SHORT COMMUNICATION

A STRATEGY AND DEMONSTRATION FOR INTEGRATED BIOTECHNOLOGY INFORMATION

SIMON B. JONES*, JACK FRANKLIN#

*CAB INTERNATIONAL, Wallingford, Oxfordshire OX10 8DE, UK. (s.jones@cabi.org)

#ASFRA B.V., Voorhaven 33, 1135 BL Edam, The Netherlands. (asfra@wx.nl)

ABSTRACT

Bioinformatics has developed as a key discipline to support science. Integrated access to the various new and established information resources is a key requirement for their future utility. A strategy for this integration has been developed and is being demonstrated to a core group of European users.

A decade ago, bioinformatics was barely recognised as a scientific discipline. Today, academic and industrial centres are searching for qualified individuals who can manage and use the growing number of databanks and databases required for ongoing biotechnology research. Data, for years seen solely as an output of research, is now accessed and manipulated as an essential part of the research process.

Because this data makes up the research story itself, many of the databases used in genetics and biotechnology today have been built by and for researchers. They were designed to solve a particular research need, and the infrastructure needed to update and distribute them has, similarly, been developed by researchers to meet their specific needs. The focal points for these databases are presently found in Bethesda – home to the American National Center for Biotechnology Information (NCBI)¹ – and Hinxton, Cambridge, where the European Bioinformatics Institute (EBI)² is sited. Europe also has a network of nationally mandated or specialist nodes which together form the European Molecular Biology Network (EMBN)³ and which serve to collect and transmit data and to mount and update the various databases⁴.

Abstracting and Indexing (A&I) services have provided comprehensive coverage of the research literature for many years. Originally in print-on-paper form, they generally started to provide on-line access to their databases in the 1970s through host services such as DIALOG, DATASTAR and DIMDI. These hosts were accessed from terminals and later personal computers over dial-up packet-switched networks. The fees charged for access and use vary according to the circumstances of the provider or publisher: some A&I services are commercial, some are independent not-for-profit organisations, and some are government funded (particularly in the US) and are available at very low cost at the point of use because of subsidies. With the advent of the Internet and World Wide Web, most A&I services and on-line hosts have implemented networked delivery, each with a different user interface.

The focus of activity for the sequence databases has been academia. However, as early as in 1988, the Swiss Chemical Industry Federation⁵ researched the importance of bioinformatics for Europe's pharmaceutical and agricultural industries, and warned their industrial colleagues that they not only needed to access this data to compete on the world stage of biotechnology research, but that there was a real danger of Europe failing to develop the databases and infrastructures needed; they could foresee a situation where European industry (and academia) had to go to the US for the relevant data.

The Swiss persuaded their colleagues in the European Chemical Industries Federation (CEFIC), and the European Commission, that European bioinformatics should be put on a stable footing, which led to a major strategic planning exercise being prepared by CEFIC and a group of publishers with support from the European Commission. This team researched and prepared two reports^{6,7} published in 1990 as the CEFIC Studies: Bioinformatics in Europe - Strategy for a European biotechnology information infrastructure; these soon became a corner-stone for further work in this field.

Among other things, the final report recommended that Europe developed its own biotechnology information centre (which led, indirectly to the EBI) and that links between the new factual databases and the literature and other traditional databases (such as patents and culture collections) should be established if the full power offered by this discipline was to be realised.

The partners who prepared this study continued to meet on an informal basis; and then formalised their meetings in December 1992 when they concluded that there was an ongoing need to monitor and examine strategic problems in the biotechnology information sector. This Biotechnology Information Strategic Forum (BTSF) made, and makes, a special point of bringing users AND producers of the databases together so that issues of strategic importance might be addressed before they become problems. The BTSF meets regularly and has handled such issues as: database access and distribution⁸, database copyright and legal issues, intellectual and product ownership contracts and issues concerning the financing of Europe's bioinformatics infrastructure. In particular, the group concentrates on ways in which users can make improved use of the databases and services on offer.

The conclusions from the first phase of the BTSF identified the need for:

- databases to be more accessible through being linked together in scientific packages;
- secure, efficient networks;
- “even playing fields” in terms of the present unbalanced situation where subsidised American databases are able to reach European users on European networks to the detriment of European producers;
- new financial infrastructures to guarantee the production of specialised databases which themselves might not generate sufficient income for their survival.

A second, more practical project, to come from the CEFIC Study was the Common Core Database (CCDB) Pilot Project, where three of Europe's leading literature database producers, CAB INTERNATIONAL, Elsevier Science and the Institut de l'Information Scientifique et Technique (INIST), examined ways of linking their databases so to improve the quality and usefulness of the material on offer. All three producers operate on a commercial basis, and one reason for doing this was to improve their competitive position in relation to subsidised literature database production in the US, by introducing economies of scale and other

enhancements so that more primary articles could be abstracted and indexed. The database manual produced from this could be extended to form the basis of a standard for data exchange between literature databases across other scientific disciplines.

The combined efforts of these two projects led to the conclusion that the two user communities (academic and industry) are not using the same kinds of information, and they are not using it to the available level of efficiency. Given the importance of information to research and development, this is obviously a serious failing, leading to inefficiencies and lost opportunities, false impressions of the patentability of products and processes, and wasted effort. Academics, in particular, instinctively turn to the apparently free literature databases subsidised by the US government, rather than the commercially produced databases, despite the better coverage offered by the latter in biotechnology (and other disciplines). Industrial users are sometimes having to forgo using the most immediate updates of the factual databases such as the EMBL Data Library or SWISS-PROT, as they are unable to access them on a secure host, or they are forced to invest in their own secure mirror systems.

A possible solution to these needs was to develop an integrated biotechnology information service covering a number of general and specialised databases in a way that would allow users to navigate through a "tank" of information with levels of security available to match user requirements. Databases and users would then benefit from there being a critical mass of relevant information with cross-support. This idea was developed into ADLIB – Advanced Database Linkages in Biotechnology⁹, now in its second year of execution as a partially funded Demonstration Project in the Biotechnology programme of the EU's Fourth Framework.

ADLIB is being developed and tested by a team of companies and public sector bodies including literature database producers (CAB INTERNATIONAL, Elsevier Science and INIST), a database host (DIMDI), large (EBI), and small (INSERM, CERDIC) factual databases, project and commercial databases (KNAW, PJB Publishing) and primary publishers (Wiley, Springer-Verlag and Kluwer plus Elsevier Science). Scientific and strategic support is provided by ASFRAB.V. The consortium has established links to important groups of users, through the European Molecular Biology Network (EMBnet), the Pharma Documentation Ring and the CEFIC Science & Technology Working Party which represents the European chemical companies, and also through existing direct public and private sector contacts of the participating companies.

The concept behind ADLIB is to develop:

- A tank of databases available as both a physically co-located collection on one or more established hosts, and as a looser federation of databases accessible over networks.
- Sophisticated parsing systems which can enable links between data of similar and dissimilar types to enable researchers to follow through a research story, for example to obtain authoritative information about a particular gene sequence (Figure 1).
- A charging model which enables both academic and commercial users access to the system, and which allows both subscription and pay-as-you-go payment types.
- Security guaranteed to commercial users through the intermediary of trusted hosts
- Sustainability of the production of the small databases through their use in this system to complete the research picture and through eased access. This process has been termed a "database nursery".
- Efficiency in database production through sharing of common aspects and knowledge.

Fig. 1. Following a research story.

If successful, ADLIB will lead to a commercial product, consistent with the near-market nature of the EU's Demonstration Project concept.

The core technology is the SRS software package developed by Thure Etzold at EMBL^{10,11}. This software has previously been used to allow users to cross search in factual, nucleic acid and protein structure databases. The CCDB work had used the software to link factual databases with small datasets from the three commercial databases mentioned above and it is now being used to link all the databases in ADLIB which are stored either on a selected group of academic hosts, or on the commercial host DIMDI or the group's own server. A link implemented using the ISO23950 information retrieval standard between DIMDI and the ADLIB server allows users to access genetic sequence data from the secure environment offered by DIMDI, the first time that sequences have been made available in this way. The architecture of the ADLIB system is shown schematically in Figure 2.

These links are now operational and users migrate from database to database. It is clear at present that network limitations will limit the number of database locations and so the group is examining how best to cluster the databases. Initially, users are being invited to duplicate searches they are carrying out as a normal part of their work, so that the project can evaluate the advantages of having the large literature databases to hand. The users are also being questioned as to what extra features they require such as patent data,

Fig. 2. Schematic architecture of ABLIB

primary articles and other services and these will be added to the “tank” whenever suitable resources are available. A key issue is: what abstracts should be covered by the ADLIB system? Further work will concentrate on examining this in the context of user needs.

ADLIB is not only a technical project, but also an experiment in approaching the market and in distributing information in this field. Earlier BTSF studies have made it clear that very few users really use all the relevant information sources they can. In addition to the above mentioned cost and security points, a lack of clarity on copyright and patent rights (for instance, does one lose the right to patent if a sequence has been submitted for checking on a public database via public networks?) has also restrained the use or utility of certain resources. Furthermore, many users are simply unaware of the databases which could benefit their work. All too often a search in, say MEDLINE, offers sufficient “hits” to whet the appetite and allow the user to move on under the mistaken impression that they have surveyed the literature. By bringing a tank of databases together in a secure environment where the user can easily check what is or is not known, ADLIB will, one hopes, enable new users to find data, reduce costs through economies of scale, and so encourage new and varied users to access old and new databases. Given that R&D is dependent upon information, it is absurd that data is not being used to the full and so new charging algorithms are also being looked at in ADLIB.

Finally, user experience is already indicating that, ultimately, a link to the located primary article will be required. The technology is now in place for many primary publishers to store the full text of their journal files in databases sometimes called

“electronic warehouses” and, as these become more available and pricing algorithms are designed and introduced to allow for on-line shopping, ADLIB will establish the links from secondary and factual databases to the target articles. ADLIB is also actively working with another European Union part-funded demonstration project – CABRI (Common Access to Biotechnological Resources & Information)¹² – which will soon allow users to search a large number of Europe’s Resource Centres for the culture or collection they require so that a common entry point to these databases is also produced.

THE ADLIB PARTNERS AND DATABASES

CAB INTERNATIONAL	publishers of CAB ABSTRACTS, the leading database on agricultural and related sciences.
Elsevier Science B.V.	publishers of EMBASE, the leading database on pharmacology and biomedicine
INIST	publishers of PASCAL, a unique multi-disciplinary databases covering science, technology and medicine
INSERM	publishers of the GENDIAG database on genetic abnormalities in man
CERDIC	suppliers of the ECACC catalogue of cell cultures
KNAW	publishers of the BIOREP database, a permanent inventory of biotechnology research projects in the European Union
PJB Publishing	publishers of BioCommerce Abstracts and Directory
European Bioinformatics Institute	an outstation of the European Molecular Biology Laboratory
Wolters Kluwer Academic Publishers	primary publishers
John Wiley & Sons Ltd	primary publishers
Springer-Verlag	primary publishers

REFERENCES

- 1 <http://www.ncbi.nlm.nih.gov/>
- 2 <http://www.ebi.ac.uk/>
- 3 <http://www.embnet.org/>
- 4 Harper, R.A. (1996). *Trends in Biochemical Sciences*, **21**, 150–152.
- 5 Swiss Chemical Industry Federation (internal document).
- 6 Bioinformatics in Europe 1 – Strategy for a European biotechnology information infrastructure. Prepared for the European Federation of Chemical Industries (CEFIC), February 1990.
- 7 Bioinformatics in Europe 2 – Strategy for a European biotechnology information infrastructure. Prepared for the European Federation of Chemical Industries (CEFIC), November 1990.
- 8 <http://www.btsf.org/ws1exec.htm>
- 9 <http://198.93.246.130:8093/adlib/index.htm>
- 10 Etzold, T. and Argos, P. (1993). *CABIOS*, **9**, 49–57.

- 11 Etzold, T. and Argos, P. (1993). *CABIOS*, **9**, 59–64.
- 12 Stacey, A., Packer, P., Doyle, A. (1996). Proceedings 3rd Internet World Congress on Biomedical Sciences, [http://www.3iwc.riken.go.jp/ CONGRESS/SYMPO/SBB0202/AD0104/TIT.HTM](http://www.3iwc.riken.go.jp/CONGRESS/SYMPO/SBB0202/AD0104/TIT.HTM)