GENOME ANNOTATION OF *SESAMUM INDICUM*

Fulya Akçimen, Serkan Sayın, Ahmet Bakırbaş, Burcu Fıratlıgil, Canan Has and Jens Allmer

Molecular Biology and Genetics Department, Izmir Institute of Technology, Urla – Izmir, Turkey

Abstract

In the growing field of life sciences, experimentally obtained raw DNA sequences have been increasing dramatically in databases. These sequences should be converted into meaningful information by annotation. Gene annotation includes the description and interpretation of various regions, features on the genome sequence and identification of the regions that could be called genes. Thus, gene annotation of a sequenced genome becomes an increasing demand for providing the bridge from the sequences to the biological meaning of the gene. In this study, annotation of a *Sesamum indicum* contigs was performed by using publicly available databases and webservers for local alignment, gene prediction, protein sub-cellular localization prediction, and protein domain prediction. Thus, by using these tools, genes and their GC contents, exons, start and stop codons, homologous protein information, protein structure, and sub-cellular localization of those homologous proteins and many more features were investigated. As a result, various corresponding genes and proteins to our contigs were identified with high confidence besides; further characterization of the proteins was carried out.

Flowchart

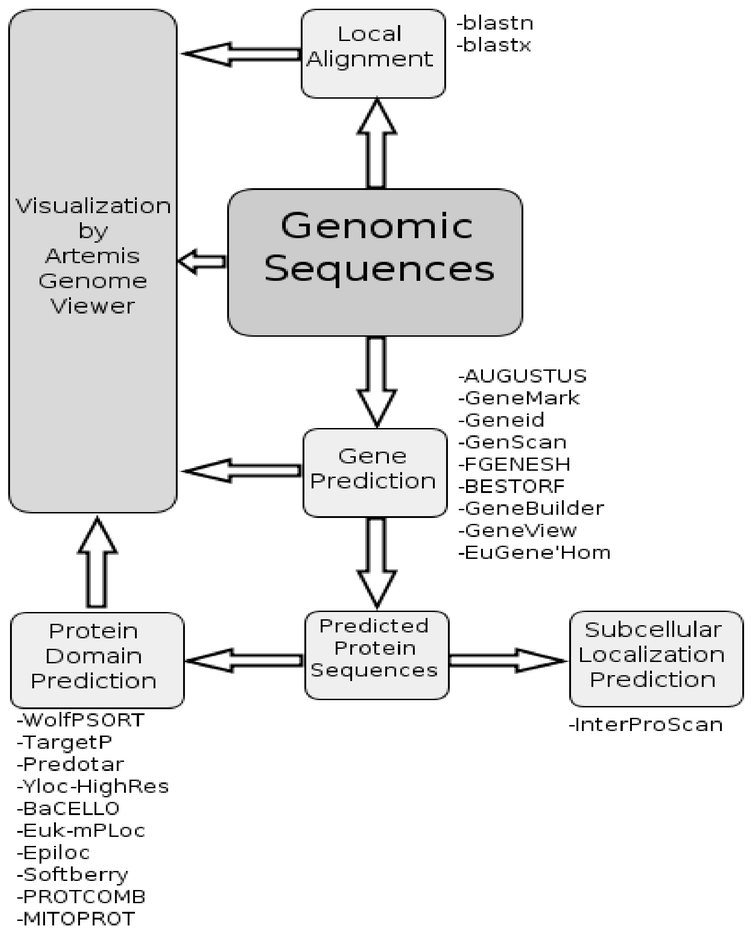


Table 1: Flowchart of the gene annotation with the applied tools

Results

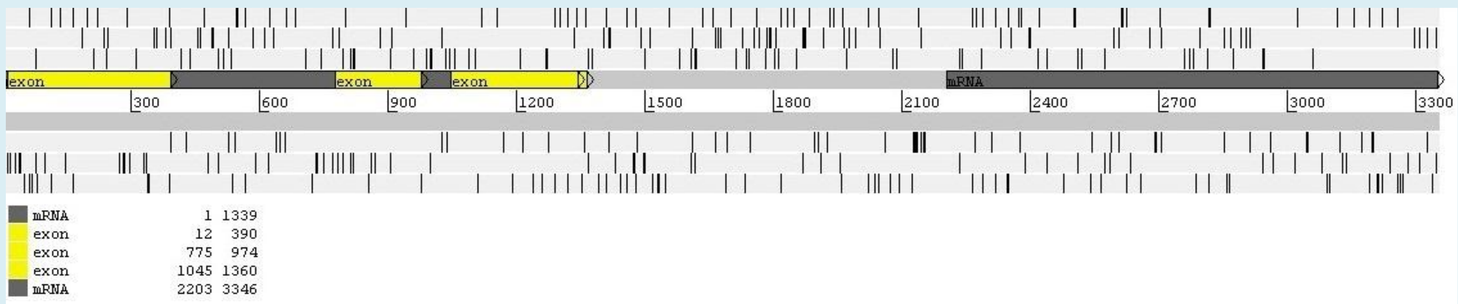
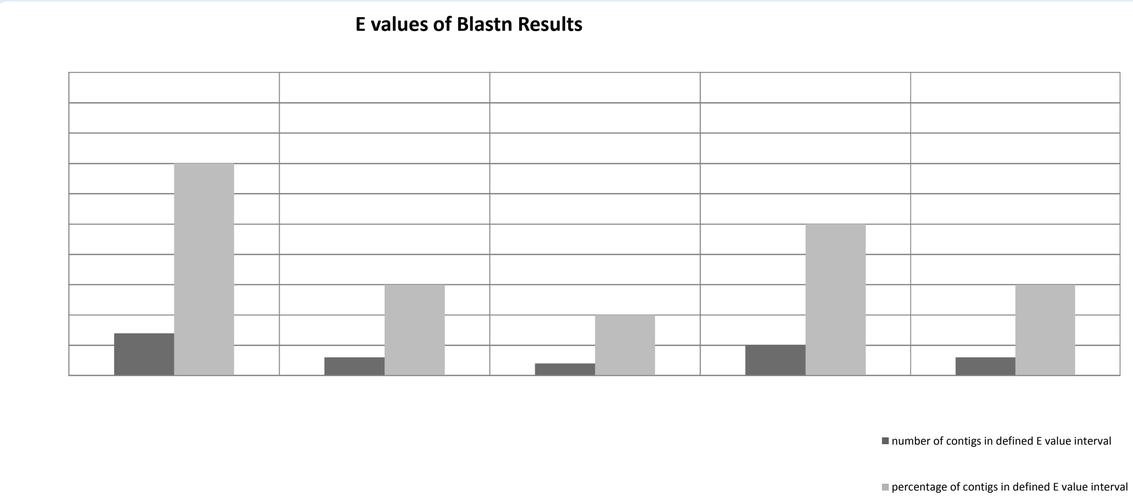


Figure 1: Visualization of a contig by Artemis



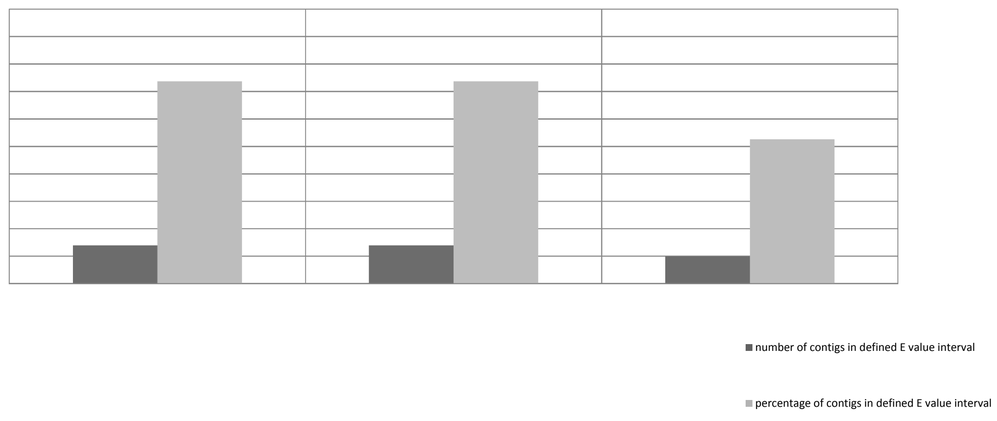


Figure 2: E value distribution of NBCI blastn and blastx in different intervals between 0 and 1. The closer E value is to 0, the less this result would be to occur by chance.

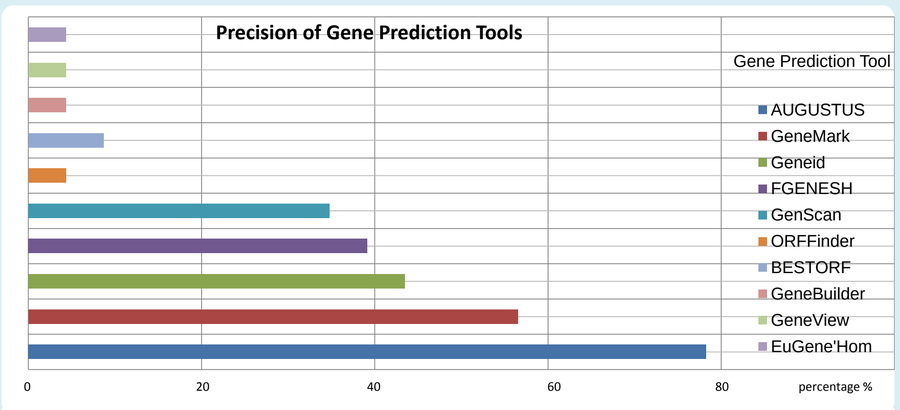


Figure 3: Precision percentage of gene prediction tools

Figure 4: Precision percentage of gene prediction tools

In Figure 3 and Figure 4, the graphs show the precision percentage of the gene prediction and protein subcellular prediction tools. Based on the gene prediction results, the tool having the highest precision results is AUGUSTUS. For protein subcellular localization prediction, various tools were used and in the figure number of each tool usage and precision percentage. Based on this result TargetP, Bacello and WolfPSORT have highest precision. As it is, unfortunately, impossible to establish a ground-truth for this comparison the precision has been established relatively by majority vote of the predictors used in this study. This does not reflect an absolute assessment of prediction accuracy but it allows us to give a soft guideline of which tools to prefer.

Discussion

For the blastn and blastx results it can be seen the number of contigs situated in different e-value intervals. The closer E value is to 0, the less it is expected to occur by chance. As it can be seen from the graphs, majority of the e-values are between 0 and E-12 for blastn / 0 and E-8 for blastx. This means that were were able to annotate most contigs with likely good homologs.

Precision of gene prediction tools was also determined by applying a statistical approach. 10 different tools were used for each contig. Among the results, we compared the intersecting results of tools and the most intersecting ones were determined as the most reliable. Graphical exhibition shows us that AUGUSTUS is the most reliable tool with above 70% intersection. The same route was followed for analysis of sub-cellular localization prediction results from 11 different tools. TargetP, Bacello and WolfPSORT are the most reliable tools that have 80 % intersection. These values may vary for other studies. We are able to choose the best tool for *Sesamum indicum* gene annotation which to us seems trustworthy for our data.