all these sequences into a file which is further given as input to mutialign().

**multialign( ):**This is mat lab function .It is used to align multiple sequences either in nucleotide format or in amino acid format. Here sequences are aligned in nucleotide format.

**multialignwrite( ):**This function is used to write the alignment obtained from multialign function to a file.

### 2.2 HINT GENERATION MODULE

This module includes the following function:

**arraymodified( )**: This function process the multiple alignment file .It step by step processes the alignment. It first processes the Dna and cDna alignment and then generates a matrix, which is known as cDNA Hint matrix. Similarly it processes the EST and Protein alignment and then generates the EST and protein hint matrices. These matrices are further refined based on rules.

### 2.3 GRAMMATICAL RULES MODULE

This module has following function:

**finalresult ( )**: This function takes as input the hint matrices generated by the previous module. It further apply the splicing and gene structure prediction rules to get the final output which consists of predicted exons.

## 3.RESULTS AND DISCUSSION

### 3.1 Dataset

To evaluate the performance of the proposed system a dataset of 18 sequences is prepared. The Dataset contain individual files for DNA ,cDNA ,EST and Protein sequences. To prepare the dataset first the Human DNA sequences are collected from the NCBI website. Only those sequences that contains boundary exons are selected for this purpose. After this Blastn is used to collect the cDNA and EST sequences with specific database selection that is nucleotide database is selected in case of cDNA and EST database is selected is selected in case of EST. Blastx is used to collect the protein sequences with the selection of amino acid database.

### 3.2Results

An interactive user friendly interface is created, so that it can be easily used by any one. It performs all the functioning by clicking on the buttons. The GUI is named as Boundary Exon Finder: Boundary Exon Detection Tool for human DNA sequences. This GUI has been designed for testing of proposed method. This front end allows the user to enter the DNA, cDNA, EST and Protein sequences and gives the indices of boundary exons as output. It gives us the Start and End indexes of the boundary exons . To test the performance of the proposed method a set of Dna , cDNA,EST and Protein sequences is given as input to the system. The system will processes the input sequences and then generates the resultant matrix, which consists of start and end indices of predicted exons. The indexes produced are then analyzed to see the accuracy of the proposed system. For this purpose the actual annotation of the gene is compared with the results produced. It results in true positives and false positives. As the focus of this work is to predict the boundary exons , In most of the cases it correctly predicts the initial exons. The accuracy of predicting terminal exons is not as good as of initial exons. Figure 2 shows the main window of GUI. Initially Four text boxes are cleared. The user can submit the sequence in by pasting the sequence in the text box.

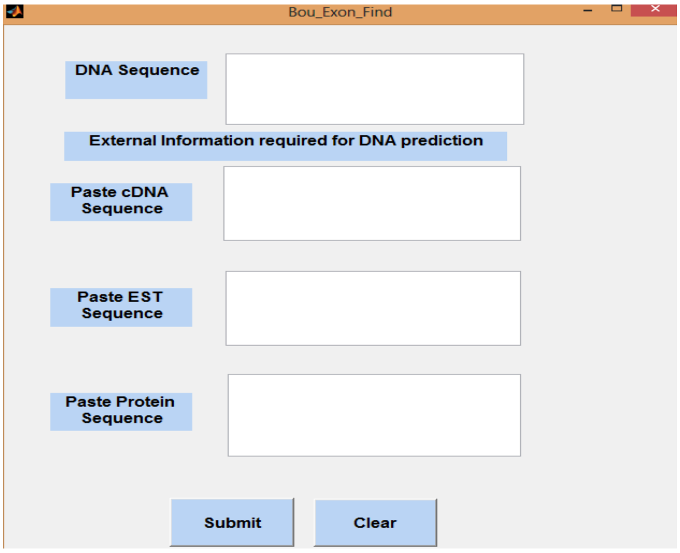
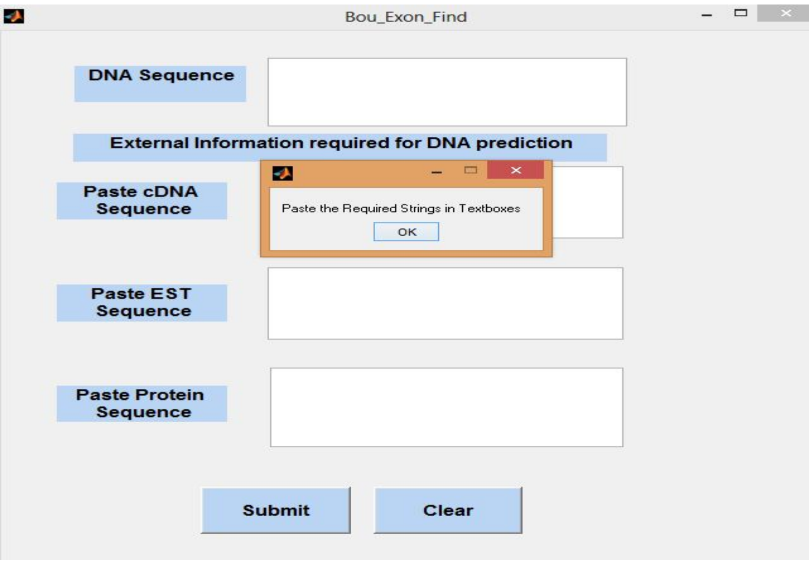


**Figure 2** BoundaryExonFinder Main window

Figure 2 shows the main window of GUI. Initially Four text boxes are cleared. The user can submit the sequence in by pasting the sequence in the text box.After pasting the sequences into corresponding textboxes click on Submit button.

## Discussion

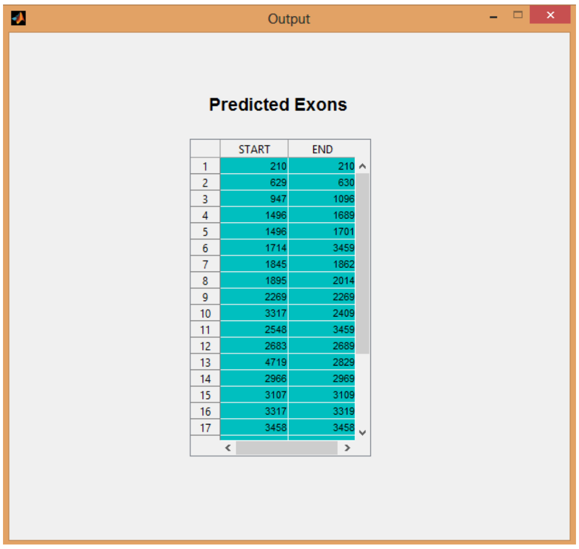
After pasting the sequences into corresponding textboxes click on Submit button.When the user clicks the Submit button the grammatical rules will be applied to the input sequences and find the resultant data.



Paste the Required Strings in Textboxes

**Figure 3** BoundaryExonFinder Alert window

Figure 3 shows the alert window of GUI.This window is basically used for validation purpose.If any user clicks on submit button without fill the required data in the textboxes then the alert message will be display on the screen. This is very important screen because it validates the user for required data. Without the input of this data it would not be possible for this tool to work.



Predicted Exons

|  | START | END |
|---|---|---|
| 1 | 210 | 210 |
| 2 | 629 | 630 |
| 3 | 947 | 1096 |
| 4 | 1496 | 1689 |
| 5 | 1496 | 1701 |
| 6 | 1714 | 3459 |
| 7 | 1845 | 1862 |
| 8 | 1895 | 2014 |
| 9 | 2269 | 2269 |
| 10 | 3317 | 2409 |
| 11 | 2548 | 3459 |
| 12 | 2683 | 2689 |
| 13 | 4719 | 2829 |
| 14 | 2966 | 2969 |
| 15 | 3107 | 3109 |
| 16 | 3317 | 3319 |
| 17 | 3458 | 3458 |

**Figure 3** BoundaryExonFinder Output window

Figure 3 shows the output window of GUI. This screen displays the boundary exon indexes of the DNA. When user enters the required strings in the textboxes and click on Submit button then according to the grammatical rules and the logics applied to the input sequences. The result will finally displayed in this screen.

## 4.CONCLUSION

In this paper, an external based boundary exon prediction method is proposed. The method uses EST,cDNA and Protein homologous of the DNA sequence. These sequences are then aligned to generate the hints. The grammar is then applied to these hints to produce the indexes of the predicted exons. The homologous sequences are searched using the blast tool. For this purpose blastn and blastx are used. From the output produced by blast tool only topmost sequences are selected. From the implementation of proposed method it is seen that blast is an efficient tool for searching homologous sequences. Among all theses homologous sequences cDNA and protein plays significant role in boundary exon prediction. The proposed method is a novel attempt in the direction of predicting boundary exons.

**FUTURE SCOPE**

This thesis is only a start of what can be done with external information for predicting boundary exons. The following list mentions some ideas that can be further examined.

- Forward and backward scanning rules further can be applied to improve the prediction results.
- The method can be combined with other signal prediction methods to enhance its accuracy.
- Different hint generation approaches can be tried to improve the results.

## REFERENCES

[1] N. Goel, S. Singh, and T. C. Aseri, “A Review of Soft Computing Techniques for Gene Prediction,” ISRN Genomics, vol. 2013, pp. 1–8, 2013.

[2] C. Burge and S. Karlin, “Prediction of Complete Gene Structures in Human Genomic DNA,” J. Mol. Biol., vol. 268, pp. 78–94, 1997.

[3] M. Pertea, X. Lin, and S. L. Salzberg, “GeneSplicer□: A New Computational Method for Splice Site Prediction,” Nucleic Acids Res., vol. 29, no. 5, pp. 1185–1190, 2001.

[4] M. M. Yin and J. T. L. Wang, “E□ffective Hidden Markov Models for Detecting Splicing Junction Sites in DNA Sequences,” Inf. Sci. (Ny)., vol. 139, pp. 139–163, 2001.

[5] S. Brunak, J. Engelbrecht, and S. Knudsen, “Prediction of Human mRNA Donor and Acceptor Sites from the DNA Sequence,” J. Mol. Biol., vol. 220, pp. 49–65, 1991.

[6] N. Tolstrup, P. Rouzé, and S. Brunak, “A Branch Point Consensus from Arabidopsis Found by Non-circular Analysis Allows for Better Prediction of Acceptor Sites,” Nucleic Acids Res., vol. 25, no. 15, pp. 3159–3163, 1997.

[7] A. Hatzigeorgiou, N. Mache, and M. Reczko, “Functional Site Prediction on the DNA Sequence by Artificial Neural Networks,” in IEEE International Joint Symposia on Intelligence and Systems, 1996, pp. 12–17.

[8] M. G. Reese, F. I. I. Eeckman, D. Kulp, and D. Haussler, “Improved Splice Site Detection in Genie,” in First annual International conference on Computational Molecular biology (RECOMB), 1997, pp. 232–240.

[9] L. S. Ho and J. C. Rajapakse, “Splice Site Detection with a Higher-Order Markov Model of Splice Sites,” Genome Informatics, vol. 14, pp. 64–72, 2003.

[10] J. C. Rajapakse and L. S. Ho, “Markov Encoding for Detecting Signals in Genomic Sequences,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 2, no. 2, pp. 131–142, 2005.

[11] T. Cai and Q. Peng, "Predicting the Splice Sites in DNA Sequences Using Neural Network Based on Complementary Encoding Method," in Proceedings of International conference on Neural Networks and Brain, 2005, vol. 146, pp. 473–476.

[12] L. Liu, Y. Ho, and S. Yau, "Prediction of Primate Splice Site Using Inhomogeneous Markov Chain and Neural Network," DNA Cell Biol., vol. 26, no. 7, pp. 477–483, 2007.

[13] Ø. Johansen, T. Ryen, T. Eftesøl, T. Kjosmoen, and P. Ruoff, "Splice Site Prediction using Artificial Neural Networks," in CIBB, 2009, pp. 102–113.

[14] N. Goel, S. Singh, and T. C. Aseri, "A comparative analysis of soft computing techniques for gene prediction," Anal. Bochemistry, vol. 438, no. 1, pp. 14–21, 2013.

[15] T. Nassa, S. Singh, and N. Goel, "Splice Site Detection in DNA Sequences using Probabilistic Neural Network," Int. J. Comput. Appl., vol. 76, no. 4, pp. 1–4, 2013.

[16] X. H.-F. Zhang, K. A. Heller, I. Hefter, C. S. Leslie, and L. A. Chasin, "Sequence information for the splicing of human pre-mRNA identified by support vector machine classification.," Genome Res., vol. 13, no. 12, pp. 2637–2650, Dec. 2003.

[17] Y.-F. Sun, X.-D. Fan, and Y.-D. Li, "Identifying splicing sites in eukaryotic RNA: support vector machine approach," Comput. Biol. Med., vol. 33, no. 1, pp. 17–29, Jan. 2003.

[18] Y. Zhang, C.-H. Chu, Y. Chen, H. Zha, and X. Ji, "Splice Site Prediction using Support Vector Machines with a Bayes kernel," Expert Syst. Appl., vol. 30, pp. 73–81, 2006.

[19] J. Huang, T. Li, K. Chen, and J. Wu, "An Approach of Encoding for Prediction of Splice Sites using SVM," Biochimie, vol. 88, pp. 923–929, 2006.

[20] A. Baten, B. C. H. Chang, S. K. Halgamuge, and J. Li, "Splice Site Identification using Probabilistic Parameters and SVM Classification," BMC Bioinformatics, vol. 7(Suppl 5), pp. 1–15, 2006.

[21] U. Kamath, J. Compton, R. Islamaj-do, K. A. De Jong, and A. Shehu, "An Evolutionary Algorithm Approach for Feature Generation from Sequence Data and Its Application to DNA Splice Site Prediction," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 9, no. 5, pp. 1387–1398, 2012.

[22] A. Y. Kashiwabara, D. C. G. Vieira, A. Machado-Lima, and A. M. Durham, "Splice Site Prediction using Stochastic Regular Grammars," Genet. Mol. Res., vol. 6, no. 1, pp. 105–115, 2007.

[23] M. Q. Zhang, "Identification of Protein Coding Regions in the Human Genome by Quadratic Discriminant Analysis," Proc. Natl. Acad. Sci., vol. 94, pp. 565–568, 1997.

[24] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[25] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Min. Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1997.

[26] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," Bioinformatics, 2003. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. [Accessed: 15-Apr-2010].

[27] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," PLoS Comput. Biol., vol. 4, no. 10, p. e1000173, Oct. 2008.

[28] M. Stanke and S. Waack, "Gene prediction with a hidden Markov model and a new intron submodel," Bioinformatics, vol. 19, no. Suppl 2, pp. ii215–ii225, Oct. 2003.

[29] S. Rogic, "HMR195 dataset." [Online]. Available: http://srogic.wordpress.com/datasets/hmr195-dataset/.

[30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, Apr. 2011.

[31] Marcel Kucharık, Jakub Kovac, and Brona Brejova, "Gene finding with complex external information" Mlynska Dolina, 842.