seminars in **CELL & DEVELOPMENTAL BIOLOGY,** Vol 8, 1997: pp 455–458

Introduction: Making and filling gene-expression developmental databases

Jonathan B. I. Bard

AT FIRST SIGHT, this issue of *Seminars in Cell & Developmental Biology* might seem to be on a rather technical point, the storing of gene-expression data in databases, but it is actually about something much more important, despair and hope in the community of developmental geneticists! 'Despair' that the sheer amount of gene-expression data that now crowds the literature on each of the major developmental model systems is so great that no normal mortal can hope even to keep in touch with it, let alone remember it. 'Hope' that the storing of all this material in databases that are accessible over the internet will not only mean that we won't have to remember or even know it, but that we will have a ready tool for analysing and integrating this data, and moreover one that can be used without our having to leave our desks.

It is to the credit of the series editors that they have chosen the topic of gene-expression databases for one of their issues, as, unlike most of their subjects, this one is in a somewhat raw state. At the time of writing, though not I hope at the time of reading, there are no whole-embryo gene-expression databases yet accessible and only a few smaller, tissue-specific ones that are up, running and comprehensive. It is, however, important that these large databases are publicly considered as early as possible because they are being established for the benefit of the public and, if the scientific community will neither stock them nor use them, their establishment is a waste of time and money. If, on the other hand, we support those involved in, as Dr Johnson considered the writing of dictionaries, this 'harmless drudge', the mutual benefits will be very great. It is thus necessary that there be a serious, early and continuing dialogue between database writers and users.

From the Department of Anatomy, Edinburgh University, Teviot Place, Edinburgh EH8 9AG, UK

Gene-expression databases

The reader with little experience in obtaining information over the internet and in using the World Wide Web might wonder why it is worth bothering with databases when there is so much information available in libraries, and even those who already use computer-searching facilities may wonder whether fully fledged databases containing gene-expression data are needed.

One property of databases make all this effort worthwhile and that is their searchability: a database with a properly constructed query system will allow a user to ask sophisticated questions based around Boolean logic (combinations of A *and* B, A *or* B and A *not* B), linkage over time and even over specific space domains (if the gene-expression data is linked to a graphical model of the embryo — see later). This searchability allows the user not merely to ask questions about, say, the expression pattern of a single gene, but also to interrogate data links. Typical examples of the sort of question to which one might expect answers would be of the form *list all the transcription factors expressed in the developing mouse forebrain during E10.5* or, once the mouse graphical database, for example, comes on line *which signal proteins are expressed within 200 μm of somite 12 at E10*. With this sort of power, it becomes possible to integrate data of whose existence the user was completely unaware.

We are not yet at this stage, but the articles here tell readers what is available now, and what they can expect to find soon. The first four of these papers consider databases for the major model systems, *C. elegans*, *Drosophila*, the zebrafish and the mouse, and it is worth noting that the creators of these databases separately realized that their client communities would need access to gene-expression data, and all will come on line at about the same time.

The first two of these papers discuss gene-expression extensions to databases that were originally established to collate sequence, genetic, mutant and other information, and, not unexpectedly are about

C. elegans (ACeDB; Martinelli *et al*) and *Drosophila* (Flyview etc.; Janning). The third paper (Westerfield *et al*) discusses the new zebrafish database in some detail. Here, the reader is given a blow-by-blow account of how to build a large-scale database and will be reassured to see how important its makers view their interaction with their potential user community. In fact, this is true for all the databases considered here, but the zebrafish article details the steps that the makers have taken to ensure that what they are producing meshes with what its users want.

The remaining papers deal with a variety of mouse databases and the first of these (Ringwald *et al*) is also text based, linking mouse gene-expression data to named anatomical tissues during the first 14 days of embryogenesis. This database, which should come on line in 1998, contains a great deal of in-situ data that has already been published and will immediately be of very great use to the mouse community.

Although all four of these databases do contain pictures, they are essentially text-based for the purposes of searching as the query languages are based on words (e.g. anatomical structures). This approach does however have a serious limitation: the expression patterns of many important genes do not respect tissue boundaries; indeed, those genes responsible for the generation of new tissues should only be expressed in the relevant parts of larger tissues.

There are two ways of handling these complex expression patterns: by annotations to written text and by mapping the expression patterns to spatial domains in the embryo. The former is fairly easy to produce but hard to interrogate, the latter is difficult to produce but searchable, and the next two papers in this issue are concerned with the latter approach. It should be emphasized that spatial mapping involves very much more than a photograph of a wholemount expression pattern: it requires a full 3-D, digital reconstruction of the embryo in which each voxel (the 3-D equivalent of the 2-D pixel) represents a small, addressable domain. If such a 3-D reconstruction is to be useful, however, it also needs its component tissues to be digitally specified so that the relationship between gene-expression and anatomical domains can be visualized and searched.

The fifth paper (Baldock *et al*) is thus something of an interlude as it discusses the techniques for making 3-D digital reconstructions and the means for delineating domains within them, and the reader may be impressed by recent improvements in this area. This paper provides the background for considering graphical gene-expression databases in general and the following article (Davidson *et al*) in particular. This describes a graphical database for mouse gene-expression data that is currently being constructed and that complements the mouse text database to which it is linked (they use the same anatomical terms, and together make up the *Mouse Gene Expression Information Resource*). The reader will note that, even before this database contains a single item of gene-expression data, it has immediate uses as a powerful (and beautiful) tool both for analysing normal and mutant mouse histology and for learning about mouse development. It will also be apparent that the database will permit questions of considerable complexity to be asked.

This sophistication comes at a considerable price, however, as constructing a graphical database requires more people and a wider range of skills than text databases which themselves need considerable resources for their establishment. The final paper (Davies *et al*) discusses databases at the opposite end of the scale spectrum as it deals with the making and running of small databases, the sort that can be handled by a single person. Such databases are produced for the benefit of a small research community and are intended to be relatively easy to create, to maintain and, most important, to use. While the article on the zebrafish database shows users how a committed and funded group can construct something large-scale, this final article shows how a single person with almost no funding can provide a tool for their own research community. It may turn out to be the most useful article in this issue!

The future

By 1999, it is likely that there will be gene-expression databases on line for all the major developmental models and they should each be of enormous value in the major task facing this and the next generation of embryologists, the unravelling of the genetic networks that control the emergence of the developmental phenotype. What more could we want?

Embryologists who see no reason to move from their desks for their library work (and why should we?) will make two very reasonable demands. The first is simply met: given that the databases already contain gene-expression data, we will want more information about our model systems so that the databases can provide a complete resource. This is already so for those extant databases onto which gene-expression data is being bolted; it would be most helpful if all

databases were as comprehensive. One easy way to do this is to provide web links to other data resources and most of the databases discussed here already do this, It is thus of key importance that anyone setting up a new database informs all other database sites so that the new link can be added (Table 1 contains a few important websites for such embryological information).

The second wish is harder to meet as it requires improvements in computer technology and involves what is known as *interoperability*. Given that gene homologues do the same thing in different organisms, we would like to be able to move seamlessly from the database of one organism to that of another. This facility requires that databases be able to communicate directly with one another, something that might seem impossible given that each database has its own internal organization. There has however been some progress on this front as database designers are slowly generating a system called CORBA (Common Object Request Broker Architecture) that allows one database to present a standard interface for communicating with another and thus permits interoperability. While this is not yet in place, it is reasonable to expect that, in perhaps 5 years time, a developmental biologist with relatively simple computer skills will be able to ask when and where a gene is expressed in one organism, and how homologues of the gene behave in other organisms. We can then look forward to enjoying the fruits of all that drudge in-situ hybridization work!

The data submission problem

It would be nice to end on this note of optimism, but it is probably more sensible to inject some realism here and finish by mentioning the most serious problem associated with these databases and that is how to ensure that users fill them with data. It is unrealistic in the longer term to expect anyone other than the creator of the data to load it into the database, and he or she will not do so if the task is too demanding.

While gene-expression databases have enormous potential both as data repositories and as working tools for those interested in elucidating molecular networks, they will only achieve their potential if the user community is prepared to spend the time needed to load them with data. Storage of gene-expression data in databases is relatively new as compared to the storage of sequence, mutation and other such information, and this new information class is harder to collect and order than the more traditional data types. Nevertheless, creators of gene-expression data must expect to e-mail their data in a standard format to the relevant databases in the same way as those who generate sequence data.

As things now stand, this will not be easy. While sequence data lends itself to a digital format and database submission is a trivial exercise, expression data is both messier and harder to archive in a format that lends itself to electronic submission. The net result is likely to be that, if those who build and

Table 1. Some key websites for developmental databases dealing with gene expression and anatomy. The links associated with these sites extend across the whole of developmental biology

Name	URL
Key development biology sites	
WWW Virtual Library–Developmental Biology (part of the SDB webpage that has many links)	http://sdb.bio.purdue.edu
Zygote [the site associated with Scott Gilbert's *Developmental Biology* (5th Edn)]	http://zygote.swarthmore.edu/
Databases for the articles discussed here	
A *C.elegans* DataBase	http://www.sanger.ac.uk/Software/Acedb/
Flyview	http://pbio07.uni-muenster.de/
The zebrafish database	http://zfish.uoregon.edu
Mouse text gene-expression database (GXD)	http://www.informatics.jax.org/gxd.html
Mouse graphical gene-expression database	http://genex.hgu.mrc.ac.uk/
Ducted-gland gene-expression databases	http://www.ana.ed.ac.uk/anatomy/database/orghome.html
Kidney gene-expression database	http://www.ana.ed.ac.uk/anatomy/database/kidbase/kidhome.html
The tooth database	http://honeybee.helsinki.fi/toothexp/toothexp.htm
A few new or related sites	
Human anatomy database	http://www.ana.ed.ac.uk/anatomy/database/humat/
Atlas of primate brain	http://rprcsgi.rprc.washington.edu/natlas/
TBASE (targeted mutations)	http://www.gdb.org/Dan/tbase/tbase.html
Xenopus molecular marker resource	http://vize222.zo.utexas.edu/

maintain these databases are not very careful, they will find that the community will not be prepared to put in the effort needed to submit data. I therefore wish to spend a moment considering what might be a sensible strategy to ensure that the data is loaded.

The key person who needs to be persuaded that data submission is worth the effort is the person who has the data, and here I expect that, as usual, he or she will need the usual mix of carrots and sticks. The only plausible stick available is that journal editors may require, with submission of a manuscript, the correct accession number to the database, but I think that this threat is currently empty: it will be some time before the databases are sufficiently established for an editor to accept that gene-expression data must be submitted to them. If proper submission procedures are not in place soon, there will then be even more such data that has not been properly stored as compared to now.

It is however a well-known experimental observation that carrots are far more effective incentives than sticks in changing human behaviour, and I can envisage two incentives that might make people actually want to submit their data to a database: ease of publication and simplicity of private data storage. People want, indeed need publications, but it is becoming harder to publish simple gene-expression data in the major journals unless the work is part of a larger, experimental study. It would therefore be sensible if those setting up databases made arrangements with the editor of an appropriate journal so that, if the gene-expression data were refereed and found acceptable, a one-page summary of the work could be published in the journal on the basis that the detailed results were accessible via the database.

Ease of publication would not however be sufficient to submit information to the database if the amount of work required for submission was deemed excessive; database builders will therefore have to make it simple for users to submit data, simpler indeed than *not* submitting data! This ideal may be achieved using the approach taken by Ringwald and his colleagues for GXD, the mouse gene expression database: they are producing an electronic *annotator* or *notebook* that can be downloaded to the user for use in storing their gene-expression data as they read it off the microscope. This filled notebook would then be the place where users could access their own data privately and would meet the need for a properly structured format for storing, retrieving and analysing one's own gene-expression data.

Submitting the data in the notebook to the database then becomes trivially easy as the user will find on the first page of the electronic notebook a small button labeled 'submit'. This would display a simple submission page on which the key details would be written and the package could then be validated and e-mailed. The data would thus only have to be written out once, and would be for the direct benefit of the user. Can I therefore put in a plea that all database designers include such a notebook as part of their database, and warn them that, if they don't, they may find themselves in the embarrassing position of holding a party to which no one comes.

This metaphor is not quite as facile as it might seem as there is a sociological aspect to entering information into a database: if database submission is considered 'normal' behaviour by the community working on that model system, then everyone will, as a matter of course, submit their new data. Such is the accepted standard for the sequence, *Drosophila* and *C. elegans* databases, while the zebrafish community already sees the establishment of its database as a communal effort.

Gene-expression databases take a long time to write, but the dividends that they can produce will be an understanding of the molecular underpinnings of development, something that we all want to have. It is therefore incumbent on us all to ensure that the database enterprise is successful, while those of us involved in their production have to make it easy for the community as a whole both to submit data to them and to interrogate their contents.

Acknowledgements

I thank the European Science Foundation for their generous support. This organization sets out to stimulate research in areas that are too new to have surfaced on the agenda of large-scale funding agencies with their heavy bureaucratic infrastructures and long lead times. Some five years ago, the ESF decided to fund a Network that brought together workers from eight European countries interested in the various disciplines of building mouse gene-expression databases (anatomy, computer science, molecular genetics, etc.). For the Network's final meeting, it seemed sensible to hold a workshop for everyone in Europe and the USA interested in this area so that we would, for the first time, be able to compare notes and, *inter alia*, try to ensure that there could be linkages between one another's data bases. These articles are part of the spin-off from this meeting, and the costs of the colour plates that accompany them were met by the ESF to whom we are most grateful.