Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Current Trends**

Silkworm nucleotide databases - Current trends and future prospects

Nicole Koshy[1], Kangayam M. Ponnuvel[1, *], Randhir K. Sinha[1] and S. M. H. Qadri[1]

[1]Biotechnology Laboratory, Central Sericulture and Germplasm research Centre, P.O. Box: 34, Thally road, Hosur-635109, Tamil Nadu, India; Kangayam M Ponnuvel* - E-mail: kmpvel@yahoo.com; * Corresponding author

received February 14, 2008; revised March 13, 2008; accepted April 07, 2008; published April 19, 2008

Abstract:
The domesticated silkworm, *Bombyx mori* serves as an ideal representative of lepidopteran species for a variety of scientific studies. As a result, databases have been created to organize information pertaining to the silkworm genome that is subject to constant updating. Of these, four main databases are important for store nucleotide information in the form of genomic data, ESTs and microsatelites. These databases also store data related to other lepidoptera and important insects, which help in insect biological research. Though a considerable amount of nucleotide data is currently available, there is a paucity of data related to silkworm and other lepidopteran proteins. Hence, the focus of this article is to present the current status of nucleotide databases of silkworm, avenues for improvement and possibilities for databases that could be created in the future.

Keywords: silkworm; *Bombyx mori*; silkbase; silk knowledgebase; lepidopterans

Background:

The silkworm *Bombyx mori*, is a monophagous insect, which feeds on the leaf of the *Morus* species, commonly known as mulberry. Bred in captivity for over 5000 years, it is now fully domesticated. Silk cocoons from this species are the source of commercial silkyarn. Its closest and only wild relative is the *Bombyx mandarina*, which is found in northern India, Southern and Southeastern parts of Asia. The *Bombyx mori* has an estimated haploid nuclear genome size of 530 Mb broken into 28 chromosomes **[1]**.

The *B. mori* genome is considered as a central model for insect species and is important in comparative and functional genomics studies for lepidopteran species, providing a solid foundation for integrating biological information of lepidopteran and even insects in general **[2]** and is second only to fruitfly (*Drosophila melanogaster*).

Many basic physiological processes of insects are conserved, in spite of the evolutionary process. Thus the biological sequences of the silkworm are of importance as, their study could help to further elucidate the function of gene homologs and facilitate studies of insect domestication, morphogenesis, endocrinology, reproduction, behavior and immunity. Genomic data helps understand the organization of genes in the genome, their clustering and regulatory patterns in common functions or pathways. The ESTs and cDNA sequences help identify and annotate all the genes that can be expressed in the silkworm at various stages of development thus aiding in elucidating its metabolic functioning and gene expression profile. In addition we can analyze the organization of exons, introns, alternate splice sites and regulatory regions in the genes, which are important in biological studies **[3]**. Information on microsattelites and non-coding regions too has its importance in studies elucidating evolutionary relationships, mapping and population genetics and forensic studies. Thus silkworm databases are essential for insect biology studies.

In light of these needs, a few databases have been created to store the data generated from genomic and cDNA sequencing operations carried out on the *B. mori*. To date, there are four major databases, that store nucleotide data, that are available in the public domain. Of these, two are genomic databases: the Silkworm KnowledgeBase (http://silkworm.genomics.org.cn) hosted by the Beijing Genomics Institute (BGI) in China and the Silkworm Genome Database (http://papilio.ab.a.u-tokyo.ac.jp/genome/index.html) hosted by the Insect Genetics and Bioscience (IGB) lab at the University of Tokyo in Japan **[4]**. The IGB lab also hosts the EST database for the silkworm namely, the Silk Base. This database contains all the EST sequences expressed in *B. mori* at various stages of growth and development, in various tissues. Another database, namely the Silk moth Microsatellite Database (SilkSatDb - http://www.cdfd.org.in/silksatdb) hosted by the Centre for DNA Fingerprinting and Diagnostics at Hyderabad, India is a relational database of microsatellites extracted from available EST and Whole Genome Shotgun (WGS) Sequences of *B. mori*. The database stores three kinds of data: the microsatellite repeats found in the *B. mori* EST, WGS sequences and details pertaining to each sequence and information on the primers developed for each of these microsatellites **[5]**.

Silkworm knowledgebase

The Silkworm KnowledgeBase (Silk Db) is a web-based repository for the curation, integration and study of silkworm genetic and genomic data maintained by many key institutes in China, mainly the Beijing Genomics Institute (BGI). After the generation of the ~6X draft genome sequence of *B. mori*, the database now provides data in the form of an integrated map of contigs, cDNAs, clusters of expressed sequence tags (ESTs), transposable elements (TEs), mutants, single nucleotide polymorphisms (SNPs) and functional annotations of genes with assignments to InterPro domains and Gene Ontology (GO) terms on a genome wide scale **[6]**. The database aims to provide a comprehensive knowledgebase about the silkworm and present the silkworm genome and related information in a systematic and graphical way for the convenience of in-depth comparative genomics studies. SilkDB is accessible at http://silkworm.genomics.org.cn in the public domain.

This database contains 16,425 EST clusters based on the sequencing of 80,470 ESTs from *B. mori* tissues, 554 lepidopteran genes, 521 *B. mori* homologs of other lepidopteran genes and SNPs identified from the ESTs **[7]**. Using the *B. mori* genome sequence as a reference, comparative analysis can be done between *B. mori* and *B. mandarina*, other lepidopterans and the sequenced dipterans (Fruit fly and Mosquito). The silkworm genome has been assembled in the form of contigs onto a scaffold using mapped genetic markers and BAC-based physical maps, complete with packages for sequence assembly, gene annotation and identification of transposable elements. This in turn serves as a framework to organize information for other lepidopterans and, along with the database's search-engine and MapView program, provides an information resource and a comparative genomics platform for genome related research of both, silkworm and other insects. The genomic data are organized into 3 modules for effective management namely; scaffold, gene and transposable elements. The scaffold module contains 23,156 scaffolds for the 28 chromosomes organized from whole genome shotgun contigs that span 428.7 Mb of the *B. mori* genome covering 90.9% of all known silkworm genes. The gene module contains the sequences of 18,510 annotated genes along with the full-length cDNA sequences of 212 known silkworm genes from GenBank. The transposable elements (TE) module catalogues the 6,01,225 TEs that have been identified in the silkworm genome into their respective classes and gives detailed information about them. The increase in silkworm genome size as compared to the fruit fly could also be explained by the identification of the transposons.

There are two tools offered by the SilkDb to users for their search. The visualization tool-MapView, which displays the *B. mori* genome as sequence contigs assembled on a scaffold that allows users to browse a series of tracks aligned with the genomic sequence. The second one, the SilkDb search engine provides access to all the major data types stored by 2 search modes: keyword-based search and BLAST-based homology search including searches for all classes of elements stored in the database. Cross-referenced links to related database entries in other databases allow access to a wider range of information related to the user's search which lends more validity to the results. Thus a comprehensive search of the database is possible to obtain results that specifically suit one's interest.

SilkBase

The SilkBase is an EST Database for the Silkworm, *B. mori* hosted by the Insect Genetics and Biosciences (IGB) Lab at the University of Tokyo in Japan. It catalogues ESTs obtained from all the different tissues at various stages of growth so as to give a complete idea of the gene expression patterns of the Silkworm. It can be accessed at www.ab.a.u-tokyo.ac.jp/silkbase **[8]**. The main aim of constructing the EST database is to help make a complete cDNA library that can be used to identify all the genes expressed by the silkworm. The genes identified were classified using protein homology search. At its inception, the database contained 35,000 ESTs from 36 cDNA libraries generated by mRNAs produced in several major tissues viz., brain, fat body, midgut, silk gland, pheromone gland, haemocytes, malphigian tubules, testis and wing discs during various developmental stages. They have been now grouped to give 11,202 non-redundant ESTs with the average length of 1.25 kb. Comparison with FlyBase showed that the database covered about >55% of all *Bombyx* genes. EST sequencing was done by randomly picking more than 1000 clones from each library and sequencing up to about 700 nucleotides from the 5' end. The representative of each group was the sequence having the maximum length in that group. Search of the database can be done using a keyword/clone name search, BLAST based searches for homologous *Bombyx* cDNAs with known amino acid sequences of other species and direct comparison of sequences with FlyBase and WormBase.

Important genes of silkworm, *B. mori*

Though the silkworm databases contain information on all genes discovered in the silkworm, there are some genes that are of greater importance as they play a major role in the biology of the silkworm and other lepidoptera **[9]**. Presently there are about 3000 silkworm genotypes being maintained in Europe and Asia. The genetic stocks consist of about 500 mutants that vary in different physical, physiological and biochemical traits. As a result the genes responsible for determination of these characteristics have been extensively studied.

Among the genes of interest, the silk gland and the chorion genes are most important. The silk-gland- because it is the best source of DNA for genomic studies as each cell accumulates between 400,000 to 800,000 copies of the haploid genome, whereas the chorion is a very effective indicator of mutation in genotype. The homeobox genes have also been studied extensively as they are responsible for embryonic tissue differentiation and normal development of all organisms. In addition, the immune genes that code for anti-bacterial and anti-viral proteins have also been extensively studied as lepidopterans constitute the most important class of biological pests **[10]**.

Current applications:

From the above discussion, the importance of databases cataloguing biological information related to lepidopterans can be understood. Different databases store various types of information that have found many applications in lepidopteran research. The silkworm genomic databases store information in various forms. Contig information allows viewing of the complete genome with necessary information at all relevant locations, so that the structural organization of genes and other genetic elements such as transposable elements can be analyzed inside the genome as a whole. Clustering of genes involved in common processes can also be studied. Gene-wise BLAST searches help to identify similar orthologous sequences in other organisms. This helps find genes in other organisms that show homology in function, which helps to identify conserved regions in the sequence, in homology modeling of putative proteins and elucidating the phylogenetic relationship between orthologous and paralogous sequences **[11, 12, 13]**. Information on transposable elements (TE) has helped to explain the difference in genome size of *Drosophila* and *B. mori*. They also help in studies of gene activation-deactivation and in recombination studies. The EST databases help in the construction of molecular linkage maps that show where the different ESTs are located in the genome and how the genes are organized and located in relation to each other. Using the same ESTs, Bacterial Artificial Chromosome (BAC) contigs have also been created, with each contig being assigned an average of 3-4 EST markers. EST data is also useful for serial analysis of gene expression (SAGE), microarray analysis in expression profile and genome studies where large amount of data are analyzed at once **[14, 15, 16, 17]**. Microsatellite data helps measure genetic distance, to carry out genetic fingerprinting of silkmoths, construct molecular linkage and single nucleotide polymorphism maps and carry out phylogenetic analysis between silkmoths and other lepidopteran species **[18, 19, 20]**. A combination of all these data has also been used in the prediction of microRNAs **[21]**.

Future prospects:

At present extensive databases are available that contain nucleotide information at different levels. In the future, protein databanks could be created that store structural information about various silkworm proteins, their structure and sequence information, family classification and several others. With a more detailed knowledge about the *B. mori* genome, greater number of ESTs could be generated that could aid in a more precise construction of BAC contigs for the silkworm genome. Currently work is being done to improve the quality of genomic data available. This in turn would be useful for generating accurate maps and representations of EST clusters which would aid in comparative genomic studies of *B. mori* with other silkmoths, lepidopterans and insects.

Efforts are also being made to complete the microsatellite database, so that linkage maps for microsatellite loci can be generated. Phylogenetic comparison of microsatellites between heterologous species can then be carried out, to elucidate a clearer evolutionary relationship. Thus, in the future, it is hoped that generation of a more complete draft of the silkworm genome will help in better understanding of lepidopteran species. Protein databases for silkworms would be useful in modeling projects looking for methods to help control lepidopteran pests and help in the study of their immune system. Hence, a lot is yet to be done to make data related to insect genomes and proteomes readily available to researchers.

References:

[01] M. Goldsmith *et al., Annual Review of Entomology*, 50: 71 (2004) [PMID: 15355234]
[02] J. Nagaraju & M. Goldsmith, *Current Science*, 83: 415 (2002)
[03] G. Nagaraju & J. Nagaraju, *Electrophoresis*, 16: 1633 (1995) [PMID: 8582347]
[04] K. Mita *et al., DNA Research*, 11: 27 (2004) [PMID: 15141943]
[05] M. Prasad *et al., Nucleic Acids Research*, 33: D403 (2005) [PMID: 15608226]
[06] Q. Xia *et al., Science*, 306: 1937 (2004) [PMID: 15591204]
[07] J. Wang *et al*., *Nucleic Acids Research*, 33: D399 (2005) [PMID: 15608225]
[08] K. Mita, *et al*., *Proc. Natl. Acad. Sci. USA*, 100: 14121 (2003) [PMID: 14614147]
[09] J. Nagaraju, *Current Science*, 78: 151 (2000)
[10] P. Schmid-Hempel, *Annual Review of Entomology*, 50: 529 (2004) [PMID: 15471530]
[11] B. Mahendran *et al., Journal of Genetics*, 85: 1 (2006) [PMID: 16809837]
[12] H. Abe *et al*., *Cytogenetics and Genome Research*, 110: 144 (2005) [PMID: 16093666]
[13] S. Toru *et al., Zoological Science*, 22: 1377 (2005)
[14] Q. Xia *et al., Genome Biology*, 8: R162 (2007) [PMID: 17683582]
[15] Y. Suetsugu *et al., BMC Genomics*, 8: 314 (2007) [PMID: 17822570]
[16] K. Yamamoto *et al., Genome Biology*, 9: R21 (2008) [PMID: 18226216]
[17] J. Huang *et al., Genomics*, 86: 233 (2005) [PMID: 15963683]
[18] S. Archak *et al., Nucleic Acids Research*, 35: D36 (2007) [PMID: 17082205]
[19] A. Yoshido *et al., Genetics*, 170: 675 (2005) [PMID: 15802516]
[20] K. Yamamoto *et al., Genetics*, 173: 151 (2006) [PMID: 16547112]
[21] T. Chuan-zhou *et al*., *J Zhejiang Univ Science B*, 10: 806 (2006)
[22] http://silkworm.genomics.org.cn
[23] http://papilio.ab.a.u-tokyo.ac.jp/genome/index. html
[24] http://www.cdfd.org.in/silksatdb
[25] http://www.ab.a.u-tokyo.ac.jp/silkbase

Edited by P. Kangueane
Citation: Koshy *et al.*, Bioinformation 2(7): 308-310 (2008)