Insect genomic resources: status, availability and future

Poonam Chilana, Anu Sharma and Anil Rai*

Recent advances in biotechnology have led to the development and evolution in the field of bioinformatics for the analysis and integration of information from genomic, transcriptomic, proteomic, metabolomic and phenomic data. Availability of whole genome sequences, expressed sequence tags, genetic linkage maps and insect transgenesis has opened up new vistas for fundamental research in entomology. This article describes the applications of bioinformatics in applied insect science and pest management. Details of insect genomes sequenced and published, and the available genomic databases along with their features are also discussed. Considering current developments in insect biotechnology and available insect genomic resources, future aspects of bioinformatics applications in insect science have been highlighted.

Keywords: Bioinformatics, bio-rational control, databases, insect genomes, pest management.

EXPONENTIAL amount of information has been generated in the field of insect science due to latest DNA sequencing technology. These data overload have, in turn, led to an absolute requirement for computerized databases to store, organize and index the data, and for specialized tools to view and analyse the data. Bioinformatics is the study of biological information using concepts and methods in computer science, statistics and engineering.

Insects constitute a remarkably diverse and largest animal group in the world, as 75% of all species are insects[1]. Insects are ecologically and economically important, as they provide an amazing diversity from being highly beneficial to harmful pests. Harmful insects can be severe agricultural pests destroying up to 30% of our potential annual harvest, and vectors for plant and human diseases such as yellow mosaic, wilt, leaf curl in plants, and malaria, elephantiasis, sleeping sickness, dengue and yellow fever in humans. Beneficial insects not only help in crop pollination and crop protection, but also provide useful materials like silk and honey to mankind.

The whole genome sequence is an invaluable resource for the insect genomic community, that allows functional genomics, comparative analysis of genomic contents and their organization, as well as functional analyses of critical parameters as insect attributes linked to their capacity to transmit disease agents, insect behaviour, the ancestral relationships between major insect groups, as well as a better understanding of many individual genes and gene families.

The authors are in the Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi 110 012, India.
*For correspondence. (e-mail: anu@iasri.res.in)

Drosophila melanogaster serves as a model system for animal and insect genetics. *D. melanogaster* is a species that is extensively studied to understand any particular biological phenomenon because it has characteristics that make it suitable for availability and traceability. A large amount of information is available from *Drosophila* that provides valuable data for the analysis of gene regulation, genetic diseases and evolutionary processes[2]. *Drosophila* research has provided insights into genetics, behaviour, development and disease systems. The *D. melanogaster* genome sequencing project was essentially completed in March 2000. *Drosophila* genome encodes approximately 15,016 genes, fewer than the smaller *Caenorhabditis elegans* genome, but with comparable functional diversity.

Advances in sequencing technologies have provided opportunities in bioinformatics for managing, processing and analysing the sequences. In this genomic era, bioinformatics is used as a bedrock of current and future biotechnology for finding new or better alternatives as designing potential target sites, safer insecticides and developing transgenic insects in applied insect science. The objective of this article is to provide comprehensive information on available insect genomic resources at one place to biotechnologists, molecular biologists, entomologists and physiologists for developing new methods in pest and disease management.

Sequencing of insect genomes

Rapid developments in genome sequencing are transforming insect biology with new avenues in the area of insect science. The genome is the set of all genetic

Table 1. Insect genomes sequenced as on July 2011

Order	Insect: Common name	Insect: Scientific name	Chromosome no.	No. of genes	No. of proteins	Genome size (Mb)	Release date	Centre/Consortium
Diptera	Fruit fly	*Drosophila melanogaster*	4	15,016	22,352	180.00	03/25/2000	Flybase Consortium, BDGP, Celera Genomics
		Drosophila pseudoobscura pseudoobscura	2	16,731	16,071	198.43	12/24/2003	Flybase
		Drosophila yakuba	6	16,891	16,082	281.98	07/07/2004	Flybase
		Drosophila simulans	7	–	–	130.98	03/23/2005	The Drosophila Simulans Sequencing Consortium
		Drosophila persimilis	5	17,573	16,878	175.58	09/28/2005	Broad Institute
		Drosophila sechellia	4	17,273	16,471	157.24	09/28/2005	Broad Institute
		Drosophila mojavensis	6	15,179	14,525	180.21	04/11/2006	Agencourt Bioscience Corporation
		Drosophila erecta	4	59	29	145.08	04/11/2006	Agencourt Bioscience Corporation
		Drosophila grimshawi	6	15,585	14,986	186.09	04/11/2006	Agencourt Bioscience Corporation
		Drosophila virilis	6	15,343	14,491	189.20	04/11/2006	Agencourt Bioscience Corporation
		Drosophila ananassae	4	–	–	213.92	04/11/2006	Agencourt Bioscience Corporation
		Drosophila willistoni	3	16,385	15,513	224.52	04/21/2006	J. Craig Venter Institute
		Drosophila elegans	–	–	–	170.52	07/14/2011	Baylor College of Medicine
		Drosophila takahashii	–	–	–	–	07/14/2011	Baylor College of Medicine
		Drosophila ficusphila	–	–	–	–	07/14/2011	Baylor College of Medicine
		Drosophila kikkawai	–	–	–	–	07/14/2011	Baylor College of Medicine
	Hessian fly	*Mayetiola destructor*	–	–	–	150.00	09/29/2010	Baylor College of Medicine
	Mosquito	*Anopheles gambiae*	5	13,227	14,086	560.00	03/22/2002	The International Consortium for the Sequencing of Anopheles Genome
		Aedes aegypti	3	16,684	16,785	259.06	02/11/2005	http://www.vectorbase.org/
		Culex quinquefasciatus	3	–	–	540.00	04/19/2007	Broad Institute
Lepidoptera	Silkworm	*Bombyx mori*	28	18,500	–	432.00	04/23/2008	The International Silkworm Genome Sequencing Consortium
Coleoptera	Flour beetle	*Tribolium castaneum*	10	10,119	9,820	339.00	08/17/2005	Baylor College of Medicine
Hymenoptera	Honey bee	*Apis mellifera*	16	11,156	10,562	218.00	12/19/2003	Human Genome Sequencing Center
		Apis florea	–	–	–	280.00	12/10/2010	Baylor College of Medicine
	Leaf-cutting bee	*Megachile rotundata*	–	–	–	250.00	07/15/2011	University of Maryland, BCUI
Hymenoptera	Bumble bee	*Bombus terrestris*	18	10,178	10,573	218	10/12/2004	Baylor College of Medicine
		Bombus impatiens	–	–	–	250	01/31/2011	Baylor College of Medicine
		Nasonia vitripennis	5	12,119	12,988	432.65	05/04/2007	Baylor College of Medicine
		Nasonia giraulti	5	–	–	180.08	11/10/2009	Baylor College of Medicine
		Nasonia longicornis	5	–	–	182.99	11/11/2009	Baylor College of Medicine
	Ants	*Camponotus floridanus*	–	16,440	14,864	220	08/27/2010	BGI–Shenzhen, China
		Atta cephalotes	–	–	–	290.02	05/11/2010	Genome Sequencing Center, Washington University
		Harpegnathos saltator	–	17,624	15,127	280	08/27/2010	BGI–Shenzhen, China
		Solenopsis invicta	–	–	–	484	02/03/2011	Université de Lausanne
		Linepithema humile		–	–	251	06/07/2011	The Ant Genomics Consortium
		Pogonomyrmex barbatus	16	–	–	250	02/04/2011	The Ant Genomics Consortium
		Acromyrmex echinatior	–	–	–	300	04/14/2011	Beijing Genomics Institute
Phthiraptera	Body louse	*Pediculus humanus*	10	10,993	10,775	108.37	04/18/2007	The Human Body Louse Genome Consortium
Hemiptera	Pea aphid	*Acyrthosiphon pisum*	4	–	–	446.6	04/01/2008	Baylor College of Medicine
	Blood-sucking bug	*Rhodnius prolixus*	11	–	–	568.77	06/17/2009	Genome Sequencing Center (GSC), Washington University

Source: http://www.ncbi.nlm.nih.gov/genome/

information of an organism encoded in the deoxyribose nucleic acid (DNA) of the nucleus and organelles. The genome contains regions that code for specific proteins, i.e. genes and non-coding regions, including some with structural and regulatory functions. In recent years, the DNA sequencing boom has made it economically feasible to obtain the entire sequence of an organism's genome, and has opened the door for the establishment of many publicly or privately funded insect genome projects. In this section, we describe the status of insect genomes sequenced so far in chronological order. A total of 39 insect genomes have been sequenced to date and numerous other insect species are in the process of being sequenced. All the insect genomes which have been sequenced and published are presented in Table 1.

Drosophila sequencing status

The first insect with published genome is *D. melanogaster*, commonly called fruit fly, the most-studied eukaryotic genome and the most-prominent model organism in molecular biology[3]. The genome of *D. melanogaster* was sequenced using a whole genome sequence (WGS) approach. The first assembly (WGS1) used only plasmid and bacterial artificial chromosome (BAC) paired-end sequences, and the second added BAC and P1-based finished and draft sequences, and then the joint assembly was submitted to GenBank as Release 1. This sequence contained many gaps and regions of low sequence quality. A second release, Release 2, corrected some errors in the order and orientation of small scaffolds present in Release 1, and filled a few hundred very small sequence gaps. Using improved WGS sequence-assembly algorithms, two additional assemblies of the WGS plasmid and BAC paired-end sequences used in WGS1 were generated in March 2001 (WGS2) and July 2002 (WGS3), roughly coinciding with the WGS assemblies of the human and mouse genomes respectively. Release 3 was generated to improve Release 2 by closing all the gaps, improving regions of low sequence quality, and extending the sequence at the telomeric and centromeric ends of each chromosome. The Release 3 euchromatic genome sequence has been reannotated using a new annotation tool, Apollo, and the complete reannotated improved genomic sequence of *Drosophila* with new expressed sequence tags (ESTs) and complementary DNA (cDNA) was then deposited in GenBank.

The improvements made to the genomic sequence in Release 3 had a large impact on the annotation of transposable elements because of the substantial corrections made in the assembly of repeated sequences. Release 3 provided a euchromatic sequence of good quality, gap-free and of high accuracy[4]. Release 4 and Release 5 are now available with Flybase and GenBank. They are considered to be of sufficient accuracy and declared to be substantially complete and support an initial analysis of genome structure and preliminary gene annotation and interpretation.

A physical map of a chromosome or a genome shows the physical locations of genes and other DNA sequences of interest. The physical map provided a benchmark for evaluating the accuracy of WGS assemblies. In *D. melanogaster*, there are five chromosomes (*X*, 2, 3, 4 and *Y*). BAC-based physical maps of chromosomes 2 and 3 of *D. melanogaster* constitute 81% of the genome. Sequence tagged site (STS) content, restriction fingerprinting, and polytene chromosome *in situ* hybridization approaches were integrated to produce a map spanning the euchromatin[5]. Major computational challenge in the construction of physical maps is to track the multiple names of markers. Similarly, multiplicity of names of a single gene, its associativity with one or more EST clusters, polymorphisms and STSs are complexities in the construction of physical maps. Further, lack of precise knowledge of markers, genes and genomic elements makes mapping more difficult.

Whole genomes of 15 other species of *Drosophila* (*ananassae*, *erecta*, *elegans*, *grimshawi*, *mojavensis*, *persimilis*, *pseudoobscura pseudoobscura*, *sechellia*, *simulans*, *virilis*, *willistoni*, *takahashii*, *ficusphila*, *kikkawai* and *yakuba*) have been sequenced and are accessible for comparative genomics through the internet[6] (http://insects.eugenes.org/DroSpeGe/). After the sequencing of *Drosophila*, much effort was focused on economically important insects of three categories: agriculturally important pests (flour beetle, aphid), beneficial insects (honey bee, parasitic wasp and silkworm) and vectors of plant and animal diseases (mosquitoes, body louse, blood-sucking bug, aphids, etc.). Mosquitoes are the second most studied insects after fruit flies, as these are vectors of some of the deadliest human diseases. Publication of the *Plasmodium falciparum* and *Anopheles gambiae* genome sequences has once again paved the way to find a permanent solution to the malaria problem[7,8]. The sequencing of *Culex pipiens*, the mosquito vector for the West Nile virus, shed further light on mosquito biology and mosquito species-specific gene functions[9]. Next to mosquito genome, an important lepidopteran silkworm has joined this group of fully sequenced insects. The silkworm, *Bombyx mori*, serves as a central model organism for the lepidoptera genomics and facilitates studies of comparative genomics and basic research leading toward new genome-based approaches for sericulture and pest control[10]. Within Hymenoptera, an international genomics effort has sequenced the honey bee (*Apis mellifera*), an economically important member of this group[11]. The genome sequences of other medically important body louse, *Pediculus humanus* (relapsing fever, trench fever and epidemic typhus) and *Rhodnius prolixus* (Chagas disease) are also available[12]. The genomes of two important agricultural pests, red flour beetle (*Tribolium castaneum*)

destroying stored grain for human consumption and the pea aphid, *Acyrthosiphon pisum*, causing severe damage to green food plants have also been sequenced[1,13]. Parasitoids are important regulators of major agricultural pests and recently *Nasonia* wasp has emerged as a genetic model for pest management genetics. The genome sequences of three closely related parasitoid wasps, i.e. *N. vitripennis*, *N. giraulti*, and *N. longicornis* have been sequenced and published[14]. The recently sequenced genomes of the Carpenter ant (*Camponotus floridanus*), Argentine ant (*Linepithema humile*), red harvester ant (*Pogonomyrmex barbatus*), leaf-cutter ant (*Atta cephalotes*), jumper ant (*Harpegnathos saltator*) and fire ant (*Solenopsis invicta*) have shown that embryos with the same genetic code develop into either queens or worker ants, and may advance our understanding of invasion biology and pest control[11,14]. The genome from the important polyphagous lepidopteran pest, tobacco budworm (*Heliothis virescens* Fab.) has already been fully sequenced, but these data are not publicly accessible[1]. Recently, the Wellcome Trust Sanger Institute, England is also engaged in the whole genome shotgun sequencing of tsetse fly (*Glossina morsitans morsitans*) (http://www.sanger.ac.uk/Projects/G_morsitans/). Many more projects are still underway for sequencing insects by different institute. Along with conventional methods, these projects are well equipped with advanced sequencing tools, to ensure maximum coverage with high-quality sequence and cost-efficient methodology. The conventional DNA sequencing method referred to as di-deoxynucleotide sequencing, or more commonly, Sanger's method of DNA sequencing, provides a large enough read length with quality sequence, but it is time-consuming and labour-intensive. With the availability of next generation sequencing (NGS) technologies for DNA sequencing, like FLX454 (Roche), Solexa (Illumina) and SOLiD (Applied Biosystems), there has been tremendous increase in the sequence database of several insects. In the near future, many more insect genomes will be sequenced as organizations involved with human health and agriculture have also developed plans to sequence several key pests and beneficial insect species.

Gene annotation in insects

Gene annotation is a process during which biological information is attached to the sequence, for example, positions of protein-coding genes, their coding regions and their regulatory elements along with the putative function of each gene. Computer algorithms are used for the identification of structural elements within the genome. Gene annotation helps in comparing the insect genomes with one another and with those of other organisms. These results are helpful in understanding the evolution of insects, the phylogenetic relationships among different orders and the molecular underpinnings of insect-specific processes. The insect genomes sequenced to date differ considerably in terms of size and gene number. The *Drosophila* sequence has been extensively annotated and a wealth of information is available about genomic organization, development, cell biology, neurobiology, behaviour and its evolution. Comparing this with human sequences suggest that the *Drosophila* coding genome is more similar to the human genome than those of yeast and nematode. The annotation summary of *D. melanogaster* is presented in Table 2. The honey bee genome appears to have evolved less rapidly than those of the fruit fly and mosquito, and it displays less similarity, for certain groups of genes, with the latter two insect genomes than with those of vertebrates[11].

Direct and indirect approaches have been used individually or in combination to identify the functions of insect genes[15]. Most indirect approaches involve the quantification of gene expression at the level of either transcript mRNA or of proteins for finding clues about the role played by a given gene product. In developmental studies, a gene that is expressed only at the onset of metamorphosis such as broad complex may be hypothesized to play a role in that process[16]. Similarly, a gene that is expressed in only one gland or tissue, such as juvenile hormone acid methyl transferase, is likely to have a function that is restricted to the metabolism of that gland or tissue[17]. Various other gene arrays have been generated for insects, including *A. mellifera*, *B. mori*, *Spodoptera frugiperda* and *Choristoneura fumiferana*, typically from cDNAs, and are used to study the differences between bee larvae raised as workers or as queens[18], the changes in gene expression during silkworm metamorphosis[19], and the modulation of gene expression following infection with a polydnavirus or with wild type and recombinant baculoviruses.

A more direct approach to the study of gene function involves the disruption of individual genes. The genomes of some microorganisms, including viruses, bacteria and yeast, can be manipulated with relative ease to generate mutants that display single-gene defects whose phenotypes allow researchers to infer the function of the disrupted gene. The application of a similar approach to more complex organisms such as insects, however, presents a greater challenge. A detailed comparison between gene content in six insect species (*A. mellifera*, *D. melanogaster*, *A. gambiae*, *T. castaneum*, *B. mori* and the migratory locust, *Locusta migratoria*) and in three non-insect eukaryotic organisms (yeast, nematode and human) led to the finding that the best represented insect-specific proteins are those associated with stress and stimulus response with cuticle formation and with pheromone or odour perception[20]. Odorant receptors are well represented in the mosquito for host seeking, whereas these type of proteins are altogether absent in the fruit fly[8]. Pheromone or odour receptor proteins are even more

Table 2. Annotation summary of *Drosophila melanogaster*[26]

Data type	Release 2 (2000)	Release 3.1 (2002)	Release 3.2b (2004)	Release 5.1 (2006)
Annotated sequence (Mb)	3.8	12.1	14.2	24
Sequence length of repeats (Mb/%)	ND	6.3/52	6.3/75	18/77
Sequence length of exons (Mb/%)	0.15/4	0.33/2.7	0.43/3.0	1.33/5.5
Repeat nest fragments (number/Mb)	ND	ND	ND	10084/10
Full-length transposable elements (TEs)	ND	ND	ND	202
Total annotations	130	447	556	11038
Protein-coding genes	130	297	472	613
Single-exon genes	43	58	195	187
Genes with finished cDNAs	48	58	92	137
Protein-coding genes with any EST/cDNA clone evidence	ND	80	142	250
Pseudogenes	0	1	7	32
ncRNAs	0	3	14	13
Recursive splice sites	ND	ND	ND	16
Miscellaneous annotations	ND	ND	ND	9
Unassembled ribsomal DNA fragments	0	6	52	67

ND, not done.

abundant in the honey bee. These proteins are hypothesized to mediate the insect's perception of pheromone blends, kin recognition signals, and diverse floral odours. The bee genome contains novel genes associated with nectar and pollen utilization. Other genes associated with innate immunity and with the detoxification of xenobiotics are less represented in the honey bee than in the two dipterans (*D. melanogaster* and *A. gambiae*)[11]. The silkworm genome contains an estimated 1793 genes for silk production, immunity, development and pheromone production, which were not found in the fruit fly or mosquito[21]. This comparative genomics approach not only helps understand the basic processes involved in insect biology but also leads to the identification of many insect or pest-specific proteins that may be potential targets for insecticide development.

EST resources in insects

EST sequencing represents an efficient alternative to whole genome sequencing by providing information of the most expressed parts of the genes at a lower cost. It is also called gene signature, which helps in cloning and characterization of full-length genes. Researchers are putting efforts to construct EST libraries for insect species, which contain collections of cDNA sequences derived from expressed genes only. With the development of ESTs in several insect species, a lot of DNA sequence information has been produced across species and deposited in on-line databases. In the NCBI EST database (dbEST; www.ncbi.nlm.nih.gov/dbEST/), there are up to 214,834, 309,472, 4,448 and 821,005 ESTs available from crop pests, beneficial insects, disease-causing pathogens and *Drosophila* species respectively, as shown in Figure 1. EST projects on *S. frugiperda*[22] and *C. fumiferana* (http://pestgenomics.org) strongly focus on important lepidopteran pests of agriculture and forestry. Many more EST projects are underway in various laboratories.

Insect genome databases

Traditionally entomologist relies on textbooks and research articles as major resources for 'omics' information on insects. With the advancements in molecular biology and genomics technologies in the insect domain, lot of genomic data have been generated in the past decade. This deluge of genomic information has led to an absolute requirement for computerized database to store, organize and index the data along with development of specialized tools to view and analyse the data. Genomic approaches are now becoming an important step for new developments in the biology and pathology of insects. The insect genomic databases contain information of all proteins, biochemical and physiological processes that occur in an insect. A detailed review of databases developed on important insects has been presented in Table 3.

Recently, efforts have begun to develop a comprehensive sequence database named Agricultural Pest Genomics Resources (Agripestbase) for storing genomic information from a broad range of pests, including insects, parasites and pathogens. Availability of genomic information on a broad range of agricultural pests will result in comparative genomics and further understanding of these species (http://agripestbase.org/). The information available from these databases will hopefully lead to better management strategies as well as new methods and targets for pest control.

Bioinformatics in applied insect science and pest management

Bioinformatics is applied in pest management in different ways, which are as follows:

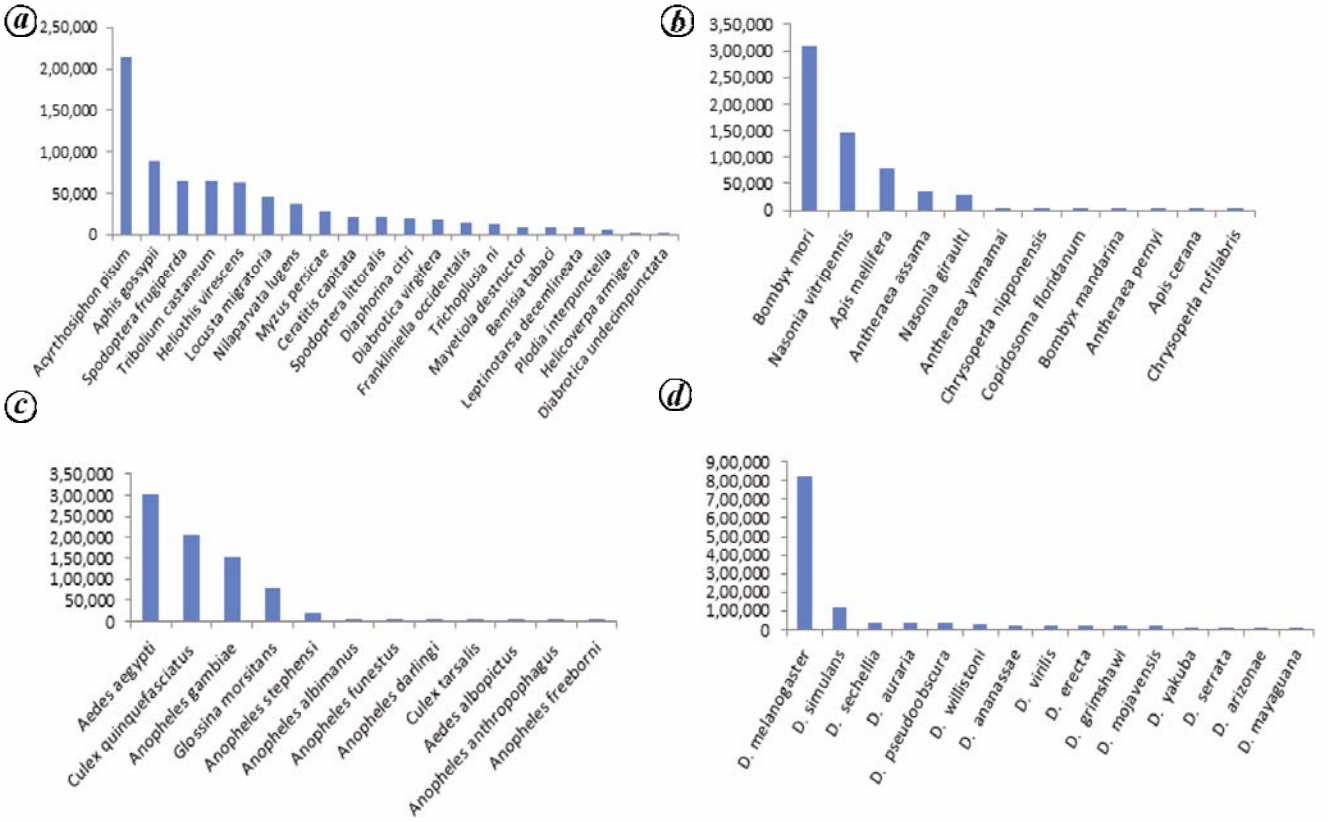

Figure 1. Expressed sequence tags available in the public domain. ***a***, Important crop pests. ***b***, Beneficial insects. ***c***, Pathogens. ***d***, *Drosophila* species. (Source: http://www.ncbi.nlm.nih.gov/dbEST/)

Insect transgenesis

Biotechnology is providing modern improvements and a range of new tools for population control of insects in crop protection. Genetic transformation of insects is another technique that will greatly affect the future role of genomics in applied entomology. Transformations may be used for gene identification and characterization and for creating strains with genes encoding lethality or sterility. Transgenic strains may be created to improve existing biocontrol programmes such as sterile insect management technique, or potentially allow new, highly efficient control strategies. The improved understanding of genome sequencing of the pest insect will stimulate the design of new classes of transgene microorganisms to be used in pest control. Baculoviruses and *Bacillus thuriengenesis*, for example, have a narrow host range, and can selectively kill certain lepidopteran larvae, but only do so at very late larval stages (when the animals have already consumed the crop). These microorganisms can be improved by introducing DNA into their genomes coding for lepidopteran proteins that stop larval phytophagy at a very early stage. These larval proteins could be hormones, or signalling chemicals, but many other possibilities may exist that would be revealed only with the upcoming sequenced lepidopteran genome.

Pest genomics and bio-rational target sites

Insect genomics has the greatest potential application in developing novel pest-control products like bio-rational insecticides. These are the chemicals that aim at disrupting a physiological function specific to insects or to a group of insects. Genomics is applied to identify the target sites (proteins) that can be exploited for the developing these bio-rationals. The active ingredients of these insecticides are synthetic compounds; their insect specificity and mode of action make them far more environment-friendly than conventional chemical insecticides. Some bio-rational insecticides are obtained from natural sources or are synthetic analogues of natural compounds. The identification of such compounds can be greatly aided by biotechnology, whereby, the genes encoding insect proteins believed to be suitable targets for enzyme inhibition or hormone receptor antagonistic interactions are cloned and used for the development of *in vitro* high-throughput screening (HTS) assays[23], where the three-dimensional structure of the target protein can be determined. Computer-assisted design can be used to help identify suitable inhibitors, agonists and antagonists in an approach similar to that currently employed for drug discovery[24].

Table 3. Insect genome databases available online

Databases	Year	Hosted by	URL adds	Contents
Hymenoptera Genome Database (HGD)	2011	Elsik Computational Genomics Laboratory, Georgetown University, Washington DC, USA	http://hymenopteragenome.org/	HGD is an informatics resource, providing access to genome sequences and annotation for the honey bee, *A. mellifera*; the parasitoid wasp *Nasonia vitripennis*, and five species of ants available through the ant genomes portal[27].
KAIKObase	2009	National Institute of Agrobiological Sciences (NIAS) Ibaraki, Japan	http://sgp.dna.affrc.go.jp/KAIKObase/	Integrated silkworm genome database and data mining tool with four map browsers, one gene viewer, and three independent databases[28].
WildSilkbase	2008	Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India	http://www.cdfd.org.in/WildSilkbase	It is a catalogue of ESTs generated from several tissues at different developmental stages of three economically important wild silkmoths, *Antheraea assama*, *Antheraea mylitta* and *Samia cynthia ricini*[29].
The Silkworm Knowledge Base (SilkDB)	2005	Beijing Genomics Institute (BGI), China	http://silkworm.genomics.org.cn	SilkDB provides an integrated representation of the large-scale, genome-wide sequence assembly, cDNAs, clusters of ESTs, TEs, mutants, single nucleotide polymorphisms (SNPs) and functional annotations of genes with assignments to InterPro domains and gene ontology (GO) terms[30].
Silkmoth Microsatellite Database (SilkSatDb)	2005	Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India	http://www.cdfd.org.in/silksatdb	SilkSatDb is an interactive online relational database that catalogues information about the microsatellite repeats of the silkworm, *B. mori*[31].
Silkworm Genome Database	2003	Insect Genetics and Bioscience (IGB) Lab, University of Tokyo, Japan	http://papilio.ab.a.u-tokyo.ac.jp/genome/index.html	This database contains EST sequences in full-length cDNA libraries of silkworm, *B. mori*.
Silk Base	2003	Insect Genetics and Bioscience (IGB) Lab, University of Tokyo, Japan	www.ab.a.u-tokyo.ac.jp/silkbase	This database presents the EST sequences of *B. mori* at various stages of growth and development, in various tissues[10].
Silkworm Genome Research Program (SGP)	1994	National Institute of Agro biological Sciences (NIAS), Ibaraki, Japan	http://sgp.dna.affrc.go.jp/index.html	SGP is currently incorporating all information from 'BombMap' and 'SilkBase' into a core database known as 'KAIKOBase', integrating all silkworm genome data including ESTs, chromosome linkage map and genome sequence data[29].
FlyExpress	2011	Arizona State University, Tempe, USA	http://www.flyexpress.net/	FlyExpress is a platform to explore expression patterns of development related genes in *Drosophila* embryogenesis. It contains a unique digital library of standardized images capturing the expression of thousands of genes at different developmental stages.
FlyBase	2009	University of Cambridge, United Kingdom	http://www.ebi.ac.uk/flybase/	FlyBase is the primary database of integrated genetic and genomic data about the *Drosophilidae*, of which *D. melanogaster* is the most extensively studied species and data-types include sequence-level gene models, molecular classification of gene product functions, mutant phenotypes, mutant lesions and chromosome aberrations, gene expression patterns, transgene insertions, and anatomical images[32].

(*Contd.*)

Table 3. (*Contd.*)

Databases	Year	Hosted by	URL adds	Contents
Berkeley Drosophila Genome Project (BDGP)	2007	Berkeley Drosophila Genome Project, Mailstop Berkeley	http://www.fruitfly.org/	The goal of the Drosophila Genome Centre is to finish the sequence of the *D. melanogaster* to high quality and to generate and maintain biological annotations of this sequence. In addition to genomic sequencing, the BDGP is producing gene disruptions using P-element-mediated mutagenesis on a scale unprecedented in metazoans, characterizing the sequence and expression of cDNAs and developing informatics tools that support the experimental process, identify features of DNA sequence, and allow us to present up-to-date information about the annotated sequence to the research community[33].
DroSpeGe	2006	Genome Informatics Lab Biology, Indiana University, Bloomington	http://insect.eugenes.org/-DroSpeGe/	DroSpeGe provides access to twelve new and old *Drosophila* genomes[34].
REDfly	2006	Center for Computational Research, State University of New York, USA	http://www.redfly.ccr.buffalo.edu	REDfly is a curated collection of known *Drosophila* transcriptional cis-regulatory modules (CRMs) and transcription factor binding sites (TFBSs)[35].
Flynet	1999	IBDM, CNRS Case, Marseille Cedex, France	http://gifts.univ-mrs.fr/FlyNets/FlyNets_home_page.html	Flynet is a specialized database which focuses on molecular interactions (protein–DNA, protein–RNA and protein–protein) involved in *Drosophila* development[36].
FlyView – A Drosophila Image Database	1997	Institutfur Neurobiologie Badestr, Munster	http://pbio07.uni-muenster.de/	A graphical atlas of expression patterns of genes and enhancer traps at all stages of development[37].
Butterflybase	2008	Max Planck Institute for Chemical Ecology, Jena, Germany	http://www.butterflybase.org	The gene database for butterflies and moths[38].
Aphidbase	2007	International Aphid Genomics Consortium, France	http://www.aphidbase.com/aphidbase/	It was developed to facilitate community annotation of the pea aphid genome by the International Aphid Genomics Consortium (IAGC)[39,40].
BeetleBase	2007	Bioinformatics Center, Kansas State University, Manhattan, USA	http://www.beetlebase.org	BeetleBase is a comprehensive sequence database and important community resource for *Tribolium* genetics, genomics and developmental biology[13].
InSatDB	2006	Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India	http://www.cdfd.org.in/InSatDB	It stores microsatellites from all the five insect genomes separately as well as carries complete annotations of these microsatellites[41].
AnoBase	2005	Institute of Molecular Biology and Biotechnology (IMBB) Hellas, Greece	http://www.anobase.org/	AnoBase is a database containing genomic/biological information on anopheline mosquitoes with emphasis on Anopheles gambiae[42].
Spodobase	2006	Integrative Biology and Virology of Insects (BIVI), INRA, France	http://bioweb.ensam.inra.fr/spodobase/	An EST database for the lepidopteran crop pest *Spodoptera*[23].
LocustDB	2006	Beijing Genomics Institute (BGI), China	http://locustdb.genomics.org.cn/jsp/about.jsp	Migratory locust EST database, including homologous/orthologous sequences, functional annotations, pathway analysis, and codon usage, based on conserved orthologous groups (COG), gene ontology (GO), protein domain (InterPro), and functional pathways (KEGG)[43].
VectorBase	2009	European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK	http://www.vectorbase.org/	VectorBase contains genome information for three mosquito species: *Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus*, a body louse, *Pediculus humanus* and a tick species, *Ixodes scapularis*[44].

Comparative genomics approaches

Comparative genomics approaches have identified genes that encode proteins unique to insects or to specific insect taxa. Genome-wide *in vivo* RNAi screens, such as the one now possible for *D. melanogaster*, would allow the selection of those insect-specific genes for which transcriptional inhibition induces lethality. Homologues of these genes could then be cloned from pest species and submitted to a similar RNAi analytical approach, thus allowing the identification of genes that are promising bio-rational target sites[23].

In this way, insect genomics, biotechnology and insect cell lines have begun to provide powerful tools for the identification of new lead compounds. Insect cell lines together with HTS procedures (HTS cell-based assays), can enable the discovery of new modes of action for insecticide candidates[23,25]. This is new emerging field in entomology and little information is available with us, but sooner or later it will enter a new era.

Future aspects

Drosophila genomic sequence is a major milestone for genomics, as it vindicates a new strategy for sequencing large eukaryotic genomes and as a model system to understand biological functions. In *Drosophila* post-genomic age, applications of genomics technology to entomology has added volume and quality to the data. At present, only two agricultural pest insect genomes, that of the red flour beetle *T. castaneum* and pea aphid, *A. pisum* have been fully sequenced, but within the next few years several lepidopteran pest sequences will be available. This information will enormously increase our knowledge for understanding the biology of insects and insecticide resistance, which poses an increasing problem for pest control. In the future, where many genomes will be sequenced, a major application of bioinformatics will be the modelling of genetic and metabolic networks, and then comparative genomics will be an increasingly useful approach for pinpointing common and different genes across species.

Genome comparisons between different organisms will be informative on several levels, and information on genomic sequence and organization will be useful to explore gene functions. Functional genomics is being applied more and more in every aspect of life sciences research, including ecology and evolution. Thus, in future, there is a growing tendency for insect molecular scientists to reach out to the broader molecular biology community with all the benefits that such interactions can have for the application of molecular tools in insect science.

Conclusion

The insect genomic databases are goldmines with information on all the proteins, biochemical and physiological processes of an insect. The newly sequenced insect genomes may harbour many surprises for biochemists, molecular biologists and insect physiologists. Insect pest control will soon enter the genomic era with all its surprises and discoveries, as pest and parasitoids genomes are now available. Thus, genomic advances during the last 10 years will revolutionize insect research.

1. Grimmelikhuijzen, C. J., Cazzamali, G., Williamson, C. M. and Hauser, F., The promise of insect genomics. *Pest Manag. Sci*., 2007, **63**, 413–416.
2. Rubin, G. *et al.*, Comparative genomics of the eukaryotes. *Science*, 2000, **287**, 2204–2215.
3. Adams, M. D. *et al*., The genome sequence of *Drosophila melanogaster*. *Science*, 2000, **287**, 2185–2195.
4. Celniker, S. E. *et al*., Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol*., 2000, **3**, research0079 (1–14).
5. Hoskins, R. A. *et al*., A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science*, 2000, **287**, 2271–2274.
6. Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P. and Gelbart, W. M., FlyBase: genomes by the dozen. *Nucleic Acids Res*., 2007, **35**, D486–D491.
7. Gardner, M. J., Tettelin, H. D., Carucci, J. D. and Cummings, L. M., Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, 1998, **282**, 1126–1132.
8. Holt, R. A. *et al*., The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 2002, **298**, 129–149.
9. Arensburger, P. *et al*., Sequencing of *Culex quinquefasciatus* establish a platform for mosquito comparative genomics. *Science*, 2010, **330**, 86–88.
10. Mita, K. *et al*., The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl. Acad. Sci. USA*, 2003, **24**, 14121–14126.
11. Honeybee Genome Sequencing Consortium, Insights into social insects from the genome of the honey bee. *Nature*, 2006, **443**, 931–949.
12. Kirkness, E. F. *et al*., Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 12168–12173.
13. Tribolium Genome Sequencing Consortium, The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, 2008, **452**, 949–955.
14. Werren, J. H. *et al.*, Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*, 2010, **327**, 343–348.
15. Cusson, M., The molecular biology toolbox and its use in basic and applied insect science. *Bioscience*, 2008, **58**(8), 691–700.
16. Zhou, B., Hiruma, K., Shinoda, T. and Riddiford, L. M., Juvenile hormone prevents ecdysteroid-induced expression of broad complex RNAs in the epidermis of the tobacco hornworm, Manduca sexta. *Dev. Biol.*, 1998, **203**, 233–244.
17. Shinoda, T. and Itoyama, K., Juvenile hormone acid methyl transferase: A key regulatory enzyme for insect metamorphosis. *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 11986–11991.
18. Evans, J. D. and Wheeler, D. E., Expression profiles during honeybee caste determination. *Genome Biol*., 2000, **2**, research0001 (1–6).
19. Kawasaki, H., Ote, M., Okano, K., Shimada, T., Guo-Xing, Q. and Mita, K., Change in the expressed gene patterns of the wing disc during the metamorphosis of *Bombyx mori*. *Gene*, 2004, **343**, 133–142.

20. Zhang, G. *et al*., Identification and characterization of insect-specific proteins by genome data analysis. *BMC Genomics*, 2007, **8**, 93.
21. Biology Analysis Group, A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, 2004, **306**, 1937–1940.
22. Negre, V. *et al*., SPODOBASE: An EST database for the lepidopteran crop pest, *Spodoptera*. *BMC Bioinf*., 2006, **7**, 322.
23. Allenza, P. and Eldridge, R., High-throughput screening and insect genomics for new insecticide leads. In *Insecticides Design using Advanced Technologies* (eds Ishaaya, I., Nauen, R. and Horowitz, A. R.), Springer, Berlin, 2007, pp. 67–68.
24. Fauman, E. B., Hopkins, A. L. and Groom, C. R., Structural bioinformatics in drug discovery. In *Structural Bioinformatics* (eds Bourne, P. E. and Weissig, H.), Hoboken Wiley-Liss, NJ, 2003, pp. 477–497.
25. Smagghe, G., Insect cell lines as tools in insecticide mode of action research. In *Insecticides Design using Advanced Technologies* (eds Ishaaya, I., Nauen, R. and Horowitz, A. R.), Springer, Berlin, 2007, pp. 263–304.
26. Smith, C. D., Shu, S. Q., Mungall, C. J. and Karpen, G. H., The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science*, 2007, **316**, 1586–1591.
27. Munoz-Torres, M. C. *et al*., Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res*., 2011, **39**, D658–D662.
28. Shimomura, M. *et al*., KAIKObase: An integrated silkworm genome database and data mining tool. *BMC Genomics*, 2009, **10**, 486–493.
29. Arunkumar, K. P., Tomar, A., Daimon, T., Shimada, T. and Nagaraju, J., WildSilkbase: EST database of wild silkmoths. *BMC Genomics*, 2008, **9**, 338.
30. Wang, J. *et al*., SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res*., 2005, **33**, D399–D402.
31. Prasad, M. D. *et al*., SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Res*., 2005, **33**, D403–D406.
32. Tweedie, S. *et al*., FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, 2009, **37**, D555–D559.
33. Bellen, H. J. *et al*., The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics*, 2004, **167**, 761–781.
34. Gilbert, D. G., DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res*., 2007, **35**, D480–D485.
35. Gallo, S. M., Li, L., Hu, Z. and Halfon, M. S., REDfly: a regulatory element database for *Drosophila*. *Bioinf. Adv.*, 2005, **22**, 381–383.
36. Sanchez, C. *et al*., Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.*, 1999, **27**, 89–94.
37. Janning, W., FlyView, a *Drosophila* image database, and other *Drosophila* databases. *Semin Cell Dev. Biol.*, 1997, **8**, 469–475.
38. Papanicolaou, A., Gebauer-Jung, S., Blaxter, M. L., McMillan, W. O. and Jiggins, C. D., ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res.*, 2008, **36**, D582–D587.
39. Legeai, F. *et al*., AphidBase: a centralized bioinformatics resource for annotation of the pea aphid genome. *Insect Mol. Biol.*, 2010, **19**, 5–12.
40. International Aphid Genomics Consortium, Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol*., 2010, **8**, e1000313 (1–24); DOI:10.1371.
41. Archak, S., Meduri, E., Kumar, P. S. and Nagaraju, J., InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Res*., 2007, **35**, D36–D39.
42. Topalis, P. *et al*., AnoBase: a genetic and biological database of anophelines. *Insect. Mol. Biol.*, 2005, **14**, 591–597.
43. Zongyuan, M., Yu, J. and Kang, L., LocustDB: a relational database for the transcriptome and biology of the migratory locust (*Locusta migratoria*). *BMC Genomics*, 2006, **7**, 11.
44. Lawson, D. *et al*., VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res.*, 2009, **37**, D583–D587.

Received 19 August 2011; revised accepted 23 December 2011