11200 Rockville Pike, Suite 302
Rockville, Maryland 20852
USA

American Society for Biochemistry and Molecular Biology, Office of Public Affairs

BENJAMIN W. CORB
Director

JULIE M. MCCLURE, PH.D.
Science Fellow

THE NEED FOR LONG-TERM PROTEOMIC DATA STORAGE

SUMMARY

The lack of a reliable and secure repository for raw data is a major problem facing science. While there are various repositories for 'processed' information these have substantial limitations and thus only serve a portion of the need (and the community), and importantly cannot store 'raw' data. Therefore there is an essential need for such an entity. This can be accomplished by providing long term fiscal support for creating an over-arching structure, actually capable of capturing not only raw data but also various forms of processed information that would provide a central storehouse.

SUPPORTING EXAMPLE: PROTEOMICS

OVERVIEW

One of the most significant hallmarks of biomedical research in this century, and perhaps one of the most unexpected, has been the size and extent of data sets that have and continue to be generated by the new technologies associated with genomic, transcriptomic, proteomic and metabolomic research (collectively the bio-omic sciences or, by some definitions, systems biology). The microarray field that underpins transcriptomics led the way but it has been supplanted by the massive outputs of next gen nucleic acid sequencing of vast numbers of human and other genomes. However, proteomic data, mostly generated by high throughput mass spectrometry (MS), will eventually dwarf both of these and when coupled with metabolomic data that will likely be collected with similar technology, is destined to create an almost unimaginable amount of information. At issue, therefore, is how to deal with this onslaught?

Clearly the problems for the individual 'omic sciences are not the same, as the types of data are quite different (excepting proteomics and metabolomics). Germane to this report is the collection and interpretation of MS data. There are several issues and several levels of data and each requires its own consideration. For ease of presentation, MS data, in support of a proteomic (or metabolic) experiment can be classified as 'raw', 'processed' or' interpreted'. The interpreted data are suitable for publication and for inclusion in searchable web-based compendia. These are outputs of search engines, which have interpreted the processed data in the form of peak lists or spectral libraries, and can involve additional software analyses including, but not limited to, quantification and functional assessments. Journals that publish proteomic data have various requirements for what information must be included in research articles and how much of the data from which the

identifications of peptides, proteins and post-translational modifications (PTM) were extracted must accompany the manuscript (during review and/or ultimately appearing in the journal, mainly as supplemental material). The extent to which the validity of these assignments can be assessed is accordingly equally variable.

To address this issue, *Molecular & Cellular Proteomics (MCP)*, starting in 2003 and culminating in 2005 (1), developed and adopted publication guidelines for reporting MS identifications and has subsequently updated them (2). As part of this evolution, in 2010 it announced (3) that it would require the deposition of the raw MS data in a public database as a requirement of publication for all accepted papers containing MS identifications. While not mandating it as a requirement, other journals publishing in this area of research supported this policy. For all practical purposes, Tranche, founded and operated out of the University of Michigan, is the only repository capable of handling this type of data submission. Unfortunately, technical problems, due mainly to inadequate fiscal support and largely manifesting themselves in the past year, have substantially curtailed the usability of Tranche and in March of the past year MCP was forced to make raw data deposition once again voluntary. Although the situation has shown signs of improving in the last six months, there is no sustained support of Tranche that has been identified. Thus, at the moment there is not a suitable and reliable repository for raw MS proteomic data available.

Why deposit raw MS data?

There are a number of reasons for why this policy should be universally adopted. First, the interpretation of MS data depends on software analyses and there is considerable variation in the search engines, how they make their determinations, and how they decide whether a result is reliable. It is important to understand that generally less than 50% of the spectra generated in an experiment are interpreted (and sometimes considerably less than that) and that assignments are given scores that indicate the probability that the identification is right after making certain assumptions about what could be in the sample. This is compounded by errors in the databases searched and in the possibility of matching a correct sequence to an incorrect protein. This is considerably exacerbated when PTMs are involved and localizing the modification sites correctly is clearly the most challenging analysis of all. The most effective way to re-examine an assignment is to have access to the raw data. Related to this, software for processing and interpreting MS data continue to improve, so re-analysis of datasets with newer software is likely to lead to the extraction of more information from previously acquired data. However, this can only performed if the raw data is available. Second, essentially all experiments are designed and executed with a purpose, i.e. there is a biological question being addressed. This means that the data will be analyzed from the orientation of this objective, and other information present in the data set will likely be ignored or simply not identified (i.e. be part of the 50% or greater of the data that was not explained during the data analysis). In addition, quantitative information present in the raw data may not have been examined (only qualitative analysis; i.e. peptide and protein identification is performed for many datasets). In fact, it may not be possible to interrogate a data set at the time it is collected for a specific question or possibility because the requisite findings that underlie it had not been previously determined. In essence, this is a manifestation of the axiom that one “sees only what one

looks for". This is particularly true for PTM analysis, as for most datasets only a very limited number of PTMs are considered during data analysis. As a result, potential large amounts of information are not analyzed and the information contained therein lost if the raw data is not made available. This is enormously wasteful from both an intellectual and financial point of view. Finally, knowledge is a continuum and all data collected adds to it. This is particularly important to the bioinformaticians and other analyzers of processed and interpreted data, who can provide the larger prospective that helps to produce the global understanding of biology and medicine, which is the real goal of the bio-omics. By not reporting the actual data collected or placing it where it can be used by others, it defeats a major part of what experimentation is supposed to be about.

It must fairly be pointed out that not everyone is in favor of raw data deposition. Some individuals, clearly recognizing that large MS data sets have unused or undiscovered potential and not wishing to have this be exploited by others, do not want to share their raw data, rather hoping to find new things in it themselves. Others are concerned that their misinterpretations and mistakes would be made plainly (and painfully?) available for others to point out and thus for all to see. And lastly, some simply don't want to be bothered with the hassle of making the necessary uploads, which can indeed be time consuming. Although in part understandable (from the human nature point of view), none of these reasons are particularly compelling or scientifically and fiscally well justified.

What is needed?

It should be made clear that the shortcomings of Tranche are basically due to lack of support rather than any inherent design flaws. It was created as an academic exercise and was largely supported initially by grant funds. When these were ultimately not renewed, it became difficult to maintain the servers and deal with user problems. Ultimately the principal designers and creators of Tranche left the project and were not appropriately replaced for financial reasons. Although data does still flow in and out of Tranche, the reliability of these activities and consequently the integrity of the data is not at earlier levels (and below the threshold that could be tolerated by *MCP*, leading to its decision not to make raw data storage mandatory until the situation is sufficiently rectified). While an infusion of money would certainly help (and there has been recently a small amount generated by the ProteomeXchange network, supported by a grant from the European Union), it is the consensus of a number of interested parties, which has been expressed at several international workshops and meetings, that either permanent support for Tranche needs to be identified or a new entity needs to be created with a reliable basis of support that would ensure the long term viability of the enterprise. The latter, which could be described as an International Repository for Proteomic Data (IRPD), would require a central facility and mirror sites placed in appropriate locales internationally and be staffed with network administration / IT staff to oversee its operation.

The stakeholders in such an IRPD would be of several varieties. First and foremost, the publishers of the main proteomic journals would be expected to be prime users. The American Society for Biochemistry and Molecular Biology, who publishes MCP, would be a strong supporter of such an activity but it can be expected that other publishers would be as well. The Nature Group is on record as actively supporting raw data

deposition. Based on the activities with other 'omic sciences, various private and public funding agencies are likely to be so as well and instrument and software vendors have a vested interest in this process (and have actively participated in workshops and discussion panels addressing this issue). Various government laboratories and agencies have also expressed support in the past. Finally, there are the end users – the scientists who create this data and then ultimately use it for different purposes. There seems to be no lack of support for the concept among any of these groups – only in the process of administration. Such a repository would presumably also become part of the ProteomeXchange consortium, which has international membership, who would be able to provide additional advice and potentially limited financial support.

REFERENCES

1). R. A. Bradshaw, A. L. Burlingame, S. Carr and R. Aebersold (2005) "Protein Identification: The Good, the Bad, and the Ugly" *Mol Cell Proteomics 4: 1221-1222.*

2). R. A. Bradshaw, A. L. Burlingame, S. Carr and R. Aebersold (2006) "Reporting Protein Identification Data: The next Generation of Guidelines" *Mol Cell Proteomics 5: 787-788. doi:10.1074/mcp.E600005-MCP200*

3). R. A. Bradshaw and A. L. Burlingame (2010) "Technological Innovation Revisited" *Mol Cell Proteomics 9: 2335-2336. doi:10.1074/mcp.E110.005447*