Comparative Analysis of Gene Prediction Tools: RAST, Genmark hmm and AMIgene

Chander Jyoti*[1], Sandeep Saini [2], Varinder Kumar[3], Kajal Abrol[4], Kanchan Pandey[5], Ankit Sharma[6]

[1-3] *Assistant Professors,* [4-6]*Undergraduates students, Department of Bioinformatics,*

GGDSD College Sector 32C, Chandigarh, INDIA.

**corresponding author*

ABSTRACT

High throughput genome sequencing made large amount of genome data available to research community. Accurate gene structure prediction and annotation is the fundamental step towards the understanding of genome function. A large number of gene prediction tool and pipeline have been developed over the past year. To understand whether the prediction tools and pipeline are providing same or different result for the same genome or not, we have compared manually the gene prediction result of RAST (Rapid Annotations using Subsystems Technology), AMIgene (Annotation of MIcrobial Genes) and Genmark hmm for organism Mycoplasm genitalium in reference to Genbank CDS (Coding Sequence) or gene. During comparative analysis we have seen the similarity as well as variation in prediction result of each tool. Variation in prediction results were also seen in total number of CDS predicted, gene coordinate and gene length. We have tried to find the reason behind the variation in prediction result and try to relate our analysis with nowadays high throughput data analysis. These types of analysis are useful to annotate a newly sequenced genome.

Key Words: *Gene Prediction, CDS, Annotation, Mycoplasm genitalium.*

I. INTRODUCTION

Human genome sequencing leads the beginning of a new era in the field of genomics. Large number of plants, animal, microbes and viruses genomes has been sequenced till date[1]. Genome annotation is most important task after sequencing any genome, which starts with prediction of protein coding genes[2]. Gene discovery by manual annotation uses the experience of expert individual who attains high degree of accuracy[3]. But the manual gene prediction is not able to keep pace with high throughput sequencing technology[4]. To solve this problem, a number of automatic annotation pipelines and gene prediction softwares were developed by bioinformatician to support and mimic the manual curation process [5,6]. These automatic annotation pipelines and gene prediction softwares rely on intrinsic (ab initio) and extrinsic (homology-based) methods of gene identification [7].

Ab initio based programs of gene identification look for the gene signals (start codon, stop codon, promoter site, ribosomal binding site etc) and gene contents (hexamer frequency, codon biased etc) to find coding region[8]. On other hand homology or sequence similarity based programs search the databases for finding homologues. Homology search based on assumption that coding region are more conserved than noncoding region. If we find any

similarity between certain genomic region regarding the gene or protein, the same information can be used to infer the function and structure of that region. But this method constrained if no statistically significant match is found in the database[9].

As most of the pathogens are prokaryotic, the accurate gene prediction among them would play a significant role in finding new therapeutic targets[10]. Being small in size, with greater gene density and 90 % coding region desired high degree of gene prediction accuracy in prokaryotes. Gene finding in prokaryotic start with searching for open reading frames (ORF), a continuous stretch of DNA having a start codon and ends with a stop codon usually after a distance of fifty codons. Most of the ab initio programs scan the genomic sequence for locating ORFs but still their results shows significant variation in prediction data[11].

A large number of gene identification methods and tools are available however existing gene prediction show variation in results[12]. Here in this study, we have done the comparative analysis of *Mycoplasm genitalium* coding sequences (CDS) by using the three *ab initio* based prokaryotic gene prediction tools named RAST (Rapid Annotations using Subsystems Technology), Genmark hmm and AMIgene (Annotation of MIcrobial Genes). Genbank CDS annotation of *Mycoplasm genitalium* genome which is done by NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) also considered during analysis.

In prokaryotes CDS and ORF terms can be used interchangeably because there are no introns in them and all the segment of ORF obtained can code for the protein. In this study we used CDS term.

II. METHODOLOGY

A. Selection of Organism

For comparative analysis we have selected the genome of *Mycoplasma genitalium* G37, the smallest genome (580076 bp) of free living organism with high gene density (one gene every 1042 base pair)[13]. This genome is training set for all comparative tools and pipeline under study. Hence it is considered as important system for exploring total number of genes for our studies

B. Genbank

Complete genome sequence of Mycoplasma genitalium G37 is given in Genbank database (accession number NC_000908) which is annotated by NCBI pipeline named PGAP. PGAP predict protein coding genes and also predict noncoding RNA, repeats, mobile elements and pseudogenes[14].But our analysis focused only on protein coding region.

C. RAST

RAST is a fully automated tool for archaeal and bacterial genome[15, 16]. RAST pipeline predicts open reading frame by GLIMMER 3, an ab initio tool for microbial gene prediction[17]. FASTA sequence of given accession number was downloaded from GenBank and analyzed by RAST server.

D. AMIgene

AMIgene tool was used for analysis of same genome. AMIgene looks for the maximum segment in frame between start and stop codon and retained the CDSs greater than sixty base pair. It identifies the most likely coding sequences (CDSs) in a large contig or a complete bacterial genome sequence[18]. Gene model parameter in AMIgene was selected mycoplasm genitalium and genetic code parameter was changed according to mycoplasm i,e. TGA for tryptophan. Rests of the parameters were kept by default.

E. Genmark hmm

Genmark hmm, a member of Genmark family is ab initio gene finder program which include DNA strand, boundaries, length and class of the gene. Gene class use typical and atypical gene model of Markov Chain[19, 20]. Select species parameter was changed to Mycoplasma_genitalium_G37 and rests were by default.

III. RESULT AND DISCUSSION

After analysing the predictions results, we have found significant differences in number of CDS, start codon used and CDS length. Highest 536, CDS were predicted by Genmark hmm, RAST and AMIgene predicted 529 and 511 CDS respectively whereas Genbank has least number, 507 as shown in figure 1.

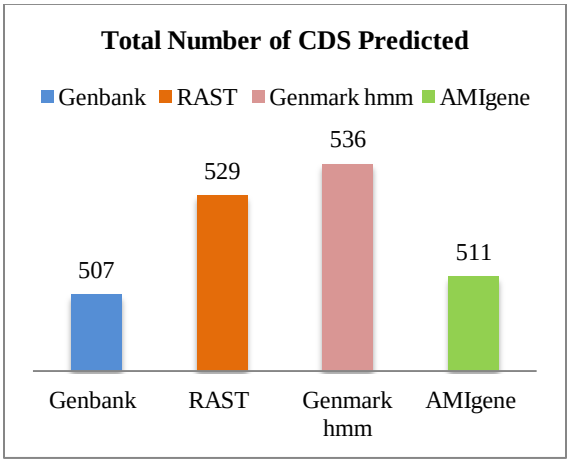

Figure 1: Total number of CDS Predicted

Next we manually analyzed the difference in the start codon called by each prediction.

Each tool preferred ATG as a start codon but other alternative start codons such as TTG, GTG were also called. RAST called ATG in 468 predictions. Genmark hmm and Genbank called ATG in 452 and 433 respectively whereas least 385 were used by AMIgene. The frequency of alternate start codon alongwith ATG is shown in figure 2.

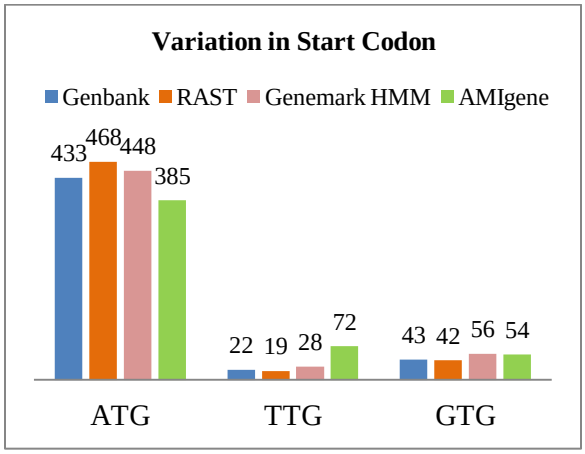

Figure 2: Frequency of different start codon called by each tool

Rather than TTG and GTG, Genbank also used CTG (3), ATA (2), ATC (2), TAC (1) and TTA (1) as alternate start codon which are absent in rest.By checking each start codon manually we have found that AMIgene prefers alternate start codon which lies upstream with respect to ATG.

Whenever GTG or TTG are upstream to ATG, AMIgene prefer either of them rather than ATG. While other preferred ATG irrespective to TTG or GTG. Due to preference of alternate start codon AMIgene generally has longest reading frame as compare to rest of tools.Each tool has predicted 379 common CDS i.e. there are same start codon and stop codon and has same gene length. Rests of the CDS are shared either by combination of three or two tools, shown in figure 3. There were some unique CDS given by single tool. Genbank, RAST and Genmark hmm predicted same CDS 59 instances, highest from all other combination.

Genbank, RAST and AMIgene predicted same CDS in 12 instances, RAST, AMIgene Genmark hmm

predicted same CDS in 19 instances. AMIgene, Genmark hmm and Genbank predicted same CDS in 7 instances.

Genbank and RAST, Genbank and Genmark hmm, Genbank and AMIgene called same CDS in 13, 12 and 6 instances.

RAST and AMIgene, RAST and Genmark hmm, AMIgene and Genmark hmm predicted same CDS in 11, 13 and 12 instances each. There were 65 instances where AMIgene predicted either different start codon or new gene as compare to rest three predictions.

In 35 instances Genmark hmm predicted unique CDS as compare to rest three predictions. Genbank and RAST predicted 19 and 23 CDS with unique start codon

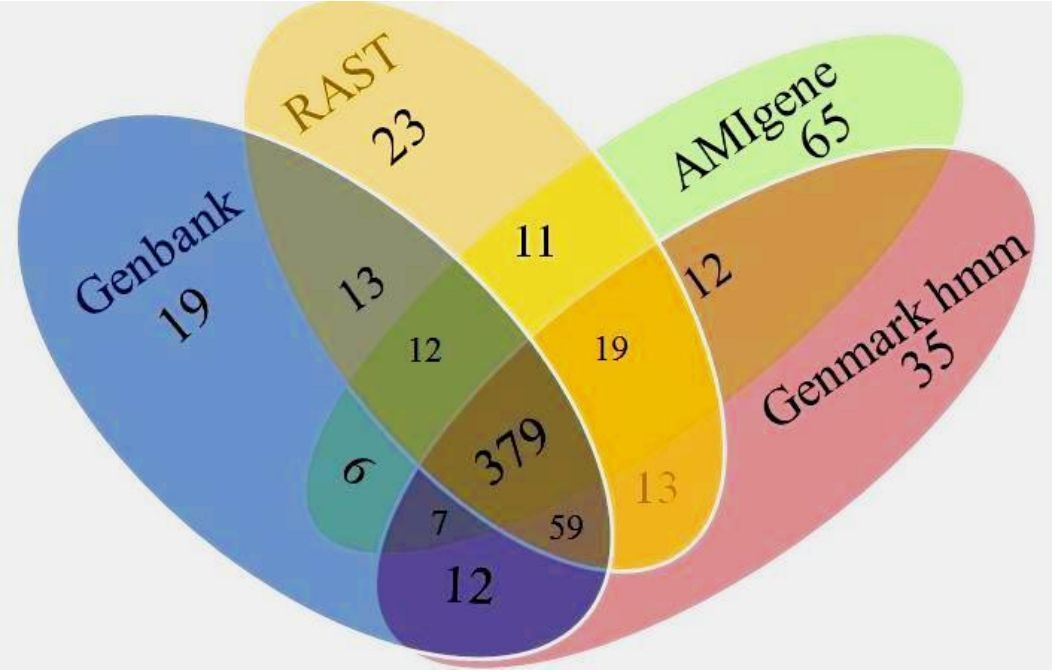

Figure 3: Vein Diagram show the variation in Prediction Result of each Tool

V. CONCLUSION

The reason behind selection of smallest free living organism was there must be very little difference in gene prediction results. Though each tool predicted exact same gene coordinate and in some instances similarity was shown by combination of two or three tools. But for smallest free living organism these tools show the variation in gene prediction. Being most important process of analysis, one should compare the data by using alternative tools, because if there is variation in prediction coordinate further feature of that gene would vary. Nowadays high throughput sequencing data are generated and it is not possible to assign function of each gene experimentally. Gene prediction tools or pipeline are used to annotate the high throughput data. To reduce the variation in gene prediction data, there must be more accurate and reliable tools for prediction analysis in future.

REFERENCES

1. Mathe C, Sagot M-F, Schiex T, Rouze P. Survey and Summary: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 2002; 30(19):4103-4117.
2. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59
   Lewis S1, Ashburner M, Reese MG. Annotating eukaryote genomes. Curr Opin Struct Biol. 2000 Jun; 10(3):349-54.
3. Aseri TC. A Review of Soft Computing Techniques for Gene Prediction. ISRN Genomics, (2013), 1–8.
4. Searls DB. Using bioinformatics in gene and drug discovery. Drug Discov Today. 2000 Apr; 5 (4):135-143.
5. Rust AG, Mongin E, Birney E. Genome annotation techniques: new approaches and challenges. Drug Discov Today. 2002 Jun 1;7 (11):S70-6.
6. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research. 2005;33 (Web Server issue):W451-W454. doi:10.1093/nar/gki487
7. Goodswen SJ, Kennedy PJ, Ellis JT (2012) Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. PLoS ONE 7(11): e50609. doi:10.1371/journal.pone.0050609
8. Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. Genomics Proteomics Bioinformatics. 2004 Nov; 2(4):216-21.
9. Beiting DP, Roos DS. A systems biological view of intracellular pathogens. Immunol Rev 2011, 240:117–128.
10. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11:119.
11. Shibuya T, Rigoutsos I. Dictionary-driven prokaryotic gene finding. Nucleic Acids Research. 2002;30(12):2710-2725.
12. Fraser CM et al. The minimal gene complement of Mycoplasma genitalium. Science. 1995 Oct 20; 270: 397-403.
13. Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research. 2016; 44(14):6614-6624. doi:10.1093/nar/gkw569.
14. Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75. doi:10.1186/1471-2164-9-75.
15. Overbeek R, Begley T, Butler RM, et al. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Research. 2005;33(17):5691-5702. doi:10.1093/nar/gki866.
16. Brettin T, Davis JJ, Disz T, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific Reports. 2015;5:8365. doi:10.1038/srep08365.
17. Bocs S, Cruveiller S, Vallenet D, Nuel G, Médigue C. AMIGENEGene: Annotation of MIcrobial Genes. Nucleic Acids Research. 2003;31(13):3723-3726.
18. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Research. 1998;26(4):1107-1115.
19. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research. 2001;29(12):2607-2618.