# EVALUATION OF EUKARYOTIC GENE PREDICTION PROGRAMMS

**DIVYA SINGHAL*, POOJA SHARMA, MOKSHA SHANDILYA**
**Deptt. of Biotechnology, Ambala College of Engg. & Applied Research, Ambala**

## ABSTRACT

Gene finding typically refers to the area of computational biology that is concern with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional. This specially includes protein coding genes but may also include other functional elements such as RNA gene and regulatory regions. Gene finding is one of the first and most steps in understanding the genome of specie once it has been sequenced. Gene prediction software's are bioinformatics tools to predict the gene structure of a given sequence in Fasta format. Gene prediction involves determining the number and location of exons (initial, intermediate or terminal), number and location of introns, CDS region, location of promoter and terminal regions (PolyA). In this study, various windows based online gene prediction software's were compared against genebank sequences for 5 different sequences. Softwares used were: HMMgene, EMBOSS, FGENESH, GENMARK and GENSCAN. Results were analyzed by calculating specificity and sensitivity at nucleotide and exon level. Correlation coefficient, average conditional probability, approximate correlations were calculated and compared to determine most efficient software for use. FgeneSH software found to be best eukaryotic gene prediction software.

**Keywords:** gene prediction, eukaryotes, software, exons, introns

## INTRODUCTION

The gene structure and the gene expression mechanism of eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of exons and introns. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called splicing takes place, in which, the intron sequences are excised and discarded from the RNA sequence. The remaining RNA segments, the ones corresponding to the exons are ligated to form the mature RNA strand. A typical multi-exon gene has the following structure. It starts with the promoter region, which is followed by a transcribed but non-coding region called 5' untranslated region (5' UTR), then follows the initial exon which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon. It is followed by another non-coding region called the 3' UTR. Ending the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signaled by specific short (2bp long) sequences. The 5'(3') end of an intron (exon) is called the donor site, and the 3'(5') end of an intron (exon) is called the acceptor site. The problem of gene identification is complicated in the case of eukaryotes by the vast variation that is found in gene structure.[1]

Many gene prediction programs are currently publicly available. Most of them are referenced in the Web site maintained by W. Li (http://linkage.rockefeller.edu/wli/gene/).[2] Gene prediction softwares for eukaryotes are Genscan,Grail,Genmark,FgeneSH,HMMgene,Emboss etc.[3,8-10]

## MATERIAL & METHODOLOGY

Sequences used for eukaryotic gene prediction are downloaded in fasta format from genebank database of NCBI. Eight different sequences of β-globin different gene of *Homo sapiens* were used for eukaryotic gene prediction. Softwares used for eukaryotic gene prediction are Genmark[4], Genscan[5], FgeneSH[6], HMMgene[7], and Emboss. Number of exons and their location predicted by all of these softwares were compared with their original genebank sequences. On the basis of comparison,

***Corresponding author:***
***Email: divysinghal@gmail.com***

**TABLE: RESULTS OF ALL EUKARYOTIC GENE PREDICTION SOFTWARES AT EXON LEVEL AS WELL AS NUCLEOTIDE LEVEL.**

| EXON LEVEL | | | | |
|---|---|---|---|---|
| **SOFTWARE** | **SN** | **SP** | **ME** | **WE** |
| Genmark | .17 | .17 | .22 | .17 |
| FgeneSH | .28 | .28 | 0.11 | .06 |
| HMMgene | .28 | .28 | 0.11 | .08 |
| Genscan | 0 | 0 | 0.19 | .11 |
| EMBOSS | .28 | .28 | 0 | .11 |

| NUCLEOTIDE LEVEL | | | | | |
|---|---|---|---|---|---|
| **SOFTWARE** | **SN** | **SP** | **CC** | **ACP** | **AC** |
| Genmark | .71 | .93 | .73 | .86 | .65 |
| FgeneSH | .72 | .99 | .69 | .90 | .80 |
| HMMgene | .68 | 1 | .76 | .88 | .77 |
| Genscan | .47 | .60 | .78 | .69 | .39 |
| EMBOSS | .68 | .99 | .75 | .87 | .75 |

SN= Sensitivity SP= Specificity ME= Missing Exons

WE= Wrong Exons CC=Correlation Coefficient AC= Accuracy

different statistics values are calculated like True-Positive (TP), False-Positive (FP), True-Negative (TN) and False-Negative (FN). Here, True positive indicates gene evaluated as genes, False positive indicates non genes evaluated as gene, False negative indicates gene evaluated as non gene, true negative indicates non genes evaluated as non genes. Results of all of these softwares are compared with each other at 2levels, exon level as well as nucleotide level. For estimation of softwares accuracy at exon level, sensitivity, specificity, missing exon and wrong exons are calculated by using, following formulas[3, 8-10]:

Actual Positive (AP) = True Positive (TP) + False Negative (FN)

Actual Negative (AN) =False Positive (FP) + True Negative (TN)

Predicted num of positive (PP) = True Positive (TP) + False Positive (FP)

Predicted num of negative (PN) = True Negative (TN) + False Negative (FN).

Sensitivity **(Sn) =** No. of Correct Exons (TP) /No. of Actual Exons (AP)

Specificity **(Sp) =** No. of Correct Exons (TP) /No. of Predicted Exons (PP)

ME = No. of Missing Exons / No. of Actual Exons

WE = No. of Wrong Exons / No. of Predicted Exons

Here, Sensitivity (Sn) is the proportion of coding nucleotides that have been correctly predicted as coding. Specificity (Sp) is the proportion of non coding nucleotides that have been correctly predicted as non-coding.

For estimation of softwares accuracy at nucleotide level, correlation coefficient and approximate correlation and then finally accuracy are calculated by using, following formulas[3, 8-10]:

Correlation Coefficient (CC) = (TP*TN)-(FN*FP)/ √ ((TP+FN)*(TN+FP)*(TP+FP)*(TN+FN))

Approximate Correlation (ACP) = ¼(TP/ (TP+FN) +TP/ (TP+FP) +TN/ (TN+FP) +TN/ (TN+FN))

Accuracy (AC) = (ACP-0.5)*2

## RESULTS & DISCUSSION:

Results of all of these gene prediction softwares both at exon level as well as nucleotide level are shown in Table 1.At exon level Sensitivity and Specificity of FgeneSH, HMMgene & Emboss is better as compared to Genmark & Genscan. In case of Missing exons and wrong exons, Emboss and FgeneSH are respectively predicted as best softwares. At Nucleotide level, sensitivity of FgeneSH is highest, in case of specificity HMMgene is highest but FgeneSH and Emboss have shown almost similar specificity to HMMgene. Genscan has shown highest correlation coefficiency, whereas FgeneSH has shown best approx. correlation as well as best accuracy. On the basis of all of these results, FgeneSH is found to be best software among Genscan,Genmark,HMMgene and Emboss.

## REFERENCES:

1. Haussler, D. et al. (1998). Computational gene finding. Trends Biochem. Sci. Suppl. Guide Bioinformatics, pp. 1215.

2. MatheÂ C, Sagot MF, Schiex T, Rouze P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 30(19): 4103-4117.

3. Do JH and Choi DK. (2006).Computaional approaches to gene prediction. J Microbiol. 44(2):137-44.

4. Borodovsky MY and McIninch JD. (1993). Genmark: Parallel gene recognition for both DNA strands. Comput. Chem. 17: 123–133.

5. Burge C and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

6. Salamov AA and Solovyev VV. (2000). *Ab initio* gene finding in Drosophila genomic DNA. Genome Res. 10: 391-393.

7. Krogh A. (2000). Using database matches with HMMgene for automated gene detection in *Drosophila*. Genome Res. 10: 523-528.

8. Singh GB and Krawetz SA. (1994). Computer based exon detection: An evaluation metric for comparison. Int. J. Genome Res. 1: 321–338.

9. Xu Y, Mural RJ and Uberbacher EC (1994). Constructing gene models from accurately predicted exons: An application of dynamic programming. Comput. Appl. Biosci. 10: 613–623.

10. Makarov V, Boulevard C and Pasadena, CA(2002). Computer programs for eukaryotic gene prediction. Briefing in Bioinformatics. 3(2): 195–199.