communications Journal of proteome research

ProteomeCommons.org Collaborative Annotation and Project Management Resource Integrated With the Tranche Repository

James A. Hill, Bryan E. Smith, Panagiotis G. Papoulias, and Philip C. Andrews*

Departments of Biological Chemistry and Bioinformatics, University of Michigan, Ann Arbor, Michigan

Received February 1, 2010

Abstract: ProteomeCommons.org has implemented a resource that incorporates concepts of Web 2.0 social networking for collaborative annotation of data sets placed in the Tranche repository. The annotation tools are part of a project management resource that is effective for individual laboratories or large distributed groups. The creation of the resource was motivated by the need for a way to encourage annotation of data sets with high accuracy and compliance rates. The system is designed to respond to the dynamic nature of research in an easy-to-use fashion through the use of a dynamic data model that does not inhibit the innovation that is important for basic research. Placing the annotation tool within a project manager allows annotation to occur over the life of the project and provides the security and monitoring capabilities needed for large or small collaborative projects. The resource effectively supports distributed groups of investigators working on common data sets and is available immediately at https://ProteomeCommons.org. In addition, a silver compliant data resource based on ProteomeCommons.org has been developed for cancer Biomedical Informatics Grid (caBIG) to allow much broader access to the annotations describing data sets in the Tranche repository.

Keywords: collaboration tool • annotation management • Proteome Commons • ProteomeCommons.org • Tranche • data service

Introduction

Annotation has been a difficult problem in the field of proteomics and other postgenome disciplines, with compliance, ease of use, and the dynamic nature of research being some of the challenges to be resolved. Addressing these concerns is necessary for the effective use of data repositories (like Tranche) used by the proteomics community. A Tranche repository is essentially a very large, unstructured, distributed RAID-like storage system, so having the ability to search the repository using the annotations describing those data sets is an important feature in locating useful data sets. Search capabilities and reuse of data sets are dependent on collection of the accurate and complete annotations.[6] Some concepts and methods used in Web 2.0 social networking were adopted in development of this resource, recognizing that they represent a particularly effective approach to collaborative efforts.

Proteomics projects have a number of challenging features that impact the informatics disciplines. They generally produce large and often complex data sets using technologies that evolve rapidly.[4] They usually involve collaboration across multiple laboratories and frequently extend for several months to multiple years in length. Interpretation of the raw data sets is still evolving, so re-evaluation of older data sets using new software may provide more robust results.[4] A further challenge to data set reuse is that not all investigators have access to all the software packages for data analysis. Most global proteomics data sets are initially mined for very specific goals and considerable further information can be gleaned from subsequent reanalyses for other scientific goals. Original data sets also offer significant value to bioinformaticians if they can obtain access to quality metadata.[6] An additional advantage of storing the raw data sets is that useful metadata embedded in the original data files is lost during conversion to standard file formats. It should also be noted that routine digital signal processing (smoothing, noise filters, peak detection, and de-isotoping) can introduce artifacts, overlook significant signals, or exhibit apparent variability in signal detection, particularly if parameters are not optimized.[4] Finally, aggregation of related data sets across multiple studies allows improved statistics—leading to more detailed interpretations.[2]

Methods

The Annotations Lexical database (ALex) was developed to store meta information in a dynamic manner. The core feature of this software is that it allows for a dynamic information model, which allows metadata stored in the database to evolve along with the field as new technologies, new terms, and new data relationships are developed. This reflects the reality of basic research where new technologies and methods lead to new terms, new concepts, and new relationships between terms. Consistent with the design of the Tranche repository, the annotation system was designed to be configurable, and because it works through a dynamic information model it can be applied to many other fields. Even though the information model is dynamic, it can be and has been made compatible with external information grid systems (e.g., caBIG).

The new ProteomeCommons.org system applies principles of Web 2.0 social networking to meet the needs of researchers, especially those involved in large collaborative projects. Proteomics projects may take years to complete, often involve several laboratories, and cross disciplinary boundaries, so

* To whom correspondence should be addressed. E-mail: andrewsp@umich.edu. Address: Rm. 1198, 300 N. Ingalls, Ann Arbor, MI 48104.

10.1021/pr1000972

Published on Web 03/31/2010

annotation of data sets by any one researcher is challenging. The individual entering the annotations at the end of a project when a data set is deposited for public access is often not the person that procured the data set. This problem is addressed in part by allowing researchers to form groups to cooperatively manage and annotate the data sets that they generated in the course of a proteomics study. Annotations may be entered over the life of a project rather than at the end when the data sets are deposited, which contributes to improved compliance. The annotation forms are divided into functional categories that allow responsibility for each category to be assigned to the domain experts by a group leader. Progress in each annotation category is monitored as percentage completion. The annotation manager supports multiple annotation standards, reflecting the high level of variation in proteomics research and technologies.

The annotation management tool is a part of the basic project management resource of ProteomeCommons.org. The management tool is designed to be consistent with multi-institutional projects, but also scales well for individual laboratories. It provides the basic information that principal investigators need to track and manage projects as well as maintaining the metrics necessary for reporting purposes (e.g., number of data set downloads). The features of project management include forming groups, inviting personnel, management of data on the Tranche repository (upload, make public, delete, etc.) and linking data sets to publications.

A typical use case for the resource might look like this: At the outset of a proteomics project, an investigator can start a new project on ProteomeCommons.org. This is a simple process of filling out a form followed by inviting project members. As each project member is invited, the investigator can specify what role and permissions the new member would have were they to join. When the invitation is accepted, the new project member will see the project added to the list of projects on their ProteomeCommons.org home page. The project can be hidden from the public at the principal investigator's discretion. Selecting the project from the project member's ProteomeCommons.org work space gives the user access to the project's data sets and associated annotations.

Data sets can be associated with the project either by uploading them to the Tranche repository or by importing data sets previously uploaded to Tranche. Data sets can remain accessible only to project members until a project member with appropriate permissions triggers public release of specific data sets (e.g., after acceptance of a manuscript). Each data set within a project has an annotation record associated with it that is accessible to all users. The MIAPE-based annotations are organized by functional categories (e.g., sample preparation, chromatography, mass spectrometry, etc.) and each category can be assigned to the domain expert on the project. That person's name is indicated next to the annotation category along with the percentage completion for that category. The percentage completion for each category and the researcher responsible for the category are visible to all project members. All members can contribute to the annotation, but the assigned domain expert is responsible for ensuring that it is completed. This allows users to complete the annotation during the course of a project rather than at the end when the data sets are submitted to the repository. It also provides a way for primary investigators to monitor annotation compliance and to send reminder emails to the domain expert as deadlines approach. Controlled vocabularies are used when appropriate and the system provides for import of standard operating procedures or materials and methods texts to supplement the MIAPE standard. Finally, investigators may link their data sets to the publication they are associated with to facilitate citation of their research.

The ProteomeCommons.org Tranche Annotations (PCTA) database has obtained caBIG silver level compatibility. The caBIG network was developed by the National Cancer Institute to enable data sharing among cancer researchers.[1] Compatibility with caBIG ensures that information stored in the annotations database will be accessible to grid users—searching and browsing can be done externally from ProteomeCommons.org. The end result is that data sets in the ProteomeCommons.org Tranche repository will become more broadly accessible to researchers through the caBIG silver compliant PCTA data service.

The ALex database software was developed in Java to be used on a MySQL database. It is open source under the Apache 2.0 License and can be found through the Tranche Project Web site at http://trancheproject.org.

Discussion

Data sets are of limited use without the appropriate information to establish the experimental and analytical contexts.[6] The ProteomeCommons.org project and annotation management resource now allows users to upload data sets to Tranche, track and share those data sets with other project members, and annotate them in collaboration with their colleagues. ProteomeCommons.org provides a mechanism for allowing annotations to be linked with original data sets and to publications. This provides a practical option for researchers to meet recently proposed requirements for publishing data.[5] The Tranche and ProteomeCommons.org systems support current data standards and complement the various centralized databases (TheGPMdb, PeptideAtlas, PRIDE, Peptidome).[3] It represents a source of data for these databases and is already being utilized by many of them. The Tranche repository currently contains 9.5 TB of data, representing 7603 total data sets. Between February and August 2009, there were 1147 downloads of 604 data sets, with total bytes downloaded of 3.98 TB. The new community resource is available immediately and can be accessed at https://ProteomeCommons.org.

Acknowledgment. The authors would like to thank Mark Gjukich and Shelly Chang for computational and graphics support for this manuscript. We also acknowledge support by NCI subcontract 27XS115 and from NCRR (1P41RR018627). Server housing support was provided by NCIBI (NIH U54 DA021519).

References

(1) National Cancer Institute, U.S. National Institutes of Health. cancer Biomedical Informatics Grid (caBIG): Fact Sheet. http://www.cancer.gov/newscenter/cabig-QA (accessed Jul 15, 2009).

(2) Craig, R.; Cortens, J. P.; Beavis, R. C. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *J. Proteome Res.* **2004**, *3* (6), 1234–1242.

(3) ProteomeXchange. Martens, L.; Deutsch, E. ; Hemjakob, H.; Omenn, G. Proteomics Data Submission Strategy For ProteomeExchange. http://proteomexchange.org/doc/ProteomExchange_data_submission_strategy_final.pdf (accessed Jul 15, 2009).

(4) Prince, J. T.; Carlson, M. W.; Wang, R.; Lu, P.; Marcotte, E. M. The Need For a Public Proteomics Repository. *Nat. Biotechnol.* **2004**, *22*, 471–472.

(5) Rodriguez, H.; Snyder, M.; Uhlén, M.; Andrews, P.; Beavis, R.; Borchers, C.; Chalkley, R. J.; Cho, S. Y.; Cottingham, K.; Dunn, M.; Dylag, T.; Edgar, R.; Hare, P.; Heck, A. J. R.; Hirsch, R. F.; Kennedy, K.; Kolar, P.; Kraus, H.; Mallick, P.; Nesvizhskii, A.; Ping, P.; Pontén, F.; Yang, L.; Yates, J. R., III; Stein, S. E.; Hermjakob, H.; Kinsinger, C. R.; Apweiler, R. Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles. *J. Proteome Res.* **2009**, *8* (7), 3689–3692.

(6) Taylor, C. F.; Paton, N. W.; Lilley, K. S.; Binz, P.; Julian, R. K.; Jones, A. R.; Zhu, W.; Apweiler, R.; Aebersold, R.; Deutsch, E. W.; Dunn, M. J.; Heck, A. J. R.; Leitner, A.; Macht, M.; Mann, M.; Martens, L.; Neubert, T. A.; Patterson, S. D.; Ping, P.; Seymour, S. L.; Souda, P.; Tsugita, A.; Vandekerckhove, J.; Vondriska, T. M.; Whitelegge, J. P.; Wilkins, M. R.; Xenarios, I.; Yates, J. R., III; Hermjakob, H. The Minimum Information About a Proteomics Experiment (MIAPE). *Nat. Biotechnol.* **2007**, *25*, 887–893.

PR1000972