FEATURE

The future of biocuration

To thrive, the field that links biologists and their data urgently needs structure, recognition and support.

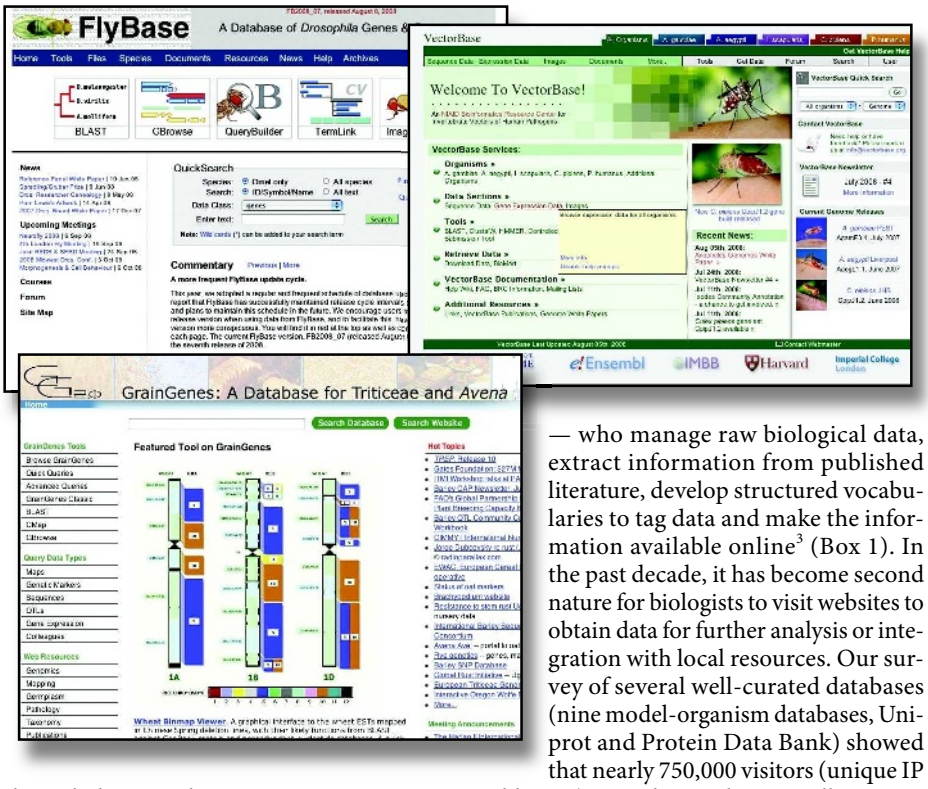
Doug Howe, Seung Yon Rhee *et al.*

The exponential growth in the amount of biological data means that revolutionary measures are needed for data management, analysis and accessibility. Online databases have become important avenues for publishing biological data. Biocuration, the activity of organizing, representing and making biological information accessible to both humans and computers, has become an essential part of biological discovery and biomedical research. But curation increasingly lags behind data generation in funding, development and recognition.

We propose three urgent actions to advance this key field. First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, researchers, academic institutions and funding agencies should, in the next ten years, increase the visibility and support of scientific curation as a professional career.

Failure to address these three issues will cause the available curated data to lag farther behind current biological knowledge. Researchers will observe an increasing occurrence of obvious gaps in knowledge. As these gaps expand, resources will become less effective for generating and testing hypotheses, and the usefulness of curated data will be seriously compromised.

When all the data produced or published are curated to a high standard and made accessible as soon as they become available, biological research will be conducted in a manner that is quite unlike the way it is done now. Researchers will be able to process massive amounts of complex data much more quickly. They will garner insight about the areas of their interest rapidly with the help of inference programs. Digesting information and generating hypotheses at the computer screen will be so much faster that researchers will get back to the bench quickly for more experiments. Experiments will be designed with more insight; this increased specificity will cause an exponential growth in knowledge, much as we are experiencing exponential growth in data today.



Data avalanche

Biology, like most scientific disciplines, is in an era of accelerated information accrual and scientists increasingly depend on the availability of each others' data. Large-scale sequencing centres, high-throughput analytical facilities and individual laboratories produce vast amounts of data such as nucleotide and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions and phenotype studies. By July 2008, more than 18 million articles had been indexed in PubMed and nucleotide sequences from more than 260,000 organisms had been submitted to GenBank[1,2]. The recently announced project to sequence 1,000 human genomes in three years to reveal DNA polymorphisms (www.1000genomes.org) is a tip of the data iceberg.

Such data, produced at great effort and expense, are only as useful as researchers' ability to locate, integrate and access them. In recent years, this challenge has been met by a growing cadre of biologists — 'biocurators' — who manage raw biological data, extract information from published literature, develop structured vocabularies to tag data and make the information available online[3] (Box 1). In the past decade, it has become second nature for biologists to visit websites to obtain data for further analysis or integration with local resources. Our survey of several well-curated databases (nine model-organism databases, Uniprot and Protein Data Bank) showed that nearly 750,000 visitors (unique IP addresses) viewed more than 20 million pages in just one month (March 2008, Eva Huala, Peter Rose, Rolf Apweiler, personal communications).

Despite the essential part that it plays in today's research, biocuration has been slow to develop. To provide a forum for the exchange of ideas and methods, and to facilitate collaborations and training, more than 150 biocurators met at two international conferences and created a mailing list and a website (www.biocurator.org). These meetings and discussions have honed in on the three actions, outlined above and elaborated on below, that must now be addressed to ensure scientists' continued access to the high-quality data on which their research depends.

Come together

Extracting, tagging with controlled vocabularies, and representing data from the literature, are some of the most important and time-consuming tasks in biocuration. Curated information from the literature serves as the gold-standard data set for computational analysis, quality assessment of high-throughput data and benchmarking of data-mining

algorithms. Meanwhile, the boundaries of the biological domain that researchers study are widening rapidly, so researchers need faster and more reliable ways to understand unfamiliar domains. This too is facilitated by literature curation.

Typically, biocurators read the full text of articles and transfer the essence into a database. For a paper about the molecular biology of a particular gene, process or pathway, such information might include gene-expression patterns, mutant phenotypes, results of biochemical assays, protein-complex membership and the authors' inferences about the functions and roles of the gene products studied. As each paper uses different experimental and analysis methods, capturing this information in a consistent fashion requires intensive thought and effort. Limited resources and staff mean that most curation groups can't keep up with all the relevant literature.

How information is presented in the literature greatly affects how fast biocurators can identify and curate it. Papers still often report newly cloned genes without providing GenBank IDs or the species from which the genes were cloned. The entities discussed in a paper, including species, genes, proteins, genotypes and phenotypes must be unambiguously identified during curation. For example, using the HUGO Gene Nomenclature Committee resource (www.genenames.org), we find that the human gene *CDKN2A* has ten literature-based synonyms. One of those, *p14*, is also a synonym for five other genes: *CDK2AP2*, *CTNNBL1*, *RPP14*, *S100A9* and *SUB1*. To confirm the identity of the gene described, curators make inferences from synonyms, reported sequences, biological context and bibliographic citations. This time-consuming and error-prone step could be eliminated by compliance with data-reporting standards[4–9].

Most recent efforts in this direction have been developed by the communities that produce large-scale genomics data. The vast majority of the peer-reviewed literature does not yet have a reporting-structure standard. As publication has become a mainly digital endeavour, however, publications and biological databases are becoming increasingly similar. Properly cross-referenced and indexed, each could serve as an access point to the other[10]. Such collaboration between databases and journals would improve researchers' access to data and make their work more visible.

We recommend that all journals and reviewers require that a distinct section of the Methods (or a supplemental document) of all published articles includes approved gene symbols (which are inherently unstable) and model-organism database IDs (which do not change) for genes discussed; nucleotide or protein accession numbers (GenBank or UniProt ID) for isoforms of each gene or protein discussed; and descriptions of species, strains, cell types and genotypes used. Examples of sources for this information are listed in Table 1. This would accelerate literature curation, uphold information integrity, facilitate the proper linkage of data to other resources and support automated mining of data from papers. Another model is for authors to provide a 'structured digital abstract' — a machine-readable XML summary of pertinent facts in the article[11] — along with a manuscript. This approach is in an experimental phase at the journal *FEBS Letters*[12].

Box 1 | The role of biocurators

- To extract knowledge from published papers
- To connect information from different sources in a coherent and comprehensible way
- To inspect and correct automatically predicted gene structures and protein sequences to provide high-quality proteomes
- To develop and manage structured controlled vocabularies that are crucial for data relations and the logical retrieval of large data sets
- To integrate knowledge bases to represent complex systems such as metabolic pathways and protein-interaction networks.
- To correct inconsistencies and errors in data representation
- To help data users to render their research more productive in a timely manner
- To steer the design of web-based resources
- To interact with researchers to facilitate direct data submissions to databases

Journals should also mandate direct submission of data into appropriate databases as a part of publication. This has been implemented by the journal *Plant Physiology* and curators of The *Arabidopsis* Information Resource (TAIR) database[13]. On acceptance of a manuscript, the corresponding author must fill out a simple web-based form to provide appropriate genetic and molecular information about the *Arabidopsis* genes in the publication. The information is sent to TAIR for integration by biocurators, who work with the authors to ensure that the data reported are of high quality and accurate.

"To date, not much of the research community is rolling up its sleeves to annotate."

As this infrastructure develops, we would like to see authors routinely tagging all aspects of the data in their publication semantically using universally agreed tag standards. Examples of such tags include the National Center for Biotechnology Information (NCBI) Taxon IDs, the Gene Ontology (GO) IDs and Enzyme Commission (EC) numbers. This information should be embedded in the electronic versions of publications or provided in a supplemental file similar to the crystallographic information file (CIF) currently required for publication of a crystal structure. The CIF file is submitted to the Protein Data Bank (www.pdb.org), which offers software to assist in preparation and validation of such crystallographic data[14]. An analogous system to help authors identify, tag and validate the crucial basic information in their research reports before publication would accelerate the automated linkage of literature to key records in existing databases and improve the accuracy of the published data.

In short, authors and publishers must use the existing publication infrastructure to facilitate literature curation much more to the benefit of all parties.

Community curation

Curation of large-scale genomics and post-genomics data enjoys no such luxury of 'an existing publication infrastructure' to leverage, although emerging standards of data reporting are promising[4–9]. Sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation. This transition will require annotation tools, standardized methods, oversight by expert curators and a combination of social infrastructure, tool development, training and feedback. Biocurators are especially important for establishing such an infrastructure and training to maintain consistency and accuracy.

To date, not much of the research community is rolling up its sleeves to annotate. What will be the tipping point? The main limitation in community annotation is the perceived lack of incentive. For example, several model-organism databases have requested that authors annotate the genes they publish. This has historically failed for one main reason: contributions by experts consist of information they already know, and do not increase the value of the resource to themselves. A mechanism tied to career or research advancement may be required before community curation can be established as a broadly accepted and productive scientific endeavour[15]. Incentives for researchers to curate data should include new information or insight for their research interests, improvement in academic reputation or impact, career advancement and better funding chances. Academic departments and funding agencies should consider community annotation as a productive contribution to the scientific research corpus and a natural extension of the publication process.

For example, in the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu) collaboration wiki, a community of more than 300 contributors took ownership of annotation of the genome while it was being sequenced at the Joint Genome Institute in Walnut Creek, California, and shared publication authorship as a consortium. Similarly, the International *Glossina* Genomics Initiative (http://iggi.sanbi.ac.za) hosted an annotation jamboree for field workers, population geneticists and molecular biologists to annotate tsetse fly molecular data as the sequence information became available. This

consortium-based publication mechanism is analogous to that used by other large-scale scientific projects such as the Sloan Digital Sky Survey (www.sdss.org). This is a viable course for communities that lack funding for dedicated curators, and offers a reward structure through consortium publication for participation and subsequent satellite papers.

The recently launched WikiProfessional Life Sciences (www.wikiprofessional.org) project links community curation with research and reputation gains. WikiProfessional indexed more than one million authors from PubMed and comparable numbers of biological concepts from authoritative databases and generated a simple way for researchers to update the information[16]. Because new potential 'facts' are mined from the network of associated concepts, the more accurate and comprehensive a particular concept is, the more chance it will have of being associated with other relevant ones, which in turn will lead to more potential new facts. All the updates researchers make are immediately publicly visible under their own name. Similarly, the Gene Wiki project generated thousands of wiki stubs in Wikipedia for human genes in an attempt to make it easier for the community to update the gene pages[17]. Although these wiki-based approaches provide an infrastructure for contributors to be recognized, there is not yet a standard practice for these contributions to be cited like a publication. It is imperative that the researchers, journal publishers and database curators start building a standard mechanism for citing annotation data sets.

Allowing anyone with a web browser, including the general public, to annotate entries would increase the number of potential annotators substantially, as pioneered in several astronomy projects. At Galaxy Zoo (www.galaxyzoo.org), 80,000 astronomers and members of the public manually classified the morphology of one million galaxies in less than three weeks. An analogous system to allow the public to contribute to biological annotation could be just as powerful if presented properly. For example, one could show a user an image of an *in situ* hybridization experiment and ask them to grade it as 'not expressed', 'restricted expression' or 'ubiquitous expression'. Even such basic information, if available for many thousands of genes, would be useful as first-pass annotation.

In sum, researchers (and even the general public) can be mobilized to provide the substantial resources needed to address the immense volume of data, if participation is appropriately rewarded. In the next five years, curators, funding agencies and academic institutions alike must find ways to consider substantial contributions to community curation efforts, much like a peer-reviewed publication, when it comes to issues of promotion, salary, hiring and funding.

Table 1 | Examples of knowledge-sharing databases

Species	Database	URL
Model organism databases		
Aedes aegypti	VectorBase	www.vectorbase.org
Anopheles gambiae	VectorBase	www.vectorbase.org
Arabidopsis thaliana	The *Arabidopsis* Information Resource	www.arabidopsis.org
Caenorhabditis elegans	WormBase	www.wormbase.org
Candida albicans	*Candida* Genome Database	www.candidagenome.org
Culex pipiens	VectorBase	www.vectorbase.org
Danio rerio	Zebrafish Information Network	http://zfin.org
Dictyostelium discoideum	dictyBase	http://dictybase.org
Drosophila sp.	FlyBase	http://flybase.org
Glycine max	SoyBase	www.soybase.org
Homo sapiens	HUGO Gene Nomenclature Committee	www.genenames.org
Hordeum vulgare	Barley Genetic Stocks Database	http://ace.untamo.net/bgs
Ixodes scapularis	VectorBase	www.vectorbase.org
Leishmania sp.	GeneDB	www.genedb.org
Mus musculus	Mouse Genome Informatics	www.informatics.jax.org
Oryza sp.	Gramene	http://gramene.org
Paramecium tetraurelia	ParameciumDB	http://paramecium.cgm.cnrs-gif.fr
Pediculus humanus	VectorBase	www.vectorbase.org
Rattus norvegicus	Rat Genome Database	http://rgd.mcw.edu
Saccharomyces cerevisiae	Saccharomyces Genome Database	www.yeastgenome.org
Schizosaccharomyces pombe	GeneDB	www.genedb.org
Solanaceae sp.	Sol Genomics Network	http://sgn.cornell.edu
Strongylocentrotus purpuratus	SpBase	http://sugp.caltech.edu/SpBase
Triticum sp.	GrainGenes	http://wheat.pw.usda.gov
Trypanosoma sp.	GeneDB	www.genedb.org
Xenopus laevis	Xenbase	www.xenbase.org
Xenopus tropicalis	Xenbase	www.xenbase.org
Zea mays	Maize Genetics and Genomics Database	www.maizegdb.org
Nucleotide, protein and structure databases		
All Species	GenBank	www.ncbi.nlm.nih.gov/Genbank
All Species	UniProt	www.pir.uniprot.org
All Species	Protein Data Bank	http://rcsb.org/pdb/home/home.do
Taxonomy		
All Species	NCBI Entrez Taxonomy	www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

Biological databases contain unique identifiers for the unambiguous identification of biological entities (scuh as genes, proteins, species and chemicals). These identifiers do not change as common biological names do. Authors should consult these databases for stable identifiers to cite in their publications.

Career path

How can biocuration mature faster as a career? Biocurators currently streamline submission to databases, automate curation, standardize data and facilitate contributions to annotation by research communities interested in the annotation process. To handle the increasing volume and types of data, journal publishers and researchers who generate data will need to be involved in the curation process and the roles of biocurators will expand to include editing and teaching. As biology moves towards more precise, quantitative science, biologists also need to adapt to thinking more quantitatively, systematically and objectively about their data; biocuration will need to become an inherent part of research and education in biology.

Biocuration requires a blend of skills and experience, including advanced scientific research and competence in database management systems, multiple operating systems and scripting languages. This type of background has typically been garnered through a combination of self-teaching and on-the-job experience, which can be narrow and spotty. Happily, formal education is becoming available. For example, the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign offers a biological information specialist master's degree and a specialization in data curation[18]. Experienced biocurators must lead the way in establishing more and better formal training programmes. In the next 5–10 years, biology curricula should include courses in biocuration as this becomes an increasingly common activity for all biological researchers. And interdisciplinary programmes that include courses in

biology, computer science and information science will be vital.

Attracting highly qualified individuals into this field has been challenging. The whole community must promote scientific curation as a professional career option. Funding agencies must assess the impact of curated data and support the development of innovative curation methods. To improve the profession, curators need a forum to share their experiences and publish their works. Oxford University Press plans to begin publishing a new journal in 2009 called *Database: The Journal of Biological Databases and Curation*. This may provide one such venue for publication of noteworthy advances in biocuration (www.database.oxfordjournals.org). Meanwhile, a committee of 20 biocurators and researchers is forming an International Society for Biocuration (www.biocurator.org/BiocuratorSociety.html) to make the discipline more visible and to promote it as an attractive career path. The official launch of the society is planned for the third International Biocuration Meeting next April in Berlin (http://projects.eml.org/Meeting2009).

Biology today needs more robust, expressive, computable, quantitative, accurate and precise ways to handle data. It is time to recognize that biocuration and biocurators are central to the future of the field. ■

1. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. *Nucl. Acid. Res.* **36,** D25–D30 (2008).
2. Wheeler, D. L. *et al. Nucl. Acid. Res.* **36,** D13–D21 (2008).
3. Salimi, N. & Vita, R. *PLoS Comput. Biol.* **2,** e125 (2006).
4. Brazma, A. *et al. Nature Genet.* **29,** 365–371 (2001).
5. Deutsch, E. W. *et al. Nature Biotechnol.* **26,** 305–312 (2008).
6. Field, D. *et al. Nature Biotechnol.* **26,** 541–547 (2008).
7. Jenkins, H. *et al. Nature Biotechnol.* **22,** 1601–1606 (2004).
8. Orchard, S. *et al. Nature Biotechnol.* **25,** 894–898 (2007).
9. Taylor, C. F. *et al. Nature Biotechnol.* **25,** 887–893 (2007).
10. Bourne, P. *PLoS Comput. Biol.* **1,** 179–181 (2005).
11. Seringhaus, M. R. & Gerstein, M. B. *BMC Bioinformatics* **8,** 17 (2007).
12. Seringhaus, M. & Gerstein, M. *FEBS Lett.* **582,** 1170 (2008).
13. Ort, D. R. & Grennan, A. K. *Plant Physiol.* **146,** 1022–1023 (2008).
14. Burkhardt, K., Schneider, B. & Ory, J. *PLoS Comput. Biol.* **2,** e99 (2006).
15. Rhee, S. Y. *Plant Physiol.* **134,** 543–547 (2004).
16. Mons, B. *et al. Genome Biol.* **9,** R89 (2008).
17. Huss, J. W. *et al. PLoS Biol.* **6,** e175 (2008).
18. Palmer, C. L., Heidorn, P. B., Wright, D. & Cragin, M. H. *Int. J. Dig. Curation* **2,** 31–40 (2007).

Author information Correspondence and requests for materials should be addressed to D.H. (e-mail: dhowe@cs.uoregon.edu) and S.Y.R. (e-mail: rhee@acoma.stanford.edu).

See Editorial, page 1.

Authorship

Doug Howe[1], Maria Costanzo[2], Petra Fey[3], Takashi Gojobori[4], Linda Hannick[5], Winston Hide[6,7], David P. Hill[8], Renate Kania[9], Mary Schaeffer[10,11], Susan St Pierre[12], Simon Twigger[13], Owen White[14] and Seung Yon Rhee[15]

[1]The Zebrafish Information Network, 5291 University of Oregon, Eugene, Oregon 97403-5291, USA. [2]*Saccharomyces* and *Candida* Genome Databases, Stanford University, Stanford, California 94305-5120, USA. [3]dictyBase, Northwestern University Biomedical Informatics Center, 750 N. Lake Shore Drive, 11-175, Chicago, Illinois 60611, USA. [4]Centre for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan. [5]J. Craig Venter Institute, Applied Bioinformatics, Rockville, Maryland 20850, USA. [6]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa. [7]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA. [8]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine 04609, USA. [9]Scientific Databases and Visualization, EML Research GmbH, Villa Bosch, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany. [10]Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA. [11]Plant Genetics Research Unit, Agricultural Research Service, United States Department of Agriculture, Columbia, Missouri 65211-7020, USA. [12]FlyBase, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Rat Genome Database, Bioinformatics Research Center, Medical College of Wisconsin, 8701 Watertown Plank Rd, Milwaukee, Wisconsin 53226, USA. [14]Department of Epidemiology and Preventative Medicine, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [15]The *Arabidopsis* Information Resource, Carnegie Institution for Science, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA.