Journal of Proteomics & Bioinformatics - Open Access

JPB/Vol.3/January 2010

Research Article OPEN ACCESS Freely available online doi:10.4172/jpb.1000114

Omics.pnl.gov: A Portal for the Distribution and Sharing of Multi-Disciplinary Pan-Omics Information

Ken J. Auberry, Gary R. Kiebel, Matthew E. Monroe, Joshua N. Adkins, Gordon A. Anderson and Richard D. Smith*

Biological Sciences Division, Mail Stop: K8-98, Pacific Northwest National Laboratory, P. O. Box 999, 3335 Q Avenue, Richland, WA 99352

Introduction

The data production of scientific studies is growing at a nearly exponential rate (Domon and Aebersold, 2006; Kiebel et al., 2006). This growth leads to challenges in disseminating primary experimental results for peer review and public access, while simultaneously providing information that enables reproducing the studies and/or analyzing the results in a proper context. Recent mandates from various public funding agencies are requiring data release plans be included as a project goal. This requirement is coupled with an increased need for transparency in complex research, as evidenced by the data release policies now being implemented by peer-reviewed journals such as Molecular & Cellular Proteomics (http://mcponline.org/misc/PhiladelphiaGuidelines.dtl). This combination of good scientific citizenship and funding requirements has brought the data distribution issue to the domain of scientific information management researchers.

Most mass spectrometry-based proteomics groups choose to utilize one of the prominent data distribution sites, such as Tranche (Falkner JA, Andrews PC, HUPO Conference 2006. Long Beach, USA, Poster presentation), PRIDE (Martens et al., 2005), NCBI's Peptidome (Slotta et al., 2009), Human ProteinPedia (Mathivanan et al., 2008), or PeptideAtlas (Desiere et al., 2006). These sites make sense for small or targeted data releases, but for large groups with diverse experimental approaches and myriad biological model systems (e.g. Callister et al., 2008; Kiebel et al., 2006), the choice may not be so clear. Additionally, these sites are aimed at managing and disseminating data that are associated with identifications and do not generally make all the raw data available. This raw data is particularly useful to developers of analysis tools, as well as in cases where the integration of multiple data sources can improve the confidence of a result. Our goal in the construction of this site is to augment these pubic repositories by making available entire sets of raw and processed results along with their associated metadata. This requires that careful considerations be made regarding the design of the site in order to render it useful to the community. Herein, we present an initial version of such a site, referred to as the Biological MS Data and Software Distribution Center, which can be visited at http://omics.pnl.gov. This site leverages vast amounts of pre-existing experimental data and metadata gathered since 2001 and stored in our purpose-built data management system, PRISM (Kiebel et al., 2006).

Design philosophy

The initial intent for the site was simply to provide local researchers with a mechanism for making large sets of experimental results available to both their collaborators and the greater scientific community. This intent was coupled with a desire to organize the data in a hierarchical structure and present results in such a way as to make them readily usable and understandable by researchers who were familiar with the field, but not necessarily experts in our particular methodologies. In addition to presenting the hierarchical metadata, another expectation was providing website users with a capability for downloading large sets of raw and processed instrumental data (greater than single Terabytes).

Omics research at Pacific Northwest National Laboratory (PNNL) involves a number of different collaborations, many of which include bioinformatics components that require large volumes of raw data at all levels of quality to produce accurate results. This system provides one model to support the current needs of these collaborations while also providing the frameworks necessary to build more advanced capabilities. In the past, the information generated by these collaborations has necessitated the shipment of hard drives full of data across the country. Streamlining this aspect of our data delivery process has driven the design of the site's initial requirements as well as many aspects of its architecture. We currently have over 150 terabytes of raw and processed data in our archives and these developments enable its dissemination.

Types of data made available

The majority of the data available on the site comes from liquid chromatography coupled mass spectrometry (LC-MS) studies of proteomes, metabolomes, etc., conducted using either traditional "shotgun" proteomics (e.g. Washburn et al., 2001; Adkins et al., 2006) or the accurate mass and time tag methodology (e.g. Smith et al., 2002; Shi et al., 2006). These data include raw LC-MS and tandem mass spectrometric results (LC-MS/MS) from multiple instrument types, ranging from benchtop linear ion traps to custom built LC-FTICR platforms with very high mass measurement accuracy. Also available are processed data in the form of peptide identifications for LC-MS/MS, peak deconvolution information for high mass accuracy LC-MS, and MASIC-generated (Monroe et al., 2008) single ion chromatograms for LC-MS(/MS) data. While the current collection of data is largely

***Corresponding author:** Richard D. Smith, PhD, Biological Sciences Division, Mail Stop: K8-98, Pacific Northwest National Laboratory, P. O. Box 999, 3335 Q Avenue, Richland, WA 99352, Tel: 509-371-6576; Fax: 509-371-6564; E-mail: rds@pnl.gov

Received December 02, 2009; **Accepted** January 05, 2010; **Published** January 06, 2010

Citation: Auberry KJ, Kiebel GR, Monroe ME, Adkins JN, Anderson GA et al. (2010) Omics.pnl.gov: A Portal for the Distribution and Sharing of Multi-Disciplinary Pan-Omics Information. J Proteomics Bioinform 3: 001-004. doi:10.4172/jpb.1000114

DAnTE	Used to perform various downstream data analysis, data reduction, and data comparison steps including normalization, hypothesis testing and clustering (Polpitiya et al., 2008).
Decon2LS	Used to de-isotope MS spectra and to detect features from MS data using isotopic signatures of expected components (Jaitly et al., 2009).
DeconMSn	Used to create .dta files or .MGF files from tandem MS/MS data (Thermo Finnigan or mzXML format) with accurate parent monoisotopic mass and charge determined using the isotopic signatures of the parent ions (Mayampurath et al., 2008).
MultiAlign	Aligns multiple LC-MS datasets to one another, after which LC-MS features can be matched to a database of peptides (e.g. an AMT tag database).
VIPER	Used to visualize and characterize the features detected during LC-MS analyses (Monroe et al., 2007).

Table 1: Selected software packages available on the site.

composed of mass spectrometric results, it is our intent to present other types of -omics data on the site as they become available.

Selected open source software packages are available on the site that allow others to process or understand the processes used to analyze the data. Some of these include tools for the manipulation and parsing of various protein database files, tools to assist in data extraction, analysis and refinement of LC-MS (/MS) data, as well as an array of programs to facilitate visualization and presentation of omics-related data. A selection of these applications is summarized in Table 1. Presentations and poster reprints describing many of these processes are also available on the site, along with a full list of available software packages (http://omics.pnl.gov/software/).

Using the site

Upon arriving at the site (http://omics.pnl.gov/), the user is presented with a menu of possible activities, including browsing or searching available data, downloading various data analysis software packages, viewing research posters and presentations, registering a new account, etc. While not needed to browse the contents of the site or download software, a minimal registration process is necessary in order to download research data. This registration enables us to gather aggregate usage data required for reporting purposes, as well as statistical information regarding how the site is used and which types of data are frequently downloaded.

Once signed in, the user can search the site for associated keywords or browse via several top-level entities that hierarchically arrange the available data into categories such as journals, associated publications, organisms, year of production, and mass spectrometer type. Either method yields a structured tree view that represents the subset of data selected. From this view, the user can descend into the hierarchy to obtain increasing levels of detail.

From the "Experiment" level down, new options are made available in the form of downloadable content icons located to the left of each entry. These icons allow collections of data to be marked for later retrieval, using a "shopping cart" metaphor familiar to anyone who has ever made an online purchase (Figure 1). A running tally of selected files and their cumulative sizes is summarized in the right hand menu column, along with estimated download times for various speeds of connectivity (Figure 2). Currently, a user could conceivably select more than 10 Terabytes of data, an amount impractical for most users to download or even store.

Once the user selects a set of data files to be retrieved, the "Download from your Cart" option can be selected from the side menu, taking them to a page that summarizes their cart contents in detail (Figure 3). From this page, individual items can be removed from the list, and entire classes of data can be enabled/disabled. This option is useful for deselecting data from a certain type of instrument, for example. The contents of the cart can then be transferred to the user's computer using a combined streaming/caching mechanism, described below.

Implementation details

The core component of the site is the metadata storage engine powered by a PostgreSQL database (PostgreSQL 8.1.3, http://www.postgresql.org/). This framework maintains all of the information necessary for the operation of the site, such as the locations of files in the archive storage hierarchy or the contents of a user's data "shopping cart". When data are to be made available on the site, metadata for the entities involved is gathered up from an internal-only PRISM/DMS server and inserted into the Postgres database on the publicly accessible server that hosts the website. This server is connected to a multi-petabyte file archive system located in EMSL, the Environmental Molecular Sciences Laboratory at PNNL (http://www.emsl.pnl.gov/) via a

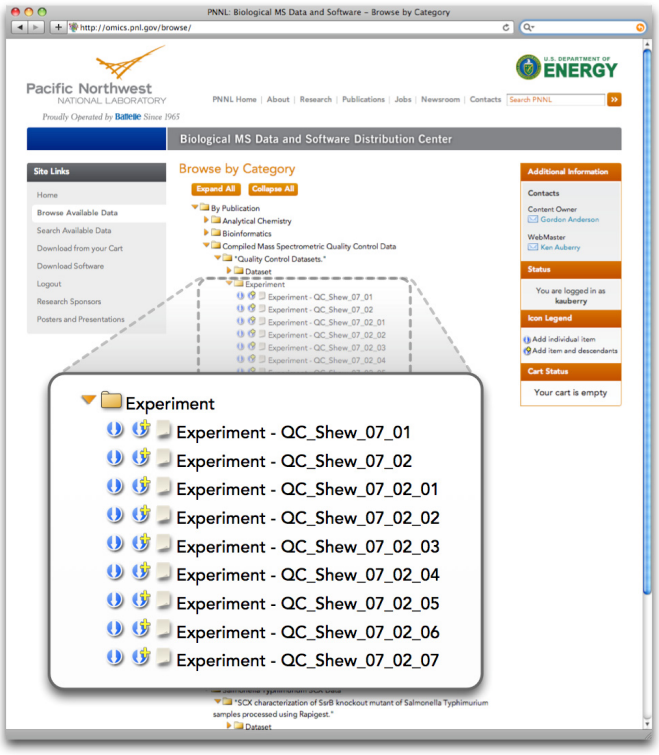

Figure 1: Browsing data by category. The icons to the left of each detail entry indicate that data are available and can be added to the user's cart. The same icon with the plus sign indicates that data from entry along with all its children should be added.

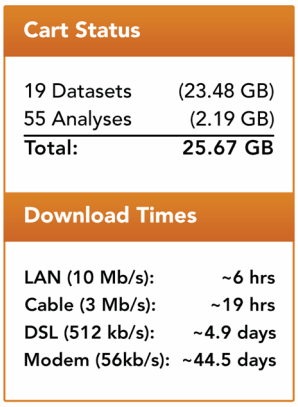

Figure 2: Cart Status and Estimated File Transfer Times panel.

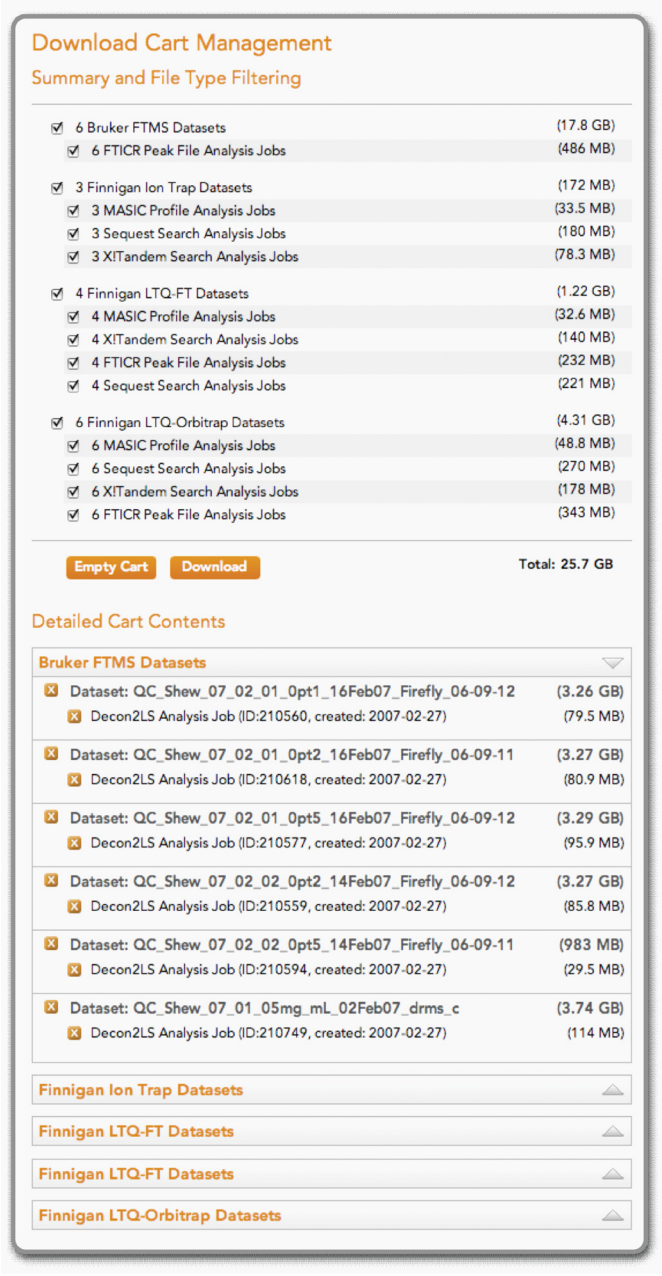

Figure 3: Inset view of the Download Cart Management interface. From here, users can remove items from their cart, or disable an entire class of results.

10Gbps Ethernet connection. Because all of our instrument and analysis data are stored in this archive system, no actual mass transfer of raw data needs to take place. The locations of the files can simply be referenced in the distribution site's database and be served directly from the archive.

The metadata storage tables are accessed using PHP (PHP 5.1.2, http://www.php.net/) as the server-side scripting language that dynamically generates page content for various types of metadata within the hierarchy. These data types include experimental data that describes the conditions under which a sample was prepared, LC-MS (/MS) data along with the parameters used to govern the operation of the instrumentation, and analysis results that describe things such as the peptides identified in a particular set of data. This content is then served to the end-user by an Apache web server (Apache HTTP Server, http://httpd.apache.org/) running under Red Hat Enterprise Linux 4 (Red Hat, Inc, http://www.redhat.com/rhel/).

To minimize page loading times, navigation elements such as the tree views used for the data browsing and search pages have the bulk of their content loaded on demand, using Ajax-style asynchronous calls (Garrett, 2005) that are triggered as a user drills down into the available data. These same types of calls are used to manage and report the contents of the user's cart, which lends a greater degree of interactivity to the site while minimizing the number of full page reloads.

When full sets of data are triggered to download from the site, a background process is invoked that steps through the contents of the user's cart in a hierarchical fashion that corresponds with the layout of the requested data. Once the manifest for the package is generated, the files themselves are collected and combined into an uncompressed Tar file (Gnu Tar, http://www.gnu.org/software/tar/). Even as the file is being constructed and cached in temporary storage, the server is already starting to stream the contents to the user, which reduces the wait time experienced by the user. The use of the cached copy of the file allows for interrupted downloads to be resumed and mitigates the possibility of having to restart a large transfer from the beginning in the event of a network failure, etc.

Future plans

The system is continually undergoing development to add new capabilities and features to expand its use to the scientific community. Currently, mass spectrometric information and analysis results are only made available in formats native to the instruments or software packages that generated them, rather than in more generic formats such as mzML (http://www.psidev.info/) and pepXML (http://tools.proteomecenter.org/wiki/). Efforts are underway to automate the production of these file formats from the existing data and display them alongside the native files. Making these interchangeable files also opens up opportunities for automating the deposit of data in other public repositories. Another planned addition to the site is tighter integration with our existing data management system (Kiebel et al., 2006), which will provide researchers with the ability to automatically push data products out to the dissemination site based on previously established matching criteria. As more and different types of data are made available on the site, additional options will be added to the system's search facility to allow deeper exploration based on the contents of the processed data (proteins ID's, gene anno-

tations, etc.) rather than solely through its associated metadata.

Acknowledgments

The research described in this paper was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U. S. Department of Energy Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory. Portions of this work were supported by the Department of Energy Office of Biological and Environmental Research at PNNL grant (ER63232-1018220-0007203), the NIH National Institute of Allergy and Infectious Diseases (interagency agreements Y1-AI-4894-01 and Y1-AI-8401-01) and the NIH National Center for Research Resources (RR18522). PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830.

References

1. Adkins JN, Mottaz HM, Norbeck AD, Gustin JK, Rue J, et al. (2006) Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. Mol Cell Proteomics 5: 1450-1461. » CrossRef » PubMed » Google Scholar
2. Callister SJ, McCue LA, Turse JE, Monroe ME, Auberry KJ, et al. (2008) Comparative bacterial proteomics: analysis of the core genome concept. PLoS One 3: e1542. » CrossRef » PubMed » Google Scholar
3. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, et al. (2006) The PeptideAtlas project. Nucleic Acids Res 34: D655-658. » CrossRef » PubMed » Google Scholar
4. Domon B, Aebersold R (2006) Challenges and opportunities in proteomics data analysis. Mol Cell Proteomics 5: 1921-1926. » CrossRef » PubMed » Google Scholar
5. Garrett JJ (2005) Ajax: A New Approach to Web Applications. (http://adaptivepath.com/ideas/essays/archives/000385.php). » CrossRef » PubMed » Google Scholar
6. Jaitly N, Mayampurath A, Littlefield K, Adkins JN, Anderson GA, et al. (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. BMC Bioinformatics 10: 87. » CrossRef » PubMed » Google Scholar
7. Kiebel GR, Auberry KJ, Jaitly N, Clark DA, Monroe ME, et al. (2006) PRISM: a data management system for high-throughput proteomics. Proteomics 6: 1783-1790. » CrossRef » PubMed » Google Scholar
8. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: the proteomics identifications database. Proteomics 5: 3537-3545. » CrossRef » PubMed » Google Scholar
9. Mathivanan S, Ahmed M, Ahn NG, Alexandre H, Amanchy R et al (2008) Human Proteinpedia enables sharing of human protein data. Nat Biotechnol 26: 164-167. » CrossRef » PubMed » Google Scholar
10. Monroe ME, Shaw JL, Daly DS, Adkins JN, Smith RD (2008) MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. Comput Biol Chem 32: 215-217. » CrossRef » PubMed » Google Scholar
11. Monroe ME, Tolic N, Jaitly N, Shaw JL, Adkins JN, et al. (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. Bioinformatics 23: 2021-2023. » CrossRef » PubMed » Google Scholar
12. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, et al. (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics 24: 1021-1023. » CrossRef » PubMed » Google Scholar
13. Polpitiya AD, Qian WJ, Jaitly N, Petyuk VA, Adkins JN, et al.(2008) DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics 24: 1556-1558. » CrossRef » PubMed » Google Scholar
14. Shi L, Adkins JN, Coleman JR, Schepmoes AA, Dohnkova A, et al. (2006) Proteomic analysis of Salmonella enterica serovar typhimurium isolated from RAW 264.7 macrophages: identification of a novel protein that contributes to the replication of serovar typhimurium inside macrophages. J Biol Chem 281: 29131-29140. » CrossRef » PubMed » Google Scholar
15. Slotta DJ, Barrett T, Edgar R (2009) NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol 27: 600-601. » CrossRef » PubMed » Google Scholar
16. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, Shen Y, et al. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. Proteomics 2:513-523. » CrossRef » PubMed » Google Scholar
17. Washburn MP, Wolters D, Yates JR 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242-247. » CrossRef » PubMed » Google Scholar