Funding high-throughput data sharing

Catherine A Ball, Gavin Sherlock & Alvis Brazma

The search for Truth is in one way hard and in another way easy. For it is evident that no one can master it fully nor miss it wholly. But each adds a little to our knowledge of Nature, and from all the facts assembled there arises a certain grandeur.

—Aristotle

Scientific progress cannot take place in a vacuum. Indeed, by its very nature, good science relies on peer review, replication of experiments and the accumulation of a body of data. Although the practice of sharing data upon publication is widely accepted as a fundamental tenet of good science[1], the application of high-throughput technologies to biological experimentation, such as large-scale DNA sequencing and microarrays, has resulted in new challenges to the implementation of this principle.

Most researchers recognize the importance of sharing data, but it is so fundamental that its advantages bear repeating. First and foremost, fully and freely available data promote and reinforce open scientific inquiry, allowing a researcher's conclusions to be validated or refuted by his or her peers. Second, it enables new analyses to be performed, which may lead to novel conclusions. This is especially important in light of the fact that rarely do researchers exploit the full potential of high-throughput data sets upon initial publication. Third, an accumulated body of public data can serve as the basis for new research and new methods of data analysis, and it provides large training and test sets for quality assessment. Fourth, access to public data can provide an excellent teaching resource. Fifth, the accumulation of public data provides all researchers with access to a data set that is larger than one that could ever be constructed by a single laboratory. It is clear that new knowledge and insight can be obtained from analyzing combined data sets, which would never be discovered examining the constituent parts. Lastly, sharing data can prevent unnecessary duplication of effort (though obviously some duplication provides rigor), and the public will benefit from a more rapid pace of scientific discovery that will be the result of decreased duplication and the creative reuse of published data.

This article outlines the central issues that face the research community in ensuring expeditious and efficient sharing of data. We start by discussing the benefits of data sharing and then describe the key components of a bioinformatics-based data-sharing infrastructure. We finish by describing the challenges in funding data-sharing initiatives and provide some potential funding solutions.

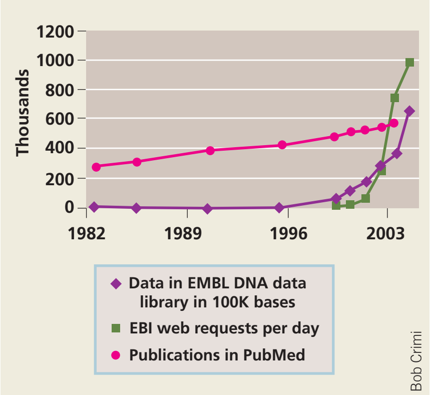
Benefits

A clear example of the benefits of data sharing comes from the explosion of available sequencing data. GenBank has experienced explosive and exponential growth over the last two decades, from just 606 sequences in 1982 to more than 30 million in 2003, comprising about 4×10^{10} base pairs (**Fig. 1**). No one would argue that the science would be better off if these data were held in private local databases—indeed, it is abundantly clear that the sum of the data is far greater than its constituent parts. Comprehensive, comparative genomic sequence analyses are possible only because all the data are available in the same format, in a single place. Furthermore, the wide availability of sequence trace reads provided large test sets for training base-calling software and for determining sequencing error rates. Once the sequencing community had tools to assess the quality of the data they were producing, they were able to produce much higher quality sequence. Additionally, the entire success of the Human Genome Project was based on data sharing using established data standards, tools and resources. An international collaborative project to profile gene expression for toxicology, involving more than 30 partners from industry and nonprofit organizations, coordinated by the International Life Sciences Institute[2], provides a preview of what should be expected in the future.

Although the need for data sharing in life sciences and biomedical research is well recognized, its actual practice can be challenging for members of a collaborative group, not to mention for an international community of researchers. The key to an effective method of sharing high-throughput data is developing and maintaining a robust, useful and dynamic informatics infrastructure. In the absence of dedicated resources, a 'quick and dirty' approach to bioinformatics is usually taken. Although such an approach may initially appear to be cost effective, it is usually a drain on resources in the long term, since it typically will not allow efficient use of the generated data, is not robust and requires large maintenance costs for little return. When a data sharing infrastructure has been developed properly, it can have a profound impact on scientific discovery, and the community of users of such a resource can hardly imagine doing research in its absence.

Examples of data-sharing infrastructures that are indispensable to scientific research include GenBank, PubMed and the *Saccharomyces* Genome Database. Although each of these databases has required signifi-

Catherine A. Ball is in the Department of Biochemistry and Gavin Sherlock is in the Department of Genetics, Stanford University School of Medicine, 269 Campus Drive, Stanford, CA 94305-5307, and Alvis Brazma is at the European Molecular Biology Laboratory (EMBL) Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. e-mail: brazma@ebi.ac.uk



Bob Crimi

Figure 1 The data explosion. There are two reasons why data sharing is becoming more and more challenging: first, the data is growing in size; and second, the data is growing in complexity. The first is easier to quantify, but the second is in fact the most important factor (see **Box 1**).

cant financial investment for both creation and maintenance, we do not wish to say that every bioinformatics effort requires a large supporting infrastructure; indeed, on the contrary, small, dynamic informatics programs by their very nature can be more innovative, creative and responsive. Even so, when the data reach a certain magnitude, and the size of the community wanting access to those data reaches a certain size, investment in a data-sharing infrastructure is not something that can be achieved cheaply if it is going to be robust.

Challenges

Once there exists either a critical mass of data, or the promise of that volume of data being generated, as well as a large group of potential users, it must be recognized that a data-sharing infrastructure is needed. There are four major components that make up the costs of a data-sharing infrastructure: development, data deposition, data management and data access.

Infrastructure development. Developing a functioning and widely accepted data-sharing infrastructure is challenging. In many ways, the developers of such an infrastructure face a chicken-and-egg situation. A common format, or standard, is needed to define the required data and their organization, such that users of the data source can retrieve the data in a single expected form. However, until such standards are adopted and widespread, data producers are unlikely to use any particular format. Thus, the architects of a data-sharing infrastructure are often faced with data being represented in many different *ad hoc* and poorly described formats (see **Box 1**). Although it is faster for a nascent resource to act individually, the involvement of many individuals and groups in a research community make the widespread adoption of a format, standard or community data repository more likely. As anyone who has ever tried to work in committee knows, accommodating input from a large community can be a slow process. Additionally, the burgeoning infrastructure for data sharing is frequently poorly funded.

Obviously, standards and data formats themselves do not enable data sharing—a supporting informatics infrastructure implementing these standards is needed. Usually the data-sharing infrastructure consists of centralized databases that store and curate the data, and a set of tools allowing users to access, retrieve, analyze and visualize the data. These databases can act either as public repositories where data are deposited (e.g., GenBank/European Molecular Biology Laboratory (EMBL)/ DNA Database of Japan (DDBJ))or as databases actively gathering or annotating their content (e.g., the UniProt protein sequence database or the *Saccharomyces* Genome Database). The complexity of the database will depend on the complexity of the respective domain and will necessarily have to respond to the dynamic and innovative nature of biological research. For instance, a database schema implementing a MAGE-ML-compatible model (see **Box 1**) for microarray data contains about 200 tables.

Data deposition. The placement of data (and metadata) in a repository has the potential to be a time-consuming exercise. Although many years of experience have simplified depositing nucleic acid sequence data, describing all the metadata necessary to interpret a microarray experiment is still far from simple. A step that would facilitate this process is the extension of the functionality of laboratory database or laboratory information management software (LIMS) that would provide a pipeline to the data-sharing infrastructure.

For instance, the Stanford Microarray Database (SMD) has constructed a direct data submission pipeline to the public repository for microarray data ArrayExpress at the European Bioinformatics Institute (EBI). The construction of this pipeline was not trivial and required the equivalent of approximately nine months of a person's time at SMD. The complexity of the MAGE-ML, and the fact that the standards and ArrayExpress were nascent, in part contributed to the difficulty of establishing this pipeline. Freely available software, developed by the MGED community, and the experience of the ArrayExpress staff likely mean that subsequent pipelines will be far easier to establish.

Data management. This component is most often overlooked, underestimated and ignored. Even with the best software and the best data entry and access interfaces, human intervention is required, preferably by someone with a high level of expertise and understanding of the biological field. Curators of databases or data repositories act fundamentally as advocates for the researchers who are submitting data to or obtaining data from a data-sharing resource. Curators can distill information from other sources, provide user assistance, build help documentation, provide a level of quality control over the data entered, design useful software and user interfaces, identify compelling new directions or exciting trends in research that the data-sharing resource should accommodate and identify tasks that would benefit from greater automation.

For a first-time submitter, completing an average-size experiment submission (data from about 30 arrays) into the ArrayExpress repository via the online submission tool may take a day to several weeks (depending on the submitters response time to the curators' questions) and requires hours of curators' time (for a returning submitter these times typically are much smaller). Disseminating incorrect (or incorrectly annotated) data can actually be harmful, and without adequate curation effort, this can easily happen. In addition to curation, there is a need for technical database administration, there are constant needs to upgrade and maintain the computer hardware and software, to increase the storage space (e.g., for microarray data, it measures in Gigabytes per array) and to upgrade and support the computer networks that house the data collections.

Data access. Providing access to the data requires intuitive and useful query interfaces and adequate standards for data communication. For instance, a database that allows only gene-by-gene based web queries would force a user to click tens of thousands of times to retrieve all the data for a single microarray experiment. On the other hand, a researcher interested in the position of splice sites within a single gene should not be forced to download and parse the sequence of an entire genome. Another confounding issue is the need to provide data in formats required for common data analysis techniques. The inability to access data in desired formats requires customized data parsing and translation software, which of course has costs as well. These costs are minimized by developing adequate data access interfaces and adequate standards.

In addition, researchers are often interested in finding answers that require data from more than one database. For example, a researcher may be studying a human transcription factor and want to know the three-dimensional structures of all proteins known to interact with the transcription factor or the three-dimensional structures of orthologs of the interacting proteins. This requires that the queried database not only store three-dimensional structures, but also store homology relationships between protein with solved structures and those whose structures have not been

Box 1 Problems posed by biological data

The two main difficulties faced by bioinformaticians developing and maintaining data-sharing infrastructures are the growing amount of data and the growing complexity of data. The first problem is illustrated by the growth of DNA sequence data (**Fig. 1**). The other data types, most notably microarray data, are experiencing the same or even faster growth, but as these are rather new and the public repositories collecting these data are even newer, no systematic statistics are available. The second problem—the growing complexity of the data—is more difficult to quantify, but in fact adequately capturing and communicating the complexity of high-throughput biological data provide the most challenging and critical problems.

We describe below some examples of successful and established data-sharing infrastructures (for nucleic acid sequence data and protein structure data) as well as some that are still developing (for protein-protein interaction data and microarray data) as illuminating examples of the complexities involved.

Nucleic acid sequence data. These can be represented as a one-dimensional string of letters that may be accompanied by the original sequencing traces. In sequence data, there is typically a high signal-to-noise ratio and often quality scores are associated with high-throughput sequences. Therefore, it may appear that the sequence data exchange standard can be very simple. Indeed, several simple formats, such as FASTA (FAST-All), have been developed and used to communicate DNA sequences. However, for many applications additional metadata need to be communicated—annotation of the location of genes, exons, introns, promoters, alternative splicing, single-nucleotide polymorphisms, among other data, which require considerably more complex standards. Such standards also need to accommodate gaps in genome sequence, the frequent lack of knowledge of relative positions of different sequence fragments and the frequent revisions of genome assemblies. A standard format to represent the metadata and the coordinate system is defined by the Ensembl database, and many of the fully or partly sequenced genomes and their annotation are now available in this format. Furthermore, there are several successful data repositories for sequence data. The strict repositories, such as GenBank/European Molecular Biology Laboratory (EMBL)/DNA Database of Japan (DDBJ) simply report sequences as reported by submitting authors, whereas others, such as Ensembl and GoldenPath, do considerable processing and analysis of raw sequence data.

Protein structure data. These are reported as atomic coordinates in three dimensions with their mappings to the protein sequence. The structures reported are largely independent of biological conditions under which those data were captured (or dependent in relatively well defined ways, such as conformation states). Measurement units (angstroms) and error models are well defined and accepted. Unlike the sequence data, where the basis of a common format is simply a sequence of letters, formats for describing the three-dimensional structures are much less intuitive. Moreover, the relationships between the crystallized monomer structures, and those of biologically active multimer molecules are far from straightforward, which further complicates the representation. The Protein Data Bank (PDB) and Macromolecular Structure Database (MSD) are accepted global repositories for three-dimensional molecular structure data. A *de facto* standard is the PDB format, but the research community has invested great energy into creating, modifying and reconciling different data formats (e.g., the mmCIF format).

Interaction data. Data describing protein-protein interactions are highly dependent on biological and experimental conditions under which those data were captured; therefore, metadata capturing this information are essential. Although the 'measurement unit' is well defined (that is, 'interact/not interact'*)*, there are complications that many interactions are not strictly binary, but potentially depend on the presence of other proteins, cofactors or protein modifications. The signal-to-noise ratio in the measurements is considerably lower than that for sequence or structure data, which again makes it important to have metadata describing the evidence for each particular interaction to be captured. A format for communicating these data has recently been developed[3] by the Proteomics Standards Initiative (PSI) working group of the Human Proteome Organization. There is an emerging collaboration between the EBI's IntAct (Open Source Molecular Interaction Database), University of California, Los Angeles' DIP (Database of Interacting Proteins), BIND (Biomolecular Interaction Network Database) and a few other databases, which has a potential to become the global federated repository for protein interaction data.

Microarray gene expression data. Not unlike protein-protein interaction data, these data depend on biological and experimental conditions. Thus, descriptions of the experimental and biological conditions, in addition to the data-processing protocols, are essential to understand the data fully. In most cases, signal-to-noise ratio is low and measurement units and/or type are often a function of the technology platform, or the software package used to analyze the microarray images. In addition, there are few widely accepted metrics for determining and communicating information about data quality. Nevertheless, the basic trends in gene expression data patterns are consistently reproduced by most microarray platforms and are demonstrably meaningful. Although the MIAME (Minimal Information About A Microarray Experiment) and MAGE-ML (Microarray Gene Expression-Markup Language) standards have been recently developed by the Microarray Gene Expression Data Society[4,5] in conjunction with community input, many gray areas remain, and much work still needs to be done in the area of data quality assessment. Community data repositories have been established for published data (ArrayExpress, CIBEX and Gene Expression Omnibus).

solved, or at least have access to that information. Thus, to be able answer such questions, databases should ideally be integrated with each other, in a fashion that is seamless to the end user, but does not simply require duplication of all database resources at every database. Web services seem to provide a promising way to avoid duplication of data and efforts in different biological database projects.

Finally, it should be noted that the development of biological data-sharing infrastructure is usually a continuous process. New requirements for data deposition and access are constantly emerging, and moreover, the databases should be continuously trying to share data transparently, to reduce duplication of bioinformatics efforts.

How much to invest in enabling data sharing?

To assess the optimal funding level for building and maintaining a data-sharing infrastructure, we should try to optimize the cost/benefit ratios. The costs, though sometimes in the millions of dollars, are still a fraction of the costs of data generation and are best measured as a percentage of the data generation costs, rather than in absolute figures. To assess the benefits, we should distinguish between the direct and indirect benefits. Direct benefits include how much is saved in terms of avoiding duplication in data generation, and indirect benefits include possibilities generated by entirely new analysis of combined data sets from different sources and possibly of different data types. One method of estimating the appropriate level of funding for a data-sharing infrastructure might be to estimate how much effort it would save directly or indirectly, and arbitrarily assign a fraction of those savings to funding the data-sharing effort. A simpler approach would be to assign an arbitrary fraction of the costs required to create the large-scale data to funding data-sharing projects. This option, although suffering from being arbitrary, is at least easier to calculate.

To assess the direct benefits, one could simply sum up the costs of generating the data *de novo* for every user accessing a chunk of data in a public database. In practice, such a calculation is difficult, as one would have to produce rather detailed user logs, a process which itself may be costly. Nevertheless, some idea can be obtained by looking at the usage of public data resources. For instance, the EBI website receives more than a million 'hits' daily, and during the first six months of 2004, over 25 Terabytes of data has been downloaded from the EBI's ftp (file transfer protocol) site. Assuming that every hit saves just 10 cents, every day of EBI's existence saves the research community $100,000. Estimating conservative sequencing costs of $0.01 per base pair, sequencing the DNA equivalent to 25 Terabytes would cost on the order of $\$500 \times 10^9 = \$500$ billion. Not all of the retrieved data would have to be regenerated, but assuming that just 0.1% would, the savings made over a six-month period would equal half a billion dollars, or one-sixth of the costs of the Human Genome Project. Moreover, it should be noted that the indirect benefits from data sharing are much higher than the direct ones, and are growing exponentially with every new data resource integrated in the common data-sharing infrastructure.

The funding required to build and maintain data-sharing funding infrastructure has been often targeted (although seldom delivered) at 20%–25% of the costs of generating the data. The results at SMD suggest that this is reasonably close to an adequate level of funding. We estimate that the labor and reagents (not including the equipment, scanners, computers or general lab supplies) to isolate RNA, amplify it (if necessary), make and label cDNA, perform the hybridization, and to scan and grid the image add up to approximately $520 per array. With 48,0000 arrays in the database, SMD supports approximately $25 million worth of data and has cost approximately 20% of that value to develop and maintain since its inception. Although it may be useful to have a rule of thumb about what level of funding should be dedicated to data-sharing efforts, it would be even more useful to have ways to define the data-sharing needs of high-throughput efforts at their inception and ways to evaluate whether they are being met or not.

Potential funding models for data sharing

Attempts to base life sciences data sharing on commercial models (e.g., subscription fees) have not enjoyed widespread success. One of the main reasons why commercial models have failed is that a subscription-based model does not allow one to download all the content of the database, combine it with other types of data and perform meta-analysis of combined data. This completely prevents all the indirect data-sharing benefits. Also, the overheads in dealing with hundreds of licenses would be unjustifiably high.

An example of a failed subscription model is provided by the experience of the Swiss-Prot database for protein sequences. Initially built as a publicly funded project, funding problems caused it to switch in 1997 to a subscription-based model (though it remained a free resource for academia). This prevented some other important genomics information resources from integrating with Swiss-Prot, and although it allowed the database to survive a funding crisis, this is regarded as an unsuccessful 'experiment.' Thanks to funding obtained from National Human Genome Research Institute (Bethesda, MD, USA), Swiss-Prot (now UniProt) has switched back to a model that provides free access for all users.

Among other reasons for limited success of a commercial approach is the highly experimental and dynamic nature of bioinformatics software projects (it is almost impossible to have a reasonable specification at the beginning of the project, as the data-generation technologies are developing all the time), the continuous nature of the development (virtually all bioinformatics databases continue the development work as long as they are used and effectively are R&D projects for their life span), limited commercial market (apart from a few large pharmaceutical companies, the main customers are from academia, which have to rely on open data, particularly to be able to publish findings), and the high risks and limited returns from the investments as a result of the combination of the above mentioned factors.

Though initial funding for the development of new data resources are often provided by the research grants, the funding agencies are much less sure how to fund the maintenance and further development of already established resources, and who should bear the costs associated with the data deposition. Committing funds to ongoing and potentially unending data-sharing projects that are perceived to be in 'maintenance mode' has little appeal to most governmental funding agencies.

It is very important to realize that costs of describing (annotating) and depositing high-throughput data are real, and that they should be budgeted for, in addition to the costs of generating these data. Funding agencies should require data-sharing plans and realistic cost estimates in all grants producing considerable amounts of high-throughput data. For instance, if the data are to be deposited in a public database, software allowing this should be either developed or purchased and staff responsible for the depositions should be funded. Currently, these costs are chronically underestimated and under-awarded. Although the US National Institutes of Health (Bethesda, MD) currently requires that most proposals have a data-sharing plan, it does not necessarily follow that the data-sharing infrastructure mentioned has stable funding or that its funding is part of the award. It should also be recognized that access to the generated data collections should be maintained long after the data generation has finished, and that maintaining access to the data cannot be accomplished for free.

The need for a data-sharing infrastructure will likely outlast any one of the various research projects that generate the data that require sharing. Accordingly, relying on small proportions of individual research grants to fund such an infrastructure is inherently unreliable, insecure and prone to fluctuation. Too many valuable components of our current data-sharing infrastructures survive on funding that is cobbled together from a variety of unstable sources: small parts of research grants, software development projects, gift monies or institutional funding. The construction and maintenance of a successful and valuable data-sharing infrastructure needs a different model to preserve long-term access to high-throughput data. Ideally, dynamic and evolving data-sharing infrastructure projects behave and produce results in a manner very much like research projects and should be funded as such in their own right. More importantly, successful endeavors should be encouraged (and funded) to provide wider access to their data-sharing infrastructure (ideally with flexible access, e.g., via web services), to prevent the proliferation and re-creation of similar projects at every research institution.

By reducing the resources spent to reinvent the wheel, the research community will surely benefit from stable, reliable funding for the widely used and essential data-sharing projects, and will also potentially reap the rewards of increased funding for innovative and high-risk projects. For those projects that provide essential infrastructure and services to a large community of researchers, this idea of stable funding could be taken one step further. Instead of a fixed-term funding period, such as three years, we believe that such resources merit rolling periods of funding. Under this model, such a resource could be granted three years funding on an annual basis, until such time as the project was deemed to be unfundable or completed. This would provide two years for essential data to be propagated elsewhere or for alternative funding sources to be identified. Under this model, researchers who rely on essential projects that serve a fundamental scientific need (e.g., UniProt or the *Saccharomyces* Genome Database) could have confidence in a resource's continued availability and would not be faced with the prospect of its loss without notice.

1. Cech, T.R. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences.* (National Academies Press, Washington, 2003) http://www.nap.edu/books/0309088593/html
2. Mattes, W.B., Pettit, S.D., Sansone, S.-A., Bushel, P.R. & Waters, M.D. *Environ Health Perspect.* **112**, 495–505 (2004).
3. Hermjakob, H., *et al. Nat. Biotechnol.* **22,** 177–183 (2004).
4. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).
5. Spellman, P. *et al. Genome Biol.* **3**, research0046.1-0046.9 (2002).