E-BioSci, Digital Archives, Databases, and the Changing Face of Publishing

Les Grivell

Paper presented at the 24th UKSG Annual Conference, Heriot-Watt, April 2001

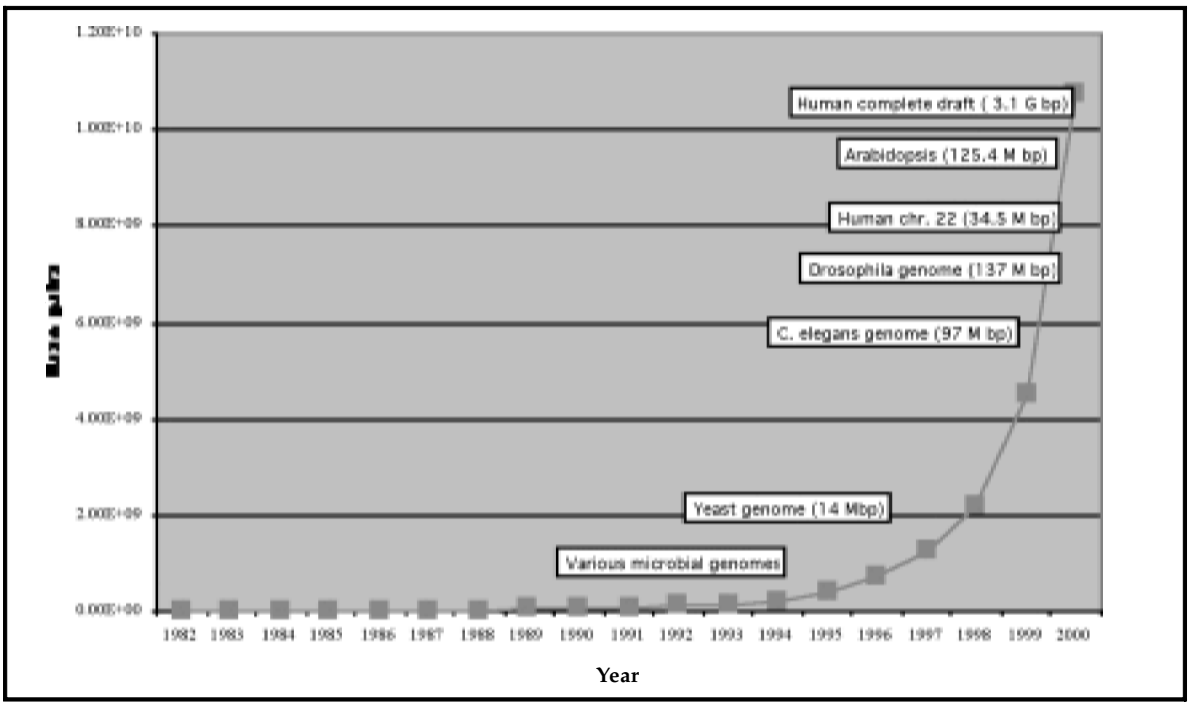
Les Grivell
European Molecular Biology Organisation (EMBO), Meyerhofstrasse 1, 69117 Heidelberg, Germany

E-BioSci is the European Molecular Biology Organisation's initiative in e-publishing. The aim of the initiative is to provide a range of services relating to the access and retrieval of information in the life sciences. However, the brevity of this mission statement possibly raises more questions than it answers, since it does not describe what E-BioSci will be, nor does it convey any impression of the full impact of the dramatic changes that are currently taking place within the world of (electronic) publishing. To do this, it is necessary to describe how such changes are being driven by quantum leaps in technology (both in information science and biology), by the growth of new disciplines, including genomics and bio-informatics and by the novel ways in which research scientists are using information in the published literature. The remainder of this article will, it is hoped, make some of these issues clear.

In order to put the concept of E-BioSci into perspective, it is necessary to take a few steps back in time and review developments in molecular biology, starting about 20 years ago with the establishment of the first international DNA sequence databanks. At that time, the labour-intensive nature of DNA-sequencing technology limited sequence reports and their interpretation to short segments of sequence information, which were usually listed in full, often alongside images of the original raw data. Readers were content to browse such data and were often able to extract interesting or useful features simply by visual inspection.

Times have changed. DNA sequencing technologies have become increasingly automated, less expensive and correspondingly more widely used. As of 28th May 2001, the number of DNA sequence records deposited in the joint EMBL-GenBank-DDBJ databanks topped 11,963,036 and amounted to a staggering total of 12,727,171,823 base pairs[1]. A plot of annual

Figure 1: Exponential growth of DNA sequence data held in the joint EMBL-GenBank-DDBJ databanks, with a number of landmarks in genomic sequencing.



totals between 1982 and the present (Fig. 1) reveals exponential growth that shows no sign of levelling off. Since 1990, sequences deposited include increasing numbers of complete genomes – the entire genetic content of an organism – and this development, at present culminating with the 3.1 G bp draft sequence of the human genome[2], has, in its turn, sparked a further revolution in the founding of new disciplines. These include genomics, bio-informatics and their related specialisations. These disciplines use genomic sequences to derive as complete a picture as possible of gene structure and activity and to integrate this information in a way that allows better understanding of living organisms at the molecular, cellular, organismal, population and evolutionary levels (Fig. 2). Common to each of them is their data-intensive character. Common, too, is the increasing dependence on the world-wide-web as a means of disseminating or acquiring data and of providing access to specialised software for analysis.

Genomics is not the only area in biology that relies heavily on information in digital form. Non-invasive spectroscopic or microscopic techniques are increasingly being used to track in real time the behaviour of, and interactions between, individual macromolecules within single cells. The resulting spectra, structures and multi-dimensional images are difficult, if not impossible, to reproduce fully within the confines of a conventional paper journal. Thus, paper journal publications reporting such research are increasingly just summary pointers to data tables that are too large to print, or to video's and multi-dimensional images that cannot be printed. Increasingly, too, users of this information are demanding it in computer-readable form to allow searching, analysis, manipulation and new forms of visualisation that aid interpretation.

For some, the availability of apparently limitless amounts of new data is seen as the death-knell for hypothesis-driven research and the dawn of an era, in which data-mining will generate novel leads and concepts for innovative research. For others, it signals just the opposite – a means of enabling biologists to construct for the first time precise, detailed and experimentally-verifiable models of cellular function. Either way, the success of data analysis depends on the ready availability of as complete a set of data as possible, together with unhindered access to the corresponding literature. Although such interlinkage of information sources is a fervent wish of researchers in general, it is particularly relevant for those involved in the intensive detective work that is required for the annotation of genes in newly sequenced genomes[3]. In the

future, interconnection of information sources and the ability to navigate between them will increasingly blur the distinction between journals and databases[4] and this process is likely to be accelerated by the wider use in web publications of structured documents that allow selected portions of text to be XML-tagged and handled as if they were fields in a database. Such tagging is hidden from the human eye. Its primary added value is in enhanced computer-searchability and readability that are instrumental in the further streamlining of the information flow. The notion of computer-understandable documents is currently being extended in the Semantic Web Activity run by the World Wide Web consortium. A new set of languages is being developed in order to make more web content accessible to machines. Their application will facilitate implementation of automated methods for information searching and retrieval[5].

It is against this background of changing practices and expectations that biomedical researchers are beginning to reconsider the nature of scientific publication and even to question established editorial, reviewing and publishing practices. It is against this background too that EMBO decided to take the lead in a collaborative effort to establish E-BioSci as a European-based information resource network with a global role. A series of discussions[6] with interested parties (including research organisations, learned societies, publishers, individual research scientists and representatives of a large number of EU member states) led to the formulation of the current initiative. This defines E-BioSci as a networked platform that will extensively combine the skills and content already present or being developed in various centres in Europe. It will work in harmony with other global initiatives, such as PubMed Central and with publishers and other information providers. Although the concepts of networked resources and distributed databases are superficially more complex than the central repository originally envisaged by PubMed Central[7-9], this setup more accurately reflects the European dimension of the project.

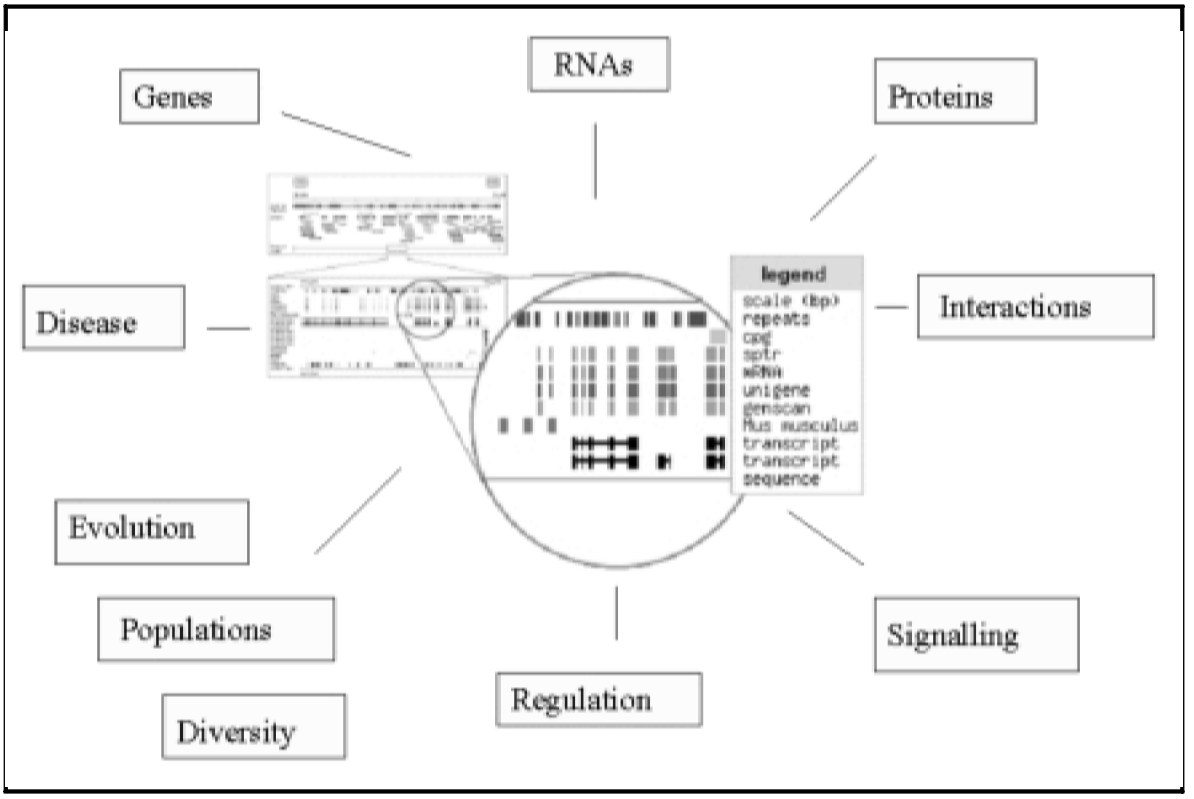


Figure 2: Some of the many facets of genomics, a rapidly developing discipline that uses DNA sequence information to derive as complete a picture as possible of gene structure and activity and to integrate this information in a way that allows better understanding of living organisms at many different levels. The Figure reproduces part of a screen shot from the Ensembl human genome server (3) that both produces and maintains automatic annotation on eukaryotic genomes and allows genomic data to be queried at different levels.

Additionally, it offers potential advantages in terms of speed of access, provision of backup or secure storage facilities and it will allow queries to be performed in different language formats.

In the course of 2000, EMBO itself put forward funds to seed activities and the European Molecular Biology Conference (EMBC) took the historic step of incorporating E-BioSci into its general programme of activities in support of molecular biology in Europe. More recently, the European Commission has lent further tangible support to the concept of a European information platform by promising financial support from within the Research Infrastructures section of the 5th Framework Quality of Life Programme. This promise of funding has been instrumental in driving the formation of a core network of institutions that will provide prototype E-BioSci services, so that there are now good expectations that a platform demonstrating proof of principle can be launched later this year. Current members of the network are listed in Table 1, together with brief descriptions of their main expertises.

When fully operational, the E-BioSci network will:

- offer multiple points of entry for end users onto as complete a set of abstract and full-text research material related to the life sciences as possible;
- allow enquiries in different language formats
- allow existing bibliographic and factual databases to link and integrate with the scientific literature;
- be European-based but will operate in harmony with other international efforts;
- be run on a not-for-profit basis and provide free access to as much abstract or other material as possible;
- protect access to commercially-produced full-text material or other types of data;
- integrate different business models for how content is collected, distributed and charged;
- allow new content providers which are fully electronic and meet the E-BioSci criteria of editorial control to participate on an equal level to existing publishing concerns;
- when content providers wish, offer complete, free, access to their content, facilitate and welcome this;
- not contain material which is non-refereed or otherwise does not meet its criteria of editorial control.

By providing an extensive set of linkages through the biological information chain, E-BioSci will:

- foster optimal pooling and use of European biological archives and data collections;

Table 1: Current members of the E-BioSci network

Organisation	**Roles**
EMBO	Coordination and management; scientific quality control
Centre Informatique National de l'Enseignement Supérieur	French node; database and repository servers; access to indexes, databases and full text document archives maintained by partner organisations; testing of cross-database search engines
Consejo Superior de Investigaciones Científicas	Spanish node; database and repository servers; implementation of (Spanish) e-journals; E-BioSci link to Latin-American countries
Deutsches Institut für Medizinische Dokumentation und Information	German node; database and repository servers; design and development of document location protocols
Edinburgh University Computing Services (EDINA)	UK node; development and test of cross-database search facilities; assessment of resource discovery models (in collaboration with BIOME, UK)
European Bioinformatics Institute	Development of E-BioSci root server, document location protocols, open client libraries; factual database management
Ingenta UK Ltd	Construction of E-BioSci presentation and access control layers
Institut National de l'Information Scientifique et Technique	French node; database and repository servers; database design and construction taking into account use of different entry languages with homogeneity of data structure and vocabulary

- stimulate the development of common protocols and methodologies for efficient searching and retrieval of information contained in bibliographic and sequence, or sequence-related databases;
- provide a framework for further research into more effective strategies for linking of bibliographic with molecular, genomic and multi-dimensional image databases.

As indicated above, E-BioSci will, besides acting as an information portal, provide hosting services for electronic publications. The aim here will be to provide a platform for the dissemination of material that has previously undergone peer review and authentication by an independent body. E-BioSci need not be the sole repository of such material and authors may choose to submit their reviewed and authenticated manuscripts to as many sites as they wish. This emphasis on stringent quality assessment and control, prior to publication, distinguishes E-BioSci from a number of other e-publishing initiatives, including those modeled on the Los Alamos Physics Archive (e.g. the eprint-based Cogprints server[10]), or commercially-based services, such as those offered by BioMed-Central[11]. One of the main issues here is that authors rely on the perceived quality of their publications as support for funding applications and career advancement and are thus usually reticent to abandon a tried and trusted model of assessment in the absence of reliable and widely-accepted alternatives. Additionally, from the reader's point of view, peer review and editorial control are, at least in part, mechanisms that guarantee that technical standards have been met, that the conclusions are adequately supported by the experimental data and that the presentation meets acceptable standards of clarity. In cases in which a submission is accompanied by significant amounts of supplementary data, the peer-review process also provides an appropriate opportunity for watermarking of both manuscript and data to protect against alterations at a later stage.

Conclusions and prospects

Just as the emerging field of genomics is changing the way in which molecular biologists plan, execute and interpret their research, so is the transition from traditional to electronic publishing technologies changing the ways in which the results of this research are disseminated to and used by other scientists. In this brief overview, I have presented a perspective largely based on that of the individual scientist, who wishes to have free, or at least unhindered access to as wide a range of electronic information sources as possible, to be able to navigate effortlessly between them and to search, select, integrate and manipulate information without leaving his or her desk. I have outlined a number of recent developments that will contribute to the achievement of this goal. E-BioSci is one of these. Much still remains to be done, however, and a brief wishlist of a typical user might include:

- conversion of as much journal back issue material as possible to digital and searchable form;
- clarification of issues relating to free access to (supplementary) image and types of data;
- the development and adoption of a common DTD that will permit faster and more accurate searching of publications than is currently possible. (The increased granularity possible with a structured document could also be used to allow easy retrieval or capture of specific portions of a manuscript, including images and corresponding metadata, datasets or methodologies.);
- more extensive crosslinking of digital objects through a wider application of digital object identifiers (DOIs);
- development of advanced search tools that will permit efficient analysis of full text, 'fuzzy' searches, document neighbouring and development of protocols for the more effective linkage of literature with factual databases.

Links and references:

1. The EMBL Nucleotide Sequence Database. Most recent statistics and release information can be viewed at www.ebi.ac.uk/embl/.
2. Both *Nature* and *Science* have devoted complete (online) issues to announcements and first surveys of human genome sequence data. These are *Nature* **409** (2001), 6822 and *Science* **291** (2001), 5507 respectively.

3. Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system that produces and maintains automatic annotation on eukaryotic genomes. It is primarily funded by the Wellcome Trust. The latest version of the database can be searched at www.ensembl.org/.
4. Gerstein M (1999) E-publishing on the web: promises, pitfalls and payoffs for bioinformatics. *Bioinformatics*, **15** 429 – 431.
5. Berners-Lee T (2001) Scientific publishing on the 'semantic web'. www.nature.com/nature/debates/e-access/Articles/bernerslee.htm.
6. For an overview of the discussions that led to formulation of the E-BioSci initiative, see www.embo.org/E_Pub_pages.html. Latest information on E-BioSci can be found at www.e-biosci.org.
7. Butler D and Wadman M (1999) Mixed response to NIH's web journal plan. *Nature* **399**, 8-9.
8. Roberts, RJ (2001) PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **98**, 381-382.
9. Sequeira E, McEntyre J and Lipman D (2001) PubMed decides to decentralize. www.nature.com/nature/debates/e-access/Articles/pubmed.html.
10. www.eprints.org/.
11. www.biomedcentral.com/info/peerreview.asp.