In silico gene Characterization and Biological Annotation of Bacillus thuringiensis Genome Sequences

Hanchipura Mallesh Mahadeva Swamy¹, Ramasamy Asokan¹
¹Indian Institute of Horticultural Research (IIHR), Hessaraghatta Lake Post, Bangalore 560089
Karnataka, INDIA
E-mail: clintonbio@gmail.com
E-mail: asokanihr@gmail.com

ABSTRACT
Genome annotation is the process of extraction, definition, and interpretation of features on the genome sequence descended by integrating computational tools and biological knowledge of a genomic Data. The target of a genome annotation is to discern the fundamental features of the genome sequence particularly, the genes and gene products. The characteristics of gene, its products, gene prediction programs of Bacillus thuringiensis are discussed. Although the number of genomes in Genomic databases are increasing day by day, genome wide analyses is afflictive depending on the quality of the genome annotations. This study illustrates the importance of integrative approaches for automatic annotations of genomes of B. thuringiensis by AMIGene (Annotation of Microbial Genes) and FgenesB computational method. Identified the Bt genes according to CDS, Transcription units and Operon. The characterized genes will stimulate the development of novel biopesticides and serve as basis for vigorous future research regarding the improvement of Bt as bioinsecticides in an integrated pest management. The interest in the development of new strategies for the improvement of Bt strains will continue to expand the knowledge in the genetic study.

Keywords: AMIGene, Annotations, Bacillus thuringiensis, FgenesB

Received 26/10/2013 Accepted 09/11/2013 ©2013 Society of Education, India

INTRODUCTION
The bacterium Bacillus thuringiensis 'Wonder insecticide’ proved to be a highly successful weapon for fighting some agricultural pests and some vectors of diseases but its use is still limited in developing countries. It is non toxic to people, most other non target insects and the environment. It can be targeted to specific pests. Though Bt is a very effective biological control agent, there are concerns over the development of resistance by insect species and also the narrow spectrum of activity of individual toxins. To address these concerns, new strains of Bt expressing novel toxins are actively sought and existing toxins are genetically modified for improved activity [1].

If an organism is sufficiently important to study in the first place, then a complete, closed genome sequence of at least one strain provides the basis for decades, perhaps centuries, of future investigations. The complete and correct sequence represents a permanent snapshot of one moment in evolutionary history, one that will always remain accurate even though the organism will continue to evolve. Identifying and annotating these genes will help investigators discern how gene activities in whole living systems are orchestrated to solve myriad life challenges [2]. Whole-genome sequencing represents the most powerful approach to identification of genomic diversity among closely related strains or isolates. Scanning whole genomes to detect genetic differences has the advantage that there is no inherent bias, in contrast to sampling methods such as multilocus sequence typing [2].

Many organisms have had their entire genome sequenced; however this is not the end of a genome project. Annotation is the process by which pertinent information about these raw DNA sequences is added to the genome databases. This involves describing different regions of the code and identifying which regions can be called genes. Genome projects produce large amounts of nucleotide sequence data, and gene annotation at the nucleotide-level is useful for interpreting the data. Nucleotide-level gene model annotation is concerned with identifying the nucleotides in a gene that are the exons and introns as well as the protein coding region. Genome annotation is a process to determine the genes, protein coding genes, and other biological features from a genome sequence [3, 4]. Here we predicted the CDS and genes.
of *B. thuringiensis* by using various computational software providers of AMIGene [5] and FgenesH Algorithm [6].

**MATERIAL AND METHODS**

Genome sequences of *B. thuringiensis* were retrieved from NCBI genome database (http://www.ncbi.nlm.nih.gov/genome/) using keyword *Bacillus thuringiensis*. Table 1 contains a comparative study by Both AMIGene and FGENESH utilizing a statistical model of gene prediction from *B. thuringiensis* for an accurate prediction. We acquired these algorithms through http://www.genoscope.cns.fr/agc/tools/amigene/Form/form.php and Softberry as well. AMIGene (Annotation of Microbial Genes) is an application for automatically identifying the most likely Coding Sequences (CDSs) in a large contig or a complete bacterial genome sequence. The first step in AMIGene is dedicated to the construction of Markov models that fit the input genomic data (i.e. the gene model), followed by the combination of well-known gene-finding methods and a heuristic approach for the selection of the most likely CDSs. The selection of the most likely CDSs consists in the elimination of the false positives according to overlapping criteria between adjacent CDSs, these overlaps being either total (they are called inclusion) or partial. FgenesB is a package developed by Softberry Inc. for automatic annotation of bacterial genomes. The gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites. The package includes options to work on sets of sequences, such as scaffolds of bacterial genomes or short sequencing reads extracted from bacterial communities. For community sequence annotation, it includes ABsplit program, which separates archebacterial and eubacterial sequences. FGENESB was used in the first published bacterial community annotation project [7].

**RESULTS AND DISCUSSION**

Unlike eukaryotes, the archaeal, bacterial and virus genomes are highly gene-dense. The protein coding regions usually represent more than 90% of the genome. Therefore the accuracy of gene predictors depends primarily on determining which of the six frames contains the real gene. The simplest approach in gene prediction is to look for Open Reading Frames (ORFs). An ORF is a DNA sequence that initiates at a start codon and ends at a stop codon, with no other intervening stop codon. One way to locate genes is to look for ORFs with the mean size of proteins (roughly 900 base pairs) [8]. Therefore, long ORFs indicate possible genes, although this methodology fails to predict small genes. The major problem in simply applying this technique is the possibility of ORF overlap in the different DNA strains. This approach must be used along with guidelines to avoid overlapping, choosing the more likely candidates. Also, numerous false positives are found in non-coding regions. Due to the high gene density, it is difficult to confidently state that any gene predicted in a non-coding region is false. This problem can be minimized by searching for homologies in closely-related organisms. If we do not find a conserved sequence in related species, it is assumed that the prediction (of a gene) is false.

Another problem faced by prediction programs in prokaryotes is how to determine the start codon of a sequence. The first initiation site in a sequence is not necessarily the true one. To solve this problem, programs can employ ribosome binding sites (RBS), which provide a strong signal, indicating the position of the true start site. In conclusion, there is a drop in prediction accuracy in high-GC-content genomes. Rich GC genomes contain fewer stop codons and more spurious ORFs. These false ORFs are often chosen by prediction programs instead of the real ones in the same DNA region. Additionally, the longer ORFs in GC-rich genomes contain more potential start codons, leading to a drop in the accuracy of translation initiation site prediction [9].

Interpretation of raw DNA sequence data involves the identification and annotation of genes, proteins, and regulatory and/or metabolic pathways. This process is typically performed using sequence annotation pipelines (i.e. a variety of software modules) and, in some cases, human expertise to handle the annotations generated automatically. The reference databases, computational methods and knowledge that form the basis of these pipelines are constantly being developed. In addition, the rapid increase in new sequence data has necessitated the evolution of software resources from functional annotation of a single genome towards simultaneous analysis of information from multiple genomes [10].
Swamy and Asokan

Table 1: Details of the *Bacillus thuringiensis* complete genome sequence data from NCBI Genome database

<table>
<thead>
<tr>
<th>Organism</th>
<th>BioProject</th>
<th>Assembly</th>
<th>Status</th>
<th>Chrs</th>
<th>Plasmids</th>
<th>Size (Mb)</th>
<th>GC%</th>
<th>Gene</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bacillus thuringiensis</em> serovar berliner ATCC 10792</td>
<td>PRJNA55229, PRJNA29723</td>
<td>ASM1616v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.26</td>
<td>34.8</td>
<td>6,338</td>
<td>6,243</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar konkukian str. 97-27</td>
<td>PRJNA58089, PRJNA10877</td>
<td>ASM850v1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>5.31</td>
<td>35.4</td>
<td>5,343</td>
<td>5,197</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> IBL 200</td>
<td>PRJNA55239, PRJNA29733</td>
<td>ASM1617v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.73</td>
<td>34.5</td>
<td>6,768</td>
<td>6,693</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar kurstaki str. HD73</td>
<td>PRJNA189188, PRJNA185468</td>
<td>ASM33875v1</td>
<td></td>
<td>1</td>
<td>7</td>
<td>5.91</td>
<td>34.8</td>
<td>6,334</td>
<td>6,194</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> BMB171</td>
<td>PRJNA49135, PRJNA46361</td>
<td>ASM9216v1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>5.64</td>
<td>35.2</td>
<td>5,513</td>
<td>5,352</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> Bt407</td>
<td>PRJNA55233, PRJNA29725</td>
<td>ASM1616v3</td>
<td></td>
<td>1</td>
<td>-</td>
<td>5.49</td>
<td>35.1</td>
<td>5,638</td>
<td>5,546</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar chinensis CT-43</td>
<td>PRJNA158151, PRJNA43737</td>
<td>ASM19335v1</td>
<td></td>
<td>1</td>
<td>10</td>
<td>6.15</td>
<td>35.1</td>
<td>6,380</td>
<td>6,206</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar kurstaki str. HD71</td>
<td>PRJNA17337, PRJNA171845</td>
<td>ASM29245v1</td>
<td></td>
<td>1</td>
<td>8</td>
<td>6.44</td>
<td>35</td>
<td>6,704</td>
<td>6,569</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> Bt407</td>
<td>PRJNA173886, PRJNA171844</td>
<td>ASM29270v1</td>
<td></td>
<td>1</td>
<td>6</td>
<td>6.33</td>
<td>35.2</td>
<td>6,626</td>
<td>6,462</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> IBL 4222</td>
<td>PRJNA55241, PRJNA29735</td>
<td>ASM16173v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.61</td>
<td>34.9</td>
<td>6,758</td>
<td>6,658</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> MC28</td>
<td>PRJNA176369, PRJNA167562</td>
<td>ASM30047v1</td>
<td></td>
<td>1</td>
<td>7</td>
<td>6.69</td>
<td>34.9</td>
<td>6,843</td>
<td>6,722</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar andalousiensis BGSC 4A1</td>
<td>PRJNA55233, PRJNA29725</td>
<td>ASM1616v3</td>
<td></td>
<td>1</td>
<td>-</td>
<td>5.49</td>
<td>35.1</td>
<td>5,638</td>
<td>5,546</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar finitimus YBT-020</td>
<td>PRJNA158875, PRJNA606447</td>
<td>ASM19051v1</td>
<td></td>
<td>1</td>
<td>2</td>
<td>5.68</td>
<td>35.4</td>
<td>5,931</td>
<td>5,782</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar baizhongensis BGSC 4B1</td>
<td>PRJNA55235, PRJNA29729</td>
<td>ASM16167v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.23</td>
<td>34.7</td>
<td>6,117</td>
<td>6,019</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar kurstaki str. T03a001</td>
<td>PRJNA55217, PRJNA29711</td>
<td>ASM16157v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>5.53</td>
<td>35</td>
<td>5,648</td>
<td>5,556</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar monterrey BGSC 4A1</td>
<td>PRJNA55215, PRJNA29709</td>
<td>ASM16159v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.49</td>
<td>34.7</td>
<td>6,563</td>
<td>6,490</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar pakistaniani str. T13001</td>
<td>PRJNA55227, PRJNA29721</td>
<td>ASM16155v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.04</td>
<td>35</td>
<td>6,117</td>
<td>6,028</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar pondicheriensis BGSC 4A1</td>
<td>PRJNA55233, PRJNA29727</td>
<td>ASM16165v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.03</td>
<td>34.9</td>
<td>6,147</td>
<td>6,053</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar puliensi BGSC 4C1</td>
<td>PRJNA55237, PRJNA29731</td>
<td>ASM16169v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.0</td>
<td>34.9</td>
<td>6,030</td>
<td>5,944</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar sotto str. T04001</td>
<td>PRJNA55221, PRJNA29715</td>
<td>ASM16153v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.11</td>
<td>34.9</td>
<td>6,615</td>
<td>6,583</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar thuringiensis str. T01001</td>
<td>PRJNA55225, PRJNA29719</td>
<td>ASM16151v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>6.32</td>
<td>34.8</td>
<td>6,414</td>
<td>6,323</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar tochigienis BGSC 4Y1</td>
<td>PRJNA55219, PRJNA29713</td>
<td>ASM16147v1</td>
<td></td>
<td>1</td>
<td>-</td>
<td>5.63</td>
<td>34.9</td>
<td>5,820</td>
<td>5,732</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar str. Al Hakam</td>
<td>PRJNA58795, PRJNA18255</td>
<td>ASM1506v1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>5.31</td>
<td>35.4</td>
<td>4,945</td>
<td>4,798</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em> serovar israelensis ATCC 35646</td>
<td>PRJNA54295, PRJNA15522</td>
<td>ASM16769v1</td>
<td></td>
<td>-</td>
<td>-</td>
<td>5.88</td>
<td>35</td>
<td>6,229</td>
<td>6,132</td>
</tr>
</tbody>
</table>
Therefore, there is a natural shift towards the creation of tools for viewing and manipulating data in a comparative genomics context. Also, genome annotations need to be reprocessed on a regular basis to take into account the identification of newly characterized functions. Furthermore, large-scale functional analyses generate additional data that contribute to the interpretation of genomic data. These considerations are driving the community to think about how to manage public collections of genomes in novel ways [11].

A wide variety of software is available to the scientific community, and can be used to identify genomic objects, before predicting their biological functions. However, only a limited number of biologically interesting features can be revealed from an isolated sequence. Comparative genomics tools, on the other hand, by bringing together the information contained in numerous genomes simultaneously, allow interesting features can be revealed from an iso...
resource of genes and other biological features from B. thuringiensis genomes. It is based on an updated version of the AMIGene and FgenesH. In AMIGene, the selection of the most likely CDSs consists in the elimination of the false positives according to overlapping criteria between adjacent CDSs, these overlaps being either total (they are called inclusion) or partial. FgenesH is based on an HMM whose parameters are genome specific. The most important difference between FgenesH and other gene finders is the use of a statistical significant measure. FgenesH takes a genome sequence as input and gives a list of statistically significant genes as output. Total number of the CDS identified using AMIGene is 143174 from 25 available genome sequence data. While 146676 genes were predicted using FgenesH (Table 2). These genes will serve as tools for the new strategies to improve the Bt based bioinsecticides research.

The genome sequence of an organism is an information resource unlike any that biologists have previously had access to. But the value of the genome is only as good as its annotation. It is the annotation that bridges the gap from the sequence to the biology of the organism. The aim of high-quality annotation is to identify the key features of the genome in particular, the genes and their products. The tools and resources for annotation are developing rapidly, and the scientific community is becoming increasingly reliant on this information for all aspects of biological research [13].

The multitude of bacterial genome sequences being determined has opened up a new field of research, that of comparative genomics. One role of bioinformatics is to assist biologists in the extraction of biological knowledge from this data flood. Software designed for the analysis and functional annotation of a single genome have, in consequence, evolved towards comparative genomics tools, bringing together the information contained in numerous genomes simultaneously [11].

The dynamic structure and functions of genomes are being revealed simultaneously with the progress of genome analyses. Evidence indicating genome regional characteristics (genome annotations in a broad sense) provide the basis for further analyses. Target listing and screening can be effectively performed in silico using such data [14].

**CONCLUSIONS**

Advances in sequencing technology now allow modern researchers to rapidly sequence multiple bacterial genomes. Automatic annotation pipelines that work via comparison to a reference database can introduce and propagate errors. In conclusion, the complete B. thuringiensis genome renders a better-defined genetic background for gene expression and regulation studies, especially crystal protein production and hybrid toxins. The genes characterized from Bt genome sequences will serve a tools for genetic improvement of Bt natural strains, in particular Bt recombination, offers a promising means of improving efficacy and cost-effectiveness of Bt-based bioinsecticides products to develop new biotechnological applications. With this data in hand, functional and comparative genomics studies can be initiated that may ultimately lead to new strategies for improving biocontrol strains as well as better understanding of genome evolution among the species within the group. These results from a number of completed genome projects have demonstrated that information on overall genome organization can provide biological insights.

**ACKNOWLEDGEMENTS**

The authors are grateful to ICAR, New Delhi for funding this study under Network project on Application of microbes in agriculture and allied sectors (AMAAS). Infrastructure facility and encouragement by The Director, Indian Institute of Horticultural Research (IIHR) are duly acknowledged.

**REFERENCES**


**Citation of This Article**