The computer and the fly – FlyBase

Michael Ashburner
m.ashburner@gen.cam.ac.uk

Department of Genetics, University of Cambridge, Cambridge,
England and European Molecular Biology Laboratory
European Bioinformatics Institute, Hinxton, Cambridge, England

Abstract

It is now almost 90 years since the common little fly Drosophila melanogaster was introduced to research in the then young science of genetics. In the period between, roughly, 1910 and 1940, research with Drosophila laid the foundations of "classical" genetics. The chromosome theory of heredity, sex-linkage, multiple allelism, the theory and practice of genetic mapping, the existence and behavior of chromosome aberrations, the induction of mutations by ionising radiation and chemicals, the specific developmental defects of particular mutations were all established by the early Drosophila work. This not only laid the foundation for the use, to this day, of Drosophila for fundamental studies in biology (witness this year's Nobel Prizes for Medicine), but also inspired the subject of genetics itself.

One reason for the "success" of Drosophila as an experimental organism was the fact that Thomas Hunt Morgan and his colleagues actively encouraged others to work with this fly, providing them with experimental material and, above all, information. In 1925 Morgan and students published a complete catalog of the mutations of Drosophila melanogaster. This was revised in 1942, by Bridges and Brehme, in 1968, by Lindsley and Grell and, finally, in 1992 by Lindsley and Zimm. Known as the 'Red Book' these catalogs have been essential for the dissemination of information, much never published in a more conventional form, for all Drosophila biologists.

About ten years ago it became clear that this form of distributing information concerning Drosophila was becoming too difficult. The number of Drosophila researchers and publications was increasing at an alarming rate (nearly 9,000 papers in the decade 1970–1979; over 16,000 in the 6 years 1990–1995). The numbers of genes, aberrations and alleles shows a similar increase (1167 genes known in 1968, over 9,000 today). Moreover, the impact of new research using the tools of molecular and cell biology meant that, as in the pre-war period, Drosophila research was of great significance to those studying other organisms, from bacteria to humans. The idea of FlyBase, a computer database for Drosophila, was born.

FlyBase is a relational database (in Sybase) devoted to the biological, genetic and molecular data of Drosophila. I will illustrate some of the problems faced by the FlyBase consortium in building this database, with the expectation that these experiences may be of greater generality. The scope of FlyBase is probably broader than that of any other similar database. This is so in two senses: FlyBase integrates both historical and contemporary data (its earliest bibliographic record is from 1684); FlyBase integrates data from a very wide range of research, from neurobiology to ecology (though, it must be said, to rather different degrees at the moment). Some of the problems faced by FlyBase come from the extraordinary sophistication of modern Drosophila genetics, for example the extent to which artificial constructs are used for mutagenesis, gene dissection and developmental studies. Some are more general to databases of this class. For example, modern biologists urgently need tools to extract data from a variety of 'single organism' databases. How is this to be done? What criteria are to be used to declare that a particular fly gene is 'homologous' (whatever that means) in fly and mouse? How are these links, once agreed by the biologists, to be made between the mouse and fly databases? How can a variety of different objects – from the single nucleotide to a large chromosome aberration – be represented in an integrated genetic map? How is gene function to be described? How can users be given tools to find all genes encoding products involved in a particular function, be it a metabolic or signal transduction pathway or the establishment of pattern in the wing imaginal disc? What tools should be developed to allow users to explore genomic sequence information, including all of the biological knowledge that often comes many years after a sequence record has been deposited in the data libraries?

FlyBase has not solved these problems, but is working towards solutions for some of them at least. I will describe the philosophy and implementation of this work.

FlyBase is supported by the NIH, Bethesda and MRC, London. The FlyBase consortium is W.M. Gelbart (Havard), T. Kaufman and K. Matthews (Bloomington) and M. Ashburner (Cambridge).