Thou shalt share your data

Starting this month, *Nature Methods* strongly recommends deposition of proteomics data to public repositories before manuscript submission.

In many fields, the routine while writing up a paper is to deposit supporting data—be it genome sequence, microarray data or protein structure—into a community-endorsed public repository, and then to reference the accession number in the manuscript. There is no such habit in the field of proteomics.

In particular, mass spectrometry results are haphazardly reported in supplementary information, often in formats making their re-analysis impossible. Even worse, the sheer volume of data associated with many proteomics papers exceeds the supplementary information capacity of most journals. In the absence of raw data, reviewers cannot fully evaluate the conclusions of the paper, and other researchers cannot reproduce the results after publication. What is more, software developers miss opportunities to access test datasets, which slows the development of tools.

This sad state of affairs is not merely a cultural difference between proteomics and other fields. Rather, the lack of appropriate repositories and the large variety of formats have gravely impaired data sharing. This situation started to change a few years ago with the emergence of repositories and standardization initiatives. These resources have now reached a stage of implementation that motivates our present recommendation, by which we follow in the steps of our colleagues at *Nature Biotechnology* (*Nat. Biotechnol.* **25**, 262; 2007).

Several proteomics data repositories are now available that differ in terms of their goals, structure and the formats they accept. They include PRIDE (http://www.ebi.ac.uk/pride), PeptideAtlas (http://www.peptideatlas.org), Global Proteome Machine Database (gpmDB; http://www.thegpm.org/GPMDB/index.html) and the file distribution system Tranche (http://www.proteomecommons.org/dev/dfs/users/index.html). The newest addition, Human Proteinpedia (http://www.humanproteinpedia.org), is a community-based annotation tool that hosts experimental data (*Nat. Biotechnol.* **26**, 164; 2008).

Importantly, the major database administrators have shown their willingness to work with users and with each other to facilitate data deposition. At this stage, the process can still be labor-intensive, but a repository like PRIDE provides extensive technical assistance. Under the umbrella of the ProteomExchange consortium, the major repositories are also devising ways to share their data in a collaborative fashion, capitalizing on their complementarities to minimize submission hassle while maximizing benefits.

We support these efforts and consider it premature to recommend a particular repository. Rather we will rely on community experience to determine which database or combination of databases emerges as the most useful. However, there are specific features that editors favor. In particular, we like the possibility currently offered by PRIDE and Human Proteinpedia to provide peer reviewers with access to datasets associated with a manuscript before public release, in an anonymous fashion, and to coordinate public release of the data with publication. The archiving of a 'frozen' version of the data, associated with each paper, should also be considered by repositories that allow submitters to modify their own data over time. While updating may be important from the database perspective, it defeats the purpose of keeping a record specifically associated with a given paper, which is particularly important for methods papers.

From this journal's perspective, the entire raw datasets that support the evaluation of a method's performance should be shared, in a format that allows the widest range of users to validate and establish the method in their own lab. In this regard, the most raw form of data, the binary output from mass spectrometry instruments, is of limited interest at this time because it can only be read by proprietary instrument-specific software.

Remarkably, several instrument vendors, spurred by the efforts of the Human Proteome Organization Proteomics Standard Initiative (HUPO PSI), have agreed to implement a standard output format, mzML, co-developed by HUPO PSI and the Institute for Systems Biology. Once the ongoing testing is completed, new instruments should start arriving on the market equipped to output mzML files, which will greatly reduce the complexity of data deposition.

Several funding agencies are now demanding data dissemination with increased emphasis. This is in line with their financial support for creating repositories and should serve as an additional incentive for data deposition. Because effective data sharing rests on stable repositories, continued funding for database maintenance and improvement of the submission process will be key.

Repository administrators and instrument vendors have realized that data deposition must be made as effortless as possible to be widely adopted. Their efforts have prompted our present recommendation. It is likely that as more researchers deposit their data, remaining difficulties will be highlighted. We hope that the dialog between database administrators and data producers will solve the problems, to a point where data deposition cannot be considered an undue burden on investigators and can be mandated as part of manuscript submission.