Advances in Bioresearch
Adv. Biores., Vol4 (4) December 2013: 93-98
©2013 Society of Education, India
Print ISSN 0976-4585; Online ISSN 2277-1573
Journal's URL:http://www.soeagra.com/abr/abr.htm
CODEN: ABRDC3

ORIGINAL ARTICLE

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

Hanchipura Mallesh Mahadeva Swamy[1], Ramasamy Asokan[1]
[1]Indian Institute of Horticultural Research (IIHR), Hessaraghatta Lake Post, Bangalore 560089 Karnataka, INDIA
E-mail: clintonbio@gmail.com
E-mail: asokaniihr@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

INTRODUCTION

The bacterium *Bacilllus 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 *Bacilllus thuringiensis* Table 1. We used a comparative study by Both AMIGene and FGENESH utilize a statistical model of gene prediction from *B.thuringiensis* for an accurate prediction. We acquired these algorithms by 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].

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

Chromosomes [24] Scaffolds or contigs [1] SRA or Traces [0]

Organism	BioProject	Assembly	Status	Chrs	Plasmids	Size (Mb)	GC%	Gene	Protein
Bacilllus thuringiensis serovar berliner ATCC 10792	PRJNA55229, PRJNA29723	ASM16161v1	●	1	-	6.26	34.8	6,338	6,243
Bacilllus thuringiensis serovar konkukian str. 97-27	PRJNA58089, PRJNA10877	ASM850v1	●	1	1	5.31	35.4	5,343	5,197
Bacilllus thuringiensis IBL 200	PRJNA55239, PRJNA29733	ASM16171v1	●	1	-	6.73	34.5	6,768	6,693
Bacilllus thuringiensis serovar kurstaki str. HD73	PRJNA189188, PRJNA185468	ASM33875v1	●	1	7	5.91	34.8	6,334	6,194
Bacilllus thuringiensis BMB171	PRJNA49135, PRJNA43631	ASM9216v1	●	1	1	5.64	35.2	5,513	5,352
Bacilllus thuringiensis Bt407	PRJNA55223, PRJNA29717	ASM16149v1	●	1	-	6.03	34.8	6,425	6,298
Bacilllus thuringiensis Bt407	PRJNA177931, PRJNA176850	ASM30674v1	●	1	9	6.13	35	6,590	6,402
Bacilllus thuringiensis HD-771	PRJNA173374, PRJNA171845	ASM29245v1	●	1	8	6.44	35	6,704	6,569
Bacilllus thuringiensis HD-789	PRJNA173860, PRJNA171844	ASM29270v1	●	1	6	6.33	35.2	6,626	6,462
Bacilllus thuringiensis IBL 4222	PRJNA55241, PRJNA29735	ASM16173v1	●	1	-	6.61	34.9	6,758	6,658
Bacilllus thuringiensis MC28	PRJNA176369, PRJNA167562	ASM30047v1	●	1	7	6.69	34.9	6,843	6,722
Bacilllus thuringiensis serovar andalousiensis BGSC 4AW1	PRJNA55231, PRJNA29725	ASM16163v1	●	1	-	5.49	35.1	5,638	5,546
Bacilllus thuringiensis serovar chinensis CT-43	PRJNA158151, PRJNA43737	ASM19335v1	●	1	10	6.15	35.1	6,380	6,206
Bacilllus thuringiensis serovar finitimus YBT-020	PRJNA158875, PRJNA60447	ASM19051v1	●	1	2	5.68	35.4	5,931	5,782
Bacilllus thuringiensis serovar huazhongensis BGSC 4BD1	PRJNA55235, PRJNA29729	ASM16167v1	●	1	-	6.23	34.7	6,117	6,019
Bacilllus thuringiensis serovar kurstaki str. T03a001	PRJNA55217, PRJNA29711	ASM16157v1	●	1	-	5.53	35	5,648	5,556
Bacilllus thuringiensis serovar monterrey BGSC 4AJ1	PRJNA55215, PRJNA29709	ASM16159v1	●	1	-	6.49	34.7	6,563	6,490
Bacilllus thuringiensis serovar pakistani str. T13001	PRJNA55227, PRJNA29721	ASM16155v1	●	1	-	6.04	35	6,117	6,028
Bacilllus thuringiensis serovar pondicheriensis BGSC 4BA1	PRJNA55233, PRJNA29727	ASM16165v1	●	1	-	6.03	34.9	6,147	6,053
Bacilllus thuringiensis serovar pulsiensis BGSC 4CC1	PRJNA55237, PRJNA29731	ASM16169v1	●	1	-	6.0	34.9	6,030	5,944
Bacilllus thuringiensis serovar sotto str. T04001	PRJNA55221, PRJNA29715	ASM16153v1	●	1	-	6.11	34.9	6,615	6,583
Bacilllus thuringiensis serovar thuringiensis str. T01001	PRJNA55225, PRJNA29719	ASM16151v1	●	1	-	6.32	34.8	6,414	6,323
Bacilllus thuringiensis serovar tochigiensis BGSC 4Y1	PRJNA55219, PRJNA29713	ASM16147v1	●	1	-	5.63	34.9	5,820	5,732
Bacilllus thuringiensis str. Al Hakam	PRJNA58795, PRJNA18255	ASM1506v1	●	1	1	5.31	35.4	4,945	4,798
Bacilllus thuringiensis serovar israelensis ATCC 35646	PRJNA54295, PRJNA15522	ASM16769v1	◑	-	-	5.88	35	6,229	6,132

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 annotators to make inferences based on the idea that evolution and natural selection are central to the definition of all biological processes [12].

Table 2: *In silico* annotation details of the *Bacillus thuringiensis* complete genome sequence using AMIGene and FgenesB Analysis

Organism	Size (Mb)	Gene	AMI Gene Analysis	FgenesB Analysis		
			predicted CDSs	Number of predicted genes	Number of transcription units	operons
Bacilllus thuringiensis serovar *berliner* ATCC 10792	6.26	6,338	6412	6729	4163	1317
Bacilllus thuringiensis serovar *konkukian* str. 97-27	5.31	5,343	5300	5386	3319	1099
Bacilllus thuringiensis IBL 200	6.73	6,768	6815	6957	4256	1382
Bacilllus thuringiensis serovar *kurstaki* str. HD73	5.91	6,334	5777	5885	3520	1184
Bacilllus thuringiensis BMB171	5.64	5,513	5313	5448	3349	1109
Bacilllus thuringiensis Bt407	6.03	6,425	5556	6149	3590	1238
Bacilllus thuringiensis Bt407	6.13	6,590	5556	6504	3836	1317
Bacilllus thuringiensis HD-771	6.44	6,704	6017	6149	3590	1238
Bacilllus thuringiensis HD-789	6.33	6,626	5533	5659	3467	1134
Bacilllus thuringiensis IBL 4222	6.61	6,758	6912	7036	4158	1393
Bacilllus thuringiensis MC28	6.69	6,843	5397	5520	3377	1111
Bacilllus thuringiensis serovar *andalousiensis* BGSC 4AW1	5.49	5,638	5688	5764	3519	1160
Bacilllus thuringiensis serovar *chinensis* CT-43	6.15	6,380	5598	5687	3401	1153
Bacilllus thuringiensis serovar *finitimus* YBT-020	5.68	5,931	5353	5470	3271	1131
Bacilllus thuringiensis serovar *huazhongensis* BGSC 4BD1	6.23	6,117	6165	6267	3944	1223
Bacilllus thuringiensis serovar *kurstaki* str. T03a001	5.53	5,648	5763	5854	3573	1174
Bacilllus thuringiensis serovar *monterrey* BGSC 4AJ1	6.49	6,563	6651	6754	4133	1374
Bacilllus thuringiensis serovar *pakistani* str. T13001	6.04	6,117	6651	6462	4074	1295
Bacilllus thuringiensis serovar *pondicheriensis* BGSC 4BA1	6.03	6,147	6184	6272	3698	1263
Bacilllus thuringiensis serovar *pulsiensis* BGSC 4CC1	6	6,030	6085	6163	3840	1245
Bacilllus thuringiensis serovar *sotto* str. T04001	6.11	6,615	6893	7009	4077	1486
Bacilllus thuringiensis serovar t*huringiensis* str. T01001	6.32	6,414	6467	6272	3698	1263
Bacilllus thuringiensis serovar *tochigiensis* BGSC 4Y1	5.63	5,820	5819	5910	3652	1204
Bacilllus thuringiensis str. Al Hakam	5.31	4,945	5269	5370	3271	1102
Bacilllus thuringiensis serovar *israelensis* ATCC 35646	5.88	6,229	-	-	-	-

The annotation process often includes a lot of meticulous inspection done by researchers; detailed biological knowledge is very valuable for this work. To analyze the vast amount of genome annotation data available today, a visual representation of genomic features in a given sequence range is required. Current study focus on Genome annotation of *B.thuringiensis*. Complete sequence has been retrieved from Genbank, the public data repository. The length of the complete genome was 5.31 to 6.73 mb. The feature prediction has not been predicted yet in any database. Hence the aim was to predict all the homogeneous

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

1. George Z and Crickmore N (2012). *Bacillus thuringiensis* Applications in Agriculture**.** E. Sansinenea (ed.), *Bacillus thuringiensis Biotechnology,* DOI 10.1007/978-94-007-3021-2_2
2. Fraser CM, Eisen JA, Nelson KE, Paulsen IT and Salzberg SL (2002). The Value of Complete Microbial Genome Sequencing (You Get What You Pay For). *J Bacteriol* 184 (23): 6403-6405
3. Keibler E, Brent MR (2003). Eval: A software package for analysis of genome annotations.BMC Bioinformatics. 4:50.
4. Ratsch G, Sonnenburg S, Srinivasan J, Witte H, Muller KRu (2007). Improving the *C. elegans* genome annotation using machine learning. *PLoS Comput Biol* 3(2):e20
5. Bocs S, Cruveiller S, Vallenet D, Nuel G and Médigue C (2003) AMIGENE: Annotation of Microbial Genes. *Nucleic Acids Res* 13(31): 3723-3726
6. Salamov A, Solovyev V (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 10: 516-522
7. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978): 37–43

8. Allen JE, Pertea, M. Salzberg, S L (2004). Computational gene prediction using multiple sources of evidence. *Genome Res* 14(1): 142–8
9. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW and Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119
10. Binnewies TT, Motro Y, Hallin PF, et al. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* 6: 165-185
11. Claudine and Ivan M (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Research in Microbio 158: 724-736
12. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C and Me´ digue C (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. Database, Article ID bap021: doi:10.1093/database/bap021
13. Lincoln Stein (2001) Genome annotation: from sequence to biology. *Nature reviews* (Genetics) 2: 493
14. Kawaji H, Hayashizaki Y (2008) Genome Annotation. Bioinformatics Methods in Mol. Bio 452: 125-139

Citation of This Article
Hanchipura Mallesh Mahadeva Swamy, Ramasamy Asokan. *In silico* gene Characterization and Biological Annotation of *Bacillus thuringiensis* Genome Sequences. Adv. Biores. Vol 4[4] December 2013: 93-98