research articles Journal of proteome research

Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles

Henry Rodriguez,*,† Mike Snyder,‡ Mathias Uhlén,§ Phil Andrews,|| Ronald Beavis,⊥ Christoph Borchers,# Robert J. Chalkley,▽ Sang Yun Cho,○ Katie Cottingham,◆ Michael Dunn,¶ Tomasz Dylag,+ Ron Edgar,□ Peter Hare,■ Albert J. R. Heck,● Roland F. Hirsch,▼ Karen Kennedy,☆ Patrik Kolar,+ Hans-Joachim Kraus,★ Parag Mallick,○ Alexey Nesvizhskii,● Peipei Ping,⊙ Fredrik Pontén,◒ Liming Yang,◐ John R. Yates,⊕ Stephen E. Stein,⊡ Henning Hermjakob,◩ Christopher R. Kinsinger,† and Rolf Apweiler◩

Center for Strategic Scientific Initiatives, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892, Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06620, KTH Biotechnology, KTH - AlbaNova University Center, Stockholm, Sweden, Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, Michigan 48109, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada, University of Victoria Proteomics Centre, Victoria, British Columbia, Canada, Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California, 94158, Yonsei Proteome Research Center, Yonsei Univerity, Seoul, Korea, Journal of Proteome Research, American Chemical Society, Washington, D.C. 20036, Wellcome Trust, London, United Kingdom, European Commission, Directorate General for Research, Brussels, Belgium, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20892, Nature Biotechnology, New York City, New York 10013, Netherlands Proteomics Centre and Utrecht University, The Netherlands, Office of Biological and Environmental Research, U.S. Department of Energy, Washington, D.C. 20585, International Genomics Program, Genome Canada, Ottawa, Ontario, Canada, Wiley-VCH Verlag, Weinheim, Germany, Spielberg Family Center for Applied Proteomics, University of California, Los Angeles, Los Angeles, California 90048, Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109, Division of Cardiology, University of California, Los Angeles, School of Medicine, Los Angeles, California 90095, Department of Genetics and Pathology, Uppsala University, Uppsala, Sweden, Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, Maryland 20892, Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California 93037, Mass Spectrometry Data Center, National Institutes of Standards and Technology, Gaithersburg, Maryland 20899, and European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

Received January 9, 2009

Policies supporting the rapid and open sharing of genomic data have directly fueled the accelerated pace of discovery in large-scale genomics research. The proteomics community is starting to implement analogous policies and infrastructure for making large-scale proteomics data widely available on a precompetitive basis. On August 14, 2008, the National Cancer Institute (NCI) convened the "International Summit on Proteomics Data Release and Sharing Policy" in Amsterdam, The Netherlands, to identify and address potential roadblocks to rapid and open access to data. The six principles agreed upon by key stakeholders at the summit addressed issues surrounding (1) timing, (2) comprehensiveness, (3) format, (4) deposition to repositories, (5) quality metrics, and (6) responsibility for proteomics data release. This summit report explores various approaches to develop a framework of data release and sharing principles that will most effectively fulfill the needs of the funding agencies and the research community.

Keywords: proteomic • data • policy • release • resource • sharing • Bermuda principles • Amsterdam principles • open • standards

Introduction

On August 14, 2008, members of the international proteomics community met for a one-day summit in Amsterdam, The Netherlands convened by the National Cancer Institute (NCI) of the U.S. National Institutes of Health (NIH).[1] This summit was undertaken to address what is seen as a considerable obstacle to accelerating the pace of discovery in proteomic research: the lack of widely followed policies governing the rapid release of large-scale proteomic data into the public domain, taking into account data quality, standards and integration, intellectual property, ethics, and

* To whom correspondence should be addressed. Henry Rodriguez, PhD, MBA; Director, Clinical Proteomic Technologies for Cancer; Center for Strategic Scientific Initiatives; National Cancer Institute; National Institutes of Health; 31 Center Drive, MS 2590; Bethesda, Maryland 20892. Tel.: 301-496-1550. E-mail: rodriguezh@mail.nih.gov.

10.1021/pr900023z CCC: $40.75

Published on Web 04/03/2009

sustainability (ensuring that there are incentives to the creation of data sets).

Rapid public data release has long been standard practice within the large-scale genomics community. It also is standard practice for the field of macromolecular structure determination. It is widely felt that this practice—made possible by the existence of universally endorsed policies governing the standards for and the availability of data in the public domain, as well as centralized repositories and portals for depositing and accessing such data—has been a driver of the rapid pace of genomic discovery. The proteomics community would benefit greatly from adopting an appropriately similar practice.

Representatives from the proteomics community were well represented at this focused meeting. Attendees included data producers, data users, databases repositories, scientific journals (Journal of Proteome Research, Molecular and Cellular Proteomics, Nature Biotech, and PROTEOMICS and PROTEOMICS—Clinical Applications), and funding agencies (National Cancer Institute, Department of Energy, European Commission, Wellcome Trust, and Genome Canada).

This postsummit document discusses the basic principles underlying data release in genomic research, the challenges to developing similar principles in the proteomics domain, and also how to synthesize the data release and sharing principles proposed by the Amsterdam summit attendees.

The Bermuda Principles. The data sharing policies of the U.S. National Human Genome Research Institute (NHGRI) and other genomic research funding bodies (i.e., those that were engaged in the International Human Genome Sequencing Consortium) are derived from a series of principles discussed and agreed upon at the First International Strategy Meeting on Human Genome Sequencing, held in Bermuda in 1996.[2] Called the "Bermuda Principles," these guidelines were intended to apply to "all human genomic sequences generated by large-scale sequencing centers, funded for the public good, in order to prevent such centers establishing a privileged position in the exploitation and control of human sequence information." As such, they state that genomic sequences should be "freely available and in the public domain as soon as possible in order to encourage research and development, and to maximize its benefit to society." In addition, the principles also called on genomic research centers to release sequence assemblies as soon as possible and to submit finished annotated sequences to public databases immediately.

Knowing that the availability of high quality data was of supreme importance to the success of the Human Genome Project and to subsequent efforts to translate the knowledge it generated, the Bermuda Principles were expanded at the Second International Strategy Meeting on Human Genome Sequencing to include standards for sequence quality and suggested standards for sequence annotation.[3] In addition, guidelines for scientific claims and etiquette were proposed so as to minimize conflict within the community regarding the rights of data producers and users.

The principles promulgated in Bermuda were reaffirmed at a meeting sponsored by the Wellcome Trust in 2003.[4] Meeting attendees also further expanded upon the Bermuda Principles in two ways. First, it was agreed that the principles of prepublication data release should be extended to "other types of data from other large-scale production centers specifically established as 'community resource projects." Such projects were defined by the meeting participants as projects "specifically devised and implemented to create a set of data, reagents, or other material whose primary utility will be as a resource for the broad scientific community."

Second, the meeting participants addressed conflicts between the interests of data producers—who desire to publish the first analyses of their own data—and those of data users—the members of the scientific community seeking rapid access to genomic data for further study. It was agreed that each of three core constituencies in large-scale biological research-data producers, data users, and funding agencies supporting and facilitating such research-shared responsibility for ensuring the growth and development of community resource projects while addressing each constituencies' interests.

Challenges Addressed at Summit to the Rapid Release of Proteomic Data into the Public Domain. The challenges to rapid proteomic data release can be divided into three categories: technical, infrastructural, and policy. Each category, however, impacts the other two. Therefore, in developing principles for proteomic data release, summit participants took a comprehensive approach, addressing each category as it applies to the overall issue of data release.

Technical Challenges. Such challenges stem from the variability that exists in nearly every aspect of proteomic data generation, interpretation, and presentation. Proteomic data can be generated using a long and growing list of experimental platforms. Mass spectra alone can be generated by MS, tandem MS, liquid chromatography-MS, and other methods. A single instrument platform can be used to produce more than one kind of data. Tandem MS, for instance, produces data on ionized peptides that can be divided into quantitative data and identification data.

Individual laboratories also tend to develop their own processes and procedures for equipment calibration; thus, mass spectra generated on the same instrument by two different laboratories using the same reagents may be incomparable because each lab calibrates its equipment differently. Also, no standardized sources exist for experimental reagents, adding to the difficulties of accurately comparing or replicating data generated across laboratories.

† National Cancer Institute, National Institutes of Health.
‡ Yale University.
§ KTH - AlbaNova University Center.
|| University of Michigan Medical School.
⊥ University of British Columbia.
University of Victoria Proteomics Centre.
▽ University of California, San Francisco.
○ Yonsei Univerity.
◆ American Chemical Society.
¶ Wellcome Trust.
+ European Commission.
□ National Library of Medicine, National Institutes of Health.
■ Nature Biotechnology.
● Netherlands Proteomics Centre and Utrecht University.
▼ U.S. Department of Energy.
☆ International Genomics Program. Current Address: Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge.
★ Wiley-VCH Verlag.
⬡ Spielberg Family Center for Applied Proteomics, University of California, Los Angeles.
⬢ University of Michigan.
⊙ Division of Cardiology, University of California, Los Angeles.
◒ Uppsala University.
◐ Center for Biomedical Informatics and Information Technology, National Cancer Institute. Current Address: Division of Biomedical Technology, National Center for Research Resources, Bethesda, Maryland 20817.
⊕ The Scripps Research Institute.
⊡ National Institutes of Standards and Technology.
◩ European Bioinformatics Institute.

Raw, unprocessed data is considered to be the best and most accurate representation of an experiment's results. However, numerous instruments have been developed for each platform, each of which produces raw data in a proprietary format developed by the instrument's manufacturer. Thus, raw data can be difficult, if not impossible, to interpret or compare across laboratories unless (1) a data user has the same instrument and software package as a data producer; or (2) the data are converted to an open format. However, data format standards have not yet been widely agreed upon within the community. Currently, the community standard format mzML[5] and the accompanying controlled vocabulary are still in the initial stages of broad community acceptance. There is also the risk that some information will be lost in the process of converting data into an open format.

In addition to variations in data generation, there also exist a variety of analytical options for peptide identification (e.g., identification based on peptide mass and retention time, comparison of fragmentation data to theoretical fragmentation, comparison of fragmentation data to previously observed data stored in a spectral library) and quantitation (e.g., spectral counting, quantitation from MS peak signal intensity or peak area, quantitation from MSMS fragment signal intensity). There are some half-dozen protein sequence databases available for searching, each varying in its level of completeness and redundancy. While search engine scoring schemes have made tremendous gains, peptide identification confidence scores can be influenced by a number of factors.

Infrastructural Challenges. The infrastructure for public deposition of proteomic data is evolving. In any given field, multiple repositories often arise more or less simultaneously. For instance, genomics researchers have a number of options for where to deposit sequence data (e.g., GenBank, EMBL, DDBJ). The availability of multiple repositories can be beneficial to the scientific community, as through collaboration and data sharing repositories can increase coverage, reduce duplication of effort, and gain some measure of security (e.g., data redundancy in the event of database failure or closure). This is already done to some extent: data deposited with the U.S. National Center for Biotechnology Information (NCBI) is mirrored at the European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ), and vice versa, to ensure long-term security.

While a few public repositories for proteomic data do exist (e.g., GPMDB, UniProtKB, Peptide Atlas, PRIDE and the newly formed NCBI Peptidome), they differ in the formats and kinds of data deposited (structural versus sequence, raw versus processed, uncurated versus curated). It appears that Tranche may be the primary mechanism that can serve the field as a repository for storing/serving raw data files. There has yet to emerge any international or centralized networks of repositories capable of reinforcing each other, although ProteomExchange is one such promising repository that is in its infancy.[6] Also, because each uses a different format, researchers may not be able to access all desired information from any single repository, and they may have to learn entirely different systems for accessing information for each.

Policy Challenges. Questions have long existed as to who holds the responsibility for setting and enforcing guidelines within the proteomics community, including guidelines for the submission of data for publication and standard metrics for assessing the quality of proteomic data (e.g., MS, protein affinity arrays) submitted for release. Such guidelines are necessary to ensure that enough information is provided to the community to explain an experiment, provide an assessment of the reliability of the data, and provide the data that support the results.

Currently, proteomics journals each develop their own guidelines for data submission. These guidelines can differ greatly in scope and stringency. The journal *Molecular & Cellular Proteomics*, for instance, has developed a set of publication guidelines based on discussions held in Paris in 2005.[7] These "Paris Guidelines" address the publication of protein sequence, quantitation, and post-translational modification data, but they have not been broadly adopted. Some journals encourage authors to adhere to these guidelines, while other journals such as PROTEOMICS have produced their own.[8]

In addition, the standards for deposition of data in centralized repositories are still evolving. For example, the HUPO Proteomics Standards Initiative (HUPO-PSI) has published standards for proteomic data representation, specifically for MS and protein−protein interaction studies, including minimal reporting requirements, standard formats, common sets of controlled vocabularies and/or ontologies, annotations, and validation guidelines.[9] These standards, though, are not yet universally accepted.

Principles for Proteomic Data Release and Sharing (Outcomes of Summit). The principles agreed upon at the Amsterdam summit are intended to address the major items required for the development of a useful and successful proteomic data release policy and to account for the challenges noted above where possible. The principles agreed upon include:

1. Timing. The timing with which proteomic data is released into the public domain should depend on the nature of the effort generating the data and should take into account the legitimate concerns of data producers-namely that prepublication release of data may jeopardize their opportunity to publish the first analyses of their own work. It is proposed that data generated by individual investigators should be released into the public domain at the latest upon publication while data generated by community resource projects should be released upon generation following appropriate QA/QC procedures.

2. Comprehensiveness. For data to be valuable to the proteomics community and other interested scientists, they must be released in a format that, as comprehensively as possible, captures the results of an experiment and the conditions under which the experiment was run. To this end, it is proposed that high quality raw data (e.g., mass spectral, protein/affinity array data) be released to the public. In order to assess the quality and value of such data, they should be well annotated with metadata, information on data quality, and identification quality control data.

3. Format. Open access to proteomic data requires community-supported standardized formats, controlled vocabularies, reasonable reporting requirements, and publicly available central repositories.

4. Deposition to repositories. Central repositories should make it attractive for depositors to use them. The processes of proteomic data submission to central repositories should be as simple and straightforward as possible but without losing essential information about the experiment. Central repositories should reduce the burden on data producers by clearly defining minimum requirements for data submission while encouraging rich annotation wherever possible, requirements

that should be enforced through curation and validation. Repositories should also provide publication-ready accession numbers and serve as access points among themselves (allowing one repository to access data housed in another).

5. Quality metrics. Central repositories should develop threshold metrics for assessing data quality. These metrics should be developed in a coordinated manner, both with the research community and among each other, so as to ensure interoperability. As data become shared through such repositories, their value will become obvious to the community, and momentum will grow to sustain data release.

6. Responsibility. Scientists, funding agencies, and journals share a joint responsibility for ensuring that all parties adhere to community standards for data release. Journals (as referees of the scientific record) and funding agencies (as supporters and facilitators of both investigator-driven research and community resource projects) hold both the power and responsibility to drive development and adoption of, and compliance with, guidelines and standards. As such, these entities should strongly encourage data and metadata deposits to centralized repositories as conditions of publication and funding, respectively.

Conclusion

The release of high quality data following standardized approaches would put the pace of proteomic research on a trajectory similar to that seen in large-scale genomics research. While numerous challenges remain in defining the policies and procedures for release of data into the public domain, the proteomics community is calling loudly for leading entities (e.g., funding agencies, journals, standards working groups, international societies) to produce the necessary guidelines. It is hoped that the Principles proposed herein will be considered and discussed by the community at large and will serve as a starting point for bringing proteomic data release practices in line with those of the genomics community as appropriate.

Acknowledgment. We thank Ruedi Aebersold and Anna D. Barker for input and advice in planning the International Summit on Proteomics Data Release and Sharing Policy.

References

(1) Rodriguez, H. International Summit on Proteomics Data Release and Sharing Policy. *J. Proteome Res.* **2008**, *7*, 4609.
(2) *Policies on Release of Human Genomic Sequence Data*; US Department of Energy Human Genome Project: Washington; http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml [accessed 30 October 2008].
(3) *Ibid.*
(4) *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*; Welllcome Trust: London, 2003; http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf [accessed 30 October 2008].
(5) Deutsch, E. mzML: A single, unifying data format for mass spectrometer output. *Proteomics* **2008**, *8* (14), 2776–7.
(6) Hermjakob, H.; Apweiler, R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: Making Proteomics Data Accessible. *Expert Rev. Proteomics* **2006**, *3* (1), 1–3.
(7) *Publication Guidelines for the Analysis and Documentation of Peptide and Protein Identifications*; American Society for Biochemistry and Molecular Biology. Molecular & Cellular Proteomics. http://www.mcponline.org/misc/ParisReport_Final.shtml [accessed 31 October 2008].
(8) Wilkins, M. R.; Appel, R. D.; Van Eyk, J. E.; Chung, M. C.; Görg, A.; Hecker, M.; Huber, L. A.; Langen, H.; Link, A. J.; Paik, Y. K.; Patterson, S. D.; Pennington, S. R.; Rabilloud, T.; Simpson, R. J.; Weiss, W.; Dunn, M. J. Guidelines for the next 10 years of proteomics. *Proteomics* **2006**, *6*, 4–8.
(9) Human Proteome Organization. Proteomics Standards Initiative. http://www.hupo.org/research/psi/ [accessed 31 October 2008].

PR900023Z