A home for raw proteomics data

A new repository for raw data from proteomics mass spectrometry experiments is available and needs community participation.

Demands for access to data behind published research are increasing, and proteomics data are no exception, but compared to other data-intensive research, such as genomics, deposition of original proteomics data is less well organized. The only repository for the storage of raw mass spectrometry data, Tranche at the University of Michigan, has run into serious funding and personnel problems, making it all but unusable. Recent developments offer a solution, but it is dependent on community support.

When investigating the proteome with mass spectrometry techniques one generates three layers of information: raw data, results and metadata. Raw data are the binary files generated by the mass spectrometer that are then processed by the researcher to yield results such as peak, peptide and protein lists. Metadata are information about experimental context and the analysis tools used. The ProteomeXchange consortium, established in 2010, seeks to enable the sharing of these three data levels by linking key pubic repositories that host them.

But whereas repositories for results and metadata are well established—Proteomics Identifications database (PRIDE), for example, stores spectra and peak lists with their associated metadata as well as protein and peptide identifications as entered by the submitter—the loss of Tranche has left no repository for raw data.

Not everybody in the proteomics community sees this as a huge problem, because the need for storage of raw data is controversial. Some argue that not many people will want to reanalyze raw data and will rather work with peptide lists generated by the original experimenter, akin to the genomics community working with sequence reads rather than the raw images generated by some sequencing platforms.

But peak or peptide lists are insufficient for three important applications, which depend on raw data. First, the reanalysis of data by third parties can be important for validating or disproving the results of a study. Second, the data can be reprocessed with new questions in mind, such as examining different post-translational modifications than the original study. Third, as discussed in a Commentary on page 455, availability of raw data is critical for the development, benchmarking and improvement of computational analysis tools that may, for example, be used to interpret more of the raw data or perform more efficient database searches for correct peptide matching and thus yield new insights that benefit everybody.

The importance of raw data deposition is supported by some leading journals in the field such as *The Journal of Molecular and Cellular Proteomics,* which mandated it in 2010, but suspended this mandate early this year due to Tranche's problems.

So the big question is, where should experimentalists make their raw data publicly available?

The answer may be different for different scales of experiments. Big consortia such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC) will set up their own data storage and retrieval system in partnership with a commercial vendor, with the goal of making the data public on a regular schedule. How the data will be maintained after the project is completed is an open question though.

The revival of Tranche may be another option, but it would need support by more than one institution and commitment from the community to maintain its distributed file-storage platform. It will also need long-term funding.

The most promising proposal comes from the European Bioinformatics institute (EBI), who offered in a recent meeting of the ProteomeXchange consortium to store raw data. The goal is to create a comprehensively annotated dataset that includes all raw data, results and metadata. As of 2 April 2012 the EBI has accepted two submissions, and it will slowly ramp up and direct people submitting data to PRIDE to include raw data.

Although the raw files are large, the storage requirements of current proteomics data are substantially less than those generated by the genomics community. EBI has an excellent track record of storing and distributing sequence information, as evidenced by their involvement in the 1000 Genomes Project (see Perspective on page 459 that discusses community access to 1000 Genomes Project data), so their offer is realistic and sustainable.

Researchers at the EBI stress their willingness to provide a service to the proteomics community, but at the same time they will monitor download activity to gauge interest. As Henning Hermjakob, team leader of Proteomics Services at EBI, puts it, "if we have terabytes of data, which are downloaded by one user per year, it is likely the service will be discontinued."

We strongly encourage individual investigators to show that such a repository is valued by taking the time and effort to upload raw data. The loss of another option for the storage of raw proteomics data would be a serious setback for the community.