nature biotechnology

Standard operating procedures

Is biological research ready for the new wave of data-reporting standards currently under development?

Understanding the complexities of standards development does not top the list of priorities for most biological researchers. Although standardized measurements (SI Units) and nomenclature systems are routinely used in research, few biologists have ever had to consider standardizing the way in which they report the experimental conditions, controls and parameters under which their data are generated. But this is just what is now being proposed by the developers of several new 'standards' initiatives that aim to facilitate the reporting and sharing of data in various fields of biology.

A standard is defined as a uniform set of specifications for some or all aspects of an activity or product that encourages cooperation and interoperation. Ideally, it should be clearly and unambiguously defined and as easy as possible to interpret and implement. In the modern world, standards have been applied to everything from light bulbs, PDFs (portable document formats), railroad track gauges and video formats to mobile phones, pharmaceutical purity and even bits of bathroom plumbing.

In the biological realm, the impetus for the current wave of data-reporting standards has primarily come from those interested in exploiting meta analysis and cross-comparisons of large data sets to provide new insights and guide new experimental directions. Ultimately, it is hoped that harmonization of data annotation will facilitate interoperability between genomic, proteomic and metabolomic data sources, enabling the modeling of comprehensive interaction networks and the elucidation of emergent system-wide properties.

One group of researchers that has been particularly forward-thinking in standards development is the DNA microarray community. In the early years of array technology, researchers struggled to harmonize the output of different platforms, identify the ancillary information needed to interpret results and even define the necessary data to enable reproduction of results. As a response, the Microarray Gene Expression Data Society (MGED) drew up the Minimum Information About a Microarray Experiment (MIAME) specification (*Nat. Genet.* **29**, 365–371, 2001), which in many respects has become the prototype for subsequent data-reporting guidelines in biology.

MIAME is now being joined by a plethora of 'minimum information' reporting standards initiatives that cover just about every large-scale technology under the biological sun. There is the Proteomics Standards Initiative's MIAPE (Minimal Information About a Proteomics Experiment), MGED's MIS-FISHIE (Minimal Information Specification for *In Situ* Hybridization and Immunohistochemistry Experiments), not to mention similar specifications for RNA interference (RNAi) experiments (MIARE), metabolomics studies (SMRS; ArMET), flow cytometry approaches (MI-FACE) and even cellular (MIACA) and enzyme activity assays (STRENDA).

Although the shared goal of these standards to promote data exchange and accessibility is laudable, achieving grassroots community 'buy-in' for several of them is unlikely to be straightforward (p. 1374). The fact is that researchers will only end up being users of a standard when they perceive that the benefits for their own work outweigh the inconvenience of compliance. In this respect, MIAME had the advantage that its developers were also its users. But in other fields the drudgery of entering data in a requisite format may offset any peripheral benefits for a researcher's own work.

In some instances—RNAi technology for example—the group of users will be relatively well defined and easily galvanized. But standards initiatives that attempt to engage much larger and diverse communities (e.g. biochemists or pathologists) are likely to face a more formidable task in getting their message out to relevant users. At the same time, those standards that attempt to address larger groups of researchers will find it more difficult to ensure broad and inclusive community consultation—a key step in the development of consensus standards (e.g., see http://www.nature.com/nbt/consult/index.html).

Standards that focus on emerging technologies where poor interoperability among instruments (e.g., mass spectrometry) or a lack of consistency in the description of experiments (e.g., RNAi) is hampering research thus seem much more likely to be rapidly adopted by the community than standards where the main beneficiaries appear to be that (albeit growing) minority interested in the collation and analysis of large-scale data sets. Of course, any standard that makes data more accessible to a broader audience is likely to result in more citations and a greater impact for a researcher's paper. And the types of large-scale analyses that would be enabled by data-reporting standards efforts may also aid users working on traditional reductionist problems; indeed, large-scale yeast two-hybrid experiments already have helped pinpoint individual proteins for further biochemical characterization.

One other lesson from MIAME is that efforts should focus on developing guidelines first, rather than formal standards. Because the language in a guideline need not be as unambiguous and rigid as a standard, user communities can resolve areas of contention before the formal standard is implemented. Paradoxically, one of the reasons that compliance with MIAME has been less than perfect (p. 1322) is that the specification's wording is more like a guideline than a standard, which has made interpretation of MIAME compliance problematic.

In the end, intervention by funding agencies may be required to provide certain end users with the incentive for standards adoption. In a manner similar to how publicly funded research findings must now be lodged in open archives (e.g., PubMed) after acceptance for publication or raw sequence data submitted to public repositories, funders could mandate implementation of certain data-reporting standards when researchers publish their work. This would prevent government agencies from wasting public funds on research that duplicates previous work because of inadequate data reporting and accessibility. And it would help realize the vision of the US National Institutes of Health and UK research councils, which have been strongly promoting high-throughput, large-scale biology.