Advances in Biochemistry
2013; 1(2): 33-42
Published online July 20, 2013 (http://www.sciencepublishinggroup.com/j/ab)
doi: 10.11648/j.ab.20130102.15

SciencePG
Science Publishing Group

Seri-bioinformatics: emerging trends and challenges in silkworm research

Punyavathi, Hosaholalu Boregowda Manjunatha*

Proteomics and Genomics Lab, Department of Studies in Sericulture, University of Mysore, Mysore 570006, Karnataka, India

Email address:
punyacoorg@gmail.com(Punyavathi), manjunathahb@yahoo.com(H. B. Manjunatha), manjunathahb@gmail.com(H. B. Manjunatha)

To cite this article:
Punyavathi, Hosaholalu Boregowda Manjunatha. Seri-Bioinformatics: Emerging Trends and Challenges in Silkworm Research. *Advances in Biochemistry*.Vol. 1, No. 2, 2013, pp. 33-42. doi: 10.11648/j.ab.20130102.15

Abstract: With the advent of genomic and proteomic research from bacteria to man an unprecedented data generated are pertinently analyzed and managed by the evolving science - bioinformatics. In scientific research, *Bombyx mori* L. is considered as a model insect for molecular studies along with the fruit fly (*Drosophila melanogaster*) and a central model species for genome studies in moths and butterflies (the insect order Lepidoptera). As a consequence, new findings in the fields of proteome, genome and bioinformatics have resulted in the exponential generation of data that are stored in assorted array of databases. These databases not only reducing the gap and time while allowing information's to be accessed also emerged as a highly valuable platform through which scientific community can use, exchange and analyze molecular data across the world on mouse click. The computational approaches in various biological disciplines including agriculture/sericulture is not merely a reflection of a general extended usage of computers and the internet, but due to the creation of useful databases coupled with appropriate software's and methods for access by the rest of the scientific community with ease. Application of bioinformatics tools and techniques not only facilitated detection of proteomic and genomic diversity among the species/strains but that resulted in finding a gap in the silkworm genome sequence of a strain that diverged during the course of domestication. In addition, bioinformatics approaches give an insight, uncovering the lineage with gene and protein count while *B. mori* and *Drosophila* encompass ~18,000 and ~16,000 (Genes) and ~9,000 and ~22,000 (Proteins) respectively, to discover their diversity and functional properties. In view of this, we have documented the innovations made in the emerging field "Seri-bioinformatics" as valuable resources aiming at feasible comparative studies among allied species and application in the field of biotechnology and biomedical sciences.

Keywords: Bombyx Mori, Database, DNA, Informatics, Lepidoptera, Protein

1. Introduction

The development of protein [1] and DNA [2,3] sequencing methods led to sequencing of several protein families and whole genomes of a variety of organisms. As a consequence, an unprecedented wealth of biological data has been generated. Dayhoff [4] was the first to assemble sequence data into a protein sequence atlas in the 1960s, and their collection eventually became known as the Protein Information Resource (PIR) and therafter translated DNA sequences are included in the PIR. Concomitantly, a DNA sequence database was assembled first at Los Alamos national laboratory by Walter Goad and colleagues and gene bank database at European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. As DNA sequences became available in the late 1970s interest also increased in developing computer programs to analyze these sequences in various ways. The application of computers in sequence analysis, from programs for large mainframe computers down to the then new microcomputers, opened a new field of innovations. Genetics Computer Group (GCG) at the University of Wisconsin developed and offered a set of programs to access, manipulate and analyze nucleotide and protein sequences that ran on a VAX computer [5]. Other companies offered microcomputer programs - Intelligenetics and DNAStar for sequence analysis and PHRED [6] and PHRAP [7] were developed to facilitate the collection and organization of data. Nowadays, massive numbers of websites are available to perform varied types of sequence analysis at free of cost for academic institutions and moderate cost to commercial users.

In recent years, DNA sequencing has become a common laboratory activity and consequently new sequences collected in the laboratory have increased exponentially but detection of coding sequences to predict the protein synthesized in the biological system hitherto unexplored. Therefore, increased demand for computer programs provides a way to access each sequence in the existing sequence database and detect their functional properties. Such searches are greatly facilitated by programs such as FASTA [8] and BLAST (Basic Local Alignment Search Tool [9]). As genetic and sequence information became available for the model organisms, interest arose in generating specific genome database that could be queried to retrieve information. As a consequence, the first genome database called the *Caenorhabditis elegans* database (AceDB) developed along with a method to access [10]. With this, bioinformatics originated as a cross-disciplinary field to cater the need for the computational solutions to research problems raised in all fields of life and biomedical sciences.

Advances in genomic research from prokaryotes to eukaryotes led to the initiation of genome sequencing in *B. mori* - second only to the fruit fly as a model insect for genetics - hoping that it might bridge the gap between the fly (180 Mb) and the human (3000 Mb) as the haploid genome size of the silkworm (432 Mb) is 2.4 times greater than that of *D. melanogaster* and one–seventh that of human genome. So, China, Japan, France and other countries have initiated an ambitious silkworm genome project [11,12] initially to establish a basic resource for comprehensive genome analysis. Consequently, a large number of insect genome projects have been initiated and whole-genome sequences for 24 insects are completed and sequencing of many more insects' is in progress under i5K project (www.arthropodgenomes.org/wiki/i5K). As copious amount of original data on gene sequence become accessible, the advances in bioinformatics gave an insight to several "omics" disciplines, including proteomics, transcriptomics, metabolomics and structural genomics.

Notably, apart from increasing silk productivity, efforts are being made to develop the silk worm as a bioreactor for the production of recombinant proteins of biomedical importance [13]. Expression of marker proteins (Luciferase and green fluorescent protein) has been successfully achieved in cell lines and caterpillars of *B. mori* employing recombinant BmNPV vector harboring reporter genes [14]. Besides, the production of silk, it has been treated as a central model organism among Lepidopterans that comprises most of the agricultural pests. Thus, the highly domesticated *B. mori* is used to unravel many mysteries involved in insect life processes including biochemistry, physiology, molecular biology and most applied fields of biotechnology, agriculture and biomedical research [15].

Keeping the progress made in proteomics, genomics and bioinformatics from bacteria to man in general and insects particularly in view, we present here a comprehensive message on the databases along with their contents and the bioinformatics in use for analysis of protein and genome sequences of mulberry and non-mulberry silkworms.

2. Silkworm Host Plant - Mulberry Resources and Repository

2.1. Mulberry Genome and Database

Unlike the silkworm, *B. mori* the chief food plant - mulberry (*Morus indica*) does not document well in any database, except the complete nucleotide sequence of the chloroplast genome (158,484 bp), which was determined using a combination of long PCR (Polymerase chain reaction) and shotgun-based approaches [16]. It comprises two identical inverted repeats of 25,678 bp for each circular double stranded molecule, separating a large and small single copy region of 87,386 bp and 19,742 bp respectively. The sequence homology of ~2000 clones with other databases facilitated to assign ESTs for functional categories like metabolism (10%), protein synthesis (10%), transport (9%), stress related proteins (7%), and energy (7%) of the total ESTs. In addition, 2,400 ESTs sequenced from the roots of *M. indica* are assembled into 148 contigs and 1,420 singletons through the CAP3 assembly program. The GENEVESTIGATOR - the *Arabidopsis* microarray database and analysis tool box (https://www.genevestigator.ethz.ch. Table-1) provides an overview of gene expression profiles of ESTs in response to environmental stresses [17]. Interestingly, ESTs matched to some genes of silkworm are of special significance that explicit their involvement in establishing *Morus* - *Bombyx* relationship and represent co-evolution of genes, which open ample scope to uncover genes involved in plant - insect interactions.

Table 1*. Databases and their URL for quick access to information on mulberry, silkworm and other allied field.*

http://www.ab.a.u-tokyo.ac.jp/silkbase/
http://www.ag.auburn.edu/enpl/hyche/saturniidae/
http://amigo.geneontology.org/cgi-bin/amigo/ blast.cgi
www.arthropodgenomes.org/wiki/i5K
http://www.bioinformaticsonline.org
http://www.butterflybase.org
http://ca.expasy.org/sprot/
http://www.cdfd.org.in/silksatdb
http://www.cdfd.org.in/wildsilkbase/team.php
http://www.ebi.ac.uk/
http://www.fruitfly.org
https://www.genevestigator. ethz.ch.
http://insects.eugenes.org/DroSpeGe/
http://www.issas.ac.cn
http://www.jassilks.com
http://kaiko2ddb.dna.affrc.go.jp
http://morus.swu.edu.cn/morusdb
http://www.naas.go.kr/
http://www.ncbi.nlm.nih.gov/
http://www.nias.affrc.go.jp
http://pir.georgetown.edu/
www.pubmedcentral.nih.gov/
http://resourcedb.nbrp.jp/resource/list.jsp
http://sgp.dna.affrc.go.jp/index.html
http://www.shigen.nig.ac.jp/silkwormbase/index.jsp
http://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi
http://www.silkgermplasm.com

http://silkworm.swu.edu.cn/silkdb/
http://silkworm.genomics.org.cn/
http://srs.ebi.ac.uk
http://us.expasy.org/sprot/
http://wego.genomics.org.cn/cgi-bin/wego/index.
http://world-2dpage.expasy.org/make2ddb/
http://www.wormbase.org/

In the light of advances in mulberry genome research, a complete repertoire of genes and their functions would facilitate phylogenetic studies to establish the relationship between other angiosperms, help in developing a long term strategy to overcome abiotic tolerance in mulberry and improve its global importance. Towards this, a proto-type database of mulberry genome is established (Table 1) consisting microsatellite markers (SSR and SCAR markers) based linkage map and DNA polymorphism data. Importantly, *Morus* genome database (MorusDB - http://morus.swu.edu.cn/morusdb) provides genomic data including gene structures, functional annotation of genes, transcriptome data, and EST of *M. notabilis*. On this platform, a search and analysis with related plant species shall be performed using onsite tools – SMS2, BLAST, EMBOSS, WebLogo, Find Motifs, Synteny Plotter and Genome Browser. This preliminary information on mulberry gave an insight for the future line of research on comparative genomics and proteomics in other host plants of Lepidopterans.

3. Silkworm Proteome And Genome Databases and Their Composition

3.1. Proteomic Resources and Repositories

3.1.1. Protein Identification, Analysis and Annotation

Cataloging of proteins from micro-organism to multi-celled organism, advances in protein and peptide separation, detection and identification necessitated a separate database for *B. mori* proteome. Eventually, silkworm proteome database (SPD) is constructed using Make2D-DB II software and annotated with a program image master (Amersham Pharmacia). It hosts tissue specific proteome (silk glands, midgut, fat bodies, haemolymph, ovaries, and malpighian tubules) of fifth instar (day 3) larvae, which were identified and analyzed by tandem mass spectrometry (MS/MS). Of the different search engines, two search engines, Sequest (Thermoquest) and Mascot are employed for identification and analysis of MS data against protein data stored in the National Center for Biotechnology Information (NCBI) data bank (http://www.ncbi.nlm.nih.gov/) and amino-acid sequence data from silkworm ESTs. The SPD contain eight kinds of proteomic analysis and 1144 kinds of spots visualized on the gels that investigated by MS/MS analysis [18]. However, still many more protein need to be identified in *B. mori* and then SPD would be a magnificent platform for comparative proteome analysis among different silkworm strains and Lepidopterans.

Make2D-DB II has user friendly environment to create, convert, interconnect and keep up-to-date 2-DE information in SPD database. This tool offers the possibility to automatically update the data related to numerous external data resources in a highly consistent manner and dynamically interconnect the user to several remote databases or projects to form a virtual global database accessible from one single entry point. The SPD opened under KAIKO (**K**ey **A**ccess to SILKWORM Genome Database - http://kaiko2ddb.dna.affrc.go.jp/cgi-bin/search_2DDB.cgi) database is also accessible through SWISS-2DPAGE (http://world-2dpage.expasy.org/make2ddb/) for comparative proteome analysis among Lepidopterans.

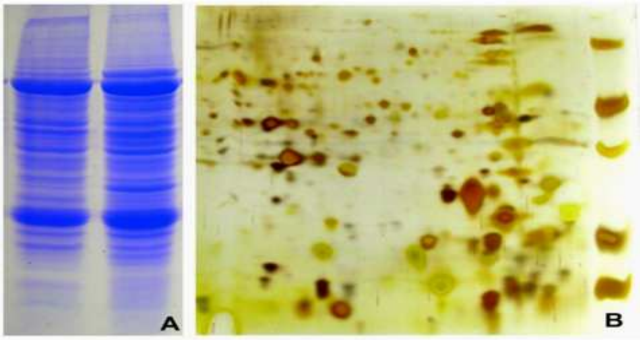

Figure 1. *The single (A) and two dimensional electrophoresis (B) gels explicit proteomic profile derived from the day 3 of fifth instar silkworm larvae of Bombyx mori strain NB4D2.*

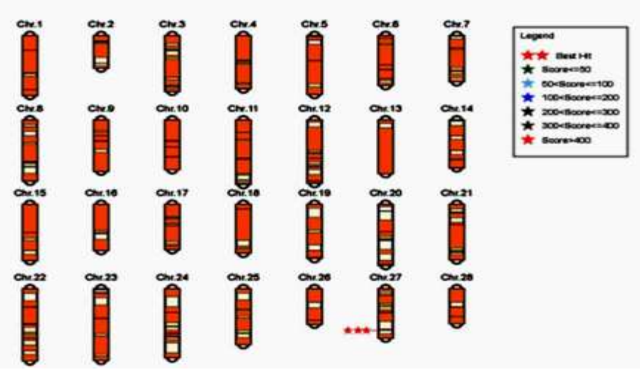

Figure 2. *Automated (based on mass spectrometry data) localization of Bombyx mori heat shock protein gene on the 27th chromosome.*

To date thousand of proteins have been identified in the whole organism [19,20; Fig.1] and various tissues, which includes fat body, midgut, hemolymph, prothoracic gland, colleterial gland, cuticle, intersegment muscle, and head of *B. mori* employing 1-DE (Fig.1A), 2-DE (Fig.1B), 1DE-LC-MS, MALDI-TOF-MS, and LTQ-Orbitrap mass spectrometer techniques. Although various bioinformatics tools and techniques MASCOT, PROTEIN PROSPECTOR, SEQUEST and X!Tandem have been used for identification and validation of these proteins against different databases but identification of the silkworm, *B. mori* specific proteins remain scarce. Although, based on MS data few proteins were identified against *Drosophila* database it did not yield same results against *B. mori* indicating protein diversity among insect groups, which offer a detailed investigation. Notably, for the first time, through a novel approach peptide mass fingerprint data used for identification of heat shock protein has been used for localization of a gene on the *B. mori* chromosome [20] through SilkMap tool (http://silkworm.swu.edu.cn/silkdb/). This novel approach is recently followed to map 87 common, 194 testis-specific and 113 ovary-specific protein genes on different chromosomes [21] envisaging the potential use of MS data for localization of protein encoding genes on the chromosome towards validation of silkworm genome map and protein-gene distribution on the chromosome (Fig. 2).

3.1.2. Expression And Post-Translational Modification

Further, for better understanding of gene expression and post-translational modifications a silkworm protein databank is constructed with 40 proteins derived from the silkworm body wall, fat body and middle intestine, which were separated by 2-DE and determined by the N-terminal amino acid sequencing. The N-terminal sequences of 27 proteins are registered in Swiss-Prot [22]. Having benefited from vital techniques such as N-terminal amino acid sequencing, MS-sequencing, WGS and functional genomics, more proteins have been identified in the domesticated silkworm. Although there is no exclusive protein database for silkworms but more than 2500 proteins have been registered to-date in protein databases, such as NCBI (http://www.ncbi.nlm.nih.gov/), Swiss-Prot (http://ca.expasy.org/sprot/) and the Protein Information Resource (http://pir.georgetown.edu/).

3.1.3. Gene Ontology and Biological Network

The exponential growth of bioinformatics with novel computational programs facilitated gene ontology (GO) annotation using protein sequence (*B. mori*) derived from MS through online BLAST search against the AmiGO database (http://amigo.geneontology.org/cgi-bin/amigo/blast.cgi). The analogous GO categories - biological process, molecular function and cellular component - extracted from the most homologous proteins using a Perl program successfully plotted with the help of Web Gene Ontology Annotation Plot (WEGO, http://wego.genomics.org.cn/cgi-bin/wego/index.). Further, the same protein sequences have been potentially used to establish biological and regulatory network of proteins using Pathway Studio software (version 7.0, Ariadne Genomics, Inc., Rockville, MD) against the *Drosophila* database, which contains the relationship of the protein interaction and the regulatory network [21,23]. Such advanced bioinformatic tools need to be employed to build a functional proteomic network for different tissues and the whole organism of *B. mori* and other allied species in the order Lepidoptera.

3.1.4. Protein Structure Modeling And Docking

Towards potential exploitation of proteins of interest in the field of biomedicine and biotechnology, MS derived amino acid sequence data shall/have been used to create structural modeling and associated ligand binding sites of a protein employing bioinformatics programs. In addition, the amino acids data of protein/s are also helpful in predicting the interaction between receptor protein and small molecules. To this, recently, amino acid sequence of *B. mori* serotonin receptor available in SWISSPROT is used to determine the physioco-chemical properties through computational tools. Further, three-dimensional structure of serotonin receptor is constructed using MODELLER and GROMOS96 programs. For interactive visualization and analysis of molecular structure established for serotonin receptor the PyMOL program is employed and validated with Ramachandran plot [24]. Realizing the significance, the MS derived amino acid sequence [20] and genomic sequence of BmHSP90 (NCBIn accession number GU324473) have been used to build a theoretical model using CPH models 3.0 web server and binding site for HSP90 inhibitor **-** Geldanamycin is determined through homology based docking studies, which is in progress in our lab (Rakesh et al - Unpublished). This *in silico* analysis opens ample prospects for functional characterization of proteins identified so far in *B. mori.* Thus, the exponential generation of proteomic data poses great challenges in the insect scientific community to detect functional properties of proteins and their biomedical and biotechnological significance.

3.2. Silkworm Genome Resources And Repositories

3.2.1. Expressed Sequence Tags (Ests)

Pursuance to the knowledge of protein profile and expression in the whole organism and different tissues, cDNA libraries are constructed for various tissues and different developmental stages to uncover and understand the entire set of *Bombyx* genes. But, efficient linkage has not been established so far. However, more than 185,000 ESTs derived from 36 cDNA libraries, which are grouped into ~11,000 non-redundant ESTs with an average length of 1.25 kb compared with FlyBase (http://www.fruitfly.org/) which revealed ~55% coverage of *Bombyx* genes stored in the EST database because the gene expression patterns deeply depend on tissues as well as developmental stages. This gap can be filled by considering the proteomics, gene expression and post-translational modification data available for *B. mori*. In addition, fraction of ESTs in each cDNA library indicates that it has not reached saturation and explicit the need of complete ESTs and its database covering all the genes. To tackle the saturation problem (to increase coverage and decrease the number of house keeping genes), subtraction and normalization methods are followed which result in a 5–11% increase in library-specific ESTs. Direct links of SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/index. cgi) with FlyBase (http://www.fruitfly.org/) and WormBase (http://www.wormbase.org) provide ready identification of Lepidoptera-specific genes [25,26] along with ESTs available for crop pests (214,834), beneficial insects (309,472), disease-causing pathogens (4,448) and *Drosophila* (821,005) species in the NCBI EST database (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/). Interestingly, this shows the inefficiency of sequencing non-normalized cDNA libraries to cover all the genes in *B. mori*. However, high EST redundancy provides us with single nucleotide polymorphisms (SNP) markers.

3.2.2. Single Nucleotide Polymorphisms (Snps)

To extract SNPs, 81,635 ESTs derived from 12 different cDNA libraries were used and of the total 12,980 contigs, 11,537 contigs assembled by PHRAP [25]. 101 candidate SNPs and 27 single base insertions/deletions were identified from 117 contigs assembled from 1576 high-quality reads with PHRED and screened on the basis of neighborhood quality standard (NQS). This analysis revealed that expressed sequences from multiple libraries may provide an abundant source of comparative reads to detect cSNPs (SNPs in coding regions) in the silkworm genome. Realizing the significance, ~16 million SNPs derived from 40 genomes of phenotypically and geographically diverse domesticated silkworm lines are used for the construction of a genetic variation map using SOAP program that revealed domestication events from Chinese wild silkworm, *B. mandarina* [27]. Similarly, development of SNPs for non-mulberry silkworms and Lepidopterans is in immediate need and these SNPs can be used as molecular markers for species differentiation and in livestock breeding program.

3.2.3. Microsatellites And Its Database

Microsatellites extracted from the available ESTs and genome sequence of *B. mori* are stored in a database – SilkSatDb (Table 1). This database was developed using PHP (hypertext preprocessor - a server side scripting language) which has simple and robust web-based search facility wherein users can retrieve the desired information on the microsatellites and the protocols used along with informative figures and polymorphism status. Information on primers and an interface coupled with an autoprimer, a primer-designing program facilitate the researcher to design the primers for the loci of interest [28]. However, the application of microsatellites in the silkworm strain improvement program although posses' limitation due to their wide distribution in the entire genome but a microsatellite tagged with respective and/or specific trait is warranted.

3.2.4. Silkworm Genome Sequence

B. mori genome sequence concurrently was determined by Chinese [11] and Japanese [12] groups following the Whole Genome Sequencing strategy (WGS) as has followed in *Drosophila* [29]. A draft sequence of *B. mori* genome covering 90.9% of all known genes with coverage of 5.9x was achieved from an inbred domesticated variety - Dazao [11]. The sequence data submitted to the DNA Data Bank (accession numbers AADK00000000, version AADK01000000**)** are accessible at http://silkworm.genomics.org.cn/ along with ESTs (Gen Bank accession numbers CK484630 to CK565104). Further, to determine similarities and differences in gene content among the fruit fly, mosquito, spider, and butterfly a gene-finder algorithm BGF (BGI Gene Finder) was developed based on *Genscan* and *FgeneSH*. Accordingly, the estimated gene count is 18,510, which exceeds the 13,379 genes reported in *D. melanogaster*. The total estimated genome size is 428.7 Mb or 3.6 and 1.54 times larger than that of the fruit fly and mosquito, respectively, including the unassembled read. The N50 contig and scaffold sizes are 12.5 and 26.9 Kb. Their assembly contains 90.9% of the 212 known silkworm genes (with full length cDNA sequence), 90.9% of ~16,425 EST clusters, and 82.7% of the 554 known genes from other lepidopterans. While 14.9% of the predicted genes confirmed by ESTs, 60.4% and 63.1% are similar to fruit fly genes and Gen Bank nonredundant proteins (*BlastP* at 10^{-6} E-value).

Concomitantly, a threefold (3x) shotgun sequence derived from another silkworm strain p50T of *B. mori* [12] was also submitted to the DNA Data Bank of Japan under accession numbers BAAB01000001 to BAAB01213289. Using the newly developed RAMEN assembler, the sequence data derived from WGS were assembled into 49,345 scaffolds that span a total length of 514 Mb including gaps and 387 Mb without gaps. Because the genome size of the silkworm is estimated to be 530 Mb, almost 97% of the genome has been organized in scaffolds, of which 75% has been sequenced. The validity of the sequence was elucidated for the majority of silkworm genes by a BLAST search for 50 characteristic *Bombyx* genes and 11,202 non-redundant ESTs in a *Bombyx* EST database against the WGS sequence data. The silkworm genome contains many repetitive sequences with an average length of <500 bp, and these repetitive sequences appear to have been derived from truncated transposons that are interspersed at 2.5-3.0 Kb intervals throughout the genome. 11 *Bombyx* gene orthologs to *Drosophila* genes controlling sex determination differ profoundly between the two species.

Apparently, the estimated genome size and the gene count that differs between two sets of sequence data for *B. mori* is due to the fact that the genomic data are derived from an inbred domesticated silkworm variety p50 (Dazao - the Beijing Genomic Institute, [11]) and the strain p50T (*Daizo* - the Silkworm Genome Research Program [12]), which diverged from each other about 30 years ago. Thus, the silkworm genome sequence has become fragmented due to the relatively shallow genome coverage using the WGS strategy making it difficult to identify and annotate the genes effectively [30]. To acquire a global view and surmount redundancy coupled with overlaps two sets of genome sequence are integrated that resulted in 10x assembly (http://silkworm.genomics.org.cn/).

The physical map of a genome denotes physical location of a gene or pool of genes distributed in different chromosomes providing a benchmark evaluating the accuracy of WGS assemblies. In *D. melanogaster*, of the five chromosomes (X, 2, 3, 4 and Y) BAC-based physical maps of chromosomes 2 and 3 constitute 81% of the genome. However, bacterial artificial chromosomes (BACs) contig based sequence and physical map of *B. mori* (Fig. 3) has not completely achieved, which contribute to more accurate sequence information as evidenced by the human genome map (International Human Genome Sequencing Consortium [31]). Towards this, construction of the silkworm genome map has been initiated using BAC clones following fluorescence in situ hybridization technique which has limitation due to acrocentric nature and smaller size of chromosomes in *B. mori* unlike human chromosome that are well differentiated based on chromatin pattern. Figure-3 depicts localization of BAC-14C11G carrying Sericin-1 and DH-PBAN gene on chromosome 11 which revealed some deviation in the position of the genes from that of linkage map (Figs. 3A and C). Figure-3B explicit the position of BAC-12B12C at the distal end of the chromosome 13 as demarcated in the respective physical map (Fig. 3D). Further, end sequencing of the two BAC libraries (an *Eco*RI- and a *Bam*HI-digested library) and *in silico* analysis were performed to characterize the sequenced data. The end sequences were deposited to GenBank with accession number DE283657-378560. The BAC end sequences (BESs) were used as queries in a BLAST similarity search of the two sets of WGS data. Interestingly, the BLAST hits of *Eco*RI and *Bam*HI BESs revealed differences, which reflect the abundance of repetitive sequences in the genome, and the possible cause of these differences, may be strain divergence [32]. The divergence observed between these two strains opened up ample scope for further investigation among the silkworm races/strains, which are geographically distinct. Eventually, 40 genomes of phenotypically and geographically diverse domesticated silkworm lines are used for sequencing to unveil domestication events from Chinese wild silkworm, *B. mandarina* [27] as has been achieved with whole genomes of 15 other species of *Drosophila* for comparative genomics (http://insects.eugenes.org/DroSpeGe/)[33].

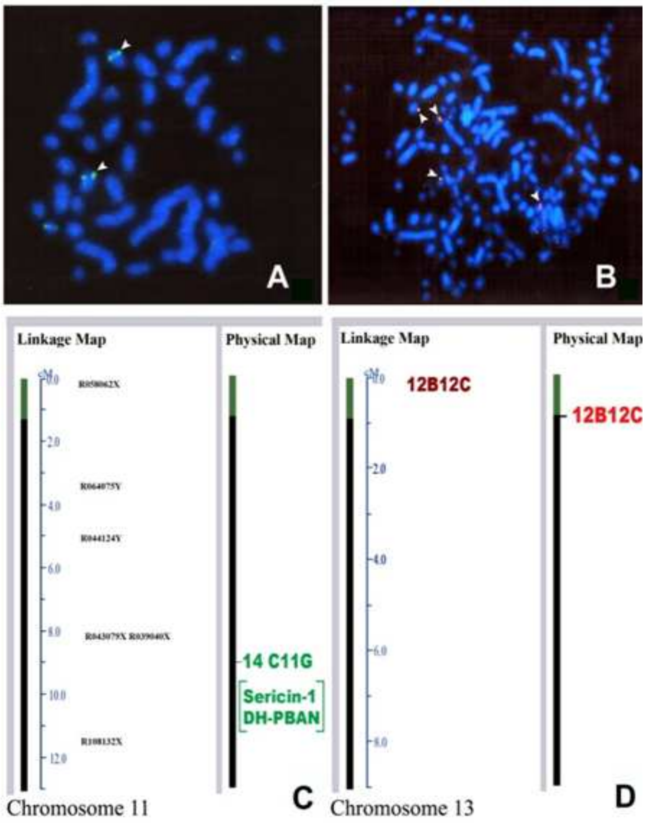

Figure 3. *Application of fluorescence in situ hybridization technique for localization of genes on the Bombyx mori chromosome. Two BAC clones labeled with FITC (A) and Texas red (B) hybridized on to the metaphase spreads revealed their position on chromosomes 11 (C) and 13 (D) respectively.*

Considering the global importance of DNA sequencing and availability of next generation sequencing (NGS) technologies - FLX454 (Roche), Solexa (Illumina) and SOLiD (Applied Biosystems) against conventional DNA sequencing - well known as di-deoxynucleotide sequencing or Sanger's method - which is time-consuming and labour intensive although it provides a large enough read length with quality sequence, a concerted efforts is being made to sequence the genome of several key pests and beneficial insect species involved with agriculture and human health [34].

3.2.5. Silkworm Genome And Knowledgebase

A silkworm genome database constructed based on the DYNACLUST system developed by DYNACOM Co. Ltd. that later updated with advanced programs (http://sgp.dna.affrc.go.jp/, http://silkbase.ab.a.u-tokyo.ac.jp/, http://papilio.ab.a.u-tokyo.ac.jp/genome/). These database hosts full length 40 cDNA libraries and ESTs representing various tissues derived from different stages and strains of *B. mori*. 26 different chromosome mutants along with their respective genetic details are available in the database. Further, KAIKO base (http://sgp.dna.affrc.go.jp/KAIKObase/) was constructed that integrate silkworm genome database having a data mining tool, 4 map browsers, one gene viewer and 3 independent databases. In addition, an integrated database (http://www.naas.go.kr/) developed using MYSQL and JAVA languages store silkworm gene resources in Oracle relational database management system (RDBMS) and ERWin Data Modeler software. It hosts 321 silkworm gene resources and 1,132 images depict the life cycle of various silkworm varieties [35].

The silkworm knowledge base (SilkDB-http://silkworm.genomics.org.cn/) is a web-based repository for the creation and integration for comparative analysis of genetic and genomic information. With 6x, SilkDB provides an integrated representation of large-scale, genome-wide sequence assembly, cDNAs, clusters of ESTs, transposable elements (TEs), mutants, SNPs and functional annotations of genes with assignments to InterPro domains and GO terms. It hosts a set of ESTs from *Bombyx mandarina*, a wild progenitor of *B. mori*, and a collection of genes from other Lepidoptera. Thus, a comparative analysis among domestic and wild silkworms, Lepidopterans and other insects (fruitfly, mosquito etc.) shall be performed on this platform using *B. mori* genome sequence as a reference framework. In addition, it also provides a comprehensive knowledge on silkworm, genome and related information in graphical ways for systematic comparative studies [36].

3.3. Bioinformatics Tools Employed In The Silkworm Genome Sequence Analysis

The RAMEN software program was employed for *B. mori* sequence analysis that basically follows the overlap layout consensus paradigm. Individual steps follow the novel or state-of-the-art software, such as look-up table, generation of seed strings for highly sensitive and rapid detection of overlapping reads, precise alignment by efficient banded dynamic programming, a repeat untangling method of transforming a repeat sub-contig flanked by two unique sub-contigs into one unique contig, and an efficient multiple alignment algorithm utilizing seeds in the look-up table. BLAST searches are carried out using BLASTN ver.2.1.2. and alignments made using the criteria of >95% identity and >50 bp in length. The coverage was calculated as the ratio of the total length of alignments in *B. mori* WGS sequence contigs to the length of the query sequence [12]. The whole *B. mori* sequence was analyzed and annotated using KAIKOBLAST and KAIKOGAAS.

KAIKOBLAST is a homology search system for analyzing silkworm sequence against nucleic acid and protein databases. It consists of nucleic acids and amino acids database search that utilizes the BLAST program for rapid search. The amino acid database search can be used for silkworm (p50T- Daizo strain) WGS sequence information. RAMEN assemble contigs and cDNA partial sequences from the KAIKO (cDNA database) confirmed gene set collection from WormBase; *D. melanogaster* full-length cDNA sequences from BDGP DGC; *D. melanogaster* EST sequences from BDGP; WGS contig, scaffold and partial cDNA sequences from the Southwest Agricultural University; silkworm cDNA partial sequences from Zhejiang Sci-Tech University; *Apis mellifera* and *A. cerana* cDNA partial sequences, and BAC End sequence data. It is also useful for accessing the WormBase amino acid sequence (http://www.wormbase.org), PIR, SwissProt and NCBI non-redundant protein databases (Table 2).

KAIKOGAAS is an automated annotation system designed for the analysis of silkworm genome that integrates programs for gene prediction and structural analysis of protein-coding regions. It includes coding region prediction programs (GENSCAN, FGENESH, MZEF), splice site prediction programs (SplicePredictor), DNA Sequence homology search analysis programs (Blast, HMMER, ProfileScan, MOTIF), a tRNA gene prediction program (tRNAscan-SE), repetitive DNA analysis programs (RepeatMasker, Printrepeats), a protein localization site prediction program (PSORT), and a membrane protein classification and secondary structure

Table 2. *Bioinformatics tools employed for construction of database and analysis of biomolecules in Mulberry and Silkworm.*

Task	Language and software used	Comments
Mulberry ESTs and DNA database	CAP3 assembly program	2,400 ESTs sequenced from the roots of *M. indica* are assembled into 148 contigs and 1,420 singletons.
Silkworm Proteome database	Make - 2DDB II SWISS-2DPAGE	Proteomes of seven major tissues of *Bombyx mori* investigated by MS/MS.
KAIKO base	Sequest & Mascot	Search engines for protein identification.
EST database	-	cDNA libraries and ESTs of *B. mori* and Lepidoptera-specific genes.
SNPs	PHRED & NQS	e-mining of single nucleotide polymorphisms from EST data of *B. mori*.
SilkSatDb	PHP, interfaced with autoprimer, primer design program	It is a relational database of microsatellites extracted from the available EST's and WGS of *B. mori*.
Genome sequence	BGI gene-finder (Genscan&FgeneSH)	Detection of genes in the genome sequence of *B. mori*.
	RAMEN-assembler	Software program - overlap layout, look-up table generation, overlapping reads, precise alignment by efficient banded dynamic programming, multiple alignment algorithm.
SGP database	DYNACLUST system	Used for construction of database.
	BLASTN or TBLASTN	Homology search.
	KAIKOBLAST	A homology search system for silkworm sequence against nucleic acid and protein databases.
	KAIKOGAAS	Automated annotation system designed specifically for analysis of *B.mori* genome.
	GENSCAN, FGENESH, MZEF	Coding region prediction programs.
	SplicePredictor	Splice site prediction programs
	Blast, HMMER, ProfileScan, MOTIF	DNA Sequence homology search analysis programs.
	tRNAscan-SE	tRNA gene prediction program.
	RepeatMasker & Printrepeats	Repetitive DNA analysis programs.
	PSORT	Protein localization and prediction program.
	SOSUI	Membrane protein classification and secondary structure prediction program
	AutoPredgeneset & Predgeneset	Annotation map
SilkDB	Web based	For curation, integration and

Task	Language and software used	Comments
	repository	study of silkworm genetic and genomic data.
Wildsilk	PAUP*	Molecular phylogeny of silk-producing insects.
	Sequencer	Sequence assembly and evaluation
	Clustal W	Manual sequences alignment
Wild Silkbase	BLAST	Searchable catalogue
	BLASTN	Compares a nucleotide query sequence against a nucleotide sequence dataset
	TBLASTX	Compares the six-frame translations of a DNA sequence to the six-frame translations of a nucleotide sequence dataset.
	TBLASTN	Compares a protein query sequence against a nucleotide sequence
Eri silkmoth database	BLAST search with BLASTN, TBLASTN, TBLASTX, BLASTX, and BLASTP programs	Comparative genome analysis
Butterfly and moth database	BLAST search	Comparative genome analysis

prediction program (SOSUI). Queries of any sequence can be performed on this web service since analysis tools and results viewer remain same as KAIKOBLAST.

The KAIKO annotation table facilitates browsing the results of annotation, which is created using KAIKOGAAS through links from the BAC-clone table and the assembled WGS contigs table. Only WGS contigs with lengths of more than 5 Kbp are sorted. In the assembled WGS contigs table, a check-mark in the ORF column indicates ORFs of autopredicted genes in the contigs. A check-mark in the BLASTp column and/or the BLASTn column indicates that the genes are the same and/or they resemble each other. The genes predicted by the system are shown in the "AutoPredgeneset" row on the annotation map. In 12L3 and 4L14 BAC clones, the predicted genes indicate the entries in GenBank and are shown in the "Predgeneset" row on the annotation map. Similarly, the positioning of all the BAC clones, which is incomplete todate, would present a complete gamet of *B. mori* genome as has been achieved in fruitfly and human genomes.

4. Non-Mulberry Silkworms and Their Databases (Wild Silk)

The non-mulberry silkworms include different species viz., *Antheraea mylitta* (Tropical tasar), *A. frithi*, *A. pernyi*, *A. roylei*, *A. proylei* (Oak tasar), *A. assama* (Muga), *Phylosamia ricini*, *Samia cynthia ricini*, (Eri), *Attacus atlas* (Fagaria) and *Gonometa postica* (Shashe). The silks produced by these insects - known as wildsilk - are commonly grown in India, China, Japan, Indonesia and Thailand for commercial purposes.

4.1. Molecular Phylogeny Of Wild Silkworms

Interestingly, bioinformatics tool PAUP* 4.01 version beta 10 [37] resolved the molecular phylogeny of silk-producing insects [38]. Towards this, DNA sequences of *A. atlas* and *B. mandarina* available in the NCBI database (AF463459 and AF395878) is used to understand the sequence evolution of the internal transcribed spacer DNA 1 (ITS 1) and phylogenetic analysis. The sequence assembly and evaluation is performed using the Sequencher (Gene Codes Corp.) and the BLAST search engine was used to determine the homologous sequences of ITS 1[39]. The sequences obtained are aligned manually and using ClustalW. Later, they were used for construction of phylogenetic trees based on maximum parsimony and maximum likelihood methods. Thus bioinformatics tools and molecular data resolved the taxonomic confusion, since a group of taxonomists assigned species status to *A. proylei*, whereas another group considered it as a hybrid (*A. pernyi* x *A. roylei*) [38]. So, based on sequence data and the structure divergence of ITS 1, *A. proylei* has been proved to be a recently derived species and ancestral position of the *A. assama*. This approach addresses the challenges exist in differentiating cryptic species of insects.

4.2. Wild Silkbase

Two databases constructed for non-mulberry silk moths are accessible at http://www.cdfd.org.in/wildsilkbase/ and http://www.ag.auburn.edu/enpl/hyche/saturniidae/. Wild silkbase basically a simple BLAST searchable catalogue of ESTs generated from several tissues of wild silk moths. It stores a total of 60,000 ESTs from three major wild silk moths, *A. assama*, *S. ricini* and *A. mylitta*. A BLAST searchable set of unigenes for all species are furnished along with cSNP discovery, homologue and SSR finder. This database also provides links to other web-resources on silk moths, EST construction, characterization and analysis methods and a tutorial. BLAST searches allow users to compare any query sequence to *A. assama*, *S. cynthia ricini* and *A. mylitta* EST sequence datasets. Further, bioinformatics tools - BLASTN (compares a nucleotide query sequence against a nucleotide dataset), TBLASTX (compares the six-frame translations of a DNA sequence to nucleotide sequence dataset), and TBLASTN (compares a protein query sequence against a nucleotide sequence dataset) dynamically translate all six reading frames.

Basically, the Eri silk moth database is housed in Silkbase (Table 1). This database has a link for a BLAST search with BLASTN, TBLASTN, TBLASTX, BLASTX, and BLASTP programs for sequence analysis and host only three cDNA libraries - IND09, IND10, and SMD06. Hence, complete Eri, Tasar and Muga silk moths databases are to be updated for comparative genome analysis among other sericigenous insects, moths and butterflies (Butterflybase - http://www.butterflybase.org/,http://butterflybase.ice.mpg/).

5. Conclusions

In insects, a considerable progress has been made through genomic and proteomic approach to understand how genes and proteins contribute to physiology, biochemistry, genetics and evolution, but, a complete picture of genomics and proteomics and their interactions is unclear. Meanwhile, rapid progress of bioinformatics uncovered such cryptic features of omics in other organisms leaving a wide gap in insects. Thus, research interest in *B. mori* has geared up in countries across the world, irrespective of whether they grow silkworm larvae for commercial cocoon production or not due to its tractability and the first Lepidopteran species to have its genome sequenced. Finding conserved synteny and gene order among Lepidoptera [40] envisages the usefulness of genomic and proteomic methods consolidated here (*B. mori*) for "comparative omics" with allied species.

Furthermore, exponential generation of proteomic and genomic data with the help bioinformatics would unravel the conventional one gene and one protein concept while the *B. mori* and *Drosophila* have ~9,000 and ~22,000 proteins reported against ~18,000 and ~16,000 genes, which represents half and more than 37% of the genes identified respectively.

In this context, the resources documented in this review are preparatory references for integration of proteomics and genomics from the point of biotechnological and biomedical applications. But, there is much more to explore and the science of "Seri-bioinformatics" is beginning to pick those threads to obtain a global view with its wide-ranging and unique features.

Acknowledgements

Ms. Punyavathi is thankful to Department of Science and Technology, New Delhi for award of DST-INSPIRE fellowship (IF-120339) and H.B.M. thanks to UGC, New Delhi for grant of project under F.No.37-458/2009 (SR).

References

[1] F. Sanger and H. Tuppy, "The amino acid sequence of the phenylalanyl chain of insulin", Biochem. J. 1951, vol. 49, pp. 481-490.

[2] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA", Proc. Natl. Acad. Sci. 1977, vol. 74, pp. 560-564.

[3] F. Sanger, S. Nicklen and A. R. Coulson, "DNA sequencing with chain terminating inhibitors", Proc. Natl. Acad. Sci. 1977, vol. 74, pp. 5463-5467.

[4] M. O. Dayhoff, (Ed). "Atlas of protein sequence and structure, Survey of new data and computer methods of analysis. In Atlas of protein sequence and structure", National Biomedical Research Foundation, Georgetown University, Washington D.C. vol.5, 1979.

[5] D. D. Womble, "GCG The Wisconsin package sequence analysis program", Methods Mol. Biol. 2000, vol.132, pp. 3-22.

[6] B. Ewing, *et al*., "Base-calling of automated sequencer traces using *Phred* I. Accuracy assessment*",* Genome Res. 1998, vol.8, pp.175-185.

[7] D. Gordon, C. Abajian, and P. Green, "*Consed* - A graphical tool for sequence finishing", Genome Res. 1998, vol.8, pp.195-202.

[8] W. R. Pearson, and D. J. Lipman, "Improved tools for biological sequence comparison", Proc. Natl. Acad. Sci. 1988, vol.85, pp.2444-2448.

[9] S. F. Altschul, W. Gish, E. W. Miller, and D. J. Lipman,"Basic local alignment search tool", J. Mol. Biol. 1990, vol.215, pp.403-410.

[10] J. M. Cherry and S. W. Cartinhour, "AceDB, a tool for biological information. In: Adams, M., *et al*. (Eds.), Automated DNA sequencing and analysis. Academic Press, New York, 1993.

[11] Q. Xia, *et al*., "A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*)", Science, 2004, vol.306, pp.1937-1940.

[12] K. Mita, *et al*., "The genome sequence of silkworm *Bombyx mori*", DNA Res. 2004, vol.1, pp.27-35.

[13] M. Tomita, *et al*., "Transgenic silkworms produce recombinant human type III procollagen in cocoons", Nature Biotech. 2003, vol.21, pp.52-56.

[14] A. Asha, S. Sriram, S. Sehrawat, M. Rahman, D. Sehgal and K. P. Gopinathan, "*Bombyx mori* nucleopolyhedrovirus: Molecular biology and biotechnological application for large scale synthesis of recombinant proteins", Current Sci. 2002, vol.83, pp.455-465.

[15] Sudhakumari, Punyavathi, Chhanda Das, M. A. Bhat, H. B. Manjunatha, "Evaluation of the medically important compounds TASKI Protasan and Combatan for its efficacy using *Bombyx mori* as a model system" , J. Pharm. Res. 2013, vol.7, pp.184-188.

[16] V. Ravi, J. P. Khurana, A. K. Tyagi, and P. Khurana, "The chloroplast genome of mulberry (*Morus indica* cv.K2): complete nucleotide sequence, gene organization and comparative analysis", Tree Gen. Geno. 2006, vol.3, pp.49-59.

[17] V. G. Checker, B. Saeed, P. Khurana, "Analysis of expressed sequence tags from mulberry (*Morus indica*) roots and implications for comparative transcriptomics and marker identification", Tree Gen.Genom. 2012, vol.8, pp.1437-1450.

[18] H. Kajiwara, *et al*., "Draft of proteome database", J. Electrophoresis, 2006, vol.50, pp.39-41.

[19] B. C. Vasudha, H. S Aparna, and H. B. Manjunatha, "Impact of heat shock on heat shock proteins expression, biological and commercial traits of *Bombyx mori*", Insect Sci. 2006, vol.13, pp.243-250.

[20] H. S. Aparna, R. R. Kundapur, and H. B. Manjunatha, "Molecular characterization of heat shock proteins 90 (HSP83?) and 70 in tropical strains of *Bombyx mori*", Proteomics, 2010, vol.10, pp.2734-2745.

[21] J. Chen *et al.,* "Proteome Analysis of Silkworm, *Bombyx mori,* Larval Gonads: Characterization of Proteins Involved in Sexual Dimorphism and Gametogenesis", J. Proteome Res. 2013, vol.12, pp. 2422-38.

[22] B. X. Zhong, "Protein databank for several tissues derived from five instar of silkworm", Acta Genetica Sin. 2001, vol.28, pp.217-224.

[23] Q. Fu *et al., "*Proteomics analysis of larval integument, trachea and adult scale from the silkworm, *Bombyx mori*", Proteomics, 2011, vol.11, pp.3761-3767.

[24] R. Sumathy, S. K. Ashwath and V. K. Gopalakrishan, "Theoretical modeling and docking studies of silkworm Serotonin receptor", J. Proteom. Bioinform. 2012, vol.5, pp.230-234.

[25] T. C. Cheng, *et al*., "Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao", Insect Biochem. Mol. Biol. 2004, vol.34, pp.523-530.

[26] K. Mita *et al.,* "The construction of an EST database Bombyx mori and its application", Proc. Natl. Acads. Sci. 2003, vol.100, pp.14121-14126.

[27] Q. Xia et al., "Complete resequencing of 40 genomes reveals domestication events and genes in silkworm", Science, 2009, vol.326, pp. 433-436.

[28] M. D. Prasad, *et al.*, "SilkSatDb: A microsatellite database of the silkworm, *Bombyx mori",* Nucleic Acids Res. 2005, vol.33, pp.403-406.

[29] M. D. Adams *et al.,* "The genome sequence of *Drosophila melanogaster"* Science, 2000, vol.287, pp.2185–2195.

[30] K. Yamamoto, *et al.,* "A BAC-based integrated linkage map of the silkworm, *Bombyx mori*", Genome Biol. 2008, vol. 9, R21.

[31] International Human Genome Sequencing Consortium. Nature, 2001,vol. 409, pp. 860-933.

[32] Y. Suetsugu, *et al., "*End Sequencing and characterization of silkworm (*Bombyx mori*) bacterial artificial chromosome libraries", BMC Genomics, 2007, vol.8, pp.314.

[33] M. A. Crosby, J. L. Goodman, V. B. Strelets, P. Zhang and W. M. Gelbart, "FlyBase: genomes by the dozen", Nucleic Acids Res. 2007, vol.35, pp.486–491.

[34] P. Chilana, A. Sharma and A. Rai, "Insect genomic resources: status, availability and future", Current Sci. 2012, vol.102, pp. 571-580.

[35] C. Kim *et al.,* "An integrated database for the enhanced identification of silkworm gene resources", Bioinformation, 2010, vol. 4 , pp.436-437.

[36] J. Wang *et al.,* "SilkDB- a knowledge base for silkworm biology and genomics", Nucleic Acids Res. 2005, vol.33, pp.399-402.

[37] D. L. Swofford, "PAUP*. Phylogenetic analysis using parsimony (* and other methods)", Sinauer Assoc, Sunderland, Massachusetts. 2003, Ver. 4.0 beta.

[38] B. Mahendran, S. K. Ghosh, and S. C. Kundu, "Molecular phylogeny of silk producing insects based on internal transcribed spacer DNA1", J. Biochem. Mol. Biol. 2006, vol.39, pp.522-529.

[39] S. F. Altschul, *et al.,* "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 1997, vol. 25, pp.3389-402.

[40] E. G. Pringle, *et al*., "Synteny and chromosome evolution in the Lepidoptera: Evidence from mapping in *Heliconius melpomene*", Genetics, 2007, vol.177, pp.417-426.