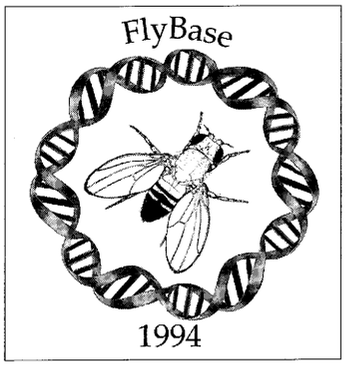
Development 120, 2077-2079 (1994)
Printed in Great Britain © The Company of Biologists Limited 1994



FlyBase — The *Drosophila* genetic database

Michael Ashburner and Rachel Drysdale,
Department of Genetics,
Downing Street,
Cambridge CB2 3EH

The first recorded scientific publication on *Drosophila* was 310 years ago (Mentzel, 1684). By 1980 about 35,000 papers on *Drosophila* had been published and at the time of writing this total had risen to over 60,000. By the year 2000, there will be over 80,000 *Drosophila* publications — and the on-going publication rate will be more than 4,000 a year. There is nothing unique in this rate of growth — it is typical for any "active" subject to double its output every fifteen years (see de Solla Price, 1986). Sooner or later, of course, the curve must plateau but, until it does, the individual scientist faces an obvious problem. Not all of the papers published will be of the standard of those in *Development*. Nevertheless, just sifting those that are worthy of reading from those that are not will be (indeed is) a daunting task. What is to be done? The answer is obvious, we must exploit the power of computers to point us to papers that we need to read. We must also exploit the power of computers to provide us with basic data about our organism. Luckily, there is every prospect that the power of engines to process and access these data will increase, and their relative cost decrease, with the growth in scientific information.

Since 1925 there have been four printed catalogues of basic genetic data about *Drosophila melanogaster*. The most recent in this worthy tradition is *The Genome of Drosophila melanogaster* by Dan Lindsley and Georgianna Zimm (Lindsley and Zimm, 1992). In the last few years of its preparation, it became clear that the rate of growth of data was far too great for us to rely on the publication of a printed guide every 25 years or so. Moreover, it is not just the *amount* of data that is increasing — it is also its complexity. Between each "edition" of the mutant catalogue there has been at least one major technical revolution - X-ray mutagenesis and the discovery of polytene chromosomes between Morgan et al. (1925) and Bridges and Brehme (1942); EMS mutagenesis and fine-structure mapping between Bridges and Brehme (1942) and Lindsley and Grell (1968); cloning, sequencing and P-element mutagenesis between Lindsley and Grell (1968) and Lindsley and Zimm (1992) and, effectively, enhancer trapping and its variants, and the PCR revolution, since the end of data collection for Lindsley and Zimm (1992) in 1989.

It is for all of these reasons that, at the 30th Annual *Drosophila* Research conference in New Orleans in 1989, various members of the fly community began to discuss among themselves what might take over from Lindsley and Zimm (1992). The outcome of these discussions was FlyBase — a computer-based database for *Drosophila*. Funded by the National Institutes of Health (USA) and the Medical Research Council (UK) since the summer of 1992, the FlyBase project involves four groups: Bill Gelbart at Harvard University (PI), Thom Kaufman and Kathy Matthews at Indiana University, John Merriam at UCLA and Michael Ashburner at Cambridge University. At each site there are both *Drosophila* biologists (responsible for the data) and computer scientists (responsible for the design and implementation of the database). FlyBase is currently stored as part of the publicly accessible Biology Archive at Indiana University.

The aim of FlyBase is to provide access for biologists to fundamental biological data about *Drosophila*. FlyBase includes the following.

(1) Genetic, genomic, molecular and phenotypic data on all genes of *Drosophila melanogaster* (and other species will be added in the future).

(2) Chromosome aberrations.

(3) Lists of mutant stocks kept in the major stock centers at Bloomington and Bowling Green in the USA and Umea in Sweden. FlyBase encourages laboratories to make stock lists public and offers help in doing this.

(4) A list of molecular clones isolated from *Drosophila*.

(5) DNA and protein sequence accession numbers. Though it would be redundant of FlyBase to store DNA and protein sequences themselves, we store the nucleic acid and protein sequence database accession numbers for these, to help users to find a sequence of interest in those databases.

(6) Transposon insertions and transformation vectors.

(7) Lists of P-element insertions and P1-phage, and lists of cosmids from the US and European genome mapping projects. FlyBase has agreements with both the Berkeley and European *Drosophila* genome projects to provide public access to their data.

(8) A bibliography of all publications on *Drosophila* — already over 60,000 records, available in a variety of formats.

(9) A directory of *Drosophila* workers, now including over 4,500 people, indexed in such a way that "useful" people such as stock keepers or principal investigators can be identified.

(10) News from the electronic publication *Drosophila* Information Newsletter and the bionet.drosophila news group.

(11) Allied databases. Several individuals in the *Drosophila* community keep more specialised databases, for example *Drosophila* codon tables, a taxonomic list of all drosophilid species, genetic data on non-*melanogaster* species and a database of polytene chromosome sites to which antibodies against specific proteins bind. FlyBase encourages these efforts by making these databases available to the public.

These data must be stored in such a way that a user can

migrate between data types — for example, from a "gene" record to a literature citation or to a stock list. They must also be stored so that "derived" information — for example a genetic map, or a list of all enhancer trap P elements that express in the male-specific muscle in abdominal segment 5 (the small but beautifully formed "muscle of Lawrence") - is readily available. A user might want to know what deletions have been isolated in region 89 of chromosome arm *3R*, then which of these are available in stock collections and then the address — or fax number — of the stock keeper. A user might want to know the absolute direction of transcription of all genes on the distal part of chromosome *2L*, and the nucleic acid sequence databank accession numbers of the sequences. A user might want to know all papers concerned with *Drosophila* jointly written by Ashburner and Lawrence since 1968 (a short list, sadly). A user might want to have a list of all Zn-finger-protein-encoding genes in *Drosophila* with the names of their mammalian "homologs" - most of these data types are (or will be) available from FlyBase.

FlyBase faces three very different, but interdependent, jobs — each vital to its success. These are the collection of the data (curation), the storage of the data (building and maintaining the database itself) and the display of the data to users. The major source of data is the published literature, both retrospective and current. The information base upon which these data are collected is Lindsley and Zimm (1992), made available to FlyBase in computer-readable form by the publishers, Academic Press. The FlyBase curators work from the published source, entering data onto a system of templates on laptop computers. The curators have the entire database available to them on their laptop computer, so that the addition of new data can take existing data into account. Frequently curators engage in correspondence with authors to clear up complexities of interpretation. In addition to curation of the literature from journals, FlyBase has arrangements with other major databases, both bibliographic (Medline, BIOSIS and EMIC) and DNA and protein sequence (EMBL, Genbank, DDBJ, SwissProt and PIR) to include automatically references and sequence accession numbers. It is vital that all curated data are entered into FlyBase in as rigorously defined way as possible. For this reason, FlyBase is constructing "controlled vocabularies" to describe the parts of a fly, mutagens and phenotypic classes and is using those already developed by others for protein domains, and motifs.

The data stored in FlyBase are not only large in extent, but also complex in terms of their inter-relationships. New information about a particular mutant allele must be reflected in many categories of data. It may have implications for molecular maps, multiple chromosome aberrations that carry that mutant allele (and therefore the stocks that carry those aberrations), the mutant phenotype of the gene in question and its relationship to other, similar phenotypes or genes. For this reason, FlyBase is being built within a well-tested commercial relational database management system, Sybase. This kind of database system was specifically designed to reflect automatically updates about a particular data-type in all relevant places in the database structure. In a rapidly advancing field such as *Drosophila* research, such a database offers a robust and economical solution to the problem of maintaining data integrity.

Users, however, will be neither expected, nor need, to understand the underlying structure of the database nor Sybase procedures to use FlyBase. Users will have several alternatives as to how they can access FlyBase. The most traditional will be by the printed word: every year or so FlyBase will publish special issues of *Drosophila Information Service* that will include tables of FlyBase data. Another alternative will be by using "flat files", simply lists of the various categories of information made available on computer servers by FlyBase, who will periodically generate these files from the central database. Users can download these into their own computer systems. In the future these files will also be available on CD-ROM, probably with software that will make them easy to query.

A popular general way of accessing FlyBase that is already in use is through a "client-server" model: a user has a "client" on his local computer (most often a Macintosh or PC) and this talks across the Internet to a "server" at a remote site which houses FlyBase (currently most users access the server at Indiana University). The great advantage of this mode of access is that users always have access to the most up-to-date version of FlyBase. The first of these client-server models is known as "Gopher". The Gopher system allows the user to access information on computers all over the world and is freely available to run on a variety of machines, from Unix workstations to PCs. A FlyBase Gopher server is installed at the publicly accessible Biology Archive at Indiana University, and through this a Gopher user can browse or query FlyBase files. This is achieved without requiring particular computer skills, using "point and click" selection mechanisms with a minimum of typing to define the subject of any particular search. Any file of interest, or any part of a file, can be returned from the server to the user's own computer file space. The new generation of Gopher clients allow users to order stocks from the Bloomington Stock Center, or to add or correct the directory of drosophilist's addresses, while directly talking to the Indiana FlyBase Gopher server.

Client/servers will continue to increase in their sophistication. Already there are different programs freely available, at least for Unix workstation or Macintosh users. One such example is provided by Expasy, a server at the University of Geneva, used with a client known as "Xmosaic" (Appel et al., 1993). This allows users to migrate between very different databases, for example from a nucleic acid sequence database into FlyBase (or vice versa). As database culture evolves and becomes more refined such links will become more useful and accessible.

Biologists working with other organisms, from prokaryotes to humans, all face similar problems of data overload. They would be best served if the many different databases currently being built had a common "look and feel", allowing them to draw on data from many organisms without having to learn as many ways of using genetic databases. A successful genetic database, "acedb", has been designed for *Caenorhabditis elegans* by Jean Thierry-Mieg and Richard Durbin, and acedb format is now being used for a variety of different organisms. A project to make FlyBase available in "acedb" format is now underway. Both FlyBase, and the user community in general, may wish to design other "front ends" that will allow users to browse and query the underlying data. Like acedb many of these front ends will largely be map based — users will be presented with a map displaying genes, chromosome aberrations, clones, transposon inserts (and, in due course, tran-

scripts and sequences). By clicking on any of these objects a window will open, providing more information — phenotypic data, information about alleles, bibliographic data, and so on. Looking to the future, subsequent additions may include windows that provide a picture of a P-element chromosome in situ or a "quick time" movie displaying a focus series through a stained embryo. For the great majority of users, knowledge of the underlying structure of the database will continue to be irrelevant.

The FlyBase group will not be able to achieve their aims alone. The *Drosophila* research community must be involved at all levels, for it is they that both provide and use the data of FlyBase. There are many ways in which the community can help directly — most importantly by not using non-standard names or systems of nomenclature in publications. FlyBase is currently preparing a document revising, updating and clarifying the system of nomenclature and this will be widely circulated to aid the community in helping us. Another is by telling FlyBase both how we could improve the database and its distribution and of any errors that are found. A third is by cooperating with FlyBase curators when they ask for your help. So far, happily, the response from *Drosophila* researchers to our queries has followed our almost century-long tradition of helpfulness and cooperation.

FlyBase is in its very early stages — the project has only been funded since October 1992 — and neither the products nor structure of FlyBase are yet stable. The main "gateway" to FlyBase is now the computer server at Indiana University, whose address is ftp.bio.indiana.edu. This is both a Gopher server and a file transfer (ftp) server. Computer Services at the vast majority of universities and research institutions will be able to guide new users in reaching FlyBase by Gopher or ftp. FlyBase users are *strongly* recommended to read the User-manual (or the more detailed Reference-manual) to get the most out of FlyBase. FlyBase has a central email address to which *any* query can be addressed. It is flybase@morgan.harvard.edu. We are very happy to give advice (or point you to advice) on obtaining an Internet connection, or a Gopher client, or to explain to you the details of making an ftp call to FlyBase. If you lack an email connection then you may write to, phone or fax FlyBase, Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA, 02138, USA, (Phone: 617-495-5668; fax: 617-495-9300).

REFERENCES

Appel, R. D., Sanchez, J.-C.,Bairoch, A., Golaz, O.,Miu, M., Vargas, J. R. and Hochstrasser, D. F. (1993). SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis* **14**, 1232-1238.

de Solla Price, D. (1986). *Little Science, Big Science ... and Beyond.* p.301. New York: Columbia University Press.

Bridges, C. B. and Brehme, K. S. (1942). *The Mutants of Drosophila melanogaster.* Carnegie Inst. Wash. Publ. 552. p.252.

Lindsley D. L. and Grell, E. H. (1968). Genetic Variations of *Drosophila melanogaster.* Carnegie Inst. Wash. Publ. 627. p.472.

Lindsley D. L. and Zimm, G. G. (1992). *The Genome of Drosophila melanogaster.* p.1133. New York: Academic Press.

Mentzel, C. (1684). De musca vini vel cerevisiae ascentic. *Miscellanea Curiosa sive ephemeridum medico-physicarum germanicum Academiae Caesareo-Leopolinae naturae curiosum.* **2,** 96-98.

Morgan, T. H., Bridges, C. B. and Sturtevant, A. H. (1925). The genetics of *Drosophila. Bibliographica Genetica* **II,** 1-262.