have been pleased with the progress of our Japanese office. The other main landmark for us is that, at the end of our first trading year, we made a small profit, which is unusual for a start-up.

What is the main obstacle you have found to this new idea taking off?
There have been no specific obstacles. It is a new way of doing development so we are expending a lot of time explaining the model to potential customers so that they become comfortable with it. However, this is just the natural evolution of any new form of business.

I think we probably get as many different concerns as we have different customers. They all have their own perspective of drug development and so they address specific questions to us on how we would deal with certain situations.

If you could set up the company again, what would you do differently?
Do it sooner!

Do you expect other similar companies to follow and are there any others that you know of?
I would be surprised if others did not follow as it has certainly been a successful model in our hands. There are occasional names that appear and disappear who might be doing something similar, and there are some groups that offer maybe segments of what we do, but we are not aware of anyone else who currently offers the whole range that we do.

What are the future goals of Fulcrum Pharma?
To grow the business we are in. We are still less than 2 years old as an independent company so we are still very much at the bottom end of our growth curve but it is turning up quite nicely. We are just working hard to maintain that growth but basically offering the same service to an increased number of customers.

Fulcrum Pharma Developments Ltd
Hamilton House
111 Marlowes
Hemel Hempstead
Hertfordshire
UK HP1 1BB

Molecular biology databases: today and tomorrow

*__Lynda B.M. Ellis__[1] and __Teresa K. Attwood__[2], [1]University of Minnesota, Minneapolis, MN 55455, USA, and [2]University of Manchester, Manchester, M13 9PL, UK. *Mayo Mail Code 609, 420 SE Delaware St, Minneapolis, MN 55455, USA. tel: +1 612 625 9122, fax: +1 612 625 1121, e-mail: lynda@tc.umn.edu

Biological scientists today can no longer afford to ignore the Internet. Whether it is used to find research funding, collaborators, current articles, data or data analysis tools, it has changed the way we work and is not a passing fad. The Internet's molecular biology research data has grown prodigiously over the past two decades. For example, GenBank began with fewer than 10^3 sequences in 1982 and had over 10^7 sequences by the end of 2000; its current doubling time is less than one year [National Center for Biotechnology Information (NCBI) (2001) GenBank statistics, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html].

GenBank, and many other web-based databases for molecular biology, are interlinked and freely available to the scientific community. Those mentioned in the text are listed in Table 1. Access to them is as important to scientific progress today as is access to a laboratory or library. But few understand how public research databases develop and are maintained. Their future is not assured.

Brief history

From the first dawn of today's now-vigorous field of bioinformatics, protein sequence information was assembled into databases and made freely available. In the 1960s and 1970s Margaret Dayhoff's pioneering work on protein evolution led to the distribution of the Protein Sequence Database, now more familiar as the international Protein Information Resource (PIR)[1]. In the early 1980s, this inspired the first public releases of Amos Bairoch's SWISS-PROT sequence database[2]. Then the first nucleic acid sequence databases began to proliferate, and ultimately to cooperate, in order to cope with the increasing quantities of sequence data being generated worldwide (e.g. GenBank[3], EMBL[4], DDBJ[5]).

At that time, there was no question that making the data freely available was both desirable and necessary, to help the research community as a whole make the best use of the data, to facilitate innovation, and to accelerate discovery. Indeed, by 1995, the European Molecular Biology Laboratory had established the European Bioinformatics Institute (EBI). Its mission was, and remains, to 'ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all facets of the scientific community in ways that promote scientific progress' [EBI (2001)

1359-6446/01/$ – see front matter PII: S1359-6446(01)01802-5

Table 1. Databases mentioned in the text (abbreviations are identified in the text)

Name	Contents	Uniform resource locator (URL)
DBcat	Biological databases	http://www.infobiogen.fr/services/dbcat/
GenBank	Nucleotide sequences	http://www.ncbi.nlm.nih.gov/Genbank/
EMBL	Nucleotide sequences	http://www.ebi.ac.uk/embl/
DDBJ	Nucleotide sequences	http://www.ddbj.nig.ac.jp/
PIR	Protein sequences	http://pir.georgetown.edu/
SWISS-PROT	Protein sequences	http://www.expasy.ch/sprot/
PRINTS	Protein sequence fingerprints	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
UM-BBD	Microbial catabolic pathways	http://umbbd.ahc.umn.edu/
ETI	Taxonomy and biodiversity information	http://www.eti.uva.nl/

About the EBI, http://www.ebi.ac.uk/Information/About_EBI/about_ebi.html].

Today, molecular biology research is dependent on publicly available databases. By 1 March 2001, 511 public resources were listed in DBcat, an online database catalog[6]. DBcat divides its databases into several subject domains, as shown in Table 2. The 'miscellaneous' domain includes both integrated databases and those of special interest.

Setting up a database

Like PIR, many databases in current use were neither conceived nor designed as databases, but grew as products to expedite particular pieces of research. Today, databases continue to emerge from local research projects in a similar fashion. When such resources grow large enough, they are often posted on the Web for the benefit of the entire scientific community. This can be done with minimum initial expense for either hardware or software.

Table 2. DBcat subject domains and number of databases in each category on 1 March 2001

Domain	Databases
DNA	87
RNA	29
Protein	94
Genomic	58
Mapping	29
Protein structure	18
Literature	43
Miscellaneous	153

Unfortunately, much more is needed to keep a public database operational. Input is required from both biological and computational perspectives. A database must be updated as knowledge in its area of expertise grows. This growth must be guided, and the software that supports it (for data entry, checking, storage, retrieval and display) must be updated to best meet the needs of the user community. For example, both PRINTS[7] and the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD)[8] started by storing their data in flat text files but evolved to use sophisticated database management systems to better serve their clientele.

How much does this cost and where do the funds come from? A financial study of several smaller biological databases for the 1997–1998 fiscal year revealed they had, on average, 2.5–3.5 FTE employees and median total income (which equals expenses) of US$200,000 (Ref. 9). For an academic database, such income might pay for computer hardware and software upgrades and maintenance, a percentage of a salary for one or two faculty directors, and a full salary for two or three data-entry technicians, programmers, postdoctoral scientists or graduate students. A larger survey of 153 biological database administrators reported that 94% of funding came from government, academic, volunteer or other non-profit agencies[10]; very little came from user fees or industry. Most non-profit agency support covered initial development, but not further maintenance.

Issues in database development and maintenance

Database development and maintenance suffer partly because of the prevailing tension in bioinformatics between service provision and research. This tension threatens to shake the foundations on which the discipline is built. Research output is the desired goal of university departments and research institutes, measured in terms of, for example, publications and grant income. Service provision is a less welcome role, because it does not constitute measurable research output, but rather places a drain on staff time that cannot easily be recompensed. The tension arises because service provision, including database support, underpins virtually all aspects of computational and molecular biology research.

As a result of the early 'free access' philosophy, biologists today take it for granted that databases are simply available for their research at the other end of their Web browsers. The question of the ready accessibility of the information just does not arise; but should it? In light of the recent controversial debates

regarding the public availability of the draft human genome sequence data and its use by a company, is the tide turning (or should the tide turn) in favour of more restricted access? Is there not something rather naïve about funnelling the fruits of tens of thousands of man-hours, funded largely by the public purse, directly into the open arms of private enterprise? In most other scientific disciplines, data generated in the course of a project are usually closely guarded, at least until publication in the literature, and sometimes until some period of exclusivity for the supporter of the research has elapsed; making results available immediately is almost never an option.

Why are results in the bioinformatics arena so different? Perhaps this is because the quantity and quality of data, often generated from a variety of sources, generally require the gathering of additional value (annotation) via the application of a range of different analytical tools to make them more useful. Bioinformatics is thus a fundamentally collaborative field that depends on the availability of large, diverse, value-added data-sets for its analyses[11]. Ultimately, greater research yield accrues from making the data available to the widest possible audience, because the research community is then free to view it from a variety of different perspectives. Without adequate mechanisms for protection, however, this involves risk of exploitation by those other than the desired target audience (e.g. Celera's use of public human genome data to create a restricted-access commercial product)[12].

Problems with public infrastructure

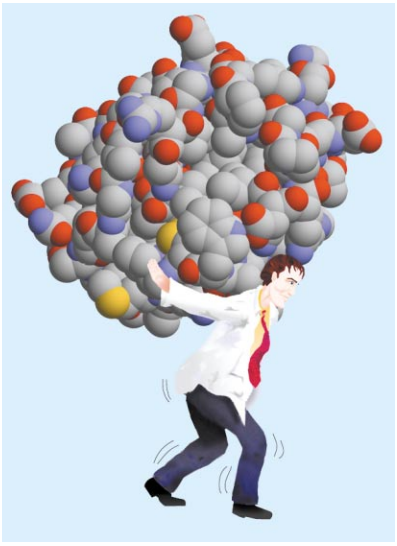
As already mentioned, hundreds of biological databases are now available in the public domain, spanning the worlds of DNA and RNA; protein sequence, family and structure; organisms and mutations; metabolic pathways; biodiversity; biodegradation; and so on. The wealth of such data resources provides an extremely rich seam of information for the biological research community to mine. Unfortunately, this seam is dangerously close to a fault line, and investigators rely on it at their peril (Fig. 1).

Many of the most successful databases depend for their maintenance on *ad hoc* income from research projects. Given the vagaries of many funding mechanisms, the average lifespan of databases is thus only about five years. Indeed, in the survey mentioned earlier, 68% of the respondents (all biological database administrators) said they faced uncertain funding for their databases within the next 1–5 years[10].

As recent funding problems of the EBI attest[13], the greatest difficulty is the ability to attract funds for infrastructure, the lifeblood of bioinformatics research. The problem is exacerbated by restrictions imposed by funding bodies. Dissemination of data and methods is mandated; thus it is now almost inevitable that bioinformatics or genomics research projects will yield databases. Such databases will be funded for the duration of the research, and no more; beyond this point, they must be self-supporting, or fold. Thus, funding bodies continue to provide support for new research projects, and hence new databases, while making work on those databases after the end of their projects difficult.

This means that, for example, to secure continuing funds for a database that originally emerged from a research project funded by the European Commission (EC), the project must be reinvented, with different aims and objectives, and with a different name. Similar devices are required to obtain funds in the UK and the US to support databases. Indeed, there is still no stable, adequately funded system of support even for large infrastructure projects and facilities, such as the EBI.

In 1999, following the publication of a new rule disqualifying the support of core funding and operational costs of infrastructure from the EC's fifth Framework Programme (FP5), the EC rejected the EBI's application for funds to support core facilities (including the EMBL database). Paradoxically, research infrastructure was made a priority in FP5, but the rule stated that such funds would only be used to support research projects[14].



Drug Discovery Today

Figure 1. The information on an enzyme molecule can include: its tertiary structure coordinates, primary sequence, signature sequences, fold, genome/chromosome location, regulation, substrates, products, cofactors, EC code, role in metabolic pathways, the names and affiliations of scientists who are and have been studying it, and in which cell compartments, cell types, tissues, and organisms it is found. All this and more is contained in hundreds of public biological databases. The financial support for most of these is uncertain; will science drop the ball?

Alternative funding sources

EMBL was not the first database of international stature to be hit by a funding crisis. Prior to this were the well-publicized problems experienced by SWISS-PROT[15]. The quality of SWISS-PROT's structured annotation files guaranteed the database a vast user community. However, the provision of in-depth manual annotation

comes at a price, one the Swiss government was increasingly unenthusiastic to meet – why should they support a resource for the rest of the world to use? Notoriety was obviously not sufficient, and the ongoing funding struggles ultimately resulted in SWISS-PROT taking the unprecedented step of partially withdrawing from the public domain; today, the database remains free for academic users, but commercial users are required to pay a license[15]. Now that the dust has settled, we are at liberty to reflect on the success of this venture, and it does seem to be working well.

However, SWISS-PROT had a major advantage that many other databases lack – it had already cornered its market. The database was well known and had generic value. But what of the countless less well-known resources, whose place in the research environment is perhaps more narrow? Most of these do not have a sufficient global audience to be able to follow SWISS-PROT's example.

An alternative approach has been taken by the Expert Center for Taxonomic Identification (ETI), who gather data from the entire biodiversity community, then process, package and sell it on CD-ROMs. However, this approach does not recover the full costs associated with producing the data[16]. As with other forms of publishing, it offers full publicity and possible royalties in exchange for data that are given away free.

A key issue in deciding whether a database can entertain a commercial relationship is ownership. If a database is to be sold, who owns it must be clear. However, two main difficulties arise from this. First, if a database results from the labor of an individual curator, as many do, and that individual relocates, this can generate layers of complexity: a database seeded at one institution might have developed shoots at a different establishment, before finally putting down roots and flourishing at a third. In such cases, the trail of 'ownership' is convoluted and defies a trouble-free step into the private sector. Second, if a database depends on input from the entire research community, it might be impossible (and counter-productive) for any of the host institutions to assert ownership of the data.

Economic models for the support of the biological database infrastructure have been reviewed[9]. In addition to commercialization, advertising has been suggested for the most general, heavily used, databases. Microcommerce, e.g. paying 'pennies' for each database search, is still in its infancy. Questions outstanding include: 'are users willing to make even micropayments for what they might have formerly received for free?' and 'can micropayments meet macro-expenses?[9] In the near future, public funding is all that most databases can hope for.

Which databases should be publicly funded?

Even if the need to provide a secure financial future for biological databases is evident, how do we decide which databases are worthy of public support? Do we count numbers of daily Web page hits, which are notoriously unreliable indicators? A protein or nucleic acid sequence database is likely to have far wider appeal than a database relating to a particular protein, or proteins; the smaller number of hits on the latter reveals nothing of the quality of the resource. Given the different target audiences, it is hard to derive generic quality control measures that provide unequivocal evidence that a particular database is worth preserving over and above any other.

Peer review has many faults, yet, considering the alternatives, it is the best system available. What is needed is the availability, at the national and international level, of public funds to maintain the public biological database infrastructure, and unbiased peer review panels to review proposals and prioritize relative value.

Future concerns

Today, an enormous number of increasingly complex databases underpin much of research in the life sciences. How to finance these databases presents an extraordinary challenge. The long-term solution will probably require both public and private funding, and will demand a suitable infrastructure to ensure that databases can be developed, maintained and exploited[17].

The research community already uses a mixture of publicly funded and commercial databases. If this combination of funding options is to continue, care must be taken both to establish what public funds will support, and to ensure that long-term funding strategies are available should public funding policies change. The role of central institutes such as the NCBI and EBI is vital. Such central services benefit from international cooperation and mutual data exchange, to which the highly successful EMBL/GenBank/DDBJ collaboration bears witness.

But what of the smaller, university-based 'specialty' databases? These need time to grow and become established as useful research tools. They require curators to validate, update and maintain them and to provide the best services to the widest group of users. In turn, this requires an assured funding structure. As current funding rules preclude support beyond the term of the research project, it is inevitable that many such databases will fold, and the scientific community will consequently have wasted an enormous amount of money and effort. In the late-1980s and early-1990s, hundreds of thousands of dollars of public funds were wasted in developing but not maintaining microbiology databases. The initial money came from research funds, but no policy was in place to support future maintenance[17]. A similar scenario might await the bioinformatics world.

Conclusions

The databases that service biological and biotechnological research have cost, and

will continue to cost, a great deal of money. Most new databases continue to be established via research grants, with little serious thought of how to finance their future.

If a database is successful, it might be integrated into larger database federations; it might spin off a commercial, proprietary version; and/or it might be mirrored around the world. Each of these growth patterns requires input from people skilled in both biology and computer science, and needs guidance from business and financial experts.

Unfortunately, many specialist databases are either too small or have too small a user community to be commercially viable. In such cases, partnership between public and private sectors might offer opportunities for sustainability. However, such relationships take time to develop. Thus, a mechanism for providing appropriate levels of continuing funding from public sources might facilitate such marriages, and consequently offer a brighter future for the databases of today and tomorrow.

Acknowledgements

TKA is a Royal Society University Research Fellow. We thank the U.S. Department of Education, National Science Foundation, and National Institutes of Health; the Biotechnology and Biological Sciences Research Council, the European Commission, GlaxoWellcome, and the Royal Society, for their support of our own database efforts. Finally, we thank the hundreds of database administrators without whom there would be no biological database infrastructure.

References

1 Barker, W.C. *et al.* (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.* 29, 29–32
2 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48
3 Benson, D.A. *et al.* (2000) GenBank. *Nucleic Acids Res.* 28, 15–18
4 Stoesser, G. *et al.* (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.* 29, 17–21
5 Tateno, Y. *et al.* (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.* 28, 24–26
6 Discala, C. *et al.* (2000) DBcat: a catalog of 500 biological databases. *Nucleic Acids Res.* 28, 8–9
7 Attwood, T.K. *et al.* (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28, 225–227
8 Ellis, L.B. *et al.* (2001) The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. *Nucleic Acids Res.* 29, 340–343
9 Ellis, L.B. and Kalumbi, D. (1999) Financing a future for public biological data. *Bioinformatics* 15, 717–722
10 Ellis, L.B. and Kalumbi, D. (1998) The demise of public data on the web? *Nat. Biotechnol.* 16, 1323–1324
11 Pool, R. and Esnayra, J. (2000) *Bioinformatics: converting data to knowledge: workshop summary*, National Academy Press
12 Powledge, T. (2001) Changing the rules? *EMBO Reports* 2, 171–172
13 Abbott, A. (2000) Europe boosts genome resource centres. *Nature* 408, 393
14 Butler, D. (1999) Life science facilities in crisis as Brussels switches off funding. *Nature* 402, 3–4
15 Bairoch, A. (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* 16, 48–64
16 Schalk, P.H. and Los, W. (1997) Electronic publication of biological databases. Does it pay? In *Financing Biotechnology Databases*, (J. Franklin, ed.), pp. 20–25, Biotechnology Information Strategic Forum
17 Franklin, J. (ed.) (1997) *Financing Biotechnology Databases*. Biotechnology Information Strategic Forum