EDITORIAL

nature cell biology

Standardizing data

Biological research is benefiting from an explosion of data. There is an urgent need to invest in bioinformatic infrastructure and education to interpret this data and guarantee its archiving.

High-throughput research has helped fuel scientific progress at an unprecedented pace and left vast amounts of digital data in its wake. Even traditional hypothesis-driven research is now published at a rate that prohibits individuals from retaining the necessary overview. Bibliographic databases, such as PubMed, are key tools to navigate the information, but do not provide access to the primary data. The value of data is only as good as its annotation and accessibility: it must be properly curated and stored in machine-readable form in public databases. Indeed, the utility of data in the high-throughput age will depend on the establishment and long-term funding of an interlinked database infrastructure. It will equally depend on researchers contributing to and using these tools, as well as developing and adopting community-wide data standards. Several recently launched projects that aim to improve the value of data in the digital era are discussed below.

Most journals require that protein structure, gene or protein sequence and microarray data be deposited into primary databases, such as GenBank, UniProt (which incorporates Swiss-Prot, TrEMBL and PIR-PSD), the Protein Data Bank (PDB) and Gene Expression Omnibus (GEO) or ArrayExpress. Researchers upload basic datasets in a standardized form, which is validated and annotated to varying degrees by professional curators before deposition in the database. However, GenBank, the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database and DNA Data Bank of Japan (DDBJ) also contain archival data, so that multiple sequences from the same genetic locus may exist. As TrEMBL protein records are automatically computed from a submitted nucleotide record, an initial poor submission can spread through the databases; thus, there is an urgent need to impart 'minimum information' standards to improve the accuracy of these resources. Furthermore, apart from genome research, structural biology and microarrays, there are few centralized databases. Indeed, efforts to standardize data and to create ontologies in areas such as light microscopy and proteomics are progressing all too slowly.

How then can we ensure that researchers record the appropriate metadata to allow for accurate interpretation and repetition of data, and that the data are appropriately annotated before being deposited into databases? Several communities have developed data standards, but there has been little effort to coordinate these to avoid redundancy and incompatibility. Recently, the European Bioinformatics Institute (EBI) spearheaded the development of a shared platform for such standardization initiatives. The Minimum Information for Biological and Biomedical Investigations (MIBBI) project currently encompasses a collection of 22 minimum information guidelines on techniques such as microarrays, RNAi, quantitative PCR or FACS analysis. MIBBI aims to be a 'one-stop shop' for so-called checklist projects including MIAME (Minimum Information About a Microarray Experiment) and MIAPE (Minimum Information About a Proteomics Experiment). The 'Portal' section of MIBBI contains a growing number of links to the checklist projects, whereas the 'Foundry' invites input aimed at creating new, non-redundant checklists. Active community participation in the MIBBI Foundry will help ensure that the minimum information checklists remain relevant and thus, high-throughput data adequately annotated.

In addition to proper annotation, data must be described systematically in unambiguous language to make them machine-readable. To achieve this goal, communities must agree on ontologies (formal vocabularies for data and concepts). Ontologies allow semantic interoperability between various bioinformatics platforms and ensure that multiple repositories are compatible with each other. Although this remains a critical roadblock, the creation of the Open Biomedical Ontologies (OBO) Foundry, which acts as an umbrella organization for ontology projects, represents progress towards this aim. Importantly, in some fields the OBO has allowed the synthesis of overlapping, redundant ontology groups into a single community standard.

Adherence to MIBBI standards and formalized ontologies goes a long way to ensuring that databases are stocked with useful information, but there are also efforts underway to update existing records. The US National Center for Biotechnology Information (NCBI), which hosts GenBank, offers several tools to improve the 'state of the data'. One such tool is the Reference Sequence (RefSeq) database, which contains extensively curated and interlinked records from public nucleotide and protein databases; importantly, there is only one set of records per genetic locus. Although GenBank records can be updated by the submitting author, they seldom are. The Third Party Annotation (TPA) database allows any researcher to annotate and revise records from GenBank, EMBL and DDBJ on the basis of experimental and inferential data. Finally, the NCBI provides the Gene Reference into Function (GeneRIF) tool, allowing users to add a short description of gene function supported by a reference.

Community annotation efforts are not just limited to existing databases. 7,500 entries from Entrez Gene have been deposited into the Gene Wiki, a portal in Wikipedia allowing community editing. Similarly, Human Proteinpedia is seeded with information from the Human Protein Reference Database (HPRD), and allows individual users to contribute information on subcellular localization, tissue distribution, protein interactions and post-translational modifications. The Protopedia Wiki presents interactive and editable three-dimensional representations of proteins, RNA, DNA and other macromolecules derived from PDB. Finally, WikiPathways is a community project to create and maintain signalling pathways and networks, and is intended as a complement to the professionally curated Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and Pathway Commons databases. Each of these resources allows users to annotate otherwise static database records and can provide context and important new findings for (in some cases) decade-old entries. It remains to be seen whether the established databases will embrace Wikis as a means to update and enhance their records. Furthermore, because Wikis are 'for the community, by the community' projects, their success fundamentally hinges on their adoption by the scientific community.

Minimal standards are also critical for large research consortia as they address the problem of sharing and archiving large amounts of data. The cancer Biomedical Informatics Grid (caBIG), an open-access, open-source site sponsored to the tune of US$20 million per year by the US National Cancer Institute (NCI) and administered by the US NCI Center for Bioinformatics (NCICB), has taken this challenge head-on. The project connects 87 NCI-designated comprehensive cancer centres with each other and external scientists, government and commercial entities. caBIG was born from the need to collect, analyse and share reagents and genomic, proteomic and clinical trial data — both new data and paper-based 'legacy' data — generated by these institutions in a common format. It also aims to provide a common set of software tools and vocabulary for interrogating, integrating and discussing data through the adoption of a metadata repository. However, as Lincoln Stein points out in *Nature Reviews Genetics*, a continuing challenge for caBIG is the lack of support from end-users at the community level. The tools, resources and reagents are freely available; however, scientists and researchers must actively use this wealth of information for the project to be successful.

NPG supports these standardization projects and the publication guidelines of the *Nature* journals aim to reflect emerging community data standards. One ongoing project, in collaboration with the 'open microscopy environment' (OME), is to enhance microscopy images with systematic metadata (see October 2004 editorial). The UCSD-*Nature* Molecule Pages aggregate data on around 4,000 signal transduction molecules from a myriad of publicly available databases; a growing number of Molecule Pages (currently 500) are supplemented by expert-authored, peer-reviewed and referenced information on protein function and regulation. The database will soon be complemented by a Wiki to provide a forum to add and update information. Separately, NPG recently launched a centralized resource for updates and information about structural biology and structural genomics with the Protein Structure Initiative (PSI).

To borrow from another web-user, Spiderman: with great data comes great responsibility. Research communities must agree on common minimum data standards and ontologies, and must be ready to annotate, curate and maintain the data, be it from high-throughput assays and large research consortia or from single-pathway studies. Furthermore, funding agencies need to provide long-term support for the necessary 'cyberinfrastructure'. To this end, the US National Science Foundation (NSF) Community-based Data Interoperability Networks (INTEROP) program will make its first awards this fall to foster the development of standards and tools that allow 're-purposing' and sharing of digital data. NSF also supports the Plant Sciences Cyberinfrastructure Collaborative (PSCIC) and it will invest a further US$100 million over five years under the 'Datanet' cyberinfrastructure program. The US National Center for Research Resources (NCRR) at the National Institutes of Health (NIH) has spent approximately US$14 million annually to develop the Biomedical Informatics Research Network (BIRN), a collaborative site designed to allow researchers to share and exchange data using a formalized ontology. Meanwhile, in Europe, FP7 funding has been earmarked for bioinformatic infrastructure developments and the EBI is taking a leading role in such endeavours. The expansion and maintenance both of centralized online databases and bioinformatic tools at the international level is absolutely essential; stewardship for data has to be the remit of dedicated long-term institutions, as it is ineffective to develop such resources through temporary grants with postdoctoral researchers. Furthermore, researchers need to become more familiar with bioinformatics and the use of ontologies to unambiguously document their data and methods so that their research is machine-readable and archiveable.

The methodical application of minimum standards for database records as set forth in MIBBI, active community participation in Wiki-style projects and dedicated funding from major research initiatives will help ensure that data produced today can be found and used by humans and computers alike, a pre-requisite of *bona fide* systems biology.

Further reading and links: Connotea.org/user/ncb/tag/datastandardization

ELSO into EMBO goes

The European Life Sciences Organization is set to close shop at the end of the year by fusing with EMBO.

The ELSO congress has established itself as the largest gathering of cell biologists this side of the Atlantic. Last month ELSO organized the seventh such meeting — for the first time, this was in conjunction with the European Molecular Biology Organization (EMBO). After decreasing attendance over the last years, the ELSO meeting seems to have reached a healthy steady state of around 1,500 participants and has successfully established itself as the European counterpart of the American Society for Cell Biology (ASCB) annual meeting.

ELSO was founded in 1999 and one of its roles has been to provide a lobbying platform for the life sciences in Europe. However, according to its founder Kai Simons, funding and staffing remained too modest for it to be effective. The fusion with EMBO will provide dedicated staff to fulfil many of ELSO's functions; no doubt EMBO will continue its usual sterling job of conference organization and public outreach, in particular, liaising with the press and educational institutions. However, EMBO is a research organization funded by its member states — it is not an independent coalition of scientists. As such, it will not find it easy to fulfil the role of a lobbyist. At a time when European research funding and policies are increasingly defined by Brussels, and when the European Research Council is coming into its own, there is an urgent need for a strong, independent and united voice for molecular and cell biologists. The European Science Foundation (ESF) should present strong and independent scientific advice, analogous to the US National Academies; however, it remains to be seen whether its new CEO Marja Makarow will develop this role to build on the strength of ESF as a facilitator of trans-European research. Furthermore, the purview of ESF is broad, representing the natural and medical sciences, as well as the humanities. As such, ESLO was set up to provide an important function in complementing ESF, much like the important role ASCB has alongside the National Academies.

It is also a time when individual European countries are developing rather diverse responses to challenges thrown up by the biosciences, including the ethics of embryonic stem cell research, the safety of genetically modified foods, plants and livestock, and the balance of basic and applied research. National societies will certainly remain important in representing scientists at the national level, but they would be more effective with support from a pan-European organization. A united European voice would undoubtedly have greater influence with Brussels. The hope is that EMBO will find a way to shoulder this crucial role of the outgoing ELSO.